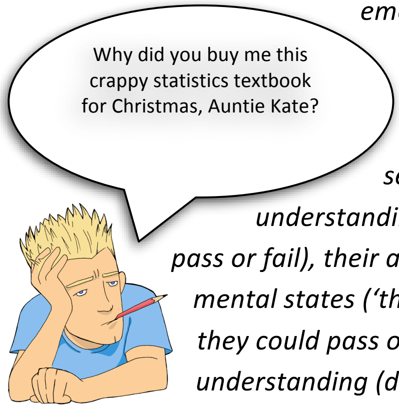
Chapter 19: Logistic regression

Smart Alex's Solutions

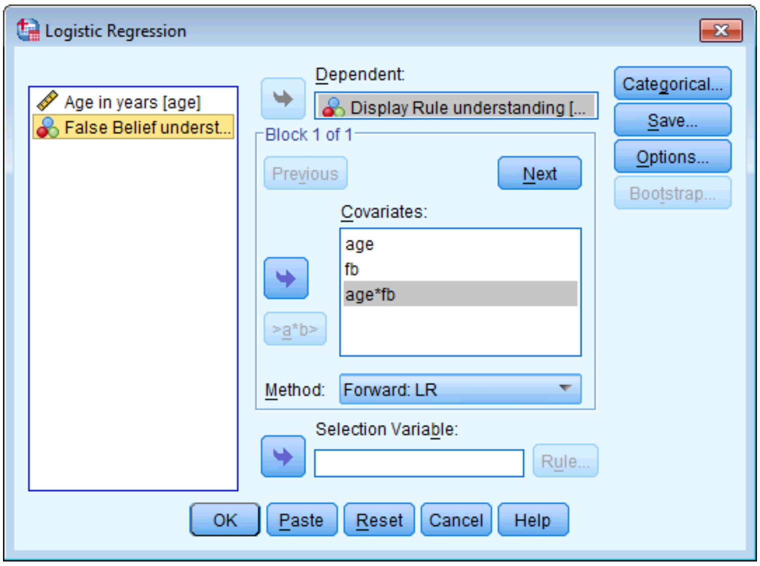
Task 1


Why did you buy me this crappy statistics textbook for Christmas, Auntie Kate?

A 'display rule' refers to displaying an appropriate emotion in a given situation. For example, if you receive a Christmas present that you don't like, the appropriate emotional display is to smile politely and say 'Thank you Auntie Kate, I've always wanted a rotting cabbage'. The inappropriate emotional display is to start crying and scream 'Why did you buy me a rotting cabbage, you selfish old bag?' A psychologist measured children's understanding of display rules (with a task that they could either pass or fail), their age (months), and their ability to understand others' mental states ('theory of mind', measured with a false-belief task that they could pass or fail). The data are in ***Display.sav****. Can display rule understanding (did the child pass the test: yes/no?) be predicted from the theory of mind (did the child pass the false-belief task: yes/no?), age and their interaction?*

The main analysis

To carry out logistic regression, the data must be entered as for normal regression: they are arranged in the data editor in three columns (one representing each variable). Open the file **Display.sav**. Looking at the data editor, you should notice that both of the categorical variables have been entered as coding variables; that is, numbers have been specified to represent categories. For ease of interpretation, the outcome variable should be coded 1 (event occurred) and 0 (event did not occur); in this case, 1 represents having display rule understanding, and 0 represents an absence of display rule understanding. For the false-belief task a similar coding has been used (1 = passed the false-belief task, 2 = failed the false-belief task). Logistic regression is accessed by selecting Analyze Regression ▸ Binary Logistic... . Following this menu path activates the main *Logistic Regression* dialog box shown in Figure 1.



Logistic Regression

Age in years [age]
False Belief underst...

Dependent: Display Rule understanding [...

Block 1 of 1

Previous | Next

Covariates:
age
fb
age*fb

>a*b>

Method: Forward: LR

Selection Variable:

Rule...

Categorical... | Save... | Options... | Bootstrap...

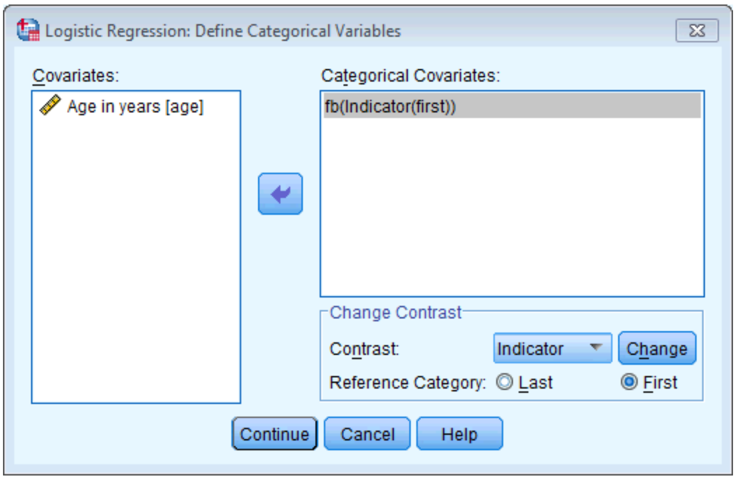
OK | Paste | Reset | Cancel | Help

Figure 1

The main dialog box is very similar to the standard regression option box. There is a space to place a dependent (or outcome) variable. In this example, the outcome was the display rule task, so we can simply click on this and transfer it to the *Dependent* box by clicking on ➡. There is also a box for specifying the covariates (the predictor variables). It is possible to specify both main effects and interactions in logistic regression. To specify a main effect, simply select one predictor (e.g. age) and then transfer this variable to the *Covariates* box by clicking on ➡. To input an interaction, click on more than one variable on the left-hand side of the dialog box (i.e. highlight two or more variables) and then click on >a*b> to move them to the *Covariates* box.

For this analysis select a *Forward:LR* method of regression.

In this example there is one categorical predictor variable. One of the great things about logistic regression is that it is quite happy to accept categorical predictors. However, it is necessary to 'tell' SPSS which variables, if any, are categorical by clicking on Categorical... in the main *Logistic Regression* dialog box to activate the dialog box in Figure 2.



Logistic Regression: Define Categorical Variables

Covariates:
Age in years [age]

Categorical Covariates:
fb(Indicator(first))

Change Contrast

Contrast: Indicator | Change

Reference Category: ○ Last ● First

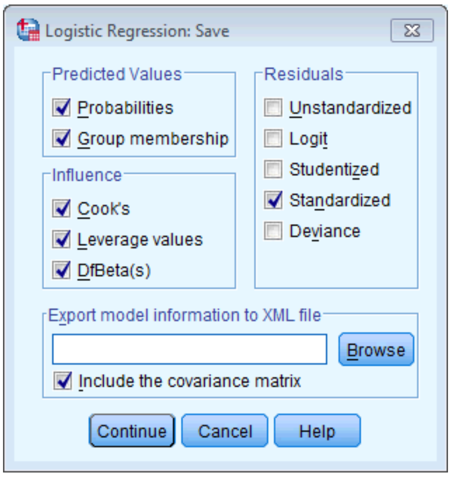
Continue | Cancel | Help

Figure 2

The covariates are listed on the left-hand side, and there is a space on the right-hand side in which categorical covariates can be placed. Simply highlight any categorical variables you have (in this example click on **fb**) and transfer them to the *Categorical Covariates* box by clicking on ➔. There are many ways in which you can treat categorical predictors. They could be incorporated into regression by recoding them using zeros and ones (known as dummy coding). Now, actually, there are different ways you can arrange this coding depending on what you want to compare, and SPSS has several 'standard' ways built into it that you can select. By default SPSS uses indicator coding, which is the standard dummy variable coding that I explained in Chapter 7 (and you can choose to have either the first or last category as your baseline). To change to a different kind of contrast click on the down arrow in the *Change Contrast* box. Select *Indicator* coding (first).

Obtaining residuals

To save residuals click on Save... in the main *Logistic Regression* dialog box. SPSS saves each of the selected variables into the data editor. The dialog box in Figure 3 gives several options, and most of these are the same as those in multiple regression. Select all of the available options, or as a bare minimum select the same options as shown in the figure.



Logistic Regression: Save

Predicted Values
- ☑ Probabilities
- ☑ Group membership

Influence
- ☑ Cook's
- ☑ Leverage values
- ☑ DfBeta(s)

Residuals
- □ Unstandardized
- □ Logit
- □ Studentized
- ☑ Standardized
- □ Deviance

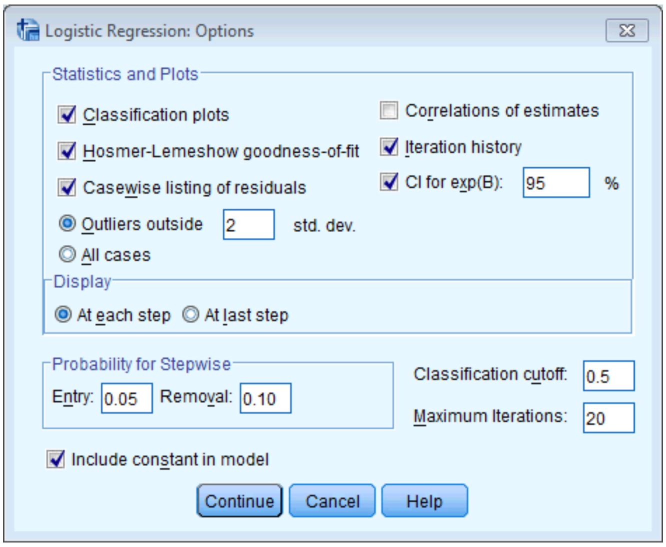
Export model information to XML file

☑ Include the covariance matrix

Figure 3

Further options

There is a final dialog box that offers further options. This box is accessed by clicking on Options... in the main *Logistic Regression* dialog box. For the most part, the default settings in this dialog box are fine. These options are explained in the chapter and so just make the selections shown in Figure 4.



Logistic Regression: Options

Statistics and Plots

☑ Classification plots

☑ Hosmer-Lemeshow goodness-of-fit

☑ Casewise listing of residuals

● Outliers outside 2 std. dev.

○ All cases

□ Correlations of estimates

☑ Iteration history

☑ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10

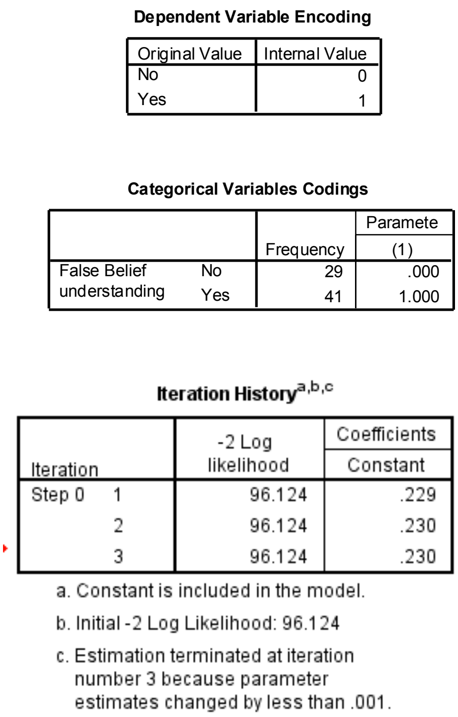
Classification cutoff: 0.5

Maximum Iterations: 20

☑ Include constant in model

Figure 4

Interpreting the output



Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Categorical Variables Codings

			Paramete
		Frequency	(1)
False Belief understanding	No	29	.000
	Yes	41	1.000

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	96.124	.229
	2	96.124	.230
	3	96.124	.230

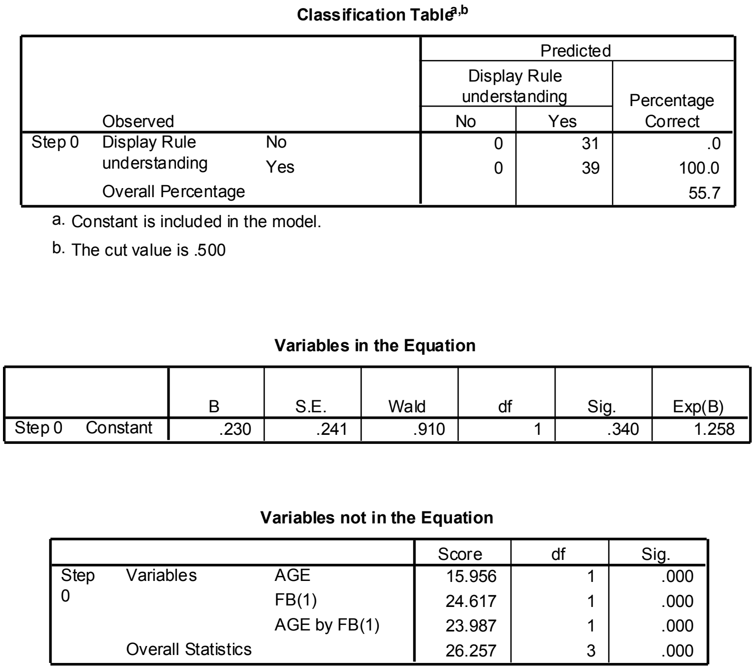
a. Constant is included in the model.

b. Initial -2 Log Likelihood: 96.124

c. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Output 1

These tables tell us the parameter codings given to the categorical predictor variable. Indicator coding was chosen with two categories, and so the coding is the same as the values in the data editor.



Classification Table[a,b]

			Predicted		
			Display Rule understanding		Percentage Correct
	Observed		No	Yes	
Step 0	Display Rule understanding	No	0	31	.0
		Yes	0	39	100.0
	Overall Percentage				55.7

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.230	.241	.910	1	.340	1.258

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	AGE	15.956	1	.000
		FB(1)	24.617	1	.000
		AGE by FB(1)	23.987	1	.000
	Overall Statistics		26.257	3	.000

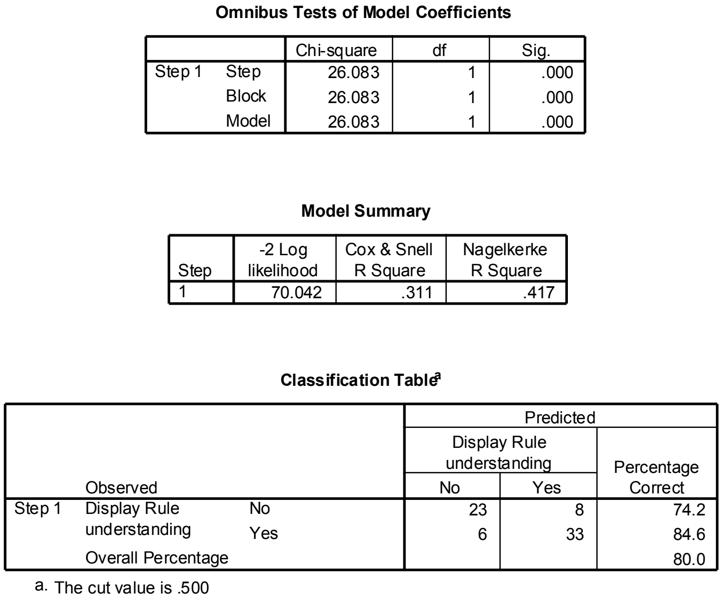
Output 1

For this first analysis we requested a forward stepwise method and so the initial model is derived using only the constant in the regression equation. The above output tells us about the model when only the constant is included (i.e. all predictor variables are omitted). The log-likelihood of this baseline model is 96.124. This represents the fit of the model when the most basic model is fitted to the data. When including only the constant, the computer bases the model on assigning every participant to a single category of the outcome variable. In this example, SPSS can decide either to predict that every child has display rule understanding, or to predict that all children do not have display rule understanding. It could make this decision arbitrarily, but because it is crucial to try to maximize how well the model predicts the observed data SPSS will predict that every child belongs to the category in which most observed cases fell. In this example there were 39 children who had display rule understanding and only 31 who did not. Therefore, if SPSS predicts that every child has display rule understanding then this prediction will be correct 39 times out of 70 (55.7%). However, if SPSS predicted that every child did not have display rule understanding, then this prediction would be correct only 31 times out of 70 (44.3%). As such, of the two available options it is better to predict that all children had display rule understanding because this results in a greater number of correct predictions. The output shows a contingency table for the model in this basic state. You can see that SPSS has predicted that all children have display rule understanding, which results in 0% accuracy for the children who were observed to have no display rule understanding, and 100% accuracy for those children observed to have passed the display rule task. Overall, the model correctly classifies 55.7% of children. The next part of the output summarizes the model, and at this stage this entails quoting the value of the constant (b_0), which is equal to 0.23.

The final table of the output is labelled *Variables not in the Equation*. The bottom line of this table reports the residual chi-square statistic as 26.257, which is significant at $p < .001$ (it labels this statistic Overall *Statistics*). This statistic tells us that the coefficients for the variables not in the model are significantly different from zero – in other words, that the addition of one or more of these variables to the model will significantly affect its predictive power. If the probability for the residual chi-square had been greater than .05 it would have meant that none of the variables excluded from the model could make a significant contribution to the predictive power of the model. As such, the analysis would have terminated at this stage.

The remainder of this table lists each of the predictors in turn with a value of Roa's efficient score statistic for each one (column labelled *Score*). In large samples when the null hypothesis is true, the score statistic is identical to the Wald statistic and the likelihood ratio statistic. It is used at this stage of the analysis because it is computationally less intensive than the Wald statistic and so can still be calculated in situations when the Wald statistic would prove prohibitive. Like any test statistic, Roa's score statistic has a specific distribution from which statistical significance can be obtained. In this example, all excluded variables have significant score statistics at $p < .001$ and so all three could potentially make a contribution to the model. The stepwise calculations are relative and so the variable that will be selected for inclusion is the one with the highest value for the score statistic that is significant at the .05 level. In this example, that variable will be **fb** because it has the highest value of the score statistic. The next part of the output deals with the model after this predictor has been added.

In the first step, false-belief understanding (**fb**) is added to the model as a predictor. As such a child is now classified as having display rule understanding based on whether they passed or failed the false-belief task.



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	26.083	1	.000
	Block	26.083	1	.000
	Model	26.083	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	70.042	.311	.417

Classification Table[a]

			Predicted		
			Display Rule understanding		Percentage Correct
	Observed		No	Yes	
Step 1	Display Rule understanding	No	23	8	74.2
		Yes	6	33	84.6
	Overall Percentage				80.0

a. The cut value is .500

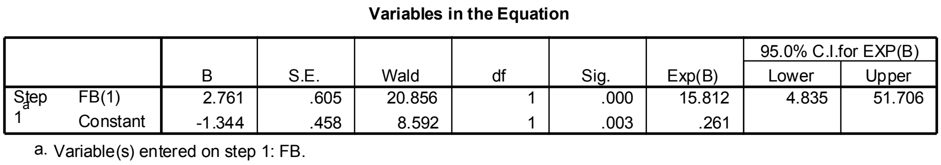
Output 3

Output 3 shows summary statistics about the new model (which we've already seen contains **fb**). The overall fit of the new model is assessed using the log-likelihood statistic. In SPSS, rather than reporting the log-likelihood itself, the value is multiplied by –2 (and sometimes referred to as –2*LL*): this multiplication is done because –2*LL* has an approximately chi-square distribution and so makes it possible to compare values against those that we might expect to get by chance alone. Remember that large values of the log-likelihood statistic indicate poorly fitting statistical models.

At this stage of the analysis the value of –2*LL* should be less than the value when only the constant was included in the model (because lower values of –2*LL* indicate that the model is predicting the outcome variable more accurately). When only the constant was included, –2*LL* = 96.124, but now **fb** has been included this value has been reduced to 70.042. This reduction tells us that the model is better at predicting display rule understanding than it was before **fb** was added. The question of how much better the model predicts the outcome variable can be assessed using the model chi-square statistic, which measures the difference between the model as it currently stands and the model when only the constant was included. We can assess the significance of the change in a model by taking the log-likelihood of the new model and subtracting the log-likelihood of the baseline model from it. The value of the model chi-square statistic works on this principle and is, therefore, equal to –2*LL* with **fb** included minus the value of –2*LL* when only the constant was in the model (96.124 – 70.042 = 26.083). This value has a chi-square distribution and so its statistical significance can be easily calculated. In this example, the value is significant at the .05 level and so we can say that overall the model is predicting display rule understanding significantly better than it was with only the constant included. The model chi-square is an analogue of the *F*-test for the linear regression sum of squares. In an ideal world we would like to see a non-significant –2*LL* (indicating that the amount of unexplained data is minimal) and a highly significant model chi-square statistic (indicating that the model including the predictors is significantly better than without those predictors). However, in reality it is possible for both statistics to be highly significant.

There is a second statistic called the step statistic that indicates the improvement in the predictive power of the model since the last stage. At this stage there has been only one step in the analysis, and so the value of the improvement statistic is the same as the model chi-square. However, in more complex models in which there are three or four stages, this statistic gives you a measure of the improvement of the predictive power of the model since the last step. Its value is equal to –2*LL* at the current step minus –2*LL* at the previous step. If the improvement statistic is significant then it indicates that the model now predicts the outcome significantly better than it did at the last step, and in a forward regression this can be taken as an indication of the contribution of a predictor to the predictive power of the model. Similarly, the block statistic provides the change in –2*LL* since the last block (for use in hierarchical or blockwise analyses).

Finally, the classification table at the end of this section of the output indicates how well the model predicts group membership. The current model correctly classifies 23 children who don't have display rule understanding but misclassifies 8 others (i.e. it correctly classifies 74.2% of cases). For children who do have display rule understanding, the model correctly classifies 33 and misclassifies 6 cases (i.e. correctly classifies 84.6% of cases). The overall accuracy of classification is, therefore, the weighted average of these two values (80%). So, when only the constant was included, the model correctly classified 56% of children, but now, with the inclusion of **fb** as a predictor, this has risen to 80%.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	FB(1)	2.761	.605	20.856	1	.000	15.812	4.835	51.706
	Constant	-1.344	.458	8.592	1	.003	.261		

a. Variable(s) entered on step 1: FB.

Output 4

The next part of the output (Output 4) is crucial because it tells us the estimates for the coefficients for the predictors included in the model. It provides the coefficients and statistics for the variables that have been included in the model at this point (namely, **fb** and the constant). The interpretation of this coefficient in logistic regression is that it represents the change in the logit of the outcome variable associated with a one-unit change in the predictor variable. The logit of the outcome is simply the natural logarithm of the odds of *Y* occurring.

The crucial statistic is the Wald statistic, which has a chi-square distribution and tells us whether the *b* coefficient for that predictor is significantly different from zero. If the coefficient is significantly different from zero then we can assume that the predictor is making a significant contribution to the prediction of the outcome (*Y*). For these data it seems to indicate that false-belief understanding is a significant predictor of display rule understanding (note the significance of the Wald statistic is less than .05).

We can calculate an analogue of *R* using the equation in the chapter (for these data, the Wald statistic and its *df* are 20.856 and 1, respectively, and the original –2*LL* was 96.124). Therefore, *R* can be calculated as:

$$\begin{aligned} R &= \sqrt{\frac{20.856-(2\times 1)}{96.124}} \\ &= .4429 \end{aligned}$$

Hosmer and Lemeshow's measure (R_{L}^{2}) is calculated by dividing the model chi-square by the original –2*LL*. In this example the model chi-square after all variables have been entered into the model is 26.083, and the original –2*LL* (before any variables were entered) was 96.124. So R_{L}^{2} = 26.083/96.124 = .271, which is different from the value we would get by squaring the value of *R* given above ($R^2 = .4429^2 = .196$).

SPSS reports Cox and Snell's R^2 as .311. This is calculated from this equation:

$$R^2_{\mathrm{CS}} = 1 - \exp\left(\frac{(-2LL(\text{new}) - (-2LL(\text{baseline})}{n}\right)$$

SPSS reports –2*LL*(new) as 70.04 and –2*LL*(baseline) as 96.124. The sample size, *n*, is 70:

$$\begin{aligned} R^2_{CS} &= 1 - \exp\left(\frac{70.04 - 96.124}{70}\right) \\ &= 1 - \exp(-0.3726) \\ &= 1 - e^{-0.3726} \\ &= .311 \end{aligned}$$

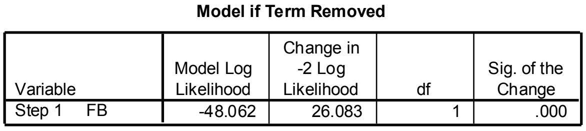
Nagelkerke's adjustment is calculated from:

$$\begin{aligned} R^2_{\mathrm{N}} &= \frac{R^2_{\mathrm{CS}}}{1 - \exp\left(-\frac{-2LL(\text{baseline})}{n}\right)} \\ &= \frac{0.311}{1 - \exp\left(-\frac{96.124}{70}\right)} \\ &= \frac{0.311}{1 - e^{-1.3732}} \\ &= \frac{0.311}{1 - 0.2533} \\ &= .416 \end{aligned}$$

As you can see, there's a fairly substantial difference between the two values!

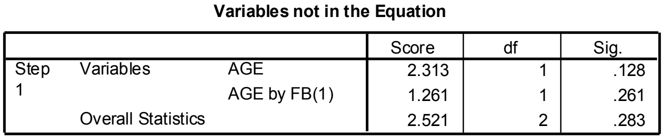
The final thing we need to look at is the odds ratio, exp(*b*) (Exp(*B*) in the SPSS output). We can interpret this in terms of the change in odds. If the value is greater than 1 then it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease. In this example, we can say that the odds of a child who has false-belief understanding also having display rule understanding are 15 times higher than those of a child who does not have false-belief understanding.

In the options, we requested a confidence interval for exp(*b*) and it can also be found in Output 4. The way to interpret this confidence interval is to say that if we ran 100 experiments and calculated confidence intervals for the value of exp(*b*), then these intervals would encompass the actual value of exp(*b*) in the population (rather than the sample) on 95 occasions. So, in this case, we can be fairly confident that the population value of exp(*b*) lies between 4.84 and 51.71. However, there is a 5% chance that a sample could give a confidence interval that 'misses' the true value.



Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	FB	-48.062	26.083	1	.000



Variables not in the Equation

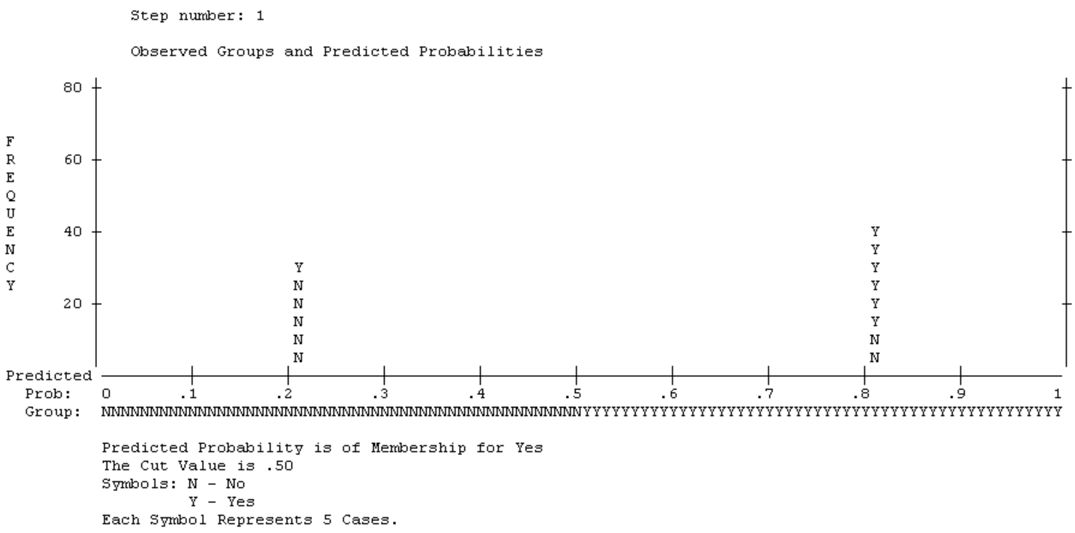
			Score	df	Sig.
Step 1	Variables	AGE	2.313	1	.128
		AGE by FB(1)	1.261	1	.261
	Overall Statistics		2.521	2	.283

Output 5

The test statistics for **fb** if it were removed from the model are reported in Output 5. The important thing to note is the significance value of the log-likelihood ratio. This is highly significant for this model ($p < .001$) which tells us that removing **fb** from the model would have a significant effect on the predictive ability of the model – in other words, it would be a very bad idea to remove it!

Finally, we are told about the variables currently not in the model. First of all, the residual chi-square (labelled *Overall Statistics* in the output), which is non-significant, tells us that none of the remaining variables have coefficients significantly different from zero. Furthermore, each variable is listed with its score statistic and significance value, and for both variables their coefficients are not significantly different from zero (as can be seen from the significance values of .128 for age and .261 for the interaction of age and false-belief understanding). Therefore, no further variables will be added to the equation.

The next part of the output (Output 6) displays the classification plot that we requested in the options dialog box. This plot is a histogram of the predicted probabilities of a child passing the display rule task. If the model perfectly fits the data, then this histogram should show all of the cases for which the event has occurred on the right-hand side, and all the cases for which the event hasn't occurred on the left-hand side. In other words, all the children who passed the display rule task should appear on the right and all those who failed should appear on the left. In this example, the only significant predictor is dichotomous and so there are only two columns of cases on the plot. If the predictor is a continuous variable, the cases are spread out across many columns. As a rule of thumb, the more that the cases cluster at each end of the graph, the better. This statement is true because such a plot would show that when the outcome did actually occur (i.e., the child did pass the display rule task) the predicted probability of the event occurring is also high (i.e., close to 1). Likewise, at the other end of the plot, it would show that when the event didn't occur (i.e., when the child failed the display rule task) the predicted probability of the event occurring is also low (i.e., close to 0). This situation represents a model that is correctly predicting the observed outcome data. If, however, there are a lot of points clustered in the centre of the plot then it shows that for many cases the model is predicting a probability of .5 that the event will occur. In other words, for these cases there is little more than a 50–50 chance that the data are correctly predicted – as such the model could predict these cases just as accurately by simply tossing a coin! Also, a good model will ensure that few cases are misclassified. In this example there are two Ns on the right of the model and one Y on the left of the model; these are misclassified cases, and the fewer of these there are, the better the model.

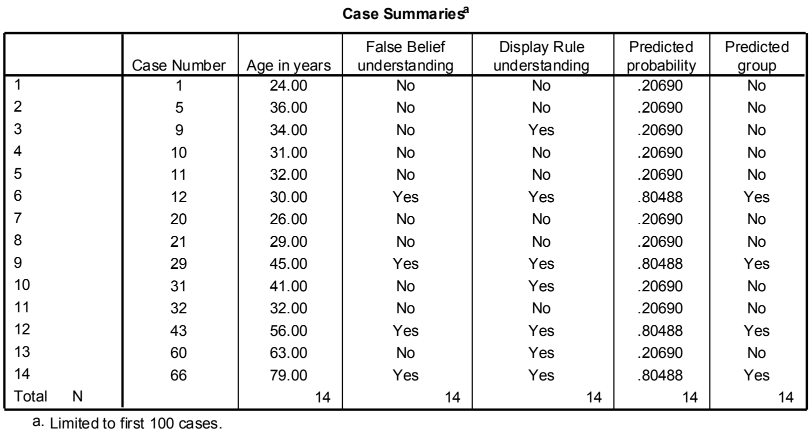


```
          Step number: 1

          Observed Groups and Predicted Probabilities

      80 +                                                                                                    +
         |                                                                                                    |
         |                                                                                                    |
F        |                                                                                                    |
R     60 +                                                                                                    +
E        |                                                                                                    |
Q        |                                                                                                    |
U        |                                                                                                    |
E     40 +                                                                                 Y                  +
N        |                                                                                 Y                  |
C        |                     Y                                                           Y                  |
Y        |                     N                                                           Y                  |
      20 +                     N                                                           Y                  +
         |                     N                                                           Y                  |
         |                     N                                                           N                  |
         |                     N                                                           N                  |
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

          Predicted Probability is of Membership for Yes
          The Cut Value is .50
          Symbols: N - No
                   Y - Yes
          Each Symbol Represents 5 Cases.
```

Output 6

Listing predicted probabilities

SPSS saved the predicted probabilities and predicted group memberships as variables in the data editor and named them PRE_1 and PGR_1 respectively. These probabilities can be listed using the *Case Summaries* dialog box (see the book chapter). Output 7 shows a selection of the predicted probabilities (because the only significant predictor was a dichotomous variable, there will be only two different probability values). It is also worth listing the predictor variables as well to clarify from where the predicted probabilities come.



Case Summaries[a]

		Case Number	Age in years	False Belief understanding	Display Rule understanding	Predicted probability	Predicted group
1		1	24.00	No	No	.20690	No
2		5	36.00	No	No	.20690	No
3		9	34.00	No	Yes	.20690	No
4		10	31.00	No	No	.20690	No
5		11	32.00	No	No	.20690	No
6		12	30.00	Yes	Yes	.80488	Yes
7		20	26.00	No	No	.20690	No
8		21	29.00	No	No	.20690	No
9		29	45.00	Yes	Yes	.80488	Yes
10		31	41.00	No	Yes	.20690	No
11		32	32.00	No	No	.20690	No
12		43	56.00	Yes	Yes	.80488	Yes
13		60	63.00	No	Yes	.20690	No
14		66	79.00	Yes	Yes	.80488	Yes
Total	N		14	14	14	14	14

a. Limited to first 100 cases.

Output 7

We found from the model that the only significant predictor of display rule understanding was false-belief understanding. This could have a value of either 1 (pass the false-belief task) or 0 (fail the false-belief task). These values tells us that when a child doesn't possess second-order false-belief understanding (**fb** = 0, No), there is a probability of .2069 that they will pass the display rule task, approximately a 21% chance (1 out of 5 children). However, if the child does pass the false-belief task (**fb** = 1, yes), there is a probability of .8049 that they will pass the display rule task, an 80.5% chance (4 out of 5 children). Consider that a probability of 0 indicates no chance of the child passing the display rule task, and a probability of 1 indicates that the child will definitely pass the display rule task. Therefore, the values obtained provide strong evidence for the role of false-belief understanding as a prerequisite for display rule understanding.

Assuming we are content that the model is accurate and that false-belief understanding has some substantive significance, then we could conclude that false-belief understanding is the single best predictor of display rule understanding. Furthermore, age and the interaction of age and false-belief understanding do not significantly predict display rule understanding. As a homework task, why not rerun this analysis using the forced entry method of analysis – how do your conclusions differ?

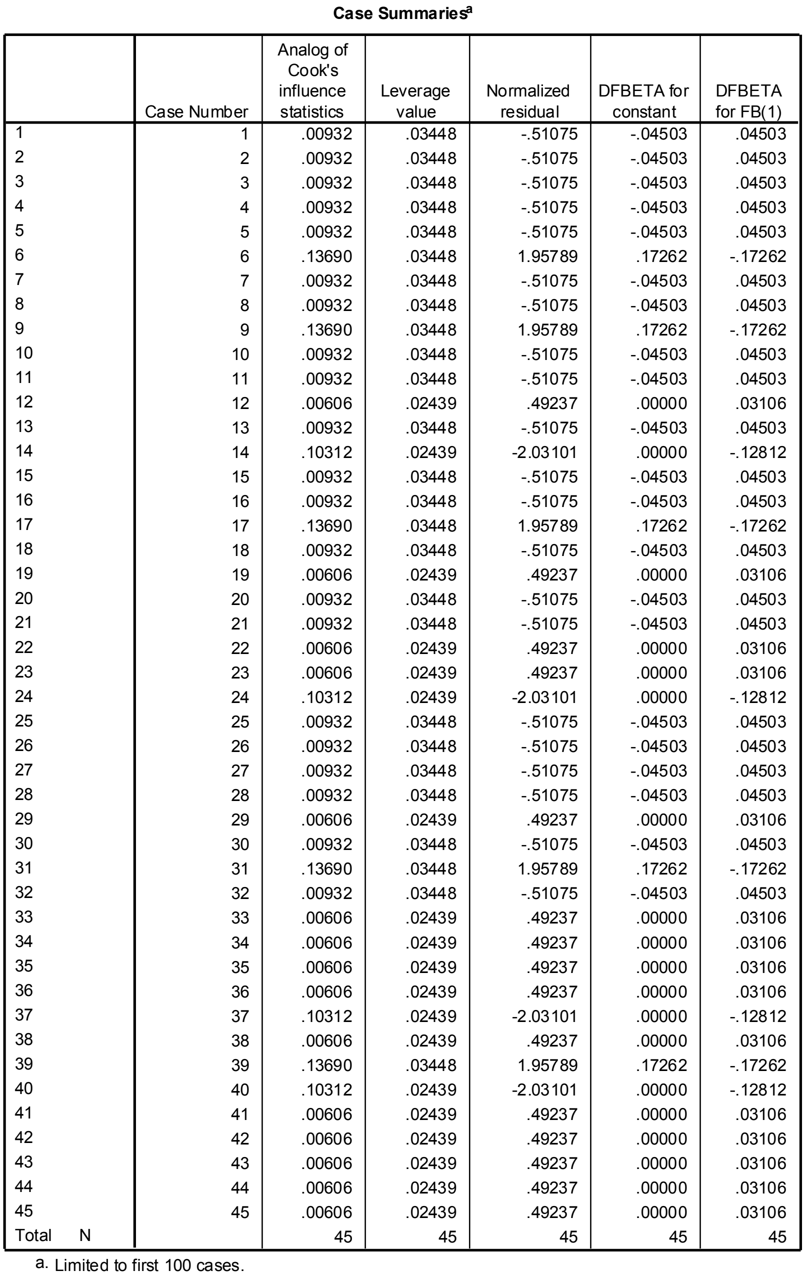
This conclusion is fine in itself, but to be sure that the model is a good one, it is important to examine the residuals.

Task 2

Are there any influential cases or outliers in the model for Task 1?

To answer this question we need to look at the model residuals. The main purpose of examining residuals in logistic regression is to (1) isolate points for which the model fits poorly, and (2) isolate points that exert an undue influence on the model. To assess the former we examine the residuals, especially the Studentized residual, standardized residual and deviance statistics. All of these statistics have the common property that 95% of cases in an average, normally distributed sample should have values which lie within ±1.96, and 99% of cases should have values that lie within ±2.58. Therefore, any values outside of ±3 are cause for concern and any outside of about ±2.5 should be examined more closely. To assess the influence of individual cases we use influence statistics such as Cook's distance (which is interpreted in the same way as for linear regression: as a measure of the change in the regression coefficient if a case is deleted from the model). Also, the value of DFBeta, which is a standardized version of Cook's statistic, tells us something of the influence of certain cases – any values greater than 1 indicate possible influential cases. Additionally, leverage statistics or hat values, which should lie between 0 (the case has no influence whatsoever) and 1 (the case exerts complete influence over the model), tell us about whether certain cases are wielding undue influence over the model. The expected value of leverage is defined as for linear regression. If you request these residual statistics, SPSS saves them in as new columns in the data editor.

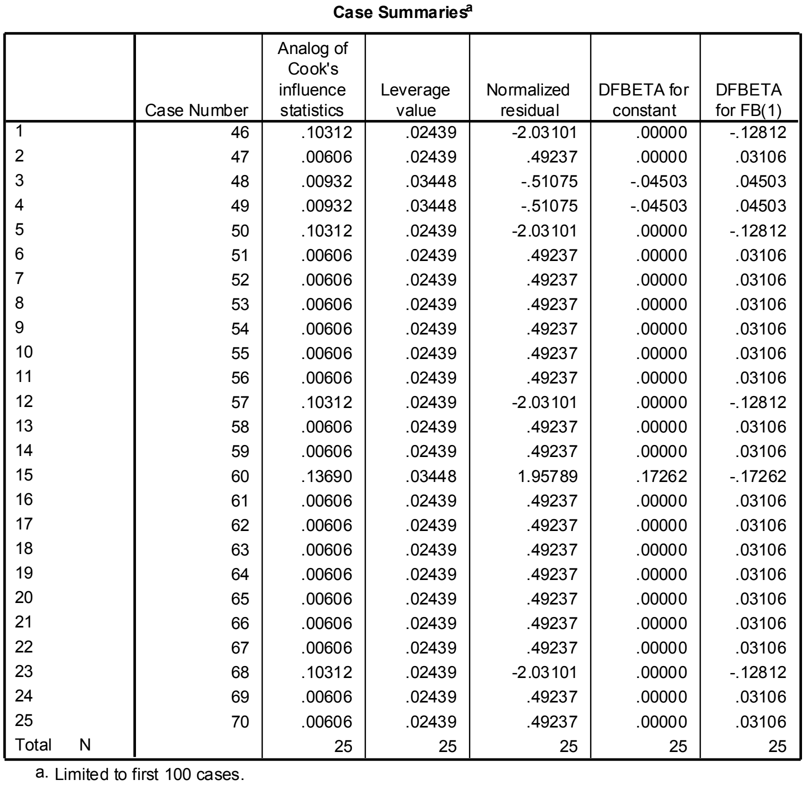
The basic residual statistics for this example (Cook's distance, leverage, standardized residuals and DFBeta values) show little cause for concern (Output 8). Note that all cases have DFBetas less than 1 and leverage statistics (LEV_1) close to the calculated expected value of 0.03. There are also no unusually high values of Cook's distance (COO_1) which, all in all, means that there are no influential cases having an effect on the model. Cook's distance is an unstandardized measure and so there is no absolute value at which you can say that a case is having an influence, Instead, you should look for values of Cook's distance which are particularly high compared to the other cases in the sample. However, Stevens (2002) suggests that a value greater than 1 is problematic. About half of the leverage values are a little high but given that the other statistics are fine, this is probably no cause for concern. The standardized residuals all have values within ±2.5 and predominantly have values within ±2, and so there seems to be very little here to concern us.



Case Summaries[a]

	Case Number	Analog of Cook's influence statistics	Leverage value	Normalized residual	DFBETA for constant	DFBETA for FB(1)
1	1	.00932	.03448	-.51075	-.04503	.04503
2	2	.00932	.03448	-.51075	-.04503	.04503
3	3	.00932	.03448	-.51075	-.04503	.04503
4	4	.00932	.03448	-.51075	-.04503	.04503
5	5	.00932	.03448	-.51075	-.04503	.04503
6	6	.13690	.03448	1.95789	.17262	-.17262
7	7	.00932	.03448	-.51075	-.04503	.04503
8	8	.00932	.03448	-.51075	-.04503	.04503
9	9	.13690	.03448	1.95789	.17262	-.17262
10	10	.00932	.03448	-.51075	-.04503	.04503
11	11	.00932	.03448	-.51075	-.04503	.04503
12	12	.00606	.02439	.49237	.00000	.03106
13	13	.00932	.03448	-.51075	-.04503	.04503
14	14	.10312	.02439	-2.03101	.00000	-.12812
15	15	.00932	.03448	-.51075	-.04503	.04503
16	16	.00932	.03448	-.51075	-.04503	.04503
17	17	.13690	.03448	1.95789	.17262	-.17262
18	18	.00932	.03448	-.51075	-.04503	.04503
19	19	.00606	.02439	.49237	.00000	.03106
20	20	.00932	.03448	-.51075	-.04503	.04503
21	21	.00932	.03448	-.51075	-.04503	.04503
22	22	.00606	.02439	.49237	.00000	.03106
23	23	.00606	.02439	.49237	.00000	.03106
24	24	.10312	.02439	-2.03101	.00000	-.12812
25	25	.00932	.03448	-.51075	-.04503	.04503
26	26	.00932	.03448	-.51075	-.04503	.04503
27	27	.00932	.03448	-.51075	-.04503	.04503
28	28	.00932	.03448	-.51075	-.04503	.04503
29	29	.00606	.02439	.49237	.00000	.03106
30	30	.00932	.03448	-.51075	-.04503	.04503
31	31	.13690	.03448	1.95789	.17262	-.17262
32	32	.00932	.03448	-.51075	-.04503	.04503
33	33	.00606	.02439	.49237	.00000	.03106
34	34	.00606	.02439	.49237	.00000	.03106
35	35	.00606	.02439	.49237	.00000	.03106
36	36	.00606	.02439	.49237	.00000	.03106
37	37	.10312	.02439	-2.03101	.00000	-.12812
38	38	.00606	.02439	.49237	.00000	.03106
39	39	.13690	.03448	1.95789	.17262	-.17262
40	40	.10312	.02439	-2.03101	.00000	-.12812
41	41	.00606	.02439	.49237	.00000	.03106
42	42	.00606	.02439	.49237	.00000	.03106
43	43	.00606	.02439	.49237	.00000	.03106
44	44	.00606	.02439	.49237	.00000	.03106
45	45	.00606	.02439	.49237	.00000	.03106
Total N		45	45	45	45	45

a. Limited to first 100 cases.



Case Summaries[a]

		Case Number	Analog of Cook's influence statistics	Leverage value	Normalized residual	DFBETA for constant	DFBETA for FB(1)
1		46	.10312	.02439	-2.03101	.00000	-.12812
2		47	.00606	.02439	.49237	.00000	.03106
3		48	.00932	.03448	-.51075	-.04503	.04503
4		49	.00932	.03448	-.51075	-.04503	.04503
5		50	.10312	.02439	-2.03101	.00000	-.12812
6		51	.00606	.02439	.49237	.00000	.03106
7		52	.00606	.02439	.49237	.00000	.03106
8		53	.00606	.02439	.49237	.00000	.03106
9		54	.00606	.02439	.49237	.00000	.03106
10		55	.00606	.02439	.49237	.00000	.03106
11		56	.00606	.02439	.49237	.00000	.03106
12		57	.10312	.02439	-2.03101	.00000	-.12812
13		58	.00606	.02439	.49237	.00000	.03106
14		59	.00606	.02439	.49237	.00000	.03106
15		60	.13690	.03448	1.95789	.17262	-.17262
16		61	.00606	.02439	.49237	.00000	.03106
17		62	.00606	.02439	.49237	.00000	.03106
18		63	.00606	.02439	.49237	.00000	.03106
19		64	.00606	.02439	.49237	.00000	.03106
20		65	.00606	.02439	.49237	.00000	.03106
21		66	.00606	.02439	.49237	.00000	.03106
22		67	.00606	.02439	.49237	.00000	.03106
23		68	.10312	.02439	-2.03101	.00000	-.12812
24		69	.00606	.02439	.49237	.00000	.03106
25		70	.00606	.02439	.49237	.00000	.03106
Total	N		25	25	25	25	25

a. Limited to first 100 cases.

Output 8

You should note that these residuals are slightly unusual because they are based on a single predictor that is categorical. This is why there isn't a lot of variability in the values of the residuals. Also, if substantial outliers or influential cases had been isolated, you would not be justified in eliminating these cases to make the model fit better. Instead these cases should be inspected closely to try to isolate a good reason why they were unusual. It might simply be an error in inputting data, or it could be that the case was one which had a special reason for being unusual: for example, the child had found it hard to pay attention to the false-belief task and you had noted this at the time of the experiment. In such a case, you may have good reason to exclude the case and duly note the reasons why.

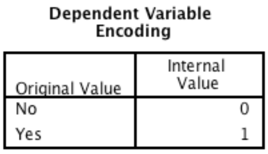
Task 3

*Piff, Stancato, Côté, Mendoza-Dentona, and Keltner (2012) showed that people of a higher social class are more unpleasant. In the first study in their paper they observed the behaviour of drivers: they classified social class by the type of car (**Vehicle**) on a 5-point scale. They then observed whether the drivers cut in front of other cars at a busy intersection (**Vehicle_Cut**). The data are in **Piff et al. (2012) Vehicle.sav**. Do a logistic regression to see whether social class predicts whether or not a driver cut in front of other vehicles.*

There is only one predictor variable (**Vehicle**) in this data set, so I selected the forced entry (*Enter*) method of regression.

Interpreting the output

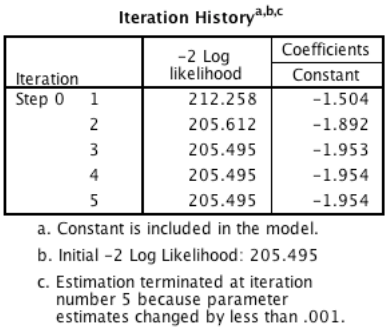
Step 0



Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Output 9



Iteration History[a,b,c]

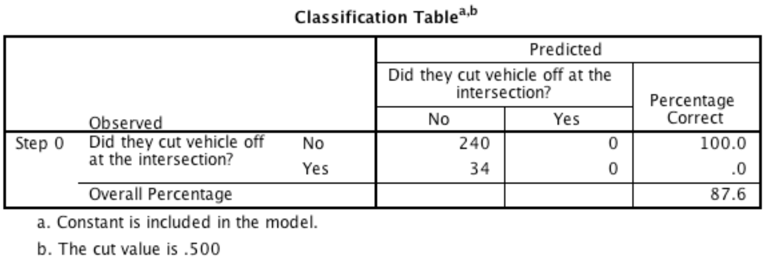
Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	212.258	-1.504
	2	205.612	-1.892
	3	205.495	-1.953
	4	205.495	-1.954
	5	205.495	-1.954

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 205.495

c. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Output 10



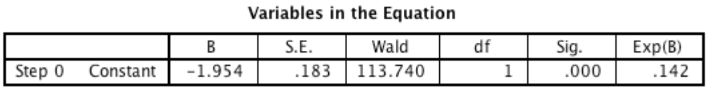
Classification Table[a,b]

Observed			Predicted		
			Did they cut vehicle off at the intersection?		Percentage Correct
			No	Yes	
Step 0	Did they cut vehicle off at the intersection?	No	240	0	100.0
		Yes	34	0	.0
	Overall Percentage				87.6

a. Constant is included in the model.

b. The cut value is .500

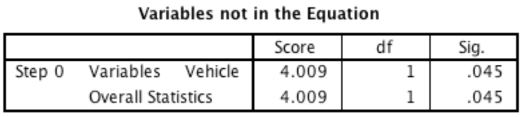
Output 11



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.954	.183	113.740	1	.000	.142

Output 12



Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Vehicle	4.009	1	.045
	Overall Statistics		4.009	1	.045

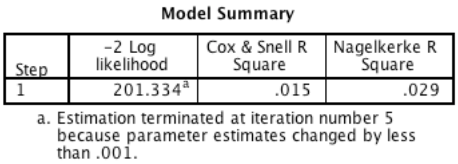
Output 13

The above output tells us about the model when only the constant is included. When including only the constant, SPSS bases the model on assigning every participant to a single category of the outcome variable. In this example, SPSS can decide either to predict that all participants cut off other vehicles at intersections, or that all participants did not cut off other vehicles at intersections. It could make this decision arbitrarily, but because it is crucial to try to maximize how well the model predicts the observed data SPSS will predict that all participants (across all status levels) belong to the category in which most observed cases fell. In this example there were 34 participants who did cut off other vehicles at intersections and 240 who did not. Therefore, if SPSS predicts that all participants did not cut off other vehicles then this prediction will be correct 240 times out of 274 (i.e. 87.6%). However, if SPSS predicted that all participants did cut off other vehicles, then this prediction would be correct only 34 times out of 274 (12.4%). As such, of the two available options it is better to predict that all participants did not cut off other vehicles because this results in a greater number of correct predictions. Output shows a contingency table for the model in this basic state. You can see that SPSS has predicted that all participants did not cut off other vehicles, which results in 0% accuracy for those who did cut off other vehicles, and 100% accuracy for those who did not. Overall, the model correctly classifies 87.6% of participants. Output summarizes the model, and at this stage this entails quoting the value of the constant (b_0), which is equal to −1.95.

The final table of the output (Output) is labelled *Variables not in the Equation*. The bottom line of this table reports the residual chi-square statistic as 4.01 which is only just significant at $p < .05$ (it labels this statistic *Overall Statistics*). This statistic tells us that the coefficient for the variable not in the model is significantly different from zero – in other words, that the addition of this variable to the model will significantly affect its predictive power. If the probability for the residual chi-square had been greater than .05 it would have meant that the variable excluded from the model could not make a significant contribution to the predictive power of the model. As such, the analysis would have terminated at this stage.

Step 1

The next part of the output deals with the model after the predictor variable (**Vehicle**) has been added to the model. As such, a person is now classified as either cutting off other vehicles at an intersection or not, based on the type of vehicle they were driving (as a measure of social status).

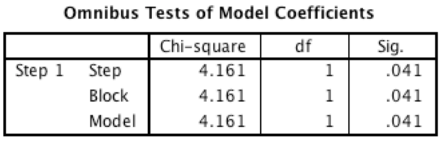


Model Summary

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	201.334[a]	.015	.029

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

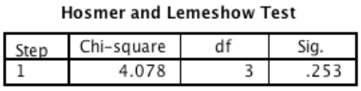
Output 2



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	4.161	1	.041
	Block	4.161	1	.041
	Model	4.161	1	.041

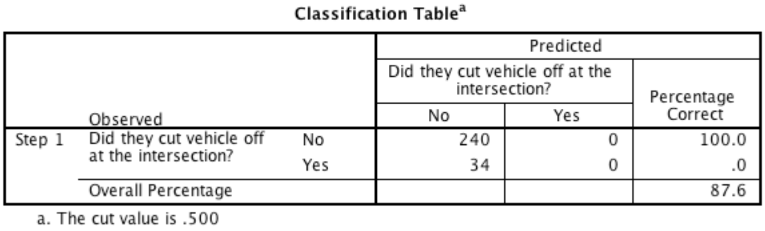
Output 3



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.078	3	.253

Output 4



Classification Table[a]

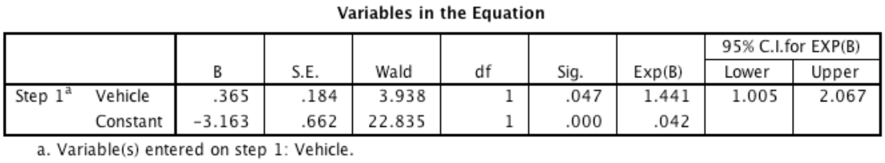
			Predicted		
			Did they cut vehicle off at the intersection?		Percentage Correct
Observed			No	Yes	
Step 1	Did they cut vehicle off at the intersection?	No	240	0	100.0
		Yes	34	0	.0
	Overall Percentage				87.6

a. The cut value is .500

Output 5

The above output shows summary statistics about the new model. The overall fit of the new model is significant because the *Model* chi-square in the table labelled *Omnibus Tests of Model Coefficients* (Output) is significant, $\chi^2(1) = 4.16$, $p = .041$. Therefore, the model that included the variable **Vehicle** predicted whether or not participants cut off other vehicles at intersections better than the model when only the constant was included.

The classification table at the end of this section of the output (Output) indicates how well the model predicts group membership. In step 1, the model correctly classifies 240 participants who did not cut off other vehicles and does not misclassify any (i.e. it correctly classifies 100% of cases). For participants who do did cut off other vehicles, the model correctly classifies 0 and misclassifies 34 cases (i.e. correctly classifies 0% of cases). The overall accuracy of classification is, therefore, the weighted average of these two values (87.6%). Therefore, the accuracy is no different than when only the constant was included in the model.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Vehicle	.365	.184	3.938	1	.047	1.441	1.005	2.067
	Constant	-3.163	.662	22.835	1	.000	.042		

a. Variable(s) entered on step 1: Vehicle.

Output 6

The significance of the Wald statistic in Output is .047, which is less than .05. Therefore, we can conclude that the status of the vehicle the participant was driving significantly predicted whether or not they cut off another vehicle at an intersection.

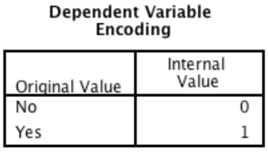
The final thing we need to look at is exp *b* (Exp(*B*) in Output). We can interpret exp *b* in terms of the change in odds. If the value is greater than 1 then it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease. In this example, the exp *b* for vehicle in step 1 is 1.441, which is greater than 1, thus indicating that as the predictor (vehicle) increases, the value of the outcome also increases, that is, the value of the categorical variable moves from 0 (did not cut off vehicle) to 1 (cut off vehicle). In other words, drivers of vehicles of a higher status were more likely to cut off other vehicles at intersections.

Task 4

*In their second study, Piff et al. (2012) again observed the behaviour of drivers and classified social class by the type of car (**Vehicle**). However, they observed whether the drivers cut off a pedestrian at a crossing (**Pedestrian_Cut**). The data are in **Piff et al. (2012) Pedestrian.sav**. Do a logistic regression to see whether social class predicts whether or not a driver prevents a pedestrian from crossing.*

As in the previous task, there is only one predictor variable (**Vehicle**), so I selected the forced entry (*Enter*) method of regression.

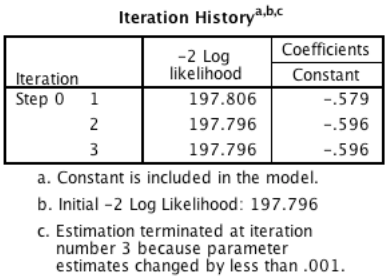
Step 0



Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Output 7



Iteration History[a,b,c]

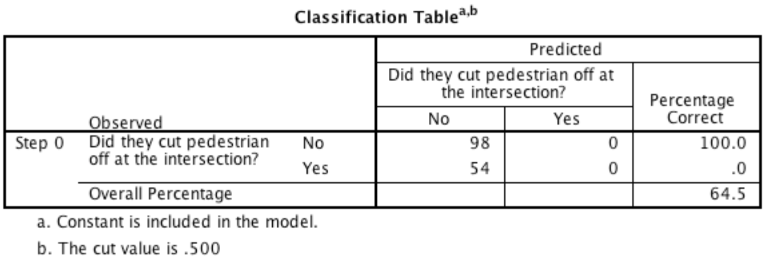
Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	197.806	-.579
	2	197.796	-.596
	3	197.796	-.596

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 197.796

c. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Output 8



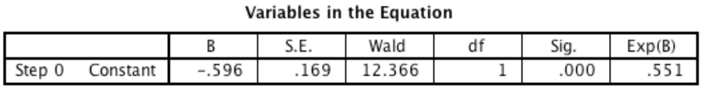
Classification Table[a,b]

Observed			Predicted		
			Did they cut pedestrian off at the intersection?		Percentage Correct
			No	Yes	
Step 0	Did they cut pedestrian off at the intersection?	No	98	0	100.0
		Yes	54	0	.0
	Overall Percentage				64.5

a. Constant is included in the model.

b. The cut value is .500

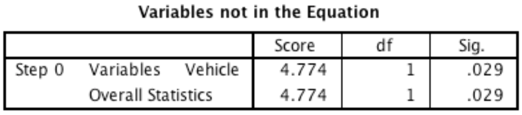
Output 9



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.596	.169	12.366	1	.000	.551

Output 10



Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Vehicle	4.774	1	.029
	Overall Statistics		4.774	1	.029

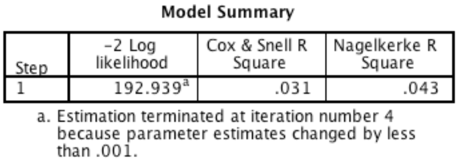
Output 11

The above output tells us about the model when only the constant is included. When including only the constant, SPSS bases the model on assigning every participant to a single category of the outcome variable. In this example there were 54 participants who did cut off pedestrians at intersections and 98 who did not. Therefore, if SPSS predicts that all participants did not cut off pedestrians then this prediction will be correct 98 times out of 152 (i.e. 64.5%). However, if SPSS predicted that all participants did cut off other vehicles, then this prediction would be correct only 54 times out of 152 (35.5%). As such, of the two available options it is better to predict that all participants did not cut off pedestrians because this results in a greater number of correct predictions. Output shows a contingency table for the model in this basic state. You can see that SPSS has predicted that all participants did not cut off pedestrians, which results in 0% accuracy for those who did cut off pedestrians, and 100% accuracy for those that did not. Overall, the model correctly classifies 64.5% of participants. Output summarizes the model, and at this stage this entails quoting the value of the constant (b_0), which is equal to –0.596.

The final table of the output (Output) is labelled *Variables not in the Equation*. The bottom line of this table reports the residual chi-square statistic as 4.77, which is significant at p < .05 (it labels this statistic *Overall Statistics*). This statistic tells us that the coefficient for the variable not in the model is significantly different from zero – in other words, that the addition of this variable to the model will significantly affect its predictive power. If the probability for the residual chi-square had been greater than .05 it would have meant that the variable excluded from the model could not make a significant contribution to the predictive power of the model. As such, the analysis would have terminated at this stage.

Step 1

The next part of the output deals with the model after the predictor variable (**Vehicle**) has been added to the model. As such, a person is now classified as either cutting off participants at an intersection or not, based on the type of vehicle they were driving (as a measure of social status).

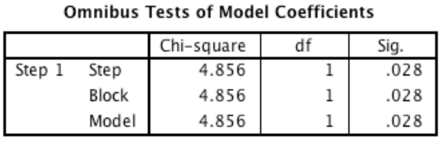


Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	192.939[a]	.031	.043

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

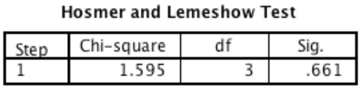
Output 12



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	4.856	1	.028
	Block	4.856	1	.028
	Model	4.856	1	.028

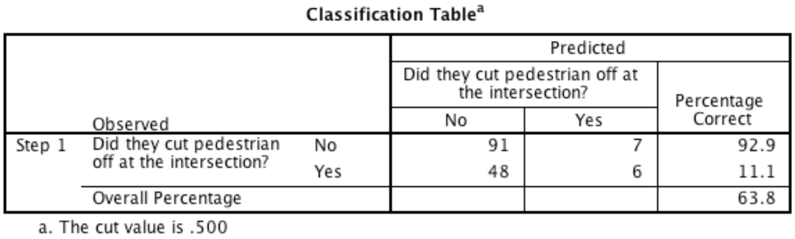
Output 13



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.595	3	.661

Output 14



Classification Table[a]

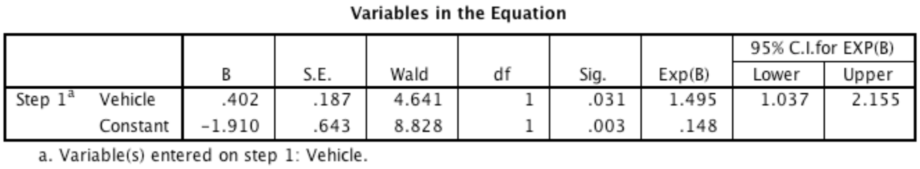
	Observed		Predicted		
			Did they cut pedestrian off at the intersection?		Percentage Correct
			No	Yes	
Step 1	Did they cut pedestrian off at the intersection?	No	91	7	92.9
		Yes	48	6	11.1
	Overall Percentage				63.8

a. The cut value is .500

Output 15

The above output shows summary statistics about the new model. The overall fit of the new model is significant because the *Model* chi-square in the table labelled *Omnibus Tests of Model Coefficients* (Output) is significant, $\chi^2(1) = 4.86$, $p = .028$. Therefore, the model that included the variable **Vehicle** predicted whether or not participants cut off pedestrians at intersections better than the model when only the constant was included.

The classification table at the end of this section of the output (Output) indicates how well the model predicts group membership. In step 1, the model correctly classifies 91 participants who did not cut off pedestrians and misclassifies 7 who did (i.e., it correctly classifies 92.9% of cases). For participants who do did cut off other vehicles, the model correctly classifies 6 and misclassifies 48 cases (i.e. correctly classifies 11.1% of cases). The overall accuracy of classification is, therefore, the weighted average of these two values (63.8%). So, when only the constant was included, the model correctly classified 64.5% participants, but now, with the inclusion of **Vehicle** as a predictor, this has actually decreased slightly to 63.8%.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Vehicle	.402	.187	4.641	1	.031	1.495	1.037	2.155
	Constant	-1.910	.643	8.828	1	.003	.148		

a. Variable(s) entered on step 1: Vehicle.

Output 16

The significance of the Wald statistic in Output is .031, which is less than .05. Therefore, we can conclude that the status of the vehicle the participant was driving significantly predicted whether or not they cut off pedestrians at intersections.

The final thing we need to look at is exp *b* (Exp(*B*) in Output). We can interpret exp *b* in terms of the change in odds. If the value is greater than 1 then it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease. In this example, the exp *b* for vehicle in step 1 is 1.495, which is greater than 1, thus indicating that as the predictor (vehicle) increases, the value of the outcome also increases, that is, the value of the categorical variable moves from 0 (did not cut off pedestrian) to 1 (cut off pedestrian). In other words, drivers of vehicles of a higher status were more likely to cut off pedestrians at intersections.

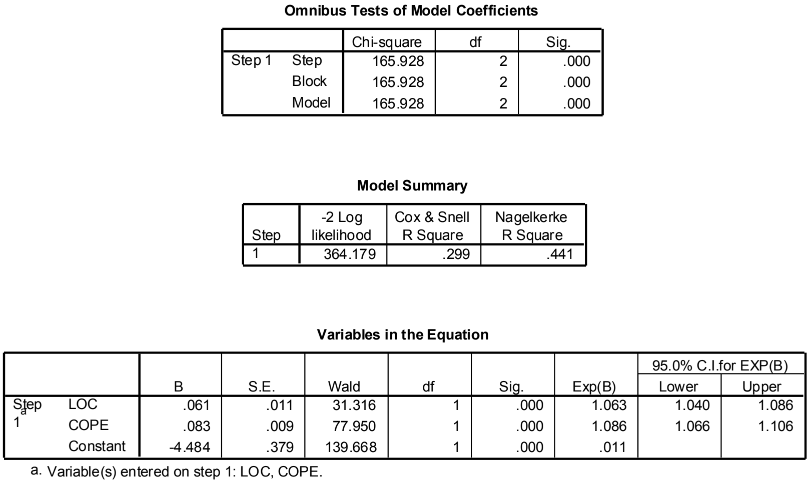
Task 5

Four hundred and sixty-seven lecturers completed questionnaire measures of ***Burnout*** *(burnt out or not),* ***Perceived Control*** *(high score = low perceived control),* ***Coping Style*** *(high score = high ability to cope with stress),* ***Stress from Teaching*** *(high score = teaching creates a lot of stress for the person),* ***Stress from Research*** *(high score = research creates a lot of stress for the person) and* ***Stress from Providing Pastoral Care*** *(high score = providing pastoral care creates a lot of stress for the person). Cooper, Sloan, and Williams' (1988) model of stress indicates that perceived control and coping style are important predictors of burnout. The remaining predictors were measured to see the unique contribution of different aspects of a lecturer's work to their burnout. Conduct a logistic regression to see which factors predict burnout. The data are in* ***Burnout.sav****.*

Test

The analysis should be done hierarchically because Cooper's model indicates that perceived control and coping style are important predictors of burnout. So, these variables should be entered in the first block. The second block should contain all other variables, and because we don't know anything much about their predictive ability, we should enter them in a stepwise fashion (I chose *Forward: LR*).

SPSS output



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	165.928	2	.000
	Block	165.928	2	.000
	Model	165.928	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	364.179	.299	.441

Variables in the Equation

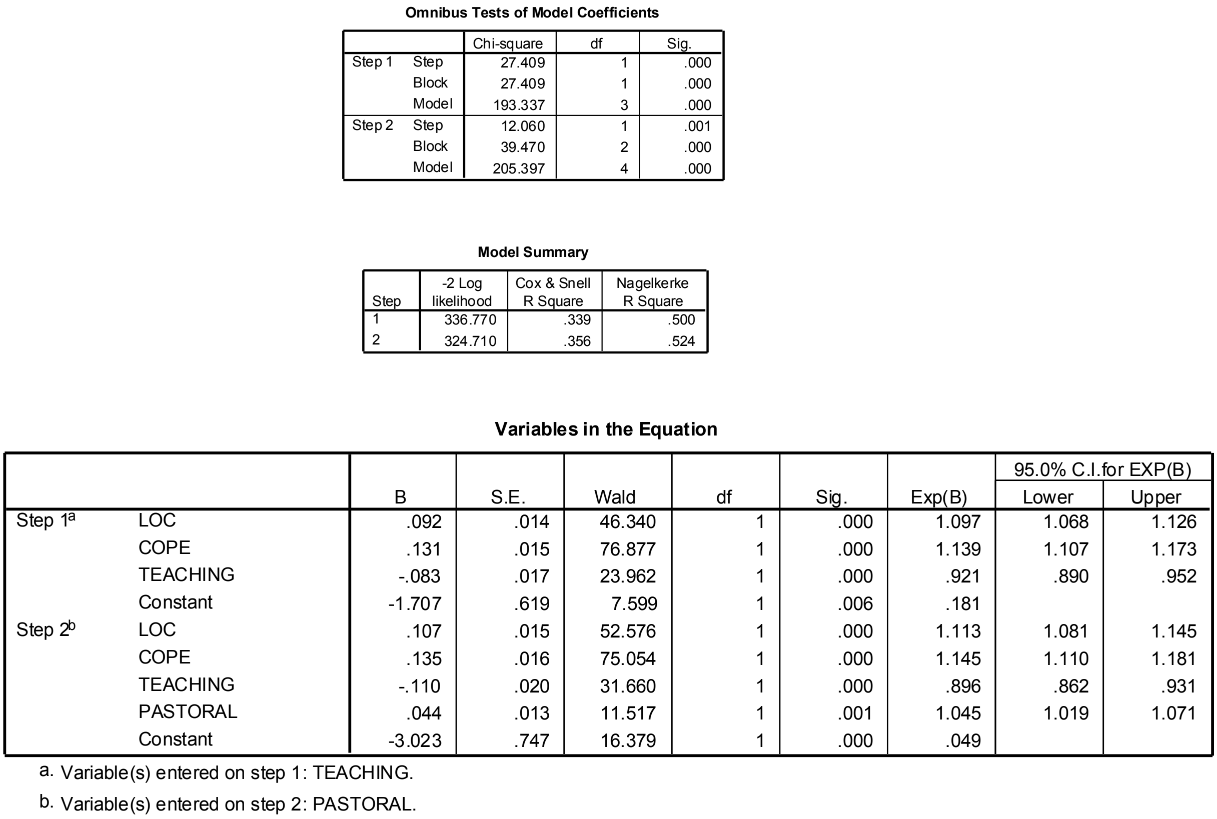
								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	LOC	.061	.011	31.316	1	.000	1.063	1.040	1.086
	COPE	.083	.009	77.950	1	.000	1.086	1.066	1.106
	Constant	-4.484	.379	139.668	1	.000	.011		

a. Variable(s) entered on step 1: LOC, COPE.

Output 17

At step 1, the overall fit of the model is significant, $\chi^2(2) = 165.93$, $p < .001$. The model accounts for 29.9% or 44.1% of the variance in burnout (depending on which measure of R^2 you use).

At step 2, the overall fit of the model is significant after both the first new variable (**teaching**), $\chi^2(3) = 193.34$, $p < .001$, and second new variable (**pastoral**) have been entered, $\chi^2(4) = 205.40$, $p < .001$. The final model accounts for 35.6% or 52.4% of the variance in burnout (depending on which measure of R^2 you use).



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	27.409	1	.000
	Block	27.409	1	.000
	Model	193.337	3	.000
Step 2	Step	12.060	1	.001
	Block	39.470	2	.000
	Model	205.397	4	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	336.770	.339	.500
2	324.710	.356	.524

Variables in the Equation

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	LOC	.092	.014	46.340	1	.000	1.097	1.068	1.126
	COPE	.131	.015	76.877	1	.000	1.139	1.107	1.173
	TEACHING	-.083	.017	23.962	1	.000	.921	.890	.952
	Constant	-1.707	.619	7.599	1	.006	.181		
Step 2[b]	LOC	.107	.015	52.576	1	.000	1.113	1.081	1.145
	COPE	.135	.016	75.054	1	.000	1.145	1.110	1.181
	TEACHING	-.110	.020	31.660	1	.000	.896	.862	.931
	PASTORAL	.044	.013	11.517	1	.001	1.045	1.019	1.071
	Constant	-3.023	.747	16.379	1	.000	.049		

a. Variable(s) entered on step 1: TEACHING.

b. Variable(s) entered on step 2: PASTORAL.

Output 18

In terms of the individual predictors we could report as follows:

	B (SE)	95% CI for Exp(*B*)		
		Lower	Exp(*B*)	Upper
Step 1				
Constant	–4.48** (0.38)			
Perceived Control	0.06** (0.01)	1.04	1.06	1.09
Coping Style	0.08** (0.01)	1.07	1.09	1.11
Final				
Constant	–3.02**			

	(0.75)			
Perceived Control	0.11** (0.02)	1.08	1.11	1.15
Coping Style	0.14** (0.02)	1.11	1.15	1.18
Teaching Stress	–0.11** (0.02)	0.86	0.90	0.93
Pastoral Stress	0.04* (0.01)	1.02	1.05	1.07

Note: R^2 = .36 (Cox and Snell), .52 (Nagelkerke). Model $\chi^2(4)$ = 205.40, *p* < .001. **p* < .01, ***p* < .001.

It seems as though burnout is significantly predicted by perceived control, coping style (as predicted by Cooper), stress from teaching and stress from giving pastoral care. The Exp(*B*) and direction of the beta values tell us that, for perceived control, coping ability and pastoral care, the relationships are positive. That is (and look back to the question to see the direction of these scales, i.e., what a high score represents), poor perceived control, poor ability to cope with stress and stress from giving pastoral care all predict burnout. However, for teaching, the relationship if the opposite way around: stress from teaching appears to be a positive thing as it predicts not becoming burnt out!

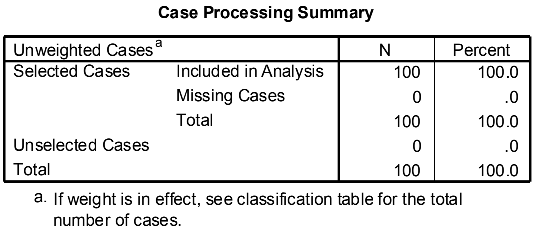
Task 6

> *An HIV researcher explored the factors that influenced condom use with a new partner (relationship less than 1 month old). The outcome measure was whether a condom was used (**Use**: condom used = 1, not used = 0). The predictor variables were mainly scales from the Condom Attitude Scale (CAS) by Sacco, Levine, Reed, and Thompson (1991): **Gender**; the degree to which the person views their relationship as 'safe' from sexually transmitted disease (**Safety**); the degree to which previous experience influences attitudes towards condom use (**Sexexp**); whether or not the couple used a condom in their previous encounter, 1 = condom used, 0 = not used, 2 = no previous encounter with this partner (**Previous**); the degree of self-control that a person has when it comes to condom use (**Selfcon**); the degree to which the person perceives a risk from unprotected sex (**Perceive**). Previous research (Sacco, Rickman, Thompson, Levine, & Reed, 1993) has shown that gender, relationship safety and perceived risk predict condom use. Carry out an analysis using **Condom.sav** to verify these previous findings, and to test whether self-control, previous usage and sexual experience predict condom use.*

The correct analysis was to run a hierarchical logistic regression entering **Perceive**, **Safety** and **Gender** in the first block and **Previous**, **Selfcon** and **Sexexp** in a second. I used forced entry on both blocks, but you could choose to run a forward stepwise method on block 2 (either strategy is justified). For the variable **Previous** I used an indicator contrast with 'No condom' as the base category.

Block 0

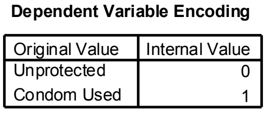
The output of the logistic regression will be arranged in terms of the blocks that were specified. In other words, SPSS will produce a regression model for the variables specified in block 1, and then produce a second model that contains the variables from both blocks 1 and 2. The results from block 1 are shown below. In this analysis we forced SPSS to enter **Perceive**, **Safety** and **Gender** into the regression model first. First, the output tells us that 100 cases have been accepted, that the dependent variable has been coded 0 and 1 (because this variable was coded as 0 and 1 in the data editor, these codings correspond exactly to the data in SPSS).



Case Processing Summary

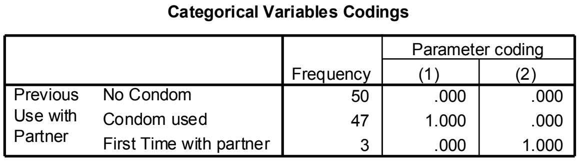
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	100	100.0
	Missing Cases	0	.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

a. If weight is in effect, see classification table for the total number of cases.



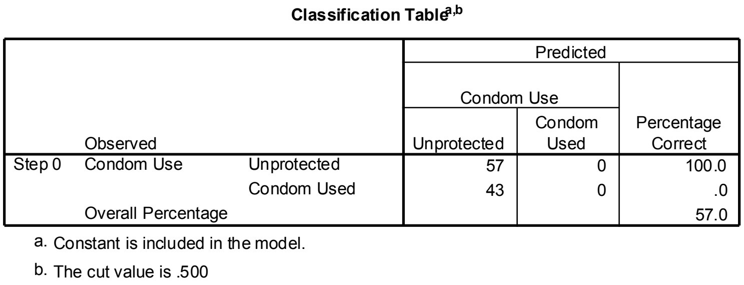
Dependent Variable Encoding

Original Value	Internal Value
Unprotected	0
Condom Used	1



Categorical Variables Codings

			Parameter coding	
		Frequency	(1)	(2)
Previous Use with Partner	No Condom	50	.000	.000
	Condom used	47	1.000	.000
	First Time with partner	3	.000	1.000



Classification Table[a,b]

			Predicted		
			Condom Use		
	Observed		Unprotected	Condom Used	Percentage Correct
Step 0	Condom Use	Unprotected	57	0	100.0
		Condom Used	43	0	.0
	Overall Percentage				57.0

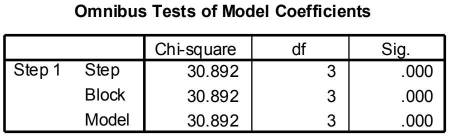
a. Constant is included in the model.

b. The cut value is .500

Output 19

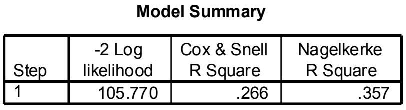
Block 1

The next part of the output tells us about block 1: as such it provides information about the model after the variables **Perceive**, **Safety** and **Gender** have been added. The first thing to note is that −2*LL* has dropped to 105.77, which is a change of 30.89 (which is the value given by the *model chi-square*). This value tells us about the model as a whole, whereas the *block* tells us how the model has improved since the last block. The change in the amount of information explained by the model is significant ($\chi^2(3) = 30.89$, $p < .001$) and so using perceived risk, relationship safety and gender as predictors significantly improves our ability to predict condom use. Finally, the classification table shows us that 74% of cases can be correctly classified using these three predictors.



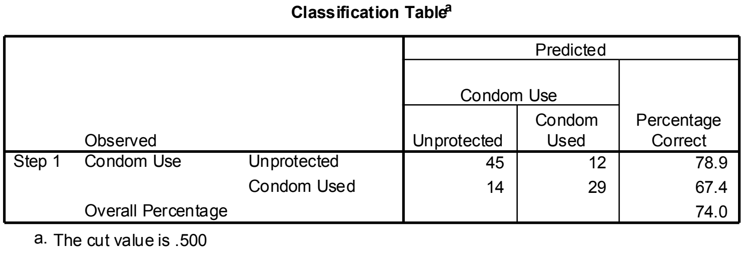
Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	30.892	3	.000
	Block	30.892	3	.000
	Model	30.892	3	.000



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	105.770	.266	.357



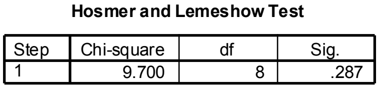
Classification Table[a]

Observed			Predicted: Condom Use: Unprotected	Predicted: Condom Use: Condom Used	Predicted: Percentage Correct
Step 1	Condom Use	Unprotected	45	12	78.9
		Condom Used	14	29	67.4
	Overall Percentage				74.0

a. The cut value is .500

Output 20

Hosmer and Lemeshow's goodness-of-fit test statistic tests the hypothesis that the observed data are significantly different from the predicted values from the model. So, in effect, we want a non-significant value for this test (because this would indicate that the model does not differ significantly from the observed data). In this case ($\chi^2(8) = 9.70$, $p =$ .287) it is non-significant, which is indicative of a model that is predicting the real-world data fairly well.

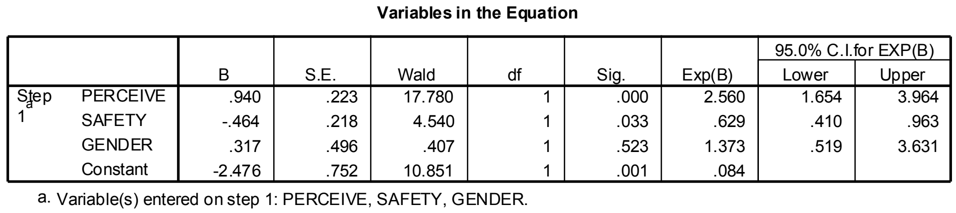


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.700	8	.287

Output 21

Output 34, labelled *Variables in the Equation*, then tells us the parameters of the model for the first block. The significance values of the Wald statistics for each predictor indicate that both perceived risk (Wald = 17.78, $p < .001$) and relationship safety (Wald = 4.54, $p <$ .05) significantly predict condom use. Gender, however, does not (Wald = 0.41, $p > .05$).



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PERCEIVE	.940	.223	17.780	1	.000	2.560	1.654	3.964
	SAFETY	-.464	.218	4.540	1	.033	.629	.410	.963
	GENDER	.317	.496	.407	1	.523	1.373	.519	3.631
	Constant	-2.476	.752	10.851	1	.001	.084		

a. Variable(s) entered on step 1: PERCEIVE, SAFETY, GENDER.

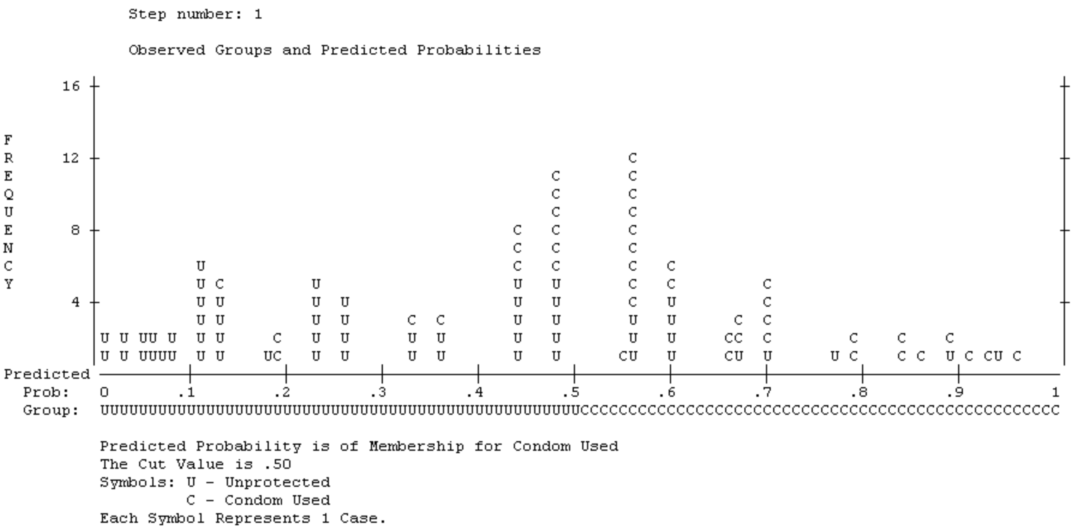
Output 22

The odds ratio for perceived risk (Exp(*B*) = 2.56, $\mathrm{CI}_{0.95}$ = [1.65, 3.96]) indicates that if the value of perceived risk goes up by 1, then the odds of using a condom also increase (because Exp(*B*) is greater than 1). The confidence interval for this value ranges from 1.65 to 3.96, so we can be very confident that the value of Exp(*B*) in the population lies somewhere between these two values. What's more, because both values are greater than 1 we can also be confident that the relationship between perceived risk and condom use found in this sample is true of the whole population. In short, as perceived risk increase by 1, people are just over twice as likely to use a condom.

The odds ratio for relationship safety (Exp(*B*) = 0.63, $\mathrm{CI}_{0.95}$ = [0.41, 0.96]) indicates that if the relationship safety increases by one point, then the odds of using a condom decrease (because Exp(*B*) is less than 1). The confidence interval for this value ranges from 0.41 to 0.96, so we can be very confident that the value of Exp(*B*) in the population lies somewhere between these two values. In addition, because both values are less than 1 we can be confident that the relationship between relationship safety and condom use found in this sample would be found in 95% of samples from the same population. In short, as relationship safety increases by one unit, subjects are about 1.6 times less likely to use a condom.

The odds ratio for gender (Exp(*B*) = 1.37, $\mathrm{CI}_{0.95}$ = [0.52, 3.63]) indicates that as gender changes from 0 (male) to 1 (female), then the odds of using a condom increase (because Exp(*B*) is greater than 1). However, the confidence interval for this value crosses 1, which limits the generalizability of our findings because the value of Exp(*B*) in other samples (and hence the population) could indicate either a positive (Exp(*B*) > 1) or negative (Exp(*B*) < 1) relationship. Therefore, gender is not a reliable predictor of condom use.

A glance at the classification plot (Output 35) brings not such good news because a lot of cases are clustered around the middle. This indicates that the model could be performing more accurately (i.e., the classifications made by the model are not completely reliable).

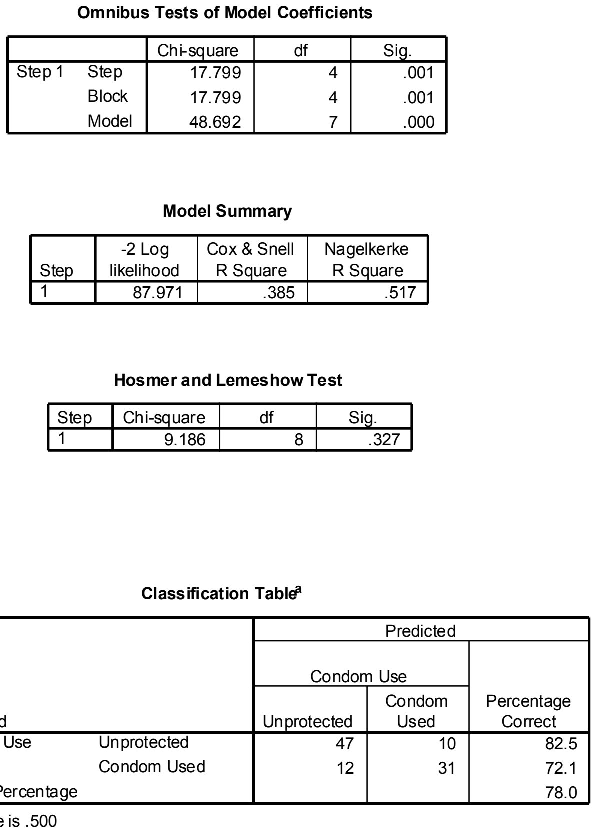


```
             Step number: 1

             Observed Groups and Predicted Probabilities

      16 +                                                                                                    +
         |                                                                                                    |
         |                                                                                                    |
F        |                                                                                                    |
R     12 +                                                       C                                            +
E        |                                               C       C                                            |
Q        |                                               C       C                                            |
U        |                                               C       C                                            |
E      8 +                                           C   C       C                                            +
N        |                                           C   C       C                                            |
C        |          U                                C   C       C   C                                        |
Y        |          U C         U                    U   U       C   C         C                              |
       4 +          U U         U  U                 U   U       C   U         C                              +
         |          U U         U  U      C  C       U   U       U   U      C  C                              |
         |U U UU U  U U     C   U  U      U  U       U   U       U   U     CC  C        C    C    C           |
         |U U UUUU  U U    UC   U  U      U  U       U   U      CU   U     CU  U      U C    C C  U C CU C    |
Predicted----------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  UUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

          Predicted Probability is of Membership for Condom Used
          The Cut Value is .50
          Symbols: U - Unprotected
                   C - Condom Used
          Each Symbol Represents 1 Case.
```

Output 23

Block 2

The output below shows what happens to the model when our new predictors are added (previous use, self-control and sexual experience). This part of the output describes block 2, which is just the model described in block 1 but with a new predictors added. So, we begin with the model that we had in block 1 and we then add **Previous**, **Selfcon** and **Sexexp** to it. The effect of adding these predictors to the model is to reduce the –2 log-likelihood to 87.97 (a reduction of 48.69 from the original model as shown in the *model chi-square* and an additional reduction of 17.80 from the reduction caused by block 1 as shown by the *block* statistics). This additional improvement of block 2 is significant ($\chi^2(4) = 17.80$, $p < .01$), which tells us that including these three new predictors in the model has significantly improved our ability to predict condom use. The classification table tells us that the model is now correctly classifying 78% of cases. Remember that in block 1 there were 74% correctly classified and so an extra 4% of cases are now classified (not a great deal more – in fact, examining the table shows us that only four extra cases have now been correctly classified).



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	17.799	4	.001
	Block	17.799	4	.001
	Model	48.692	7	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	87.971	.385	.517

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.186	8	.327

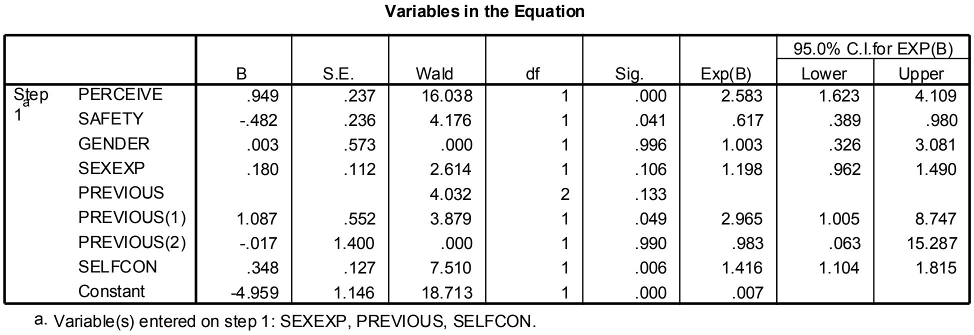
Classification Table[a]

			Predicted		
			Condom Use		
Observed			Unprotected	Condom Used	Percentage Correct
Step 1	Condom Use	Unprotected	47	10	82.5
		Condom Used	12	31	72.1
	Overall Percentage				78.0

a. The cut value is .500

Output 24

The section labelled *Variables in the Equation* (Output 37) now contains all predictors. This part of the output represents the details of the final model. The significance values of the Wald statistics for each predictor indicate that both perceived risk (Wald = 16.04, *p* < .001) and relationship safety (Wald = 4.17, *p* < .05) still significantly predict condom use and, as in block 1, gender does not (Wald = 0.00, *p* > .05). We can now look at the new predictors to see which of these has some predictive power.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1[a]	PERCEIVE	.949	.237	16.038	1	.000	2.583	1.623	4.109
	SAFETY	-.482	.236	4.176	1	.041	.617	.389	.980
	GENDER	.003	.573	.000	1	.996	1.003	.326	3.081
	SEXEXP	.180	.112	2.614	1	.106	1.198	.962	1.490
	PREVIOUS			4.032	2	.133			
	PREVIOUS(1)	1.087	.552	3.879	1	.049	2.965	1.005	8.747
	PREVIOUS(2)	-.017	1.400	.000	1	.990	.983	.063	15.287
	SELFCON	.348	.127	7.510	1	.006	1.416	1.104	1.815
	Constant	-4.959	1.146	18.713	1	.000	.007		

a. Variable(s) entered on step 1: SEXEXP, PREVIOUS, SELFCON.

Output 25

Previous use has been split into two components (according to whatever contrasts were specified for this variable). Looking at the very beginning of the output, we are told the parameter codings for **Previous(1)** and **Previous(2)**. You can tell by remembering the rule from contrast coding in ANOVA which groups are being compared: that is, we compare groups with codes of 0 against those with codes of 1. From the output we can see that **Previous(1)** compares the condom used group against the no condom used group, and **Previous(2)** compares the base category of first time with partner against the other two categories. Therefore we can tell that previous use is not a significant predictor of condom use when it is the first time with a partner compared to when it is not the first time (Wald = 0.00, *p* < .05). However, when we compare the condom used category to the other categories we find that using a condom on the previous occasion does predict use on the current occasion (Wald = 3.88, *p* < .05).

Of the other new predictors we find that self-control predicts condom use (Wald = 7.51, *p* < .01) but sexual experience does not (Wald = 2.61, *p* > .05).

The odds ratio for perceived risk (Exp(*B*) = 2.58, $\text{CI}_{0.95}$ = [1.62, 4.106]) indicates that if the value of perceived risk goes up by 1, then the odds of using a condom also increase. What's more, because the confidence interval doesn't cross 1 we can also be confident that the relationship between perceived risk and condom use found in this sample is true of the whole population. As perceived risk increases by 1, people are just over twice as likely to use a condom.

The odds ratio for relationship safety (Exp(*B*) = 0.62, $\text{CI}_{0.95}$ = [0.39, 0.98]) indicates that if the relationship safety decreases by one point, then the odds of using a condom increase. The confidence interval does not cross 1 so we can be confident that the relationship between relationship safety and condom use found in this sample would be found in 95% of samples from the same population. As relationship safety increases by one unit, subjects are about 1.6 times less likely to use a condom.

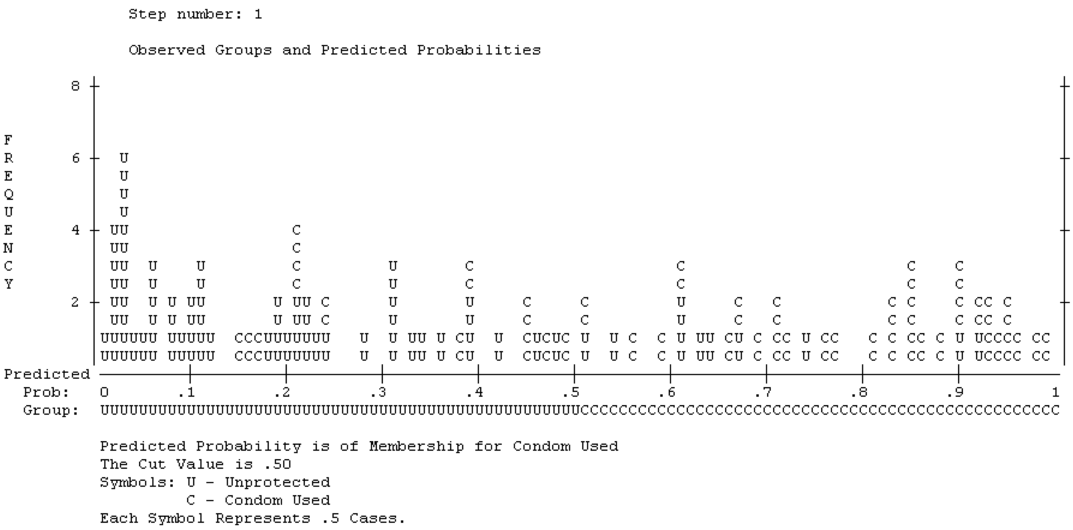
The odds ratio for gender (Exp(*B*) = 1.00, $\text{CI}_{0.95}$ = [0.33, 3.08]) indicates that as gender changes from 0 (male) to 1 (female), then the odds of using a condom do not change (because Exp(*B*) is equal to 1). The confidence interval crosses 1, therefore gender is not a reliable predictor of condom use.

The odds ratio for previous use (1) (Exp(*B*) = 2.97, $\text{CI}_{0.95}$ = [1.01, 8.75]) indicates that if the value of previous usage goes up by 1 (i.e., changes from not having used one or being the first time to having used one), then the odds of using a condom also increase. What's more, because the confidence interval doesn't cross 1 we can also be confident that this relationship is true in the whole population. If someone used a condom on their previous encounter with this partner (compared to if they didn't use one, or if it is their first time) then they are three times more likely to use a condom. For previous use (2) the odds ratio (Exp(*B*) = 0.98, $\text{CI}_{0.95}$ = [0.06, 15.29]) indicates that if the value of previous usage goes up by 1 (i.e., changes from not having used one or having used one to being their first time with this partner), then the odds of using a condom do not change (because the value is very nearly equal to 1). What's more, because the confidence interval crosses 1 we can tell that this is not a reliable predictor of condom use.

The odds ratio for self-control (Exp(*B*) = 1.42, $\text{CI}_{0.95}$ = [1.10, 1.82]) indicates that if self-control increases by one point, then the odds of using a condom increase also. The confidence interval does not cross 1, so we can be confident that the relationship between relationship safety and condom use found in this sample would be found in 95% of samples from the same population. As self-control increases by one unit, subjects are about 1.4 times more likely to use a condom.

The odds ratio for sexual experience (Exp(*B*) = 1.20, $CI_{0.95}$ = [0.95, 1.49]) indicates that as sexual experience increases by one unit, then the odds of using a condom increase slightly. However, the confidence interval crosses 1, therefore sexual experience is not a reliable predictor of condom use.

A glance at the classification plot (Output 38) brings good news because a lot of cases that were clustered in the middle are now spread towards the edges. Therefore, overall this new model is more accurately classifying cases compared to block 1.



```
          Step number: 1

          Observed Groups and Predicted Probabilities

       8 +                                                                                                    +
         |                                                                                                    |
         |                                                                                                    |
F        |                                                                                                    |
R      6 +  U                                                                                                 +
E        |  U                                                                                                 |
Q        |  U                                                                                                 |
U        |  U                                                                                                 |
E      4 + UU                 C                                                                               +
N        | UU                 C                                                                               |
C        | UU  U    U         C         U       C                     C                       C    C          |
Y        | UU  U    U         C         U       C                     C                       C    C          |
       2 + UU  U U UU       U UU C      U       U     C     C         U     C   C           C C    C CC C     +
         | UU  U U UU       U UU C      U       U     C     C         U     C   C           C C    C CC C     |
         |UUUUUU UUUUU  CCCUUUUUUU   U  U UU U CU  U  CUCUC U  U C  C U UU CU C CC U CC   C C CC C U UCCCC CC |
         |UUUUUU UUUUU  CCCUUUUUUU   U  U UU U CU  U  CUCUC U  U C  C U UU CU C CC U CC   C C CC C U UCCCC CC |
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  UUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

          Predicted Probability is of Membership for Condom Used
          The Cut Value is .50
          Symbols: U - Unprotected
                   C - Condom Used
          Each Symbol Represents .5 Cases.
```

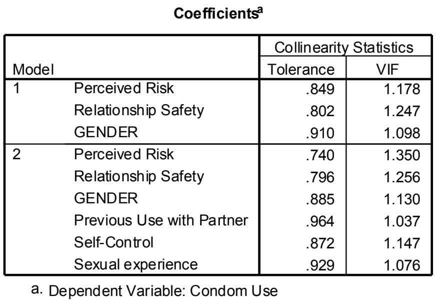
Output 26

Task 7

How reliable is the model for Task 6?

Multicollinearity can affect the parameters of a regression model. Logistic regression is equally prone to the biasing effect of collinearity, and it is essential to test for collinearity following a logistic regression analysis (see the book for details of how to do this). The results of the analysis are shown below. Menard (1995) suggests that tolerance values below 0.1 indicate a serious collinearity problem; from the first table in Output 39 we can see that the tolerance values for all variables are close to 1, much larger than this criterion. Myers (1990) also suggests that a VIF value greater than 10 is cause for concern, and in these data the values are all less than this criterion.

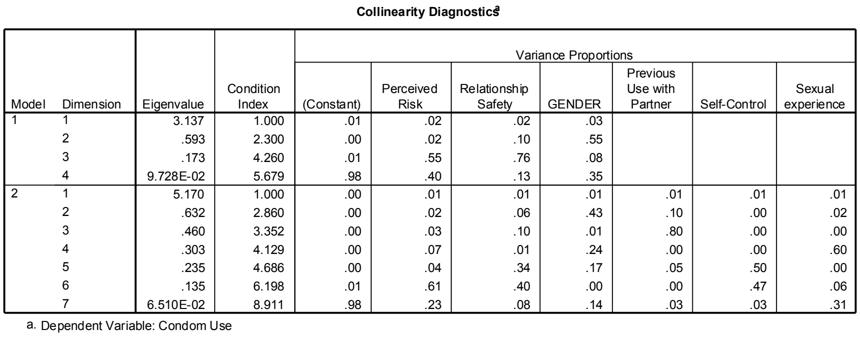
Output 39 also shows a table labelled *Collinearity Diagnostics*. In this table, we are given the eigenvalues of the scaled, uncentred cross-products matrix, the condition index and the variance proportions for each predictor. If any of the eigenvalues in this table are much larger than others then the uncentred cross-products matrix is said to be ill-conditioned, which means that the solutions of the regression parameters can be greatly affected by small changes in the predictors or outcome. In plain English, these values give us some idea as to how accurate our regression model is: if the eigenvalues are fairly similar then the derived model is likely to be unchanged by small changes in the measured variables. The *condition indexes* are another way of expressing these eigenvalues and represent the square root of the ratio of the largest eigenvalue to the eigenvalue of interest (so, for the dimension with the largest eigenvalue, the condition index will always be 1). For these data the condition indexes are all relatively similar, showing that a problem is unlikely to exist.



Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Perceived Risk	.849	1.178
	Relationship Safety	.802	1.247
	GENDER	.910	1.098
2	Perceived Risk	.740	1.350
	Relationship Safety	.796	1.256
	GENDER	.885	1.130
	Previous Use with Partner	.964	1.037
	Self-Control	.872	1.147
	Sexual experience	.929	1.076

a. Dependent Variable: Condom Use



Collinearity Diagnostics[a]

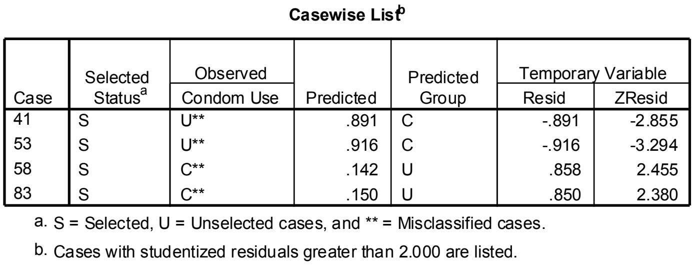
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	Perceived Risk	Relationship Safety	GENDER	Previous Use with Partner	Self-Control	Sexual experience
1	1	3.137	1.000	.01	.02	.02	.03			
	2	.593	2.300	.00	.02	.10	.55			
	3	.173	4.260	.01	.55	.76	.08			
	4	9.728E-02	5.679	.98	.40	.13	.35			
2	1	5.170	1.000	.00	.01	.01	.01	.01	.01	.01
	2	.632	2.860	.00	.02	.06	.43	.10	.00	.02
	3	.460	3.352	.00	.03	.10	.01	.80	.00	.00
	4	.303	4.129	.00	.07	.01	.24	.00	.00	.60
	5	.235	4.686	.00	.04	.34	.17	.05	.50	.00
	6	.135	6.198	.01	.61	.40	.00	.00	.47	.06
	7	6.510E-02	8.911	.98	.23	.08	.14	.03	.03	.31

a. Dependent Variable: Condom Use

Output 27

The final step in analysing this table is to look at the variance proportions. The variance of each regression coefficient can be broken down across the eigenvalues, and the variance proportions tell us the proportion of the variance of each predictor's regression coefficient that is attributed to each eigenvalue. These proportions can be converted to percentages by multiplying them by 100 (to make them more easily understood). In terms of collinearity, we are looking for predictors that have high proportions on the same *small* eigenvalue, because this would indicate that the variances of their regression coefficients are dependent (see the main textbook for more detail). Again, no variables appear to have similarly high variance proportions for the same dimensions. The result of this analysis is pretty clear-cut: there is no problem of collinearity in these data.

Residuals should be checked for influential cases and outliers. As a brief guide, the output lists cases with standardized residuals greater than 2. In a sample of 100, we would expect around 5–10% of cases to have standardized residuals with absolute values greater than this. For these data we have only four cases and only one of these has an absolute value greater than 3 (Output 40). Therefore, we can be fairly sure that there are no outliers.



Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable	
		Condom Use			Resid	ZResid
41	S	U**	.891	C	-.891	-2.855
53	S	U**	.916	C	-.916	-3.294
58	S	C**	.142	U	.858	2.455
83	S	C**	.150	U	.850	2.380

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

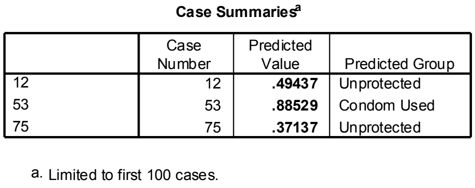
b. Cases with studentized residuals greater than 2.000 are listed.

Output 28

Task 8

Using the final model of condom use in Task 6, what are the probabilities that participants 12, 53 and 75 will use a condom?

The values predicted for these cases will depend on exactly how you ran the analysis (and the parameter coding used on the variable **Previous**). Therefore, your answers might differ slightly from mine.



Case Summaries[a]

	Case Number	Predicted Value	Predicted Group
12	12	**.49437**	Unprotected
53	53	**.88529**	Condom Used
75	75	**.37137**	Unprotected

a. Limited to first 100 cases.

Output 29

Task 9

A female who used a condom in her previous encounter with her new partner scores 2 on all variables except perceived risk (for which she scores 6). Use the model to estimate the probability that she will use a condom in her next encounter.

Step 1. Logistic regression equation:

$$P(Y) = \frac{1}{1 + e^{-Z}}$$

where

$$Z = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Step 2. Use the values of *b* from the SPSS output (final model and the values of *X* for each variable (from the question) to construct the following table:

Variable	b_i	X_i	$b_i X_i$
Gender	0.0027	1	0.0027
Safety	−0.4823	2	−0.9646
Sexexp	0.1804	2	0.3608
Previous (1)	1.0870	1	1.0870
Previous (2)	−.0167	0	0
Selfcon	0.3476	2	0.6952
Perceive	0.9489	6	5.6934

Step 3. Place the values of $b_i X_i$ into the equation for *z* (remembering to include the constant):

$$\begin{aligned} z &= -4.6009 + 0.0027 - 0.9646 + 0.3608 + 1.0870 + 0 + 0.6952 + 5.6934 \\ &= 2.2736 \end{aligned}$$

Step 4. Replace this value of *z* into the logistic regression equation:

$$\begin{aligned} P(Y) &= \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-2.2736}} \\ &= .9067 \end{aligned}$$

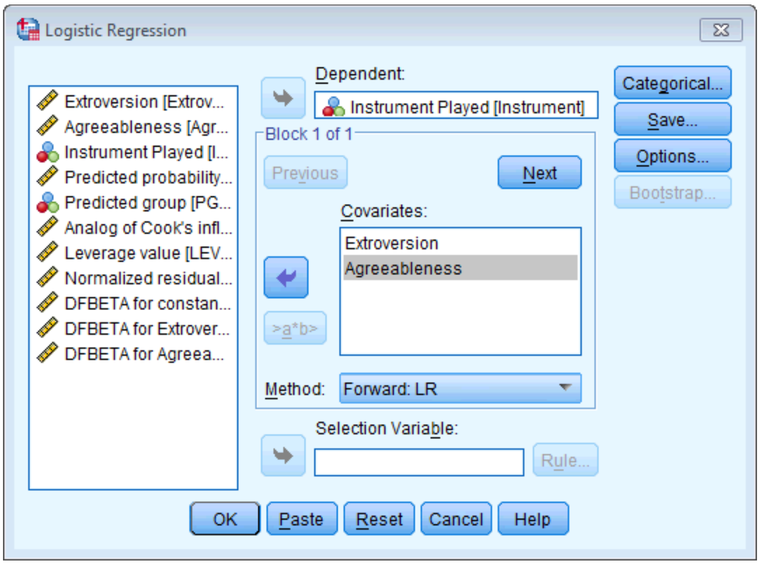
Therefore, there is a 91% chance that she will use a condom on her next encounter.

Task 10

At the start of the chapter we looked at whether the type of instrument a person plays is connected to their personality. A musicologist got 200 singers and guitarists from bands. She noted the ***Instrument*** *they played (Singer, Guitar), and measured two personality variables in each:* ***Extroversion*** *and* ***Agreeableness****. Conduct a logistic*

> *regression to see which of these variables (ignore the interaction) predicts which instrument a person plays. Data are in* ***Sing or Guitar.sav****.*

Logistic regression is located in the regression menu accessed by selecting Analyze Regression ▸ Binary Logistic... . Following this menu path activates the main *Logistic Regression* dialog box shown in Figure 5. Complete the dialog boxes as shown in Figures 5–7 (see the book chapter for more information on how to complete these boxes).



Extroversion [Extrov...
Agreeableness [Agr...
Instrument Played [I...
Predicted probability...
Predicted group [PG...
Analog of Cook's infl...
Leverage value [LEV...
Normalized residual...
DFBETA for constan...
DFBETA for Extrover...
DFBETA for Agreea...

Dependent: Instrument Played [Instrument]

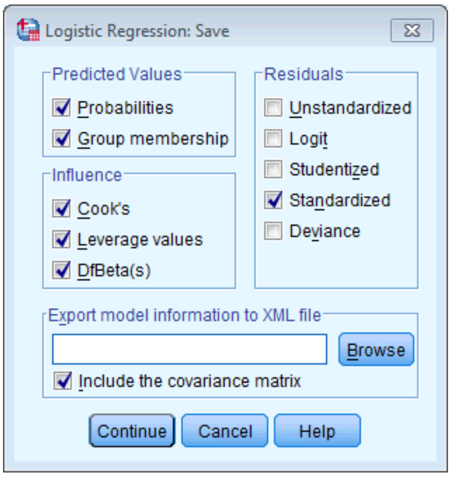
Block 1 of 1

Covariates:
Extroversion
Agreeableness

Method: Forward: LR

Selection Variable:

Figure 5



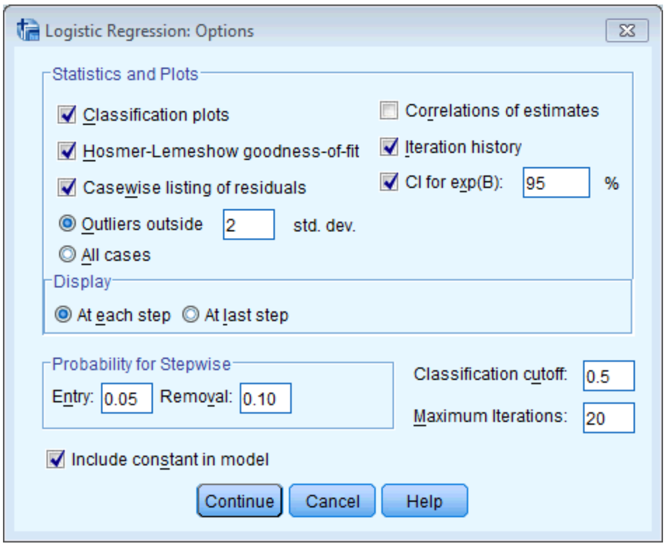
Predicted Values
☑ Probabilities
☑ Group membership

Influence
☑ Cook's
☑ Leverage values
☑ DfBeta(s)

Residuals
□ Unstandardized
□ Logit
□ Studentized
☑ Standardized
□ Deviance

Export model information to XML file
☑ Include the covariance matrix

Figure 6



Logistic Regression: Options

Statistics and Plots

☑ Classification plots

☑ Hosmer-Lemeshow goodness-of-fit

☑ Casewise listing of residuals

● Outliers outside 2 std. dev.

○ All cases

☐ Correlations of estimates

☑ Iteration history

☑ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10

Classification cutoff: 0.5

Maximum Iterations: 20

☑ Include constant in model

Figure 7

Interpreting the output

Dependent Variable Encoding

Original Value	Internal Value
Singer	0
Guitar	1

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	271.957	-.152
	2	271.957	-.153

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 271.957

c. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001.

Classification Table[a,b]

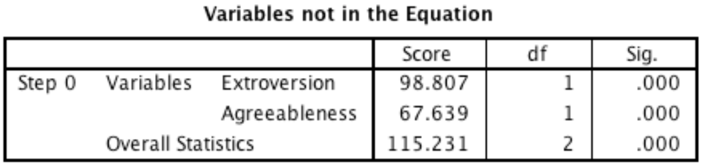
Observed			Predicted		
			Instrument Played		Percentage Correct
			Singer	Guitar	
Step 0	Instrument Played	Singer	106	0	100.0
		Guitar	91	0	.0
	Overall Percentage				53.8

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.153	.143	1.140	1	.286	.858



Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Extroversion	98.807	1	.000
		Agreeableness	67.639	1	.000
	Overall Statistics		115.231	2	.000

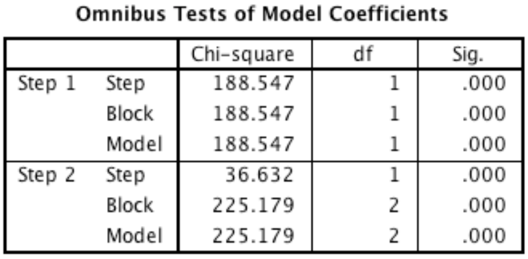
Output 30

For this first analysis we requested a forward stepwise method and so the initial model is derived using only the constant in the regression equation. Output 42 tells us about the model when only the constant is included (i.e., all predictor variables are omitted). The log-likelihood of this baseline model is 271.957. This represents the fit of the model when the most basic model is fitted to the data. When including only the constant, the computer bases the model on assigning every participant to a single category of the outcome variable. The output shows a contingency table for the model in this basic state. You can see that SPSS has predicted that all participants are singers, which results in 0% accuracy for the participants who played the guitar, and 100% accuracy for those participants who were singers. Overall, the model correctly classifies 53.8% of participants. The next part of the output summarizes the model, and at this stage this entails quoting the value of the constant (b_0), which is equal to –0.153.

The final table of the output is labelled *Variables not in the Equation*. The bottom line of this table reports the residual chi-square statistic as 115.231 which is significant at p < .001 (it labels this statistic *Overall Statistics*). This statistic tells us that the coefficients for the variables not in the model are significantly different from zero – in other words, that the addition of one or more of these variables to the model will significantly affect its predictive power. If the probability for the residual chi-square had been greater than .05 it would have meant that none of the variables excluded from the model could make a significant contribution to the predictive power of the model. As such, the analysis would have terminated at this stage.

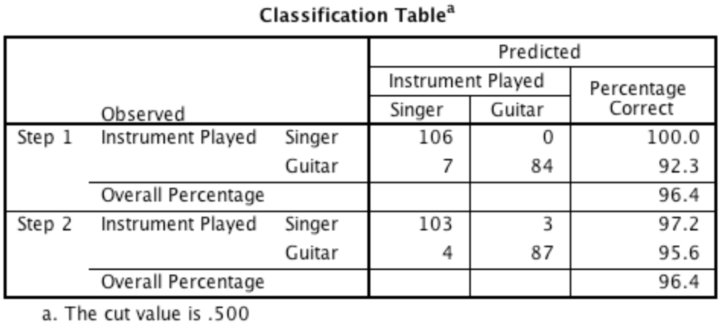
The remainder of this table lists both of the predictors in turn with a value of Roa's efficient score statistic for each one (column labelled *Score*). In large samples when the null hypothesis is true, the score statistic is identical to the Wald statistic and the likelihood ratio statistic. It is used at this stage of the analysis because it is computationally less intensive than the Wald statistic and so can still be calculated in situations when the Wald statistic would prove prohibitive. Like any test statistic, Roa's score statistic has a specific distribution from which statistical significance can be obtained. In this example, both excluded variables have significant score statistics at p < .001 and so both could potentially make a contribution to the model. The stepwise calculations are relative and so the variable that will be selected for inclusion is the one with the highest value for the score statistic that is significant at the .05 level. In this example, that variable will be **Extroversion** because it has the highest value of the score statistic. The next part of the output deals with the model after this predictor has been added.

In the first step, **Extroversion** is added to the model as a predictor. As such, a participant is classified as being a singer based on how extroverted they are. In the second step, **Agreeableness** is added to the model as a predictor. As such a participant is classified as being a singer based on how agreeable they are.



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	188.547	1	.000
	Block	188.547	1	.000
	Model	188.547	1	.000
Step 2	Step	36.632	1	.000
	Block	225.179	2	.000
	Model	225.179	2	.000



Classification Table[a]

	Observed		Predicted		
			Instrument Played		Percentage Correct
			Singer	Guitar	
Step 1	Instrument Played	Singer	106	0	100.0
		Guitar	7	84	92.3
	Overall Percentage				96.4
Step 2	Instrument Played	Singer	103	3	97.2
		Guitar	4	87	95.6
	Overall Percentage				96.4

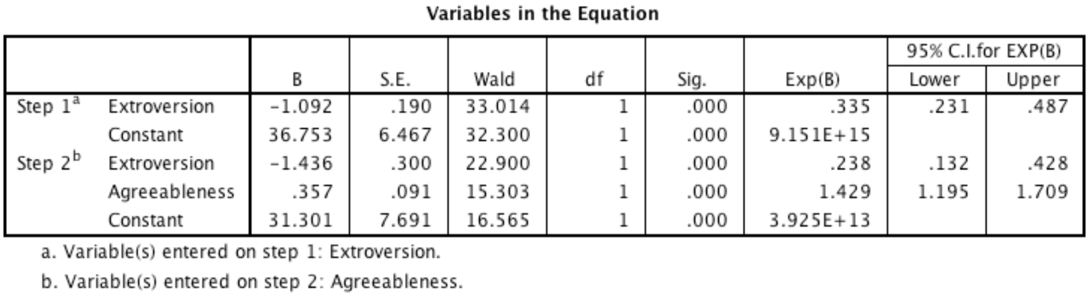
a. The cut value is .500

Output 31

Output 43 shows summary statistics about the two new models. The overall fit of the new models is assessed using the log-likelihood statistic. In SPSS, rather than reporting the log-likelihood itself, the value is multiplied by –2 (and sometimes referred to as –2*LL*): this multiplication is done because –2*LL* has an approximately chi-square distribution and so makes it possible to compare values against those that we might expect to get by chance alone. Remember that large values of the log-likelihood statistic indicate poorly fitting statistical models.

The value of –2 log-likelihood for each new model should be less than the value for the previous model (because lower values of –2*LL* indicate that the model is predicting the outcome variable more accurately). When only the constant was included, –2*LL* = 271.96. When extroversion was included (step 1) this value was reduced to 83.41, which tells us that the model is better at predicting instrument played than it was before **Extroversion** was added. Finally, when **Agreeableness** was added to the model (step 2), this value of –2*LL* was reduced even further to 46.78, which tells us that the model was better at predicting which instrument participants' played when both predictors were included. The question of how much better the models predicts the outcome variable can be assessed using the model chi-square statistic, which measures the difference between the model as it currently stands and the model when only the constant was included. We can assess the significance of the change in a model by taking the log-likelihood of the new model and subtracting the log-likelihood of the baseline model from it. The value of the model chi-square statistic when only **Extroversion** is included is 271.96 – 83.41 = 188.55 and the value of the model chi-square statistic when **Agreeableness** is added to the model is 83.41 – 46.78 = 36.63; these values are both significant. Therefore, the model that included extroversion predicted instrument played significantly better than the model when only the constant was included, and the model that included both agreeableness and extroversion predicted instrument played better than when only extroversion and the constant were included in the model.

The classification table in Output 43 indicates how well the model predicts group membership. In step 1, the model correctly classifies 106 participants who are singers and does not misclassify any (i.e., it correctly classifies 100% of cases). For participants who play guitar, the model correctly classifies 84 and misclassifies 7 cases (i.e., it correctly classifies 92.3% of cases). The overall accuracy of classification is, therefore, the weighted average of these two values (96.4%). So, when only the constant was included, the model correctly classified 53.8% of children, but now, with the inclusion of **extroversion** as a predictor, this has risen to 96.4%. In step 2, the model correctly classifies 103 participants who are singers but misclassifies 3 others (i.e., it correctly classifies 97.2% of cases). For participants who play guitar, the model correctly classifies 87 and misclassifies 4 cases (i.e., it correctly classifies 95.6% of cases). The overall accuracy of classification is, therefore, the weighted average of these two values (96.4%). So, when only the constant was included, the model correctly classified 53.8% of children, but now, with the inclusion of **extroversion** as a predictor, this has risen to 96.4%.



Variables in the Equation

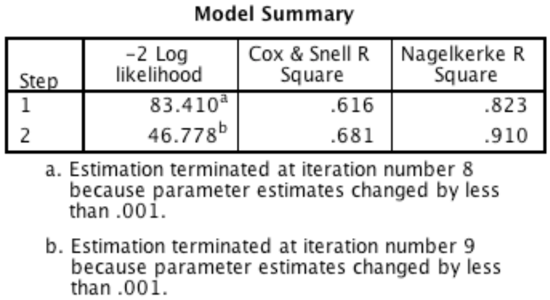
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Extroversion	-1.092	.190	33.014	1	.000	.335	.231	.487
	Constant	36.753	6.467	32.300	1	.000	9.151E+15		
Step 2[b]	Extroversion	-1.436	.300	22.900	1	.000	.238	.132	.428
	Agreeableness	.357	.091	15.303	1	.000	1.429	1.195	1.709
	Constant	31.301	7.691	16.565	1	.000	3.925E+13		

a. Variable(s) entered on step 1: Extroversion.

b. Variable(s) entered on step 2: Agreeableness.

Output 32

The next part of the output is crucial because it tells us the estimates for the coefficients for the predictors included in the model (Output 44). The interpretation of this coefficient in logistic regression is that it represents the change in the logit of the outcome variable associated with a one-unit change in the predictor variable. The logit of the outcome is simply the natural logarithm of the odds of *Y* occurring. The crucial statistic is the Wald statistic, which has a chi-square distribution and tells us whether the *b* coefficient for that predictor is significantly different from zero. If the coefficient is significantly different from zero then we can assume that the predictor is making a significant contribution to the prediction of the outcome (*Y*). For these data it seems to indicate that both extroversion and agreeableness are significant predictors of instrument played (note the significance of the Wald statistics are both less than .05).



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	83.410[a]	.616	.823
2	46.778[b]	.681	.910

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

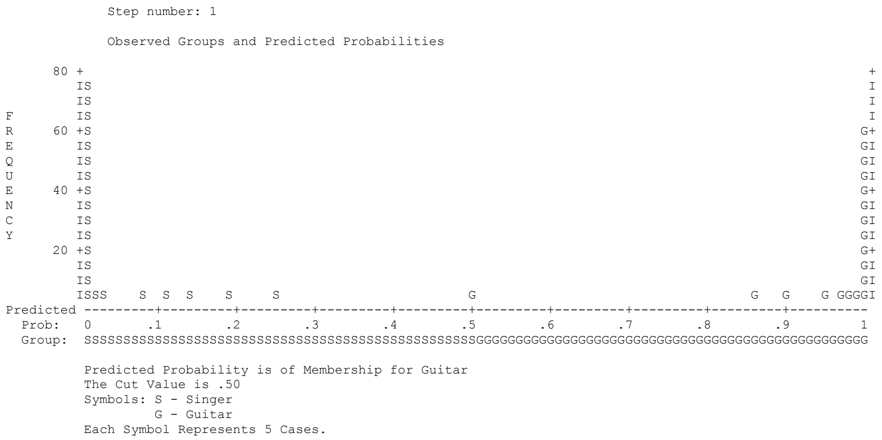
Output 33

The final thing we need to look at is exp *b* (Exp(*B*) in the SPSS output). We can interpret exp *b* in terms of the change in odds. If the value is greater than 1 then it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease. In this example, the exp *b* for extroversion in step 2 is 0.238, which is less than 1 thus indicating that as the predictor (extroversion) increases, the value of the outcome decreases, that is, the value of the categorical variable moves from 1 (guitarist) to 0 (singer). In other words, more extroverted participants are more likely to be singers. The exp *b* for agreeableness is 1.43, which is greater than 1, thus indicating that as the predictor (agreeableness) increases, the value of the outcome also increases (i.e., the value of the categorical variable moves from 0 (singer) to 1 (guitarist)). In other words, more agreeable participants are more likely to be guitarists. Essentially, singers are predicted from less agreeableness and more extroversion whereas guitarists are predicted from more agreeableness and less extroversion.

Task 11

Which problem associated with logistic regression might we have in the analysis for Task 10?

Looking at the classification plot, it looks as though we might have complete separation. The model almost perfectly predicts group membership.



```
             Step number: 1

             Observed Groups and Predicted Probabilities

      80 +                                                                                                    +
         IS                                                                                                   I
         IS                                                                                                   I
F        IS                                                                                                   I
R     60 +S                                                                                                  G+
E        IS                                                                                                  GI
Q        IS                                                                                                  GI
U        IS                                                                                                  GI
E     40 +S                                                                                                  G+
N        IS                                                                                                  GI
C        IS                                                                                                  GI
Y        IS                                                                                                  GI
      20 +S                                                                                                  G+
         IS                                                                                                  GI
         IS                                                                                                  GI
         ISSS    S  S  S    S     S                        G                                   G   G    G GGGGI
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0       .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

          Predicted Probability is of Membership for Guitar
          The Cut Value is .50
          Symbols: S - Singer
                   G - Guitar
          Each Symbol Represents 5 Cases.
```

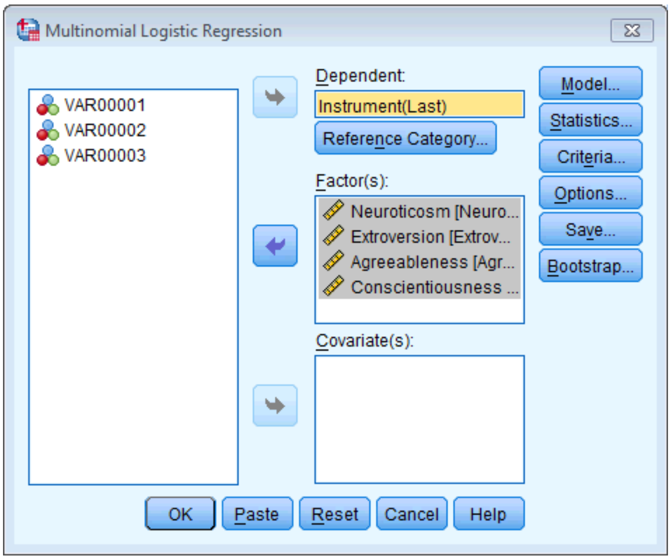
Output 34

Task 12

*The musicologist extended her study by collecting data from 430 musicians. Again she noted the **Instrument** a person played (Singer, Guitar, Bass, Drums), and the same personality variables. However, she also measured **Conscientiousness**. Conduct a multinomial logistic regression to see which of these three variables (ignore interactions) predicts which instrument a person plays (use drums as the reference category). The data are in **Band Personality.sav**.*

To run multinomial logistic regression in SPSS, first select the main dialog box by selecting Analyze Regression ▸ Multinomial Logistic... . In this dialog box (Figure) there are spaces to place the outcome variable (*Instrument*), any categorical predictors (*Factor(s)*) and any continuous predictors (*Covariate(s)*). In this example, the outcome variable is **Instrument**, so select this variable from the list and transfer it to the box labelled *Dependent* by dragging it there or clicking on ➡. We also have to tell SPSS whether we want to compare categories against the first category or the last, and we do this by clicking on Reference Category... . By default SPSS uses the last category, which is perfect for us because *drums* is the last category and is also the category that we want to use as our reference category.

Next we have to specify the predictor variables. We have three continuous predictors or covariates (**Agreeableness**, **Extroversion** and **Conscientiousness**). Drag all three to the box labelled *Covariate(s)* or click on ➡. For a basic analysis in which all of these predictors are forced into the model, this is all we really need to do.



VAR00001
VAR00002
VAR00003

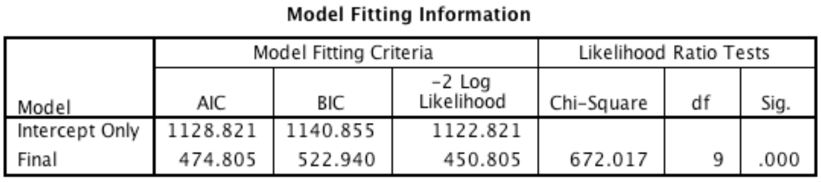
Dependent: Instrument(Last)

Factor(s):
Neuroticosm [Neuro...
Extroversion [Extrov...
Agreeableness [Agr...
Conscientiousness ...

Covariate(s):

Figure 8

Output

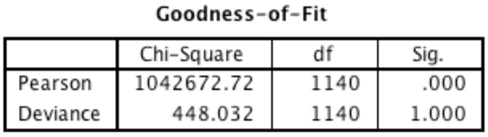


Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1128.821	1140.855	1122.821			
Final	474.805	522.940	450.805	672.017	9	.000

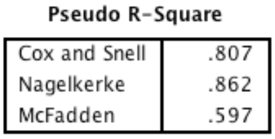
Output 35

Output 47 displays the results of the log-likelihood, which is a measure of how much unexplained variability there is in the data; therefore, the difference or change in log-likelihood indicates how much new variance has been explained by the model. The chi-square test tests the decrease in unexplained variance from the baseline model (1122.82) to the final model (450.91), which is a difference of 1149.53–871 = 672.02. This change is significant, which means that our final model explains a significant amount of the original variability (in other words, it's a better fit than the original model).



Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1042672.72	1140	.000
Deviance	448.032	1140	1.000



Pseudo R-Square

Cox and Snell	.807
Nagelkerke	.862
McFadden	.597

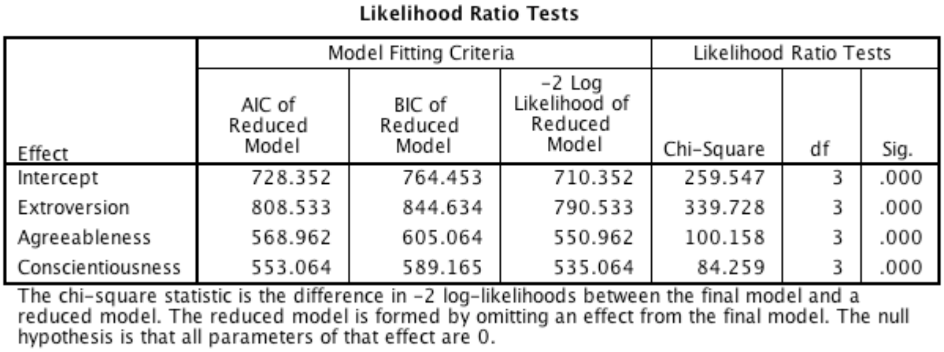
Output 36

The next part of the output relates to the fit of the model to the data (Output 48). We know that the model is significantly better than no model, but is it a good fit of the data? The Pearson and deviance statistics test the same thing, which is whether the predicted values from the model differ significantly from the observed values. If these statistics are not significant then the model is a good fit. Here we have contrasting results: the deviance statistic says that the model is a good fit of the data (p = 1.00, which is much higher than .05), but the Pearson test indicates the opposite, namely that predicted values are significantly different from the observed values (p < .001). Oh dear. Differences between these statistics can be caused by overdispersion. This is a possibility that we need to look into. We can compute the dispersion parameters from both statistics:

$$\phi_{\text{Pearson}} = \frac{\chi^2_{\text{Pearson}}}{df} = \frac{1042672.72}{1140} = 914.63$$

$$\phi_{\text{Deviance}} = \frac{\chi^2_{\text{Deviance}}}{df} = \frac{448.032}{1140} = 0.39$$

As we can see, the dispersion parameter based on the Pearson statistic is 914.63, which is ridiculously high compared to the value of 2, which I cited in the chapter as being a threshold for ‘problematic’. Conversely, the value based on the deviance statistic is below 1, which we saw in the chapter indicated underdispersion. Again, these values contradict, so all we can really be sure of is that there’s something pretty weird going on. Large dispersion parameters can occur for reasons other than overdispersion, for example omitted variables or interactions and predictors that violate the linearity of the logit assumption. In this example there were several interaction terms that we could have entered but chose not to, which might go some way to explaining these strange results. The output also shows us the two other measures of R^2. The first is Cox and Snell’s measure, which SPSS reports as .81, and the second is Nagelkerke’s adjusted value, which SPSS reports as .86. As you can see, they are reasonably similar values and represent very large effects.



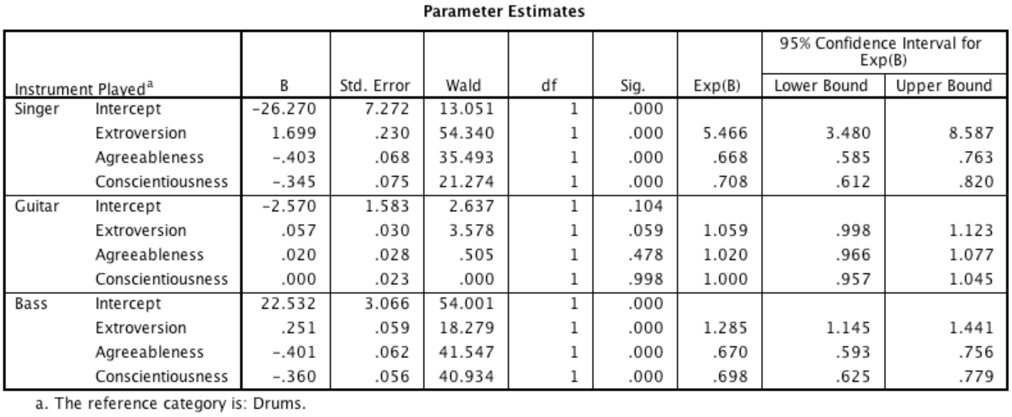
Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	728.352	764.453	710.352	259.547	3	.000
Extroversion	808.533	844.634	790.533	339.728	3	.000
Agreeableness	568.962	605.064	550.962	100.158	3	.000
Conscientiousness	553.064	589.165	535.064	84.259	3	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Output 37

The next part of the output shows the results of the likelihood ratio tests (Output 49), and these can be used to ascertain the significance of predictors to the model. This table tells us that extroversion had a significant main effect on type of instrument played, $\chi^2(3)$ = 339.73, *p* < .001, as did agreeableness, $\chi^2(3)$ = 100.16, *p* < .001, and conscientiousness, $\chi^2(3)$ = 84.26, *p* < .001. These likelihood statistics can be seen as sorts of overall statistics that tell us which predictors significantly enable us to predict the outcome category, but they don't really tell us specifically what the effect is. To see this we have to look at the individual parameter estimates (Output 50).



Parameter Estimates

Instrument Played[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Singer	Intercept	-26.270	7.272	13.051	1	.000			
	Extroversion	1.699	.230	54.340	1	.000	5.466	3.480	8.587
	Agreeableness	-.403	.068	35.493	1	.000	.668	.585	.763
	Conscientiousness	-.345	.075	21.274	1	.000	.708	.612	.820
Guitar	Intercept	-2.570	1.583	2.637	1	.104			
	Extroversion	.057	.030	3.578	1	.059	1.059	.998	1.123
	Agreeableness	.020	.028	.505	1	.478	1.020	.966	1.077
	Conscientiousness	.000	.023	.000	1	.998	1.000	.957	1.045
Bass	Intercept	22.532	3.066	54.001	1	.000			
	Extroversion	.251	.059	18.279	1	.000	1.285	1.145	1.441
	Agreeableness	-.401	.062	41.547	1	.000	.670	.593	.756
	Conscientiousness	-.360	.056	40.934	1	.000	.698	.625	.779

a. The reference category is: Drums.

Output 38

We specified the last category (drums) as our reference category; therefore, each section of Output 50 is comparing one of the instrument categories against the *drums* category. To do this SPSS recodes the instrument variable using standard dummy coding, so for each comparison, *drums* will be coded 0 and the instrument that it is being compared to (singer, guitar, or bass) will be coded as 1. Let's look at the effects one by one; because we are just comparing two categories the interpretation is the same as for binary logistic regression (so if you don't understand my conclusions reread the book chapter):

- **Extroversion**. How extroverted the participant was significantly predicted whether they were a drummer or a singer, *b* = 1.70, Wald $\chi^2(1)$ = 54.34, *p* < .001. The odds ratio tells us that as extroversion increases by one unit, the change in the odds of being a singer (rather than being a drummer) is 5.47. The odds ratio (5.47) is greater than 1, therefore we can say that as participants move up the extroversion scale, they were more likely to be a singer (coded 1) than they were to be a drummer (coded 0). Similarly, how extroverted the participant was also significantly predicted whether they were a drummer or a bass player, *b* = 0.25, Wald $\chi^2(1) = 18.28$, *p* < .001. The odds ratio tells us that as extroversion increases by one unit, the change in the odds of being a bass player (rather than being a drummer) is 1.29, so the more extroverted the participant was, the more likely they were to be a bass player than they were to be a drummer. However, how extroverted the participant was did not significantly predict whether they were a drummer or a guitarist, *b* = .06, Wald $\chi^2(1) = 3.58$, *p* = .06.

- **Agreeableness**. How agreeable the participant was significantly predicted whether they were a drummer or a singer, *b* = −0.40, Wald $\chi^2(1) = 35.49$, *p* < .001. The odds ratio tells us that as agreeableness increases by one unit, the change in the odds of being a singer (rather than being a drummer) is 0.67, so the more agreeable the participant was, the more likely they were to be a drummer than they were to be a singer. Similarly, how agreeable the participant was also significantly predicted whether they were a drummer or a bass player, *b* = −0.40, Wald $\chi^2(1) = 41.55$, *p* < .001. The odds ratio tells us that as agreeableness increases by one unit, the change in the odds of being a bass player (rather than being a drummer) is 0.67, so, the more agreeable the participant was, the more likely they were to be a drummer than they were to be a bass player. However, how agreeable the participant was did not significantly predict whether they were a drummer or a guitarist, *b* = .02, Wald $\chi^2(1) = 0.51$, *p* = .48.
- **Conscientiousness**. How conscientious the participant was significantly predicted whether they were a drummer or a singer, *b* = −0.35, Wald $\chi^2(1) = 21.27$, *p* < .001. The odds ratio tells us that as conscientiousness increases by one unit, the change in the odds of being a singer (rather than being a drummer) is 0.71, so the more conscientious the participant was, the more likely they were to be a drummer than they were to be a singer. Similarly, how conscientious the participant was also significantly predicted whether they were a drummer or a bass player, *b* = −0.36, Wald $\chi^2(1) = 40.93$, *p* < .001. The odds ratio tells us that as conscientiousness increases by one unit, the change in the odds of being a bass player (rather than being a drummer) is 0.70, so the more conscientious the participant was, the more likely they were to be a drummer than they were to be a bass player. However, how conscientious the participant was did not significantly predict whether they were a drummer or a guitarist, *b* = 0.00, Wald $\chi^2(1) = 0.00$, *p* = 1.00.