* **Task 5:** A study was carried out to explore the relationship between **Aggression** and several potential predicting factors in 666 children who had an older sibling. Variables measured were **Parenting Style** (high score = bad parenting practices), **Computer Games** (high score = more time spent playing computer games), **Television** (high score = more time spent watching television), **Diet** (high score = the child has a good diet low in harmful additives), and **Sibling Aggression** (high score = more aggression seen in their older sibling). Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child. All other variables were treated in an exploratory fashion. The data are in the file **Child Aggression.sav.** Analyse them with multiple regression. 

8.6.4. Saving regression diagnostics 

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

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

• **pre\_1:** unstandardized predicted value;

• **zpr\_1:** standardized predicted value;

• **adj\_1:** adjusted predicted value;

• **sep\_1:** standard error of predicted value;

• **res\_1:** unstandardized residual;

• **zre\_1:** standardized residual;

• **sre\_1:** Studentized residual;

• **dre\_1:** deleted residual;

• **sdr\_1:** Studentized deleted residual;

• **mah\_1:** Mahalanobis distance;

• **coo\_1:** Cook's distance;

• **lev\_1:** centred leverage value;

• **sdb0\_1:** standardized DFBETA (intercept);

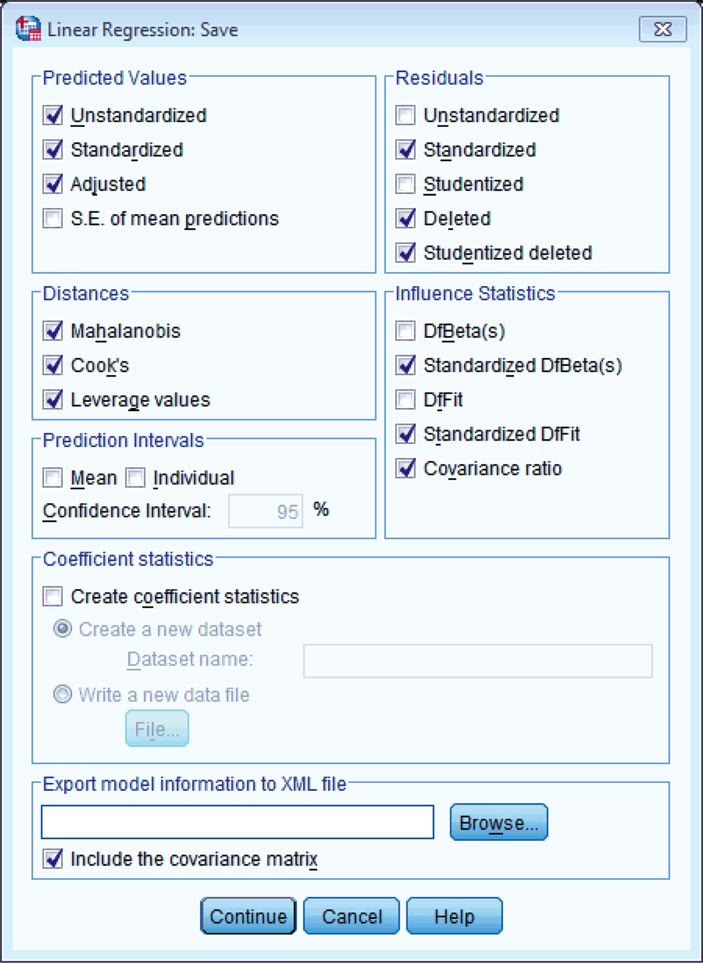
• **sdb1\_1:** standardized DFBETA (predictor 1);

• **sdb2\_1:** standardized DFBETA (predictor 2);

• **sdf\_1:** standardized DFFIT;

• **cov\_1:** covariance ratio.

FIGURE 8.18 Dialog box for regression diagnostics

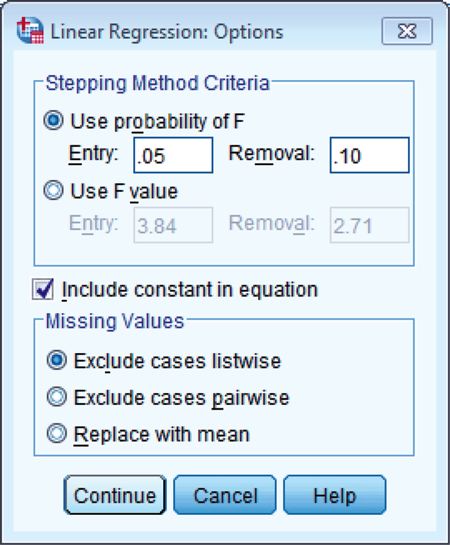


8.6.5. Further options 

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

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

FIGURE 8.19 Options for linear regression



8.6.6. Robust regression 

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

 ODITI'S LANTERN *Regression*

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

8.7. Interpreting multiple regression 

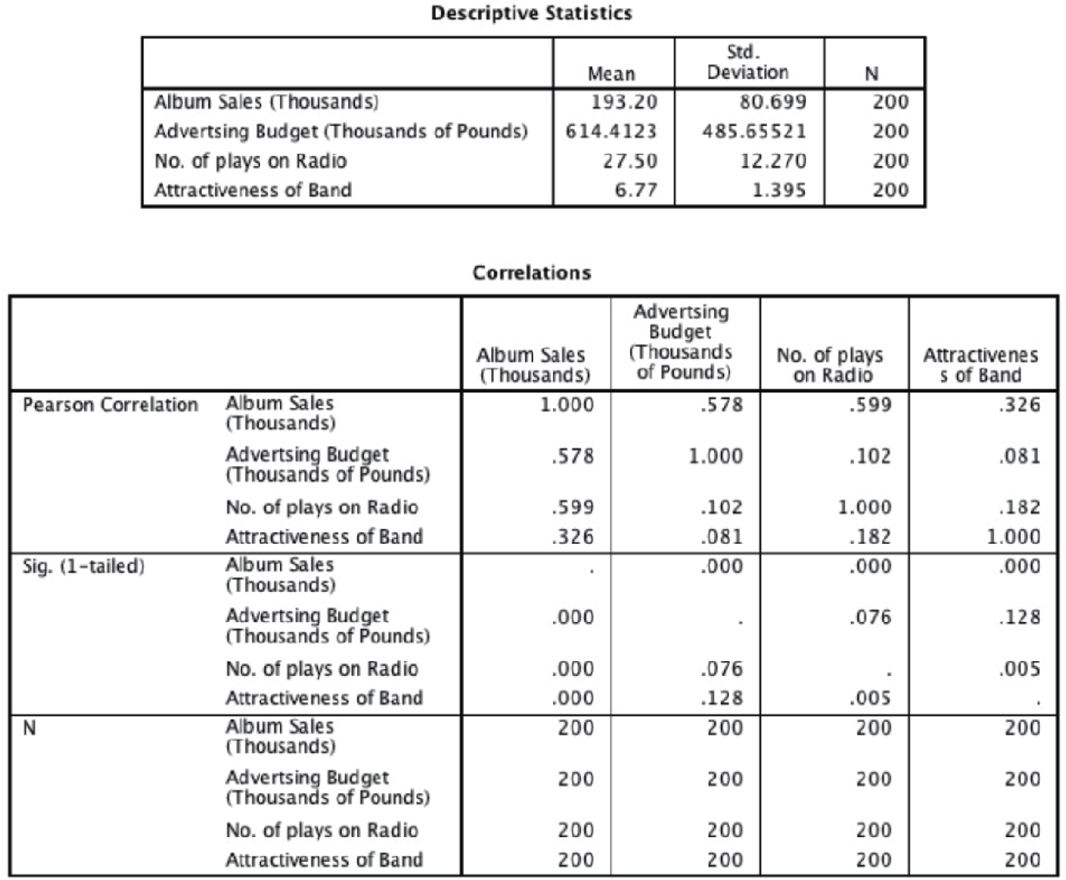
Having selected all of the relevant options and returned to the main dialog box, we need to click on  to run the analysis. SPSS will spew out copious amounts of output in the viewer window, and we now turn to look at how to make sense of this information.

8.7.1. Descriptives 

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

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

OUTPUT 8.4 Descriptive statistics for regression analysis



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

 CRAMMING SAM'S TIPS Descnptive statistics

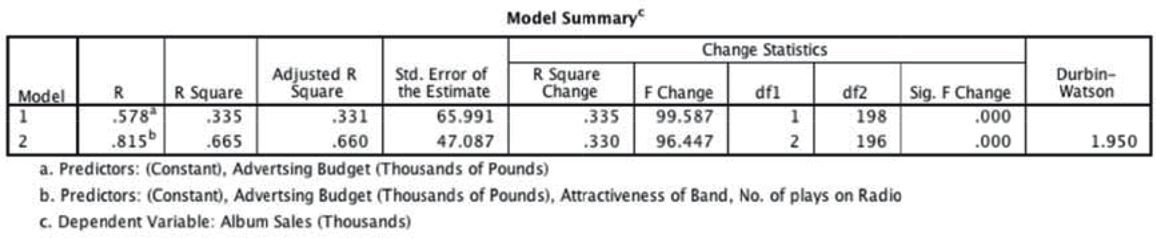
• Use the descriptive statistics to check the correlation matrix for multicollinearity – that is, predictors that correlate too highly with each other, r > .9.

8.7.2. Summary of model 

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

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

OUTPUT 8.5 Regression model summary



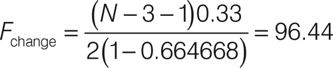
The adjusted *R*2 gives us some idea of how well our model generalizes and ideally we would like its value to be the same as, or very close to, the value of *R*2. In this example the difference for the final model is small (in fact the difference between the values is .665 – .660 = .005 or 0.5%). This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

 JANE SUPERBRAIN 8.2 *Maths frenzy* 

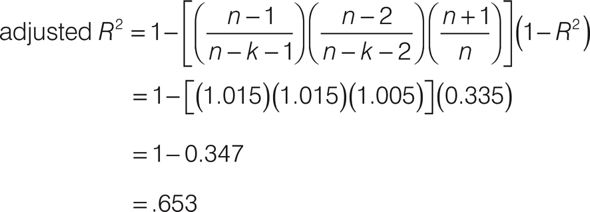
We can have a look at how some of the values in the output are computed by thinking back to the theory part of the chapter. For example, looking at the change in *R*2 for the first model, we have only one predictor (so *k* = 1) and 200 cases (*N* = 200), so the *F* comes from Equation (8.10):15



In model 2 in Output 8.5 two predictors have been added (attractiveness and radio play), 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 .330 (*R*2change), making the *R*2 of the new model .665 (*R*2new).16 The *F*-ratio for this change comes from Equation (8.15):



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

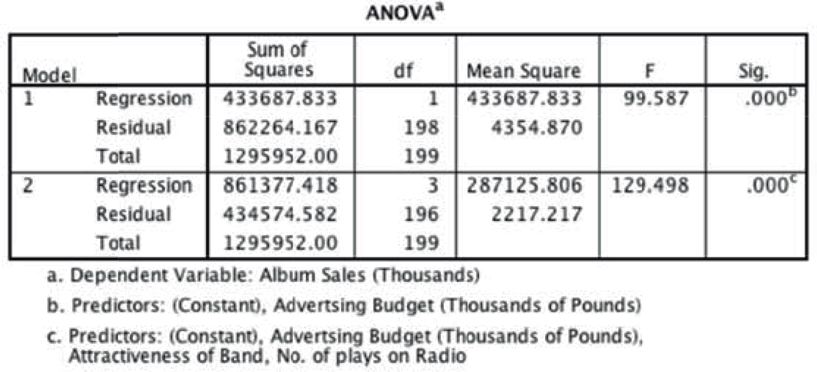


The change statistics are provided only if requested, and these tell us whether the change in *R*2 is significant. In Output 8.5, the change is reported for each block of the hierarchy. So, model 1 causes *R*2 to change from 0 to .335, and this change in the amount of variance explained gives rise to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

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

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

OUTPUT 8.6



 CRAMMING SAM'S TIPS The model summary

• The fit of the regression model can be assessed using the *Model Summary* and *ANOVA* tables from SPSS.

• Look for the *R*2 to tell you the proportion of variance explained by the model.

• If you have done a hierarchical regression then assess the improvement of the model at each stage of the analysis by looking at the change in *R*2 and whether this change is significant (look for values less than .05 in the column labelled *Sig F Change*).

• The ANOVA also tells us whether the model is a significant fit of the data overall (look for values less than .05 in the column labelled *Sig.*).

• The assumption that errors are independent is likely to be met if the Durbin–Watson statistic is close to 2 (and between 1 and 3).

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

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

16 The more precise value is 0.664668.

8.7.3. Model parameters 

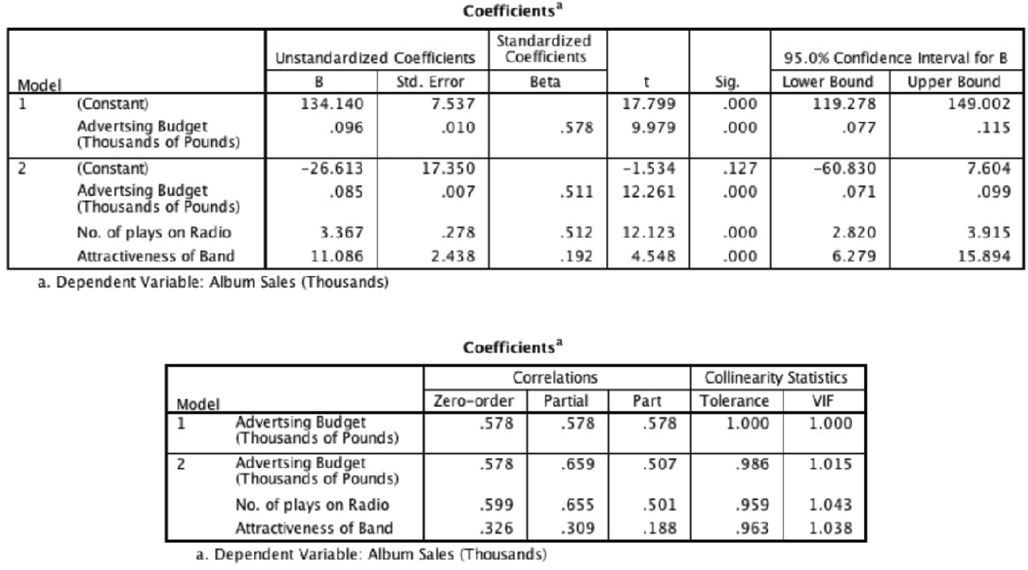
So far we have looked at whether or not the model has improved our ability to predict the outcome variable. The next part of the output is concerned with the parameters of the model. Output 8.7 shows the model parameters for both steps in the hierarchy. Now, the first step in our hierarchy was to include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

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



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

OUTPUT 8.7 Coefficients of the regression model17



• **Advertising budget** (*b* = 0.085): This value indicates that as advertising budget increases by one unit, album sales increase by 0.085 units. Both variables were measured in thousands; therefore, for every £1000 more spent on advertising, an extra 0.085 thousand albums (85 albums) are sold. This interpretation is true only if the effects of attractiveness of the band and airplay are held constant.

• **Airplay** (*b* = 3.367): This value indicates that as the number of plays on radio in the week before release increases by one, album sales increase by 3.367 units. Therefore, every additional play of a song on radio (in the week before release) is associated with an extra 3.367 thousand albums (3367 albums) being sold. This interpretation is true only if the effects of attractiveness of the band and advertising are held constant.

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

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

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

• **Advertising budget** (*standardized β* = .511): This value indicates that as advertising budget increases by one standard deviation (£485,655), album sales increase by 0.511 standard deviations. The standard deviation for album sales is 80,699 and so this constitutes a change of 41,240 sales (0.511 × 80,699). Therefore, for every £485,655 more spent on advertising, an extra 41,240 albums are sold. This interpretation is true only if the effects of attractiveness of the band and airplay are held constant.

• **Airplay** (*standardized β* = .512): This value indicates that as the number of plays on radio in the week before release increases by one standard deviation (12.27), album sales increase by 0.512 standard deviations. The standard deviation for album sales is 80,699 and so this constitutes a change of 41,320 sales (0.512 × 80,699). Therefore, if Radio 1 plays the song an extra 12.27 times in the week before release, 41,320 extra album sales can be expected. This interpretation is true only if the effects of attractiveness of the band and advertising are held constant.

• **Attractiveness** (*standardized β* = .192): This value indicates that a band rated one standard deviation (1.40 units) higher on the attractiveness scale can expect additional album sales of 0.192 standard deviations units. This constitutes a change of 15,490 sales (0.192 × 80,699). Therefore, a band with an attractiveness rating 1.40 higher than another band can expect 15,490 additional sales. This interpretation is true only if the effects of radio airplay and advertising are held constant.

SELF-TEST

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

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

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

 CRAMMING SAM'S TIPS Model parameters

• The individual contribution of variables to the regression model can be found in the *Coefficients* table from SPSS. If you have done a hierarchical regression then look at the values for the final model.

• For each predictor variable, you can see if it has made a significant contribution to predicting the outcome by looking at the column labelled *Sig.* (values less than .05 are significant).

• The standardized beta values tell you the importance of each predictor (bigger absolute value = more important).

• The tolerance and VIF values will also come in handy later on, so make a note of them.

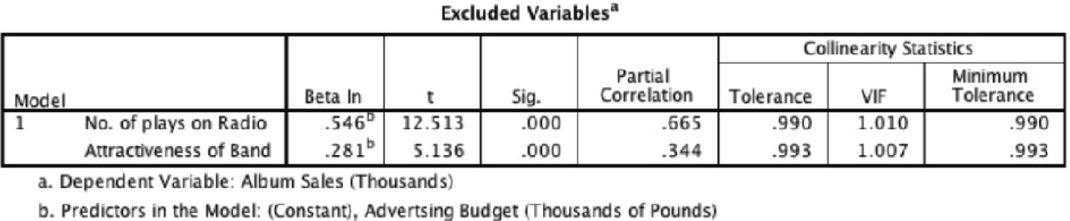
17 To spare your eyesight I have split this part of the output into two tables; however, it should appear as one long table in the SPSS viewer.

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

8.7.4. Excluded variables 

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

OUTPUT 8.8



8.7.5. Assessing multicollinearity 

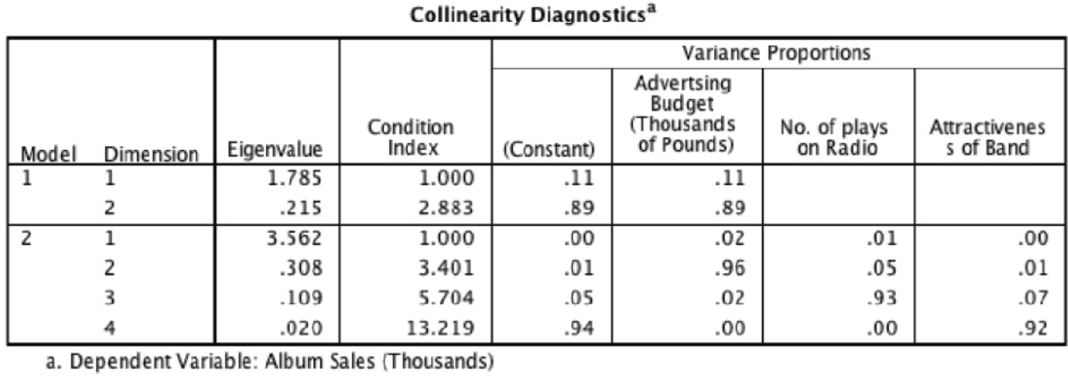
Output 8.7 provided some measures of whether there is collinearity in the data. Specifically, it provided the VIF and tolerance statistics (with tolerance being 1 divided by the VIF). We can apply the guidelines from Section 8.5.3 to our model. The VIF values are all well below 10 and the tolerance statistics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



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

SPSS also produces a table of eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions. There is a lengthy discussion, and example, of collinearity in Section 19.8.2 and how to detect it using variance proportions, so I will limit myself now to saying that we are looking for large variance proportions on the same *small* eigenvalues (Jane Superbrain Box 8.3). Therefore, in Output 8.9 we look at the bottom few rows of the table (these are the small eigenvalues) and look for any variables that both have high variance proportions for that eigenvalue. The variance proportions vary between 0 and 1, and for each predictor should be distributed across different dimensions (or eigenvalues). For this model, you can see that each predictor has most of its variance loading onto a different dimension (advertising has 96% of variance on dimension 2, airplay has 93% of variance on dimension 3 and attractiveness has 92% of variance on dimension 4). These data represent a classic example of no multicollinearity. For an example of when collinearity exists in the data and some suggestions about what can be done, see Chapters 19 (Section 19.8.2) and 17 (Section 17.3.3.3).

OUTPUT 8.9



 CRAMMING SAM'S TIPS Multicollinearity

• To check for multicollinearity, use the VIF values from the table labelled *Coefficients* in the SPSS output.

• If these values are less than 10, then there probably isn't cause for concern.

• If you take the average of VIF values, and it is not substantially greater than 1, then there's also no cause for concern.

 JANE SUPERBRAIN 8.3 *What are eigenvectors and eigenvalues?* 

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

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

FIGURE 8.20 A scatterplot of two variables forms an ellipse

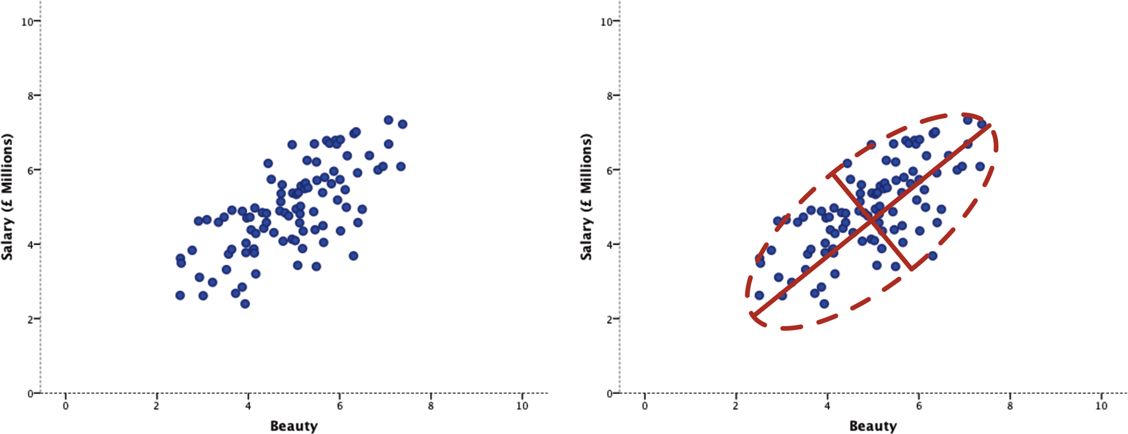
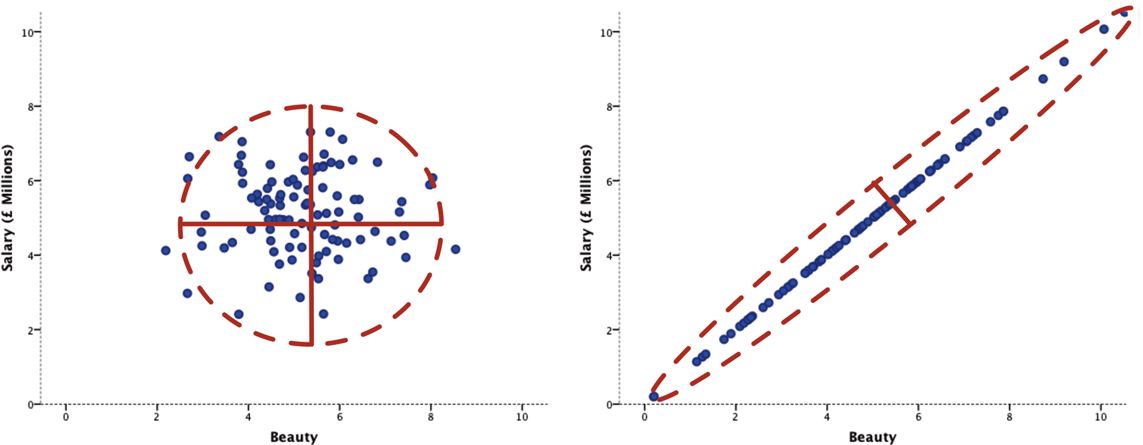


FIGURE 8.21 Perfectly uncorrelated (left) and correlated (right) variables

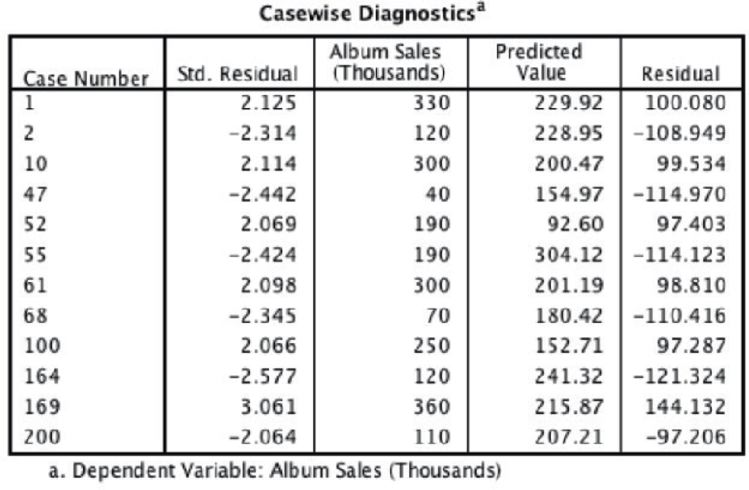


In the case of two variables, the *condition* of the data is related to the ratio of the larger eigenvalue to the smaller. Figure 8.21 shows the two extremes: when there is no relationship at all between variables (left), and when there is a perfect relationship (right). When there is no relationship,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

8.7.6. Bias in the model: casewise diagnostics 

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

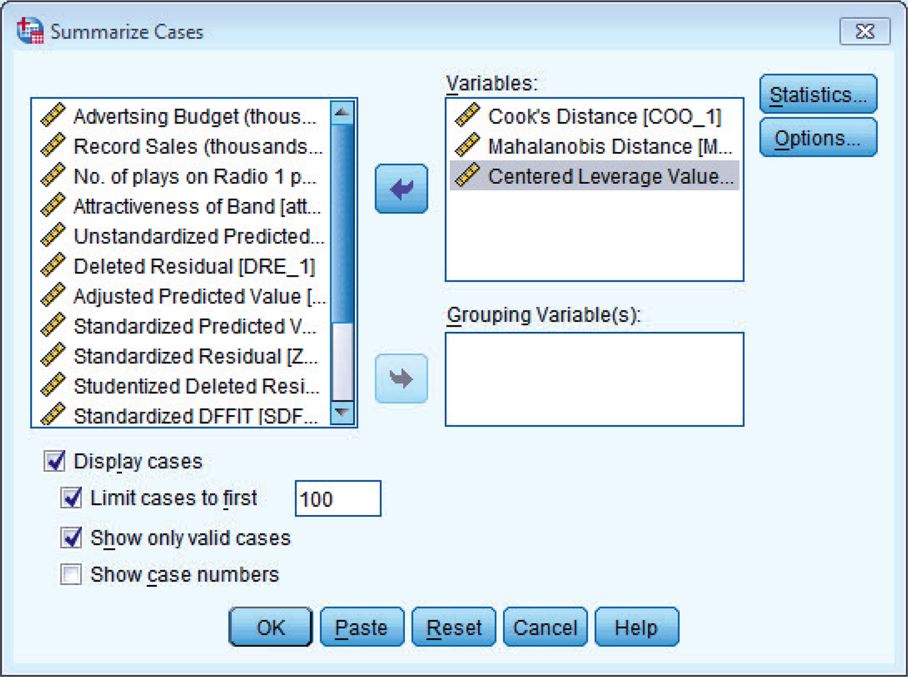
OUTPUT 8.10



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

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

FIGURE 8.22 The *Summarize Cases* dialog box



 SPSS TIP 8.1 Selecting cases 

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

USE ALL.

COMPUTE cook\_problem = (COO\_1 > 1).

VARIABLE LABELS cook\_problem ‘Cooks distance greater than 1'.

VALUE LABELS cook\_problem 0 ‘Not Selected’ 1 ‘Selected'.

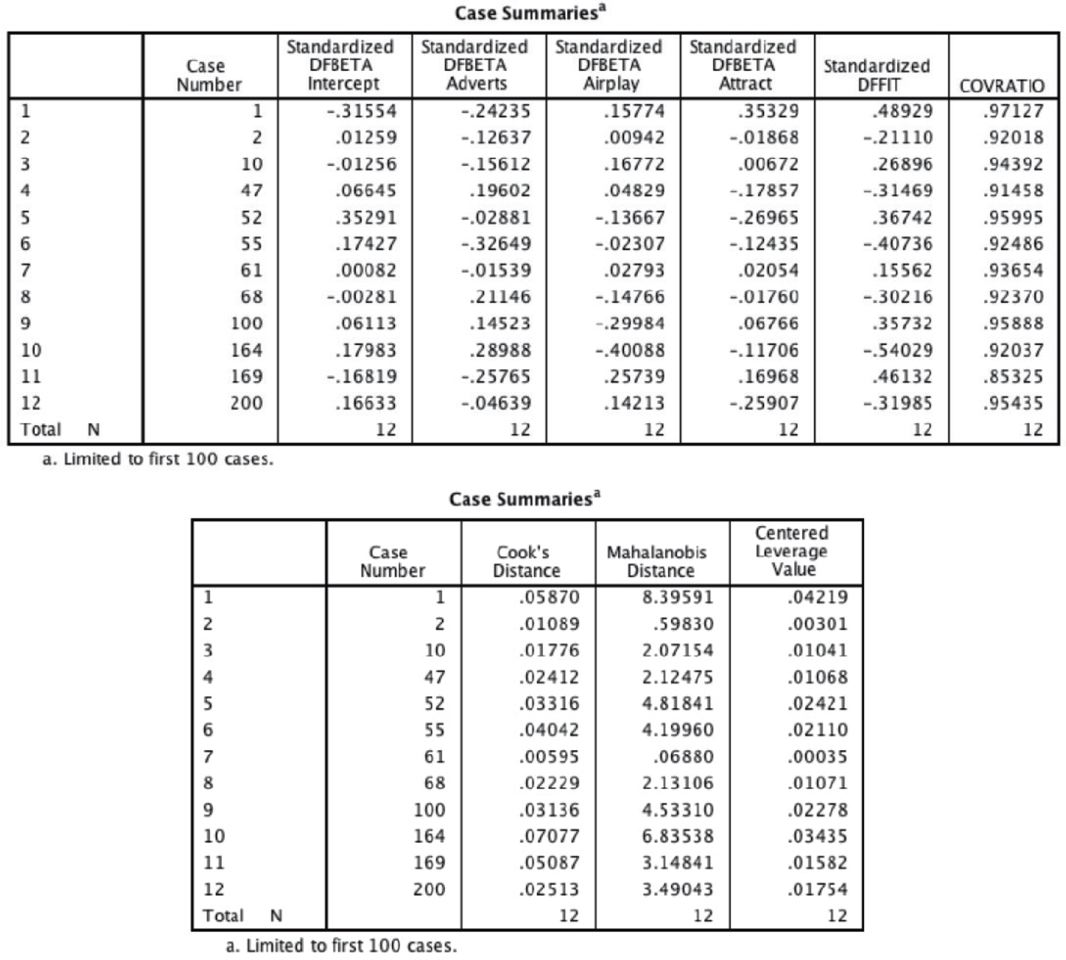
FILTER BY cook\_problem.

EXECUTE.

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

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

OUTPUT 8.11



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

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

• CVR*i* > 1 + [3(*k* + 1)/*n*] = 1 + [3(3 + 1)/200] = 1.06,

• CVR*i* < 1 – [3(*k* + 1)/*n*] = 1 – [3(3 + 1)/200] = 0.94.

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

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

 CRAMMING SAM'S TIPS Residuals

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

• Look at standardized residuals and check that no more than 5% of cases have absolute values above 2, and that no more than about 1% have absolute values above 2.5. Any case with a value above about 3 could be an outlier.

• Look in the data editor for the values of Cook's distance: any value above 1 indicates a case that might be influencing the model.

• Calculate the average leverage (the number of predictors plus 1, divided by the sample size) and then look for values greater than twice or three times this average value.

• For Mahalanobis distance, a crude check is to look for values above 25 in large samples (500) and values above 15 in smaller samples (100). However, Barnett and Lewis (1978) should be consulted for more detailed analysis.

• Look for absolute values of DFBeta greater than 1.

• Calculate the upper and lower limit of acceptable values for the covariance ratio, CVR. The upper limit is 1 plus three times the average leverage, while the lower limit is 1 minus three times the average leverage. Cases that have a CVR that falls outside these limits may be problematic.

8.7.7. Bias in the model: assumptions 

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

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

FIGURE 8.23 Plot of standardized predicted values against standardized residuals (top left), and partial plots of album sales against advertising (top right), airplay (bottom left) and attractiveness of the band (bottom right)

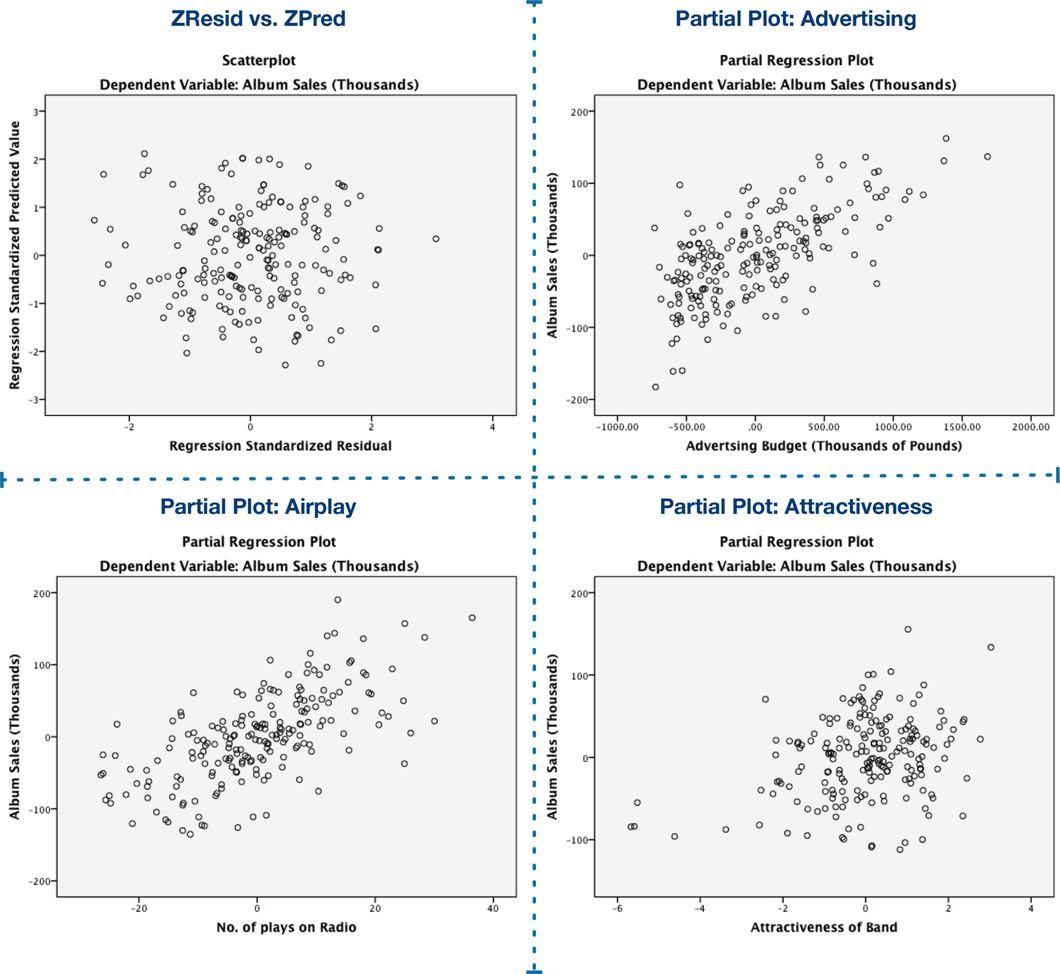
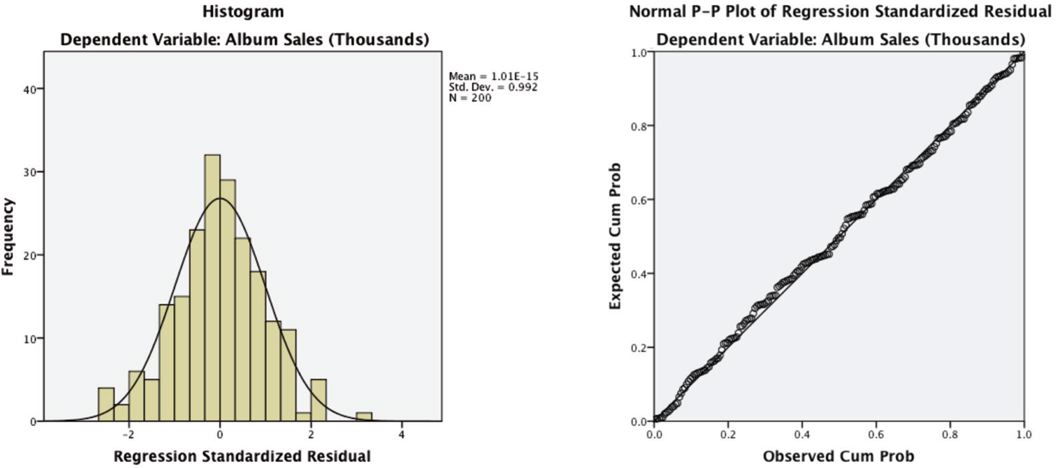


FIGURE 8.24 Histograms and normal P-P plots of normally distributed residuals (left-hand side) and non-normally distributed residuals (right-hand side)



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

 CRAMMING SAM'S TIPS Model assumptions

• Look at the graph of ZRESID\* plotted against ZPRED\*. If it looks like a random array of dots then this is good. If the dots seem to get more or less spread out over the graph (look like a funnel) then this is probably a violation of the assumption of homogeneity of variance. If the dots have a pattern to them (i.e., a curved shape) then this is probably a violation of the assumption of linearity. If the dots seem to have a pattern and are more spread out at some points on the plot than others then this probably reflects violations of both homogeneity of variance *and* linearity. Any of these scenarios puts the validity of your model into question. Repeat the above for all partial plots too.

• Look at histograms and P-P plots. If the histograms look like normal distributions (and the P-P plot looks like a diagonal line), then all is well. If the histogram looks non-normal and the P-P plot looks like a wiggly snake curving around a diagonal line then things are less good. Be warned, though: distributions can look very non-normal in small samples even when they are normal.

8.8. What if I violate an assumption? Robust regression 

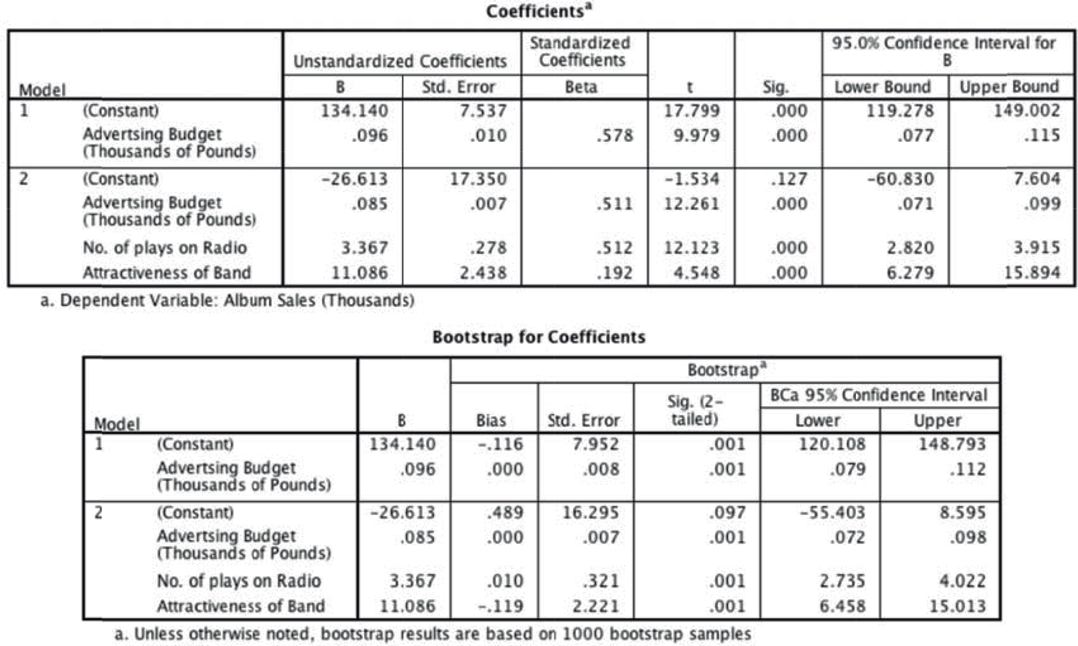
We could summarize by saying that our model appears, in most senses, to be both accurate for the sample and generalizable to the population. The only slight glitch is some concern over whether attractiveness ratings had violated the assumption of homoscedasticity. Therefore, we could conclude that in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

 LABCOAT LENI'S REAL RESEARCH 8.1 *I want to be loved (on Facebook)* 

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

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

OUTPUT 8.12



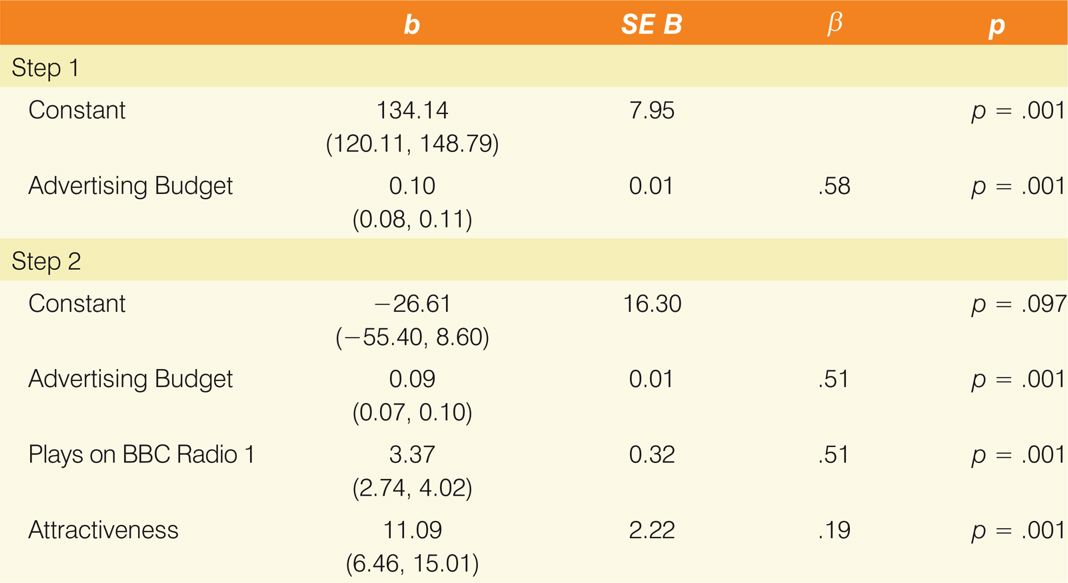
19 Remember that because of how bootstrapping works the values in your output will be slightly different than mine, and different again if you rerun the analysis.

8.9. How to report multiple regression 

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

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

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



*Note. R*2 = .34 for Step 1; *AR*2 = .33 for Step 2 (*p*s < .001).

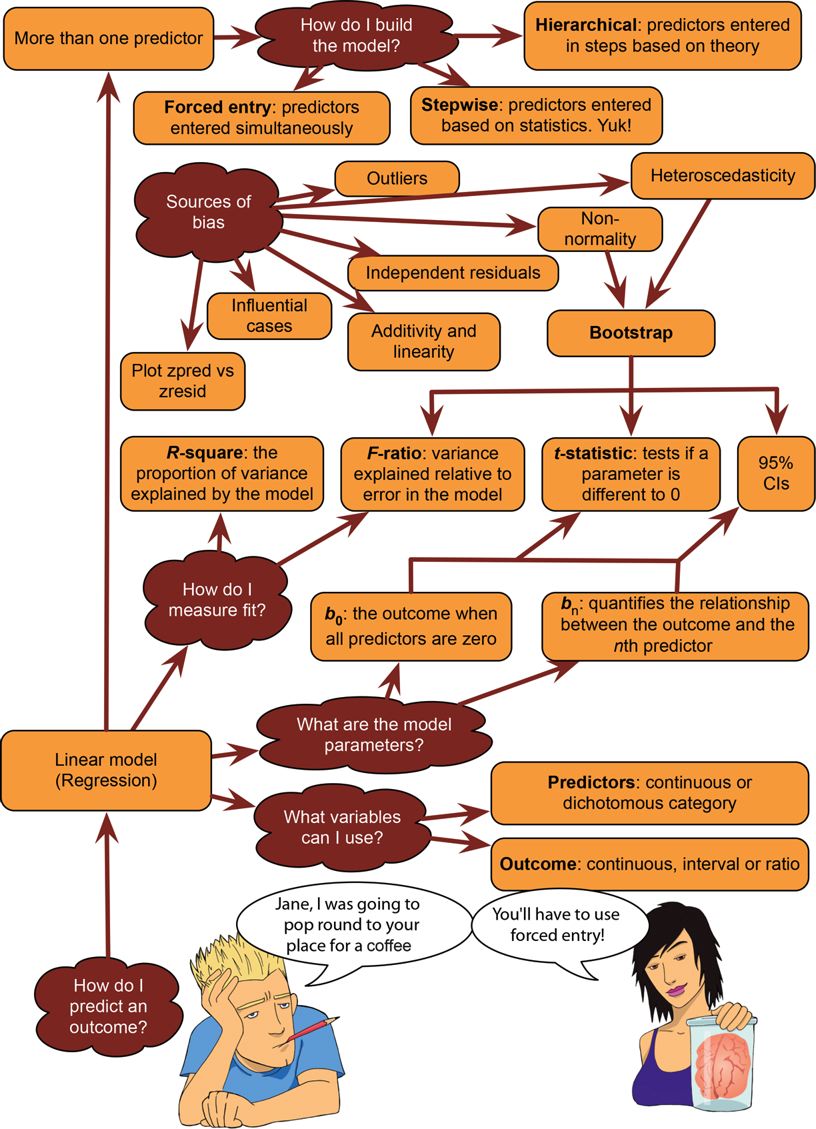
 LABCOAT LENI'S REAL RESEARCH 8.2 *Why do you like your lecturers?* 

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

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

8.10. Brian's attempt to woo Jane 

FIGURE 8.25 What Brian learnt from this chapter



8.11. 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 ever need to know about statistics. Really, it's true; you'll discover in the coming chapters that everything else we discuss is basically a variation of this chapter. So, although you may be near death having spent your life reading this chapter (and I'm certainly near death having written it) you are officially a stats genius – well done!

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

8.12. Key terms that I've discovered

**Adjusted predicted value**

**Adjusted *R*2**

**Autocorrelation**

***bi***

***βi***

**Cook's distance**

**Covariance ratio (CVR)**

**Cross-validation**

**Deleted residual**

**DFBeta**

**DFFit**

**Dummy variables**

**Durbin–Watson test**

***F*-ratio**

**Generalization**

**Goodness of fit**

**Hat values**

**Heteroscedasticity**

**Hierarchical regression**

**Homoscedasticity**

**Independent errors**

**Leverage**

**Mahalanobis distances**

**Mean squares**

**Model sum of squares**

**Multicollinearity**

**Multiple *r***

**Multiple regression**

**Ordinary least squares (OLS)**

**Outcome variable**

**Perfect collinearity**

**Predicted value**

**Predictor variable**

**Residual**

**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)**