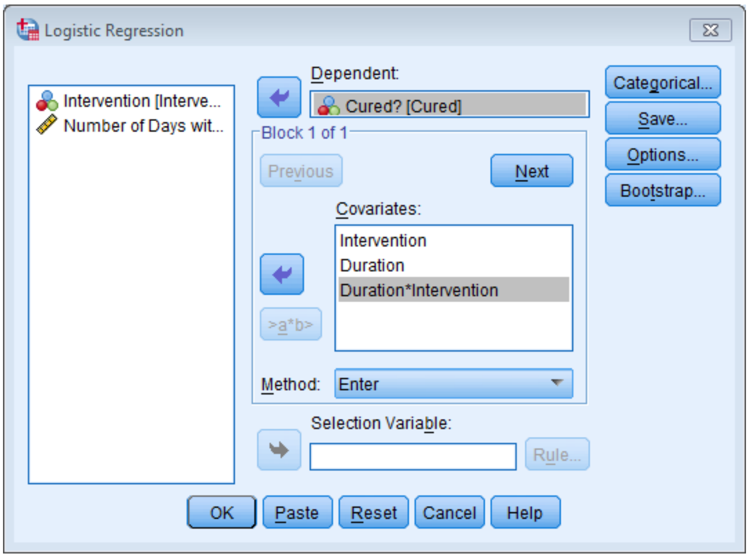
Chapter 19: Logistic regression

Self-test answers

SELF-TEST Rerun this analysis using a stepwise method (*Forward: LR*) entry method of analysis.

The main analysis

To open the main *Logistic* Regression dialog box select Analyze Regression ▸ Binary Logistic... .



Dependent: Cured? [Cured]

Covariates: Intervention, Duration, Duration*Intervention

Method: Enter

Figure 1: *Logistic Regression* main dialog box

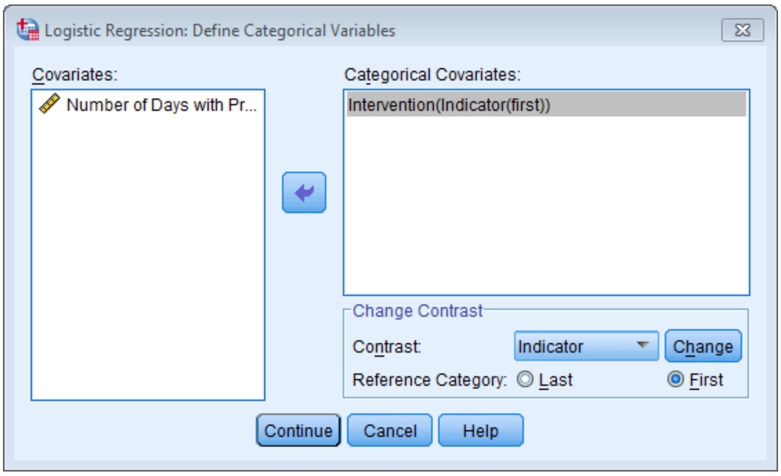
In this example, the outcome was whether or not the patient was cured, so we can simply drag **Cured** from the variable list to the *Dependent* box (or select it and click on ➡). There is also a box for specifying the covariates (the predictor variables). It is possible to specify the main effect of a predictor variable (remember, this is the effect on an outcome variable of a variable *on its own*). You can also specify an interaction effect, which is the *combined* effect (on an outcome variable) of two or more variables. To specify a main effect, select one predictor (e.g., **Duration**) and then drag it to the *Covariates* box (or click on ➡). To input an interaction, click on more than one variable on the left-hand side of the dialog box (i.e., click on several variables while holding down the *Ctrl* key, or *Cmd* on a Mac) and then click on >a*b> to move them to the *Covariates* box. In this example there are only two predictors and therefore there is only one possible interaction (the **Duration** × **Intervention** interaction), but if you have three predictors then you can select interactions using two predictors, and an interaction involving all three. In Figure 1, I have selected the two main effects of **Duration**, **Intervention** and the **Duration** × **Intervention** interaction. Select these variables too.

Method of regression

You can select a particular method of regression by clicking on Enter ▾ and then clicking on a method in the resulting drop-down menu. You were asked to do a forward stepwise analysis so select the *Forward: LR* method of regression.

Categorical predictors

SPSS needs to know which, if any, predictor variables are categorical. Click on Categorical... in the *Logistic Regression* dialog box to activate the dialog box in Figure 2. Notice that 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. Select any categorical variables you have (in this example we have only one, so click on **Intervention**) and drag them to the *Categorical Covariates* box (or click on ➡).



Logistic Regression: Define Categorical Variables

Covariates:
Number of Days with Pr...

Categorical Covariates:
Intervention(Indicator(first))

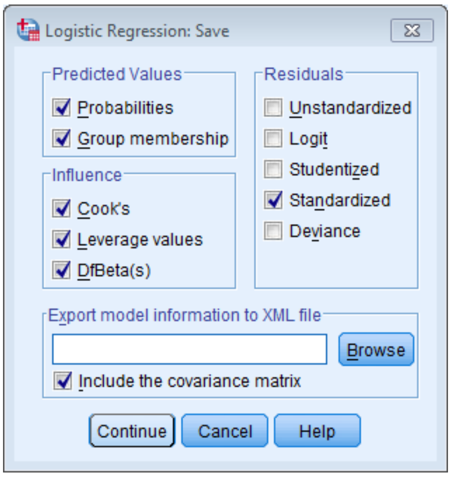
Change Contrast
Contrast: Indicator ▾
Reference Category: ○ Last ● First

Figure 2: Defining categorical variables in logistic regression

Let's use standard dummy coding (*indicator*) for this example. In our data, I coded 'cured' as 1 and 'not cured' (our control category) as 0; therefore, select the contrast, then click on ● First and then Change so that the completed dialog box looks like Figure 2.

Obtaining residuals

To save residuals click on Save... in the main *Logistic Regression* dialog box. Select the same options as in Figure 3.



Logistic Regression: Save

Predicted Values
- ☑ Probabilities
- ☑ Group membership

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

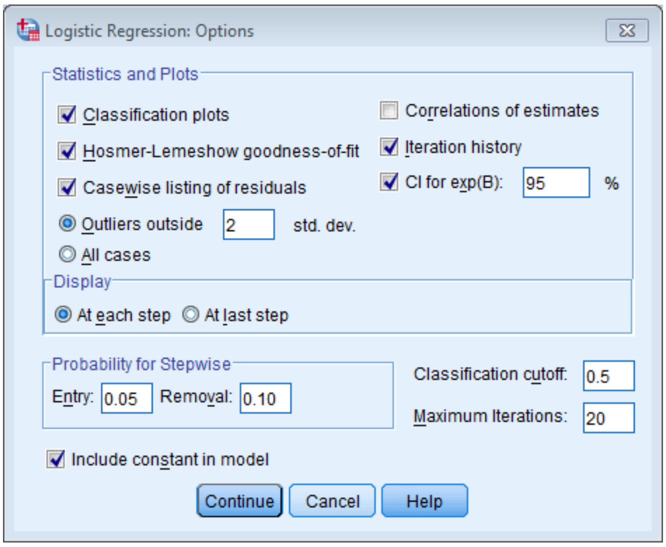
Residuals
- □ Unstandardized
- □ Logit
- □ Studentized
- ☑ Standardized
- □ Deviance

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

Figure 3: Dialog box for obtaining residuals for logistic regression

Further options

Finally, click on Options... in the main *Logistic Regression* dialog box to obtain the dialog box in Figure 4. Select the same options as in the figure.



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

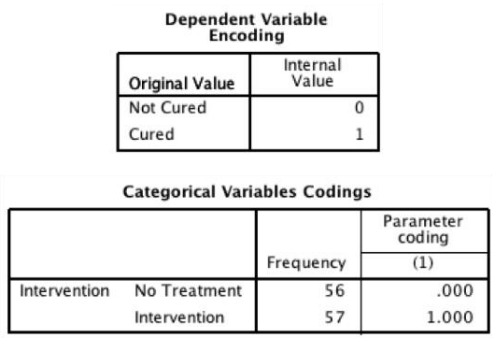
Figure 4: Dialog box for logistic regression options

Interpretation

Initial output

Output 1 tells both how we coded our outcome variable (it reminds us that 0 = not cured and 1 = cured) and how it has coded the categorical predictors (the parameter codings for **Intervention**). We chose indicator coding and so the coding is the same as the values in the data editor (0 = no treatment, 1 = treatment). If *deviation* coding had been chosen then the coding would have been −1 (treatment) and 1 (no treatment). With a *simple* contrast, if ● First had been selected as the reference category the codes would have been −0.5 (**Intervention** = no treatment) and 0.5 (**Intervention** = treatment). and if

● Last had been selected as the reference category then the value of the codes would have been the same but with their signs reversed. The parameter codes are important for calculating the probability of the outcome variable ($P(Y)$), but we will come to that later.



Dependent Variable Encoding

Original Value	Internal Value
Not Cured	0
Cured	1

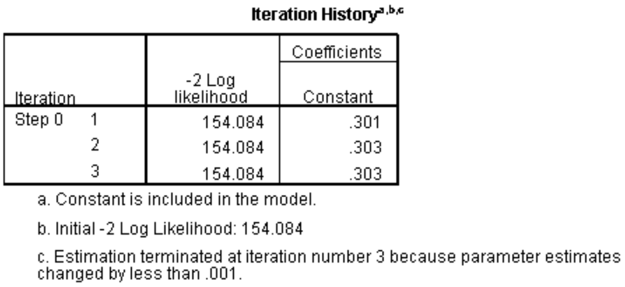
Categorical Variables Codings

		Frequency	Parameter coding
			(1)
Intervention	No Treatment	56	.000
	Intervention	57	1.000

Output 1

For this first analysis we requested a forward stepwise method[1] and so the initial model is derived using only the constant in the regression equation. Output 2 tells us about the model when only the constant is included (i.e., all predictor variables are omitted). The table labelled *Iteration History* tells us that the log-likelihood of this baseline model is 154.08. This represents the fit of the most basic model 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 the patient was cured, or that every patient was not cured. 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 patient belongs to the category in which most observed cases fell. In this example there were 65 patients who were cured, and only 48 who were not cured. Therefore, if SPSS predicts that every patient was cured then this prediction will be correct 65 times out of 113 (i.e., about 58% of the time). However, if SPSS predicted that every patient was not cured, then this prediction would be correct only 48 times out of 113 (42% of the time). As such, of the two available options it is better to predict that all patients were cured 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 patients are cured, which results in 0% accuracy for the patients who were not cured, and 100% accuracy for those observed to be cured. Overall, the model correctly classifies 57.5% of patients.

[1] Actually, this is a *really* bad idea when you have an interaction term because to look at an interaction you need to include the main effects of the variables in the interaction term. I chose this method *only* to illustrate how stepwise methods work.



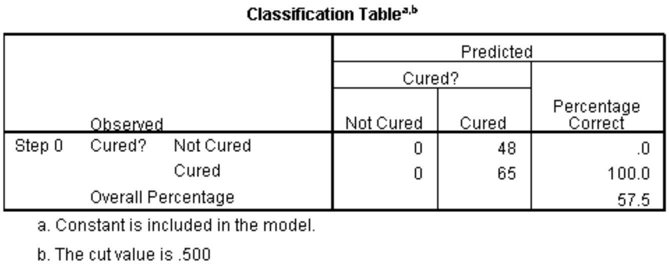
Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	154.084	.301
	2	154.084	.303
	3	154.084	.303

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 154.084

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



Classification Table[a,b]

Observed			Predicted Cured? Not Cured	 Cured	 Percentage Correct
Step 0	Cured?	Not Cured	0	48	.0
		Cured	0	65	100.0
	Overall Percentage				57.5

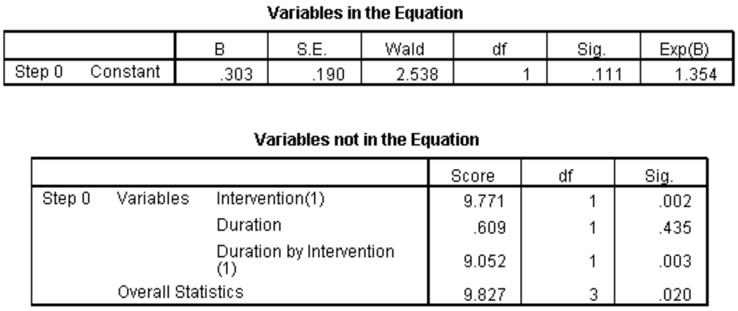
a. Constant is included in the model.

b. The cut value is .500

Output 2

Output 3 summarizes the model (*Variables in the Equation*), and at this stage this entails quoting the value of the constant (b_0), which is equal to 0.30. The table labelled *Variables not in the Equation* tells us that the residual chi-square statistic is 9.83 which is significant at $p < .05$ (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 forcing all of the variables excluded from the model into the model would not have made a significant contribution to its predictive power.

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, **Intervention** and the **Intervention × Duration** interaction both have significant score statistics at $p < .01$ and could potentially make a contribution to the model, but **Duration** alone does not look likely to be a good predictor because its score statistic is non-significant, $p > .05$. As mentioned earlier, 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 has a significance below .05. In this example, that variable will be **Intervention** because its score statistic (9.77) is the biggest.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.303	.190	2.538	1	.111	1.354

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Intervention(1)	9.771	1	.002
		Duration	.609	1	.435
		Duration by Intervention (1)	9.052	1	.003
	Overall Statistics		9.827	3	.020

Output 3

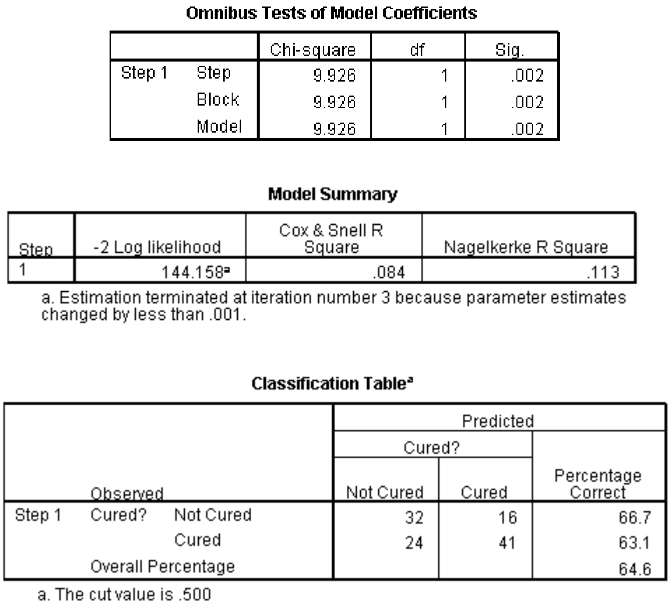
Step 1: Intervention

As I predicted in the previous section, whether or not an intervention was given to the patient (**Intervention**) is added to the model in the first step. As such, a patient is now classified as being cured or not based on whether they had an intervention or not (waiting list). This can be explained easily if we look at the crosstabulation for the variables **Intervention** and **Cured**. The model will use whether a patient had an intervention or not to predict whether they were cured or not by applying the crosstabulation table shown in Table 1.

Table 1: Crosstabulation of intervention with outcome status (cured or not)

		Intervention or Not (Intervention)	
		No Treatment	**Intervention**
Cured? (**Cured**)	**Not Cured**	32	16
	Cured	24	41
	Total	**56**	**57**

The model predicts that all of the patients who had an intervention were cured. There were 57 patients who had an intervention, so the model predicts that these 57 patients were cured; it is correct for 41 of these patients, but misclassifies 16 people as 'cured' who were not cured – see Table 1. In addition, this new model predicts that all of the 56 patients who received no treatment were not cured; for these patients the model is correct 32 times but misclassifies as 'not cured' 24 people who were.



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.926	1	.002
	Block	9.926	1	.002
	Model	9.926	1	.002

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	144.158[a]	.084	.113

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

Classification Table[a]

			Predicted		
			Cured?		
Observed			Not Cured	Cured	Percentage Correct
Step 1	Cured?	Not Cured	32	16	66.7
		Cured	24	41	63.1
	Overall Percentage				64.6

a. The cut value is .500

Output 4

Output 4 shows summary statistics about the new model (which we've already seen contains **Intervention**). 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 $-2LL$): this multiplication is done because $-2LL$ has an approximately chi-square distribution and so it 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.

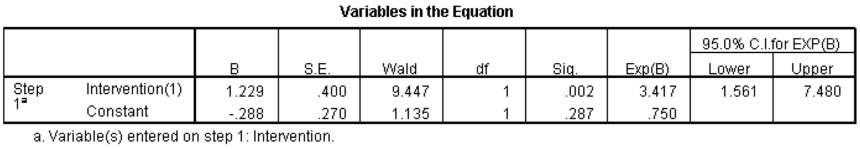
At this stage of the analysis the value of $-2LL$ should be less than the value when only the constant was included in the model (because lower values of $-2LL$ indicate that the model is predicting the outcome variable more accurately). When only the constant was included, $-2LL$ = 154.08, but now **Intervention** has been included this value has been reduced to 144.16. This reduction tells us that the model is better at predicting whether someone was cured than it was before **Intervention** 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 $-2LL$ with **Intervention** included minus the value of $-2LL$ when only the constant was in the model (154.08 − 144.16 = 9.92). This value has a chi-square distribution and so its statistical significance can be calculated easily.[2] In this example, the value is significant

[2] The degrees of freedom will be the number of parameters in the new model (the number of predictors plus 1, which in this case with one predictor, means 2) minus the number of parameters in the baseline model (which is 1, the constant). So, in this case, *df* = 2 – 1 = 1.

at the .05 level and so we can say that overall the model is predicting whether a patient is cured or not 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. In an ideal world we would like to see a non-significant overall $-2LL$ (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 a measure of the improvement of the predictive power of the model since the last step. Its value is equal to $-2LL$ at the current step minus $-2LL$ 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 $-2LL$ since the last block (for use in hierarchical or blockwise analyses).

Output 4 also tells us the values of Cox and Snell's and Nagelkerke's R^2, but we will discuss these a little later. There is also a classification table that indicates how well the model predicts group membership; because the model is using **Intervention** to predict the outcome variable, this classification table is the same as Table 1. The current model correctly classifies 32 patients who were not cured but misclassifies 16 others (it correctly classifies 66.7% of cases). The model also correctly classifies 41 patients who were cured but misclassifies 24 others (it correctly classifies 63.1% of cases). The overall accuracy of classification is, therefore, the weighted average of these two values (64.6%). So, when only the constant was included, the model correctly classified 57.5% of patients, but now, with the inclusion of **Intervention** as a predictor, this has risen to 64.6%.



Variables in the Equation

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Intervention(1)	1.229	.400	9.447	1	.002	3.417	1.561	7.480
	Constant	-.288	.270	1.135	1	.287	.750		

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

Output 5

The next part of the output (Output 5) is crucial because it tells us the estimates for the coefficients for the predictors included in the model. This section of the output gives us the coefficients and statistics for the variables that have been included in the model at this point (namely **Intervention** and the constant). The b-value is the same as the b-value in linear regression: they are the values that we need to replace in the regression equation to establish the probability that a case falls into a certain category. We saw in linear regression that the value of *b* represents the change in the outcome resulting from a unit change in the predictor variable. The interpretation of this coefficient in logistic regression is very similar in 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[3] 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*). The Wald statistic should be used cautiously because when the regression coefficient (*b*) is large, the standard error tends to become inflated, resulting in the Wald statistic being underestimated (see Menard, 1995). However, for these data it seems to indicate that having the intervention (or not) is a significant predictor of whether the patient is cured (note that the significance of the Wald statistic is less than .05).

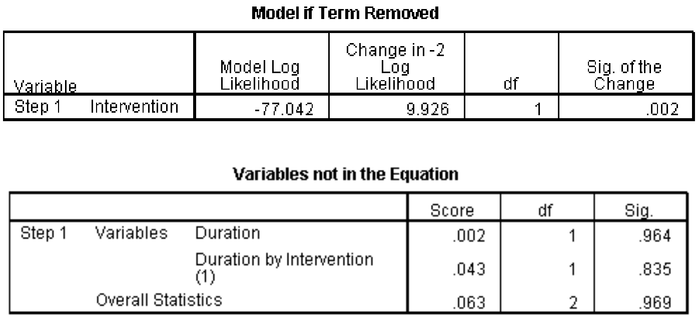
You should notice that the odds ratio is what SPSS reports as Exp(*B*). The odds ratio is the change in odds; if the value is greater than 1 then it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease. In this example, we can say that the odds of a patient who is treated being cured are 3.41 times higher than those of a patient who is not treated.

In the options, we requested a confidence interval for the odds ratio and it can also be found in the output. As with any confidence interval it is computed such that if we calculated confidence intervals for the value of the odds ratio in 100 different samples, then these intervals would include value of the odds ratio in the population in 95 of those samples. Assuming the current sample is one of the 95 for which the confidence interval contains the true value, then we know that the population value of the odds ratio lies between 1.56 and 7.48. However, our sample could be one of the 5% that produces a confidence interval that 'misses' the population value.

The important thing about this confidence interval is that it doesn't cross 1 (both values are greater than 1). This is important because values greater than 1 mean that as the predictor variable increases, so do the odds of (in this case) being cured. Values less than 1 mean the opposite: as the predictor variable increases, the odds of being cured decrease. The fact that both limits of our confidence interval are above 1 gives us confidence that the direction of the relationship that we have observed is true in the population (i.e. it's likely that having an intervention compared to not increases the odds of being cured). If the lower limit had been below 1 then it would tell us that there is a chance that in the population the direction of the relationship is the opposite to what we

[3] As we have seen, this is simply *b* divided by its standard error (1.229/0.40 = 3.0725); however, SPSS actually quotes the Wald statistic squared. For these data $3.0725^2 = 9.44$ as reported (within rounding error) in the table.

have observed. This would mean that we could not trust that our intervention increases the odds of being cured.



Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	Intervention	-77.042	9.926	1	.002

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	Duration	.002	1	.964
		Duration by Intervention (1)	.043	1	.835
	Overall Statistics		.063	2	.969

Output 6

The test statistics for **Intervention** if it were removed from the model are in Output 6. Now, remember that earlier on I said how the regression would place variables into the equation and then test whether they then met a removal criterion. Well, the *Model if Term Removed* part of the output tells us the effects of removal. The important thing to note is the significance value of the log-likelihood ratio. The log-likelihood ratio for this model is significant (p < .01) which tells us that removing **Intervention** 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 .964 for **Duration** and .835 for the **Duration×Intervention** interaction). Therefore, no further variables will be added to the model.

SELF-TEST Calculate the values of Cox and Snell's and Nagelkerke's R^2 reported by SPSS. (Remember the sample size, N, is 113.)

Cox and Snell's R^2 is calculated from this equation:

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

SPSS reports $-2LL$(new) as 144.16 and $-2LL$(baseline) as 154.08. The sample size, N, is 113. So

$$R^2_{CS} = 1 - \exp\left(\frac{144.16 - 154.08}{113}\right)$$

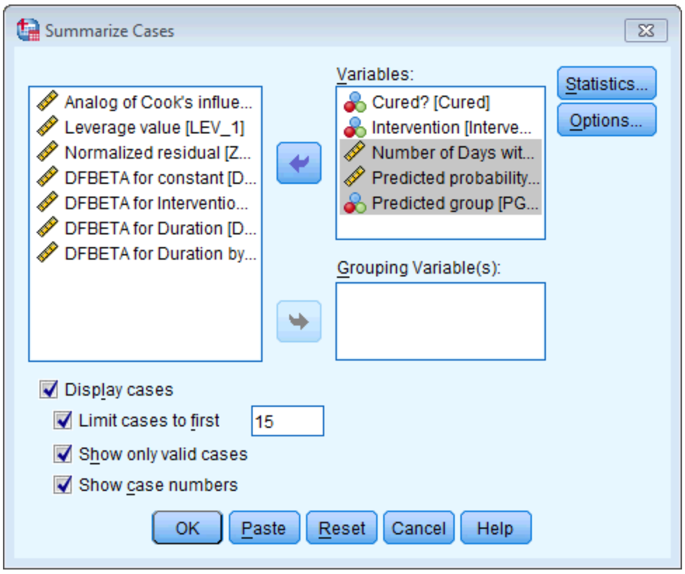
$$
\begin{aligned}
&= 1 - \exp(-0.0878) \\
&= 1 - e^{-0.0878} \\
&= .084
\end{aligned}
$$

Nagelkerke's adjustment is calculated from:

$$
\begin{aligned}
R_N^2 &= \frac{R_{CS}^2}{1 - \exp\left(-\frac{-2LL(\text{baseline})}{n}\right)} \\
&= \frac{0.084}{1 - \exp\left(-\frac{154.08}{113}\right)} \\
&= \frac{0.084}{1 - e^{-1.3635}} \\
&= \frac{0.084}{1 - 0.2558} \\
&= .113
\end{aligned}
$$

SELF-TEST Use the *case summaries* function in SPSS to create a table for the first 15 cases in the file **Eel.sav** showing the values of **Cured**, **Intervention**, **Duration**, the predicted probability (**PRE_1**) and the predicted group membership (**PGR_1**) for each case.

The completed dialog box should look like this:



Variables:
- Cured? [Cured]
- Intervention [Interve...
- Number of Days wit...
- Predicted probability...
- Predicted group [PG...

Grouping Variable(s):

☑ Display cases

☑ Limit cases to first 15

☑ Show only valid cases

☑ Show case numbers

Summarize Cases
Analog of Cook's influe...
Leverage value [LEV_1]
Normalized residual [Z...
DFBETA for constant [D...
DFBETA for Interventio...
DFBETA for Duration [D...
DFBETA for Duration by...
Statistics...
Options...
OK
Paste
Reset
Cancel
Help

Figure 5

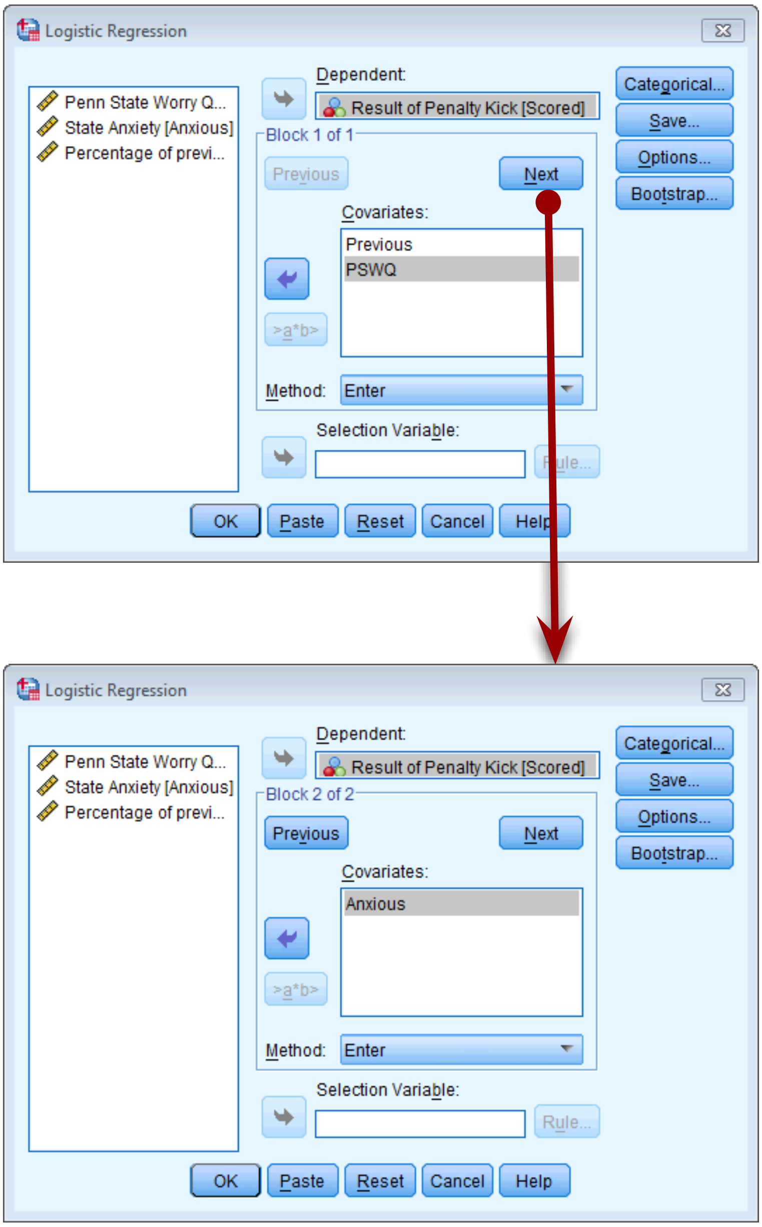
SELF-TEST Conduct a hierarchical logistic regression analysis on these data. Enter **Previous** and **PSWQ** in the first block and **Anxious** in the second (forced entry).

Running the analysis: block entry regression

To run the analysis, we must first bring up the main *Logistic Regression* dialog box, by selecting Analyze Regression ▸ Binary Logistic... . In this example, we know of two previously established predictors and so it is a good idea to enter these predictors into the model in a single block. Then we can add the new predictor in a second block (by doing this we effectively examine an old model and then add a new variable to this model to see whether the model is improved). This method is known as block entry and Figure 6 shows how it is specified.

It is easy to do block entry regression. First you should use the mouse to select the variable **scored** from the variables list and then transfer it to the box labelled *Dependent* by clicking on ➡. Second, you should select the two previously established predictors. So, select **PSWQ** and **Previous** from the variables list and transfer them to the box labelled *Covariates* by clicking on ➡. Our first block of variables is now specified. To specify the second block, click on Next to clear the *Covariates* box, which should now be labelled *Block 2 of 2*. Now select **Anxious** from the variables list and transfer it to the box labelled *Covariates* by clicking on ➡. We could at this stage select some interactions to be included in the model, but unless there is a sound theoretical reason for believing that the predictors should interact there is no need. Make sure that *Enter* is selected as the method of regression (this method is the default and so should be selected already).

Once the variables have been specified, you should select the options described in the chapter, but because none of the predictors are categorical there is no need to use the Categorical... option. When you have selected the options and residuals that you want you can return to the main *Logistic Regression* dialog box and click on OK.



Logistic Regression

Dependent: Result of Penalty Kick [Scored]

Block 1 of 1

Covariates:
Previous
PSWQ

Method: Enter

Logistic Regression

Dependent: Result of Penalty Kick [Scored]

Block 2 of 2

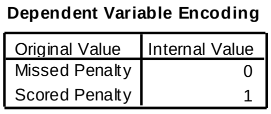
Covariates:
Anxious

Method: Enter

Figure 6

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.

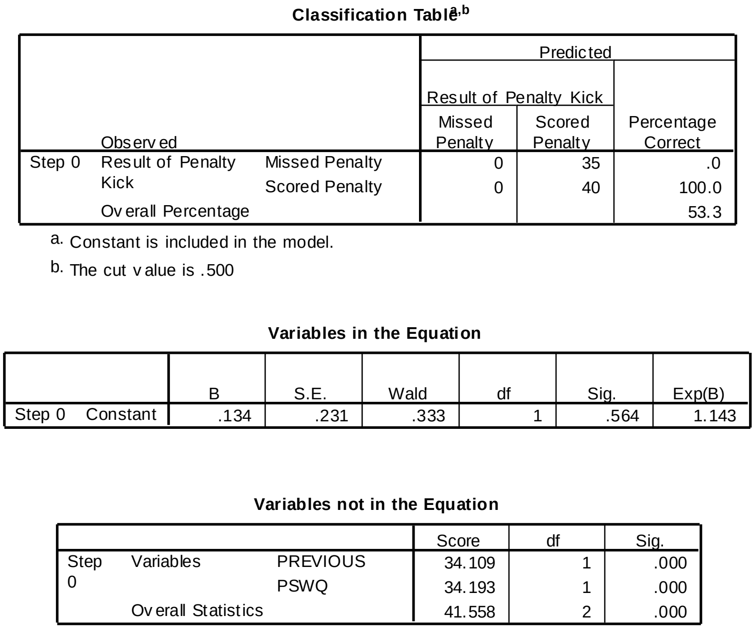
First, the output shows the results from block 0: the output tells us that 75 cases have been accepted, and that the dependent variable has been coded 0 and 1 (because this variable was coded as 0 and 1 in the data editor, these codings correspond exactly to the data in SPSS). We are then told about the variables that are in and out of the equation. At this point only the constant is included in the model, and so to be perfectly honest none of this information is particularly interesting!



Dependent Variable Encoding

Original Value	Internal Value
Missed Penalty	0
Scored Penalty	1

Output 7



Classification Table[a,b]

			Predicted		
			Result of Penalty Kick		
Observed			Missed Penalty	Scored Penalty	Percentage Correct
Step 0	Result of Penalty Kick	Missed Penalty	0	35	.0
		Scored Penalty	0	40	100.0
	Overall Percentage				53.3

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	.134	.231	.333	1	.564	1.143

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	PREVIOUS	34.109	1	.000
		PSWQ	34.193	1	.000
	Overall Statistics		41.558	2	.000

Output 8

The results from block 1 are shown next, and in this analysis we forced SPSS to enter **Previous** and **PSWQ** into the regression model. Therefore, this part of the output provides information about the model after the variables **Previous** and **PSWQ** have been added. The first thing to note is that $-2LL$ is 48.66, which is a change of 54.98 (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 (p < .001), and so using previous experience and worry as predictors significantly improves our ability to predict penalty success. A bit further down, the classification table shows us that 84% of cases can be correctly classified using **PSWQ** and **Previous**.

In the intervention example, Hosmer and Lemeshow's goodness-of-fit test was 0. The reason is that this test can't be calculated when there is only one predictor and that predictor is a categorical dichotomy! However, for this example the test can be calculated. The important part of this test is the test statistic itself (7.93) and the significance value (.3388). This 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). We have a non-significant value here, which is indicative of a model that is predicting the real-world data fairly well.

The part of the output labelled *Variables in the Equation* then tells us the parameters of the model when **Previous** and **PSWQ** are used as predictors. The significance values of the Wald statistics for each predictor indicate that both **PSWQ** and **Previous** significantly predict penalty success ($p < .01$). The value of the odds ratio (Exp(B)) for **Previous** indicates that if the percentage of previous penalties scored goes up by one, then the odds of scoring a penalty also increase (because the odds ratio is greater than 1). The confidence interval for this value ranges from 1.02 to 1.11, so we can be very confident that the value of the odds ratio 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 **Previous** and penalty success found in this sample is true of the whole population of footballers. The odds ratio for **PSWQ** indicates that if the level of worry increases by one point along the Penn State worry scale, then the odds of scoring a penalty decrease (because it is less than 1). The confidence interval for this value ranges from .68 to .93 so we can be very confident that the value of the odds ratio 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 **PSWQ** and penalty success found in this sample is true of the whole population of footballers. If we had found that the confidence interval ranged from less than 1 to more than 1, then this would limit the generalizability of our findings because the odds ratio in the population could indicate either a positive (odds ratio > 1) or negative (odds ratio < 1) relationship.

A glance at the classification plot also brings us good news because most cases are clustered at the ends of the plot and few cases lie in the middle of the plot. This reiterates what we know already: that the model is correctly classifying most cases. We can, at this point, also calculate R^2 by dividing the model chi-square by the original value of $-2LL$. The result is:

$$R^2 = \frac{\text{model chi-square}}{\text{original} - 2LL} = \frac{54.977}{103.6385} = .53$$

We can interpret the result as meaning that the model can account for 53% of the variance in penalty success (so, roughly half of what makes a penalty kick successful is still unknown).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	54.977	2	.000
	Block	54.977	2	.000
	Model	54.977	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	48.662	.520	.694

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.931	7	.339

Contingency Table for Hosmer and Lemeshow Test

		Result of Penalty Kick = Missed Penalty		Result of Penalty Kick = Scored Penalty		
		Observed	Expected	Observed	Expected	Total
Step 1	1	8	7.904	0	.096	8
	2	8	7.779	0	.221	8
	3	8	6.705	0	1.295	8
	4	4	5.438	4	2.562	8
	5	2	3.945	6	4.055	8
	6	2	1.820	6	6.180	8
	7	2	1.004	6	6.996	8
	8	1	.298	7	7.702	8
	9	0	.108	11	10.892	11

Classification Table[a]

			Predicted		
			Result of Penalty Kick		
	Observed		Missed Penalty	Scored Penalty	Percentage Correct
Step 1	Result of Penalty Kick	Missed Penalty	30	5	85.7
		Scored Penalty	7	33	82.5
	Overall Percentage				84.0

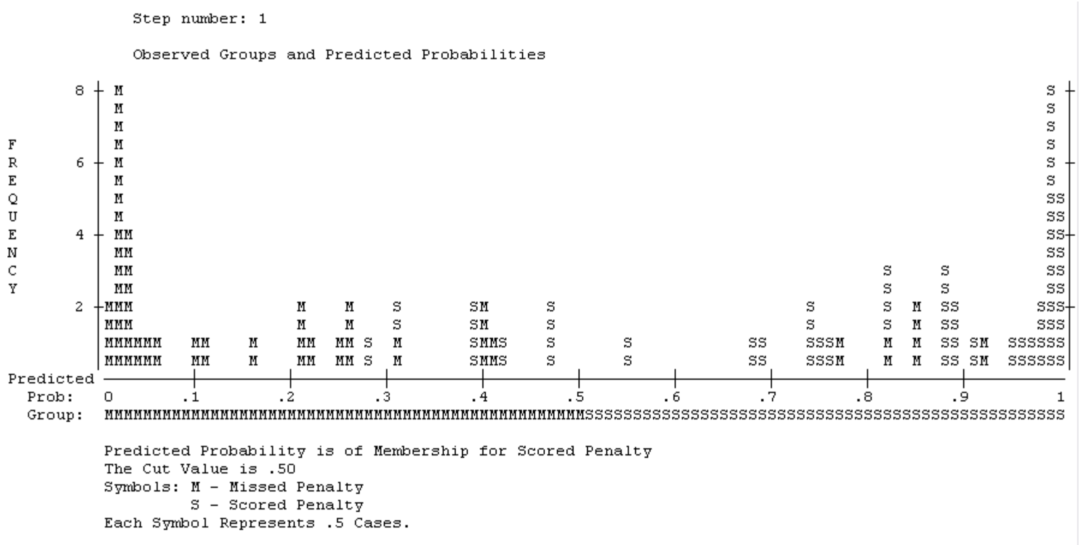
a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PREVIOUS	.065	.022	8.609	1	.003	1.067	1.022	1.114
	PSWQ	-.230	.080	8.309	1	.004	.794	.679	.929
	Constant	1.280	1.670	.588	1	.443	3.598		

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

Output 9

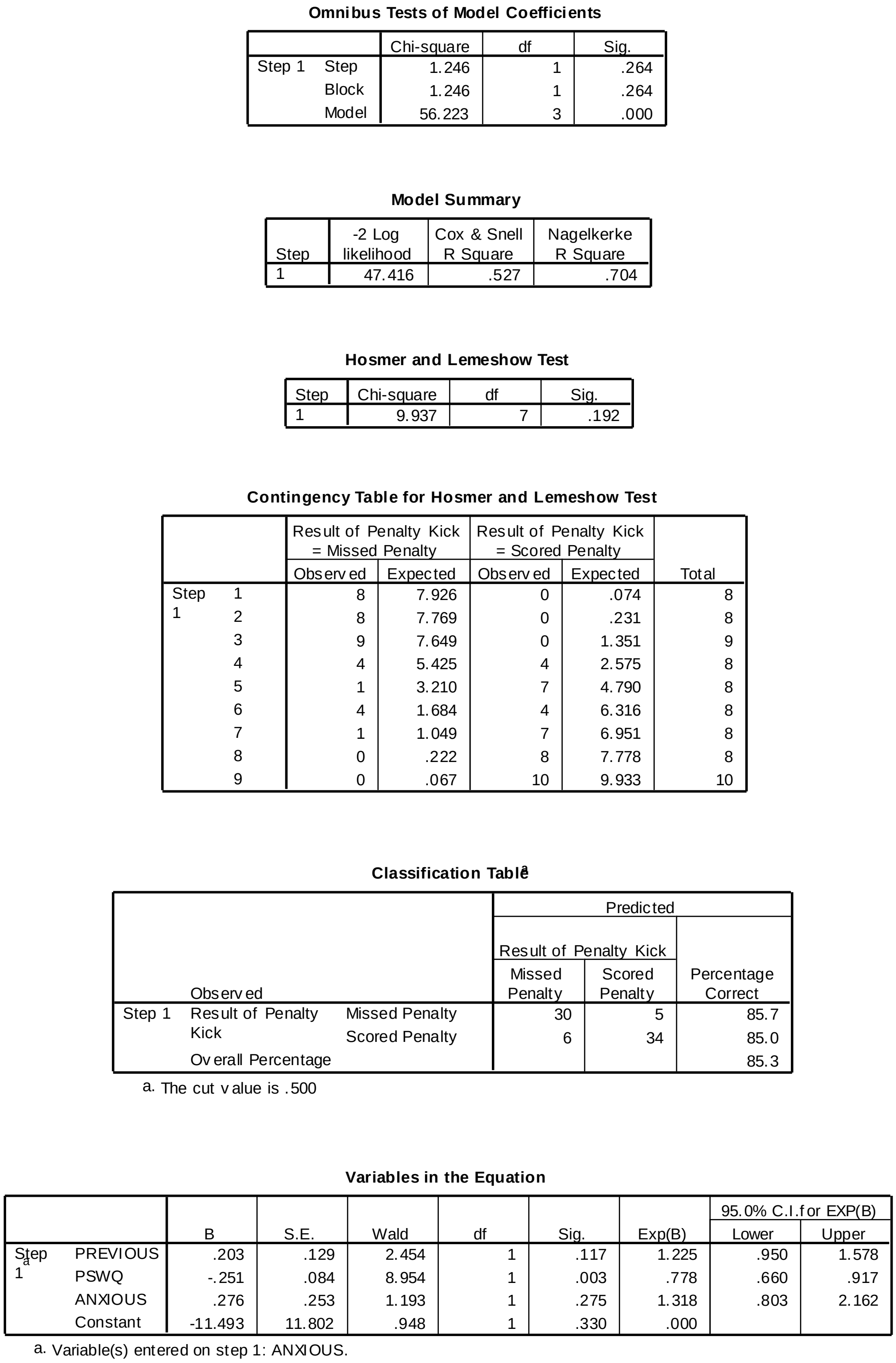


```
             Step number: 1

             Observed Groups and Predicted Probabilities

       8 + M                                                                                                S +
         | M                                                                                                S |
         | M                                                                                                S |
F        | M                                                                                                S |
R      6 + M                                                                                                S +
E        | M                                                                                                S |
Q        | M                                                                                                SS|
U        | M                                                                                                SS|
E      4 + MM                                                                                               SS+
N        | MM                                                                                               SS|
C        | MM                                                                              S     S          SS|
Y        | MM                                                                              S     S          SS|
       2 +MMM                 M    M    S       SM      S                          S       S  M  SS        SSS+
         |MMM                 M    M    S       SM      S                          S       S  M  SS        SSS|
         |MMMMMM   MM    M    MM  MM S  M       SMMS    S       S            SS    SSSM    M  M  SS SM  SSSSSS|
         |MMMMMM   MM    M    MM  MM S  M       SMMS    S       S            SS    SSSM    M  M  SS SM  SSSSSS|
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0       .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS

          Predicted Probability is of Membership for Scored Penalty
          The Cut Value is .50
          Symbols: M - Missed Penalty
                   S - Scored Penalty
          Each Symbol Represents .5 Cases.
```

Output 10

The output for block 2 shows what happens to the model when our new predictor is added (**Anxious**). So, we begin with the model that we had in block 1 and we add **Anxious** to it. The effect of adding **Anxious** to the model is to reduce −2*LL* to 47.416 (a reduction of 1.246 from the model in block 1 as shown in the *model chi-square* and *block* statistics). This improvement is non-significant, which tells us that including **Anxious** in the model has not significantly improved our ability to predict whether a penalty will be scored or missed. The classification table tells us that the model is now correctly classifying 85.33% of cases. Remember that in block 1 there were 84% correctly classified and so an extra 1.33% of cases are now classified (not a great deal more – in fact, examining the table shows us that only one extra case has now been correctly classified).

The table labelled *Variables in the Equation* now contains all three predictors and something very interesting has happened: **PSWQ** is still a significant predictor of penalty success; however, **Previous** experience no longer significantly predicts penalty success. In addition, state anxiety appears not to make a significant contribution to the prediction of penalty success. How can it be that previous experience no longer predicts penalty success, and neither does anxiety, yet the ability of the model to predict penalty success has improved slightly?



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1.246	1	.264
	Block	1.246	1	.264
	Model	56.223	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	47.416	.527	.704

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.937	7	.192

Contingency Table for Hosmer and Lemeshow Test

		Result of Penalty Kick = Missed Penalty		Result of Penalty Kick = Scored Penalty		
		Observed	Expected	Observed	Expected	Total
Step 1	1	8	7.926	0	.074	8
	2	8	7.769	0	.231	8
	3	9	7.649	0	1.351	9
	4	4	5.425	4	2.575	8
	5	1	3.210	7	4.790	8
	6	4	1.684	4	6.316	8
	7	1	1.049	7	6.951	8
	8	0	.222	8	7.778	8
	9	0	.067	10	9.933	10

Classification Tableª

			Predicted		
			Result of Penalty Kick		
	Observed		Missed Penalty	Scored Penalty	Percentage Correct
Step 1	Result of Penalty Kick	Missed Penalty	30	5	85.7
		Scored Penalty	6	34	85.0
	Overall Percentage				85.3

a. The cut value is .500

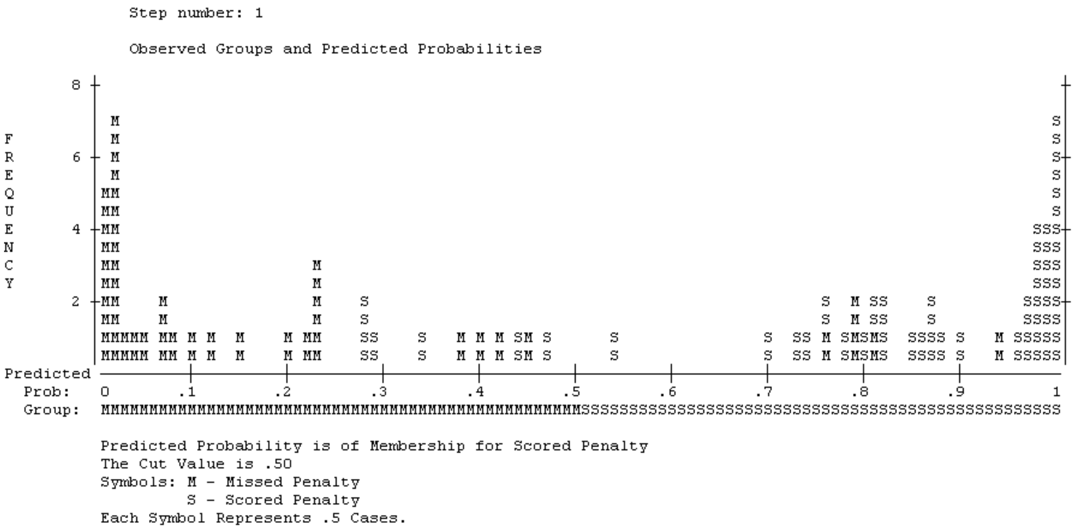
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1ª	PREVIOUS	.203	.129	2.454	1	.117	1.225	.950	1.578
	PSWQ	-.251	.084	8.954	1	.003	.778	.660	.917
	ANXIOUS	.276	.253	1.193	1	.275	1.318	.803	2.162
	Constant	-11.493	11.802	.948	1	.330	.000		

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

Output 11

The classification plot is similar to before and the contribution of **PSWQ** to predicting penalty success is relatively unchanged. What has changed is the contribution of previous experience. If we examine the values of the odds ratio for both **Previous** and **Anxious** it is clear that they both potentially have a positive relationship to penalty success (i.e., as they increase by a unit, the odds of scoring improve). However, the confidence intervals for these values cross 1, which indicates that the direction of this relationship may be unstable in the population as a whole (i.e., the value of the odds ratio in our sample may be quite different from the value if we had data from the entire population).



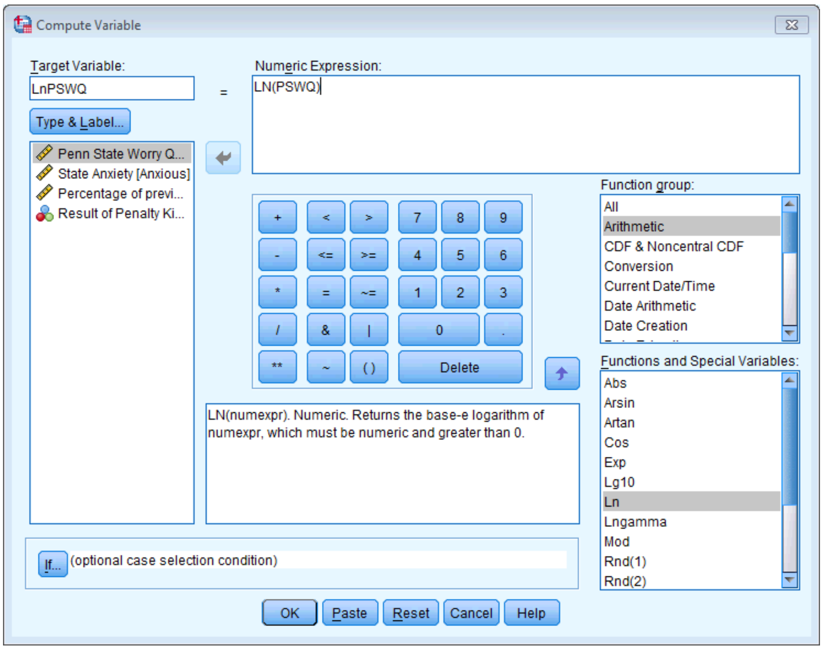
```
             Step number: 1

             Observed Groups and Predicted Probabilities

       8 +                                                                                                    +
         |                                                                                                    |
         | M                                                                                               S|
F        | M                                                                                               S|
R      6 + M                                                                                               S+
E        | M                                                                                               S|
Q        |MM                                                                                               S|
U        |MM                                                                                               S|
E      4 +MM                                                                                             SSS+
N        |MM                                                                                             SSS|
C        |MM                    M                                                                        SSS|
Y        |MM                    M                                                                        SSS|
       2 +MM    M               M    S                                               S  M SS    S         SSSS+
         |MM    M               M    S                                               S  M SS    S         SSSS|
         |MMMMM MM M M  M    M MM    SS    S   M M M SM S      S               S  SS M SMSMS  SSSS S   M SSSSS|
         |MMMMM MM M M  M    M MM    SS    S   M M M SM S      S               S  SS M SMSMS  SSSS S   M SSSSS|
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0       .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS

          Predicted Probability is of Membership for Scored Penalty
          The Cut Value is .50
          Symbols: M - Missed Penalty
                   S - Scored Penalty
          Each Symbol Represents .5 Cases.
```

Output 12

You may be tempted to use this final model to say that, although worry is a significant predictor of penalty success, the previous finding that experience plays a role is incorrect. This would be a dangerous conclusion to draw, and if you read the section on multicollinearity in the book you'll see why.

SELF-TEST Try creating two new variables that are the natural logs of **Anxious** and **Previous**.

First of all, the completed dialog box for **PSWQ** is given in Figure 7 to give you some idea of how this variable is created (following the instructions in the chapter).



Compute Variable

Target Variable: LnPSWQ

Numeric Expression: LN(PSWQ)

Penn State Worry Q...
State Anxiety [Anxious]
Percentage of previ...
Result of Penalty Ki...

Function group: All, Arithmetic, CDF & Noncentral CDF, Conversion, Current Date/Time, Date Arithmetic, Date Creation

Functions and Special Variables: Abs, Arsin, Artan, Cos, Exp, Lg10, Ln, Lngamma, Mod, Rnd(1), Rnd(2)

LN(numexpr). Numeric. Returns the base-e logarithm of numexpr, which must be numeric and greater than 0.

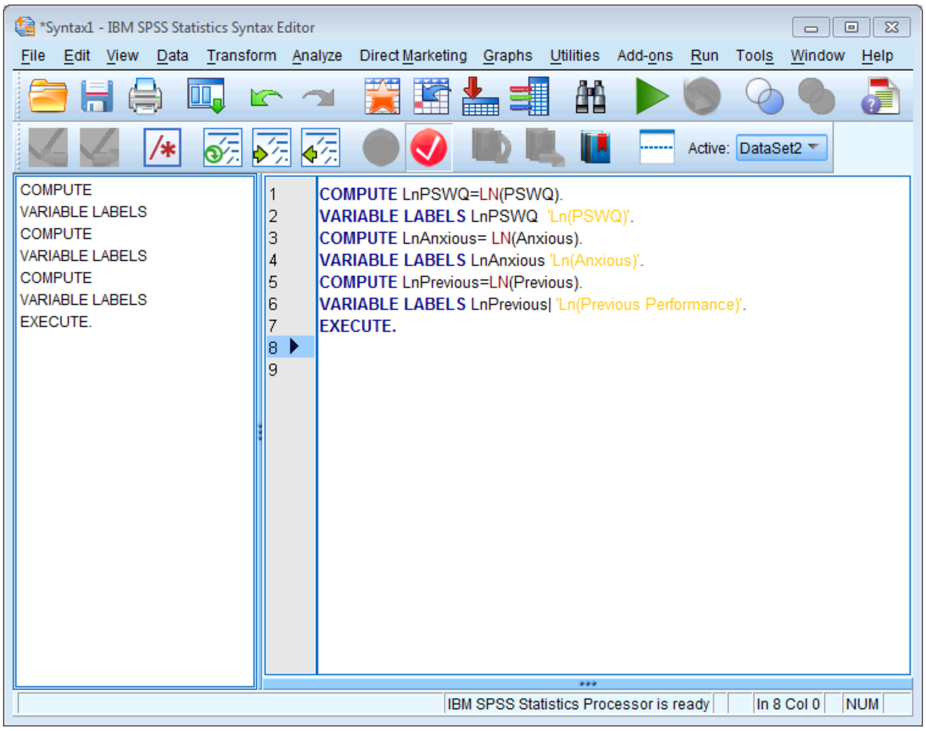
If... (optional case selection condition)

Figure 7

For **Anxious**, create a new variable called **LnAnxious** by entering this name into the box labelled *Target Variable* and then click on Type & Label... and give the variable a more descriptive name such as *Ln(anxiety)*. In the list box labelled *Function group*, click on *Arithmetic* and then in the box labelled *Functions and Special Variables* click on *Ln* (this is the natural log transformation) and transfer it to the command area by clicking on ↑. Replace the question mark with the variable **Anxious** by either selecting the variable in the list and clicking on → or just typing 'Anxious' where the question mark is. Click on OK to create the variable.

For **Previous**, create a new variable called **LnPrevious** by entering this name into the box labelled *Target Variable* and then click on Type & Label... and give the variable a more descriptive name such as *Ln(previous performance)*. In the list box labelled *Function group*, click on *Arithmetic* and then in the box labelled *Functions and Special Variables* click on *Ln* and transfer it to the command area by clicking on ↑. Replace the question mark with the variable **Previous** by either selecting the variable in the list and clicking on → or just typing 'Previous' where the question mark is. Click on OK to create the variable.

Alternatively, you can create all three variables in one go using the syntax shown in Figure 8.

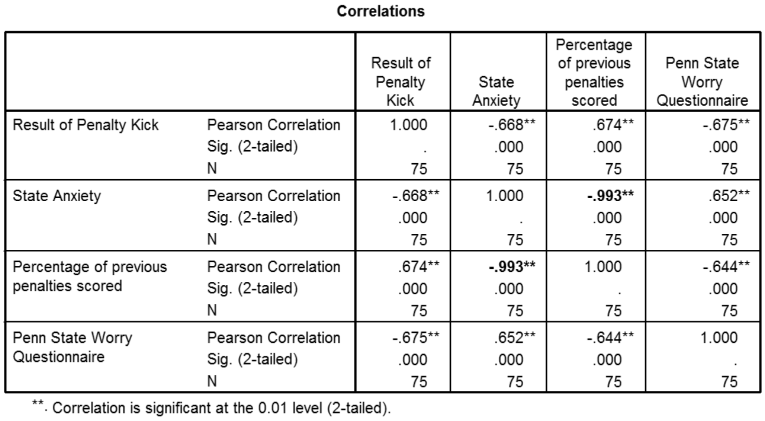

*Syntax1 - IBM SPSS Statistics Syntax Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet2
COMPUTE
VARIABLE LABELS
COMPUTE
VARIABLE LABELS
COMPUTE
VARIABLE LABELS
EXECUTE.
IBM SPSS Statistics Processor is ready
In 8 Col 0
NUM

```
COMPUTE LnPSWQ=LN(PSWQ).
VARIABLE LABELS LnPSWQ 'Ln(PSWQ)'.
COMPUTE LnAnxious= LN(Anxious).
VARIABLE LABELS LnAnxious 'Ln(Anxious)'.
COMPUTE LnPrevious=LN(Previous).
VARIABLE LABELS LnPrevious| 'Ln(Previous Performance)'.
EXECUTE.
```

Figure 8

SELF-TEST Using what you learned in Chapter 6, carry out a Pearson correlation between all of the variables in this analysis. Can you work out why we have a problem with collinearity?

The results of your analysis should look like this:



Correlations

		Result of Penalty Kick	State Anxiety	Percentage of previous penalties scored	Penn State Worry Questionnaire
Result of Penalty Kick	Pearson Correlation Sig. (2-tailed) N	1.000 . 75	-.668** .000 75	.674** .000 75	-.675** .000 75
State Anxiety	Pearson Correlation Sig. (2-tailed) N	-.668** .000 75	1.000 . 75	**-.993**** .000 75	.652** .000 75
Percentage of previous penalties scored	Pearson Correlation Sig. (2-tailed) N	.674** .000 75	**-.993**** .000 75	1.000 . 75	-.644** .000 75
Penn State Worry Questionnaire	Pearson Correlation Sig. (2-tailed) N	-.675** .000 75	.652** .000 75	-.644** .000 75	1.000 . 75

**. Correlation is significant at the 0.01 level (2-tailed).

Output 13

From this output we can see that **Anxious** and **Previous** are highly negatively correlated (r = –0.99); in fact they are nearly perfectly correlated. Both **Previous** and **Anxious** correlate with penalty success[4] but because they are correlated so highly with

[4] If you think back to Chapter 6, these correlations with penalty success (a dichotomous variable) are point-biserial correlations.

each other, it is unclear which of the two variables predicts penalty success in the regression. As such our multicollinearity stems from the near-perfect correlation between **Anxious** and **Previous**.

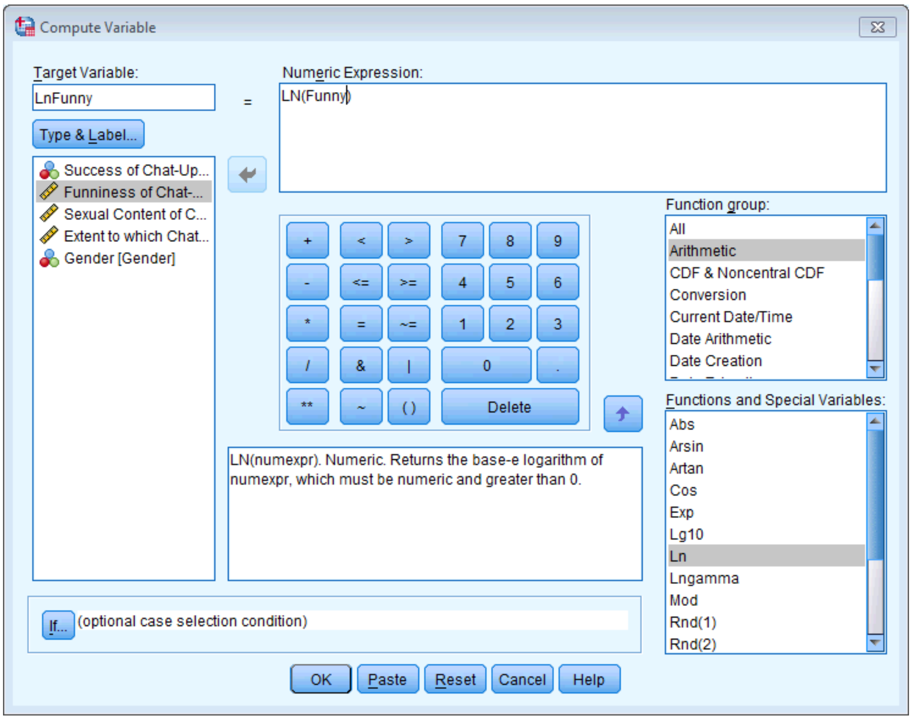
SELF-TEST What does the log-likelihood measure?

The log-likelihood statistic is analogous to the residual sum of squares in multiple regression in the sense that it is an indicator of how much unexplained information there is after the model has been fitted. It follows, therefore, that large values of the log-likelihood statistic indicate poorly fitting statistical models, because the larger the value of the log-likelihood, the more unexplained observations there are.

SELF-TEST Use what you learnt earlier in this chapter to check the assumptions of multicollinearity and linearity of the logit.

Testing for linearity of the logit

In this example we have three continuous variables (**Funny**, **Sex**, **Good_Mate**), therefore we have to check that each one is linearly related to the log of the outcome variable (**Success**). To test this assumption we need to run the logistic regression but include predictors that are the interaction between each predictor and the log of itself. For each variable create a new variable that is the log of the original variable. For example, for **Funny**, create a new variable called **LnFunny** by entering this name into the box labelled *Target Variable* and then click on Type & Label... and give the variable name such as *Ln(Funny)*. In the list box labelled *Function group*, click on *Arithmetic* and then in the box labelled *Functions and Special Variables* click on *Ln* and transfer it to the command area by clicking on ⬆. When the command is transferred, it appears in the command area as 'LN(?)' and the question mark should be replaced with a variable name (which can be typed manually or transferred from the variables list). So replace the question mark with the variable **Funny** by either selecting the variable in the list and clicking on ➡, or just typing 'Funny' where the question mark is. Click on OK to create the variable.



Target Variable: LnFunny

Numeric Expression: LN(Funny)

LN(numexpr). Numeric. Returns the base-e logarithm of numexpr, which must be numeric and greater than 0.

Function group: Arithmetic

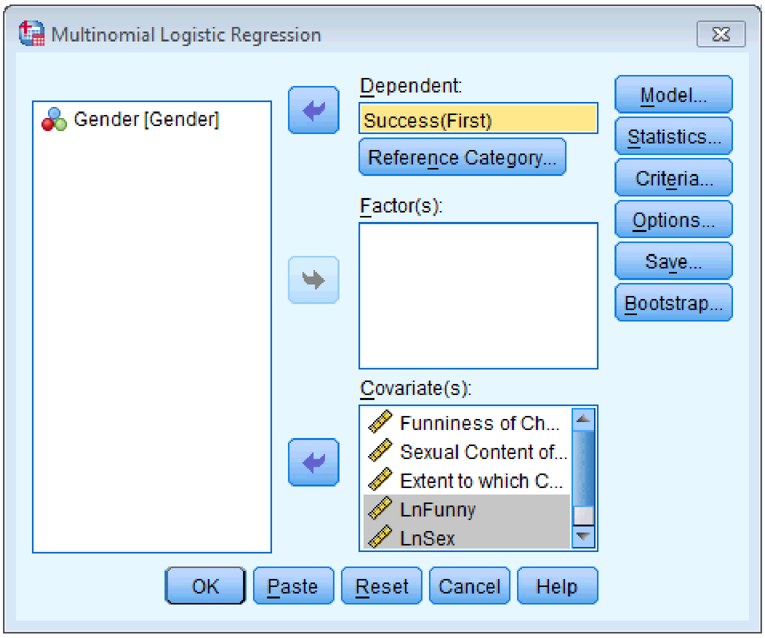
Functions and Special Variables: Ln

Figure 9

Repeat this process for **Sex** and **Good_Mate**. Alternatively, do all three at once using this syntax:

```
COMPUTE LnFunny=LN(Funny).

COMPUTE LnSex=LN(Sex).

COMPUTE LnGood_Mate=LN(Good_Mate).

EXECUTE.
```

To test the assumption we need to redo the analysis but putting in our three covariates, and also the interactions of these covariates with their natural logs. So, as with the main example in the chapter, we need to specify a custom model:



Multinomial Logistic Regression

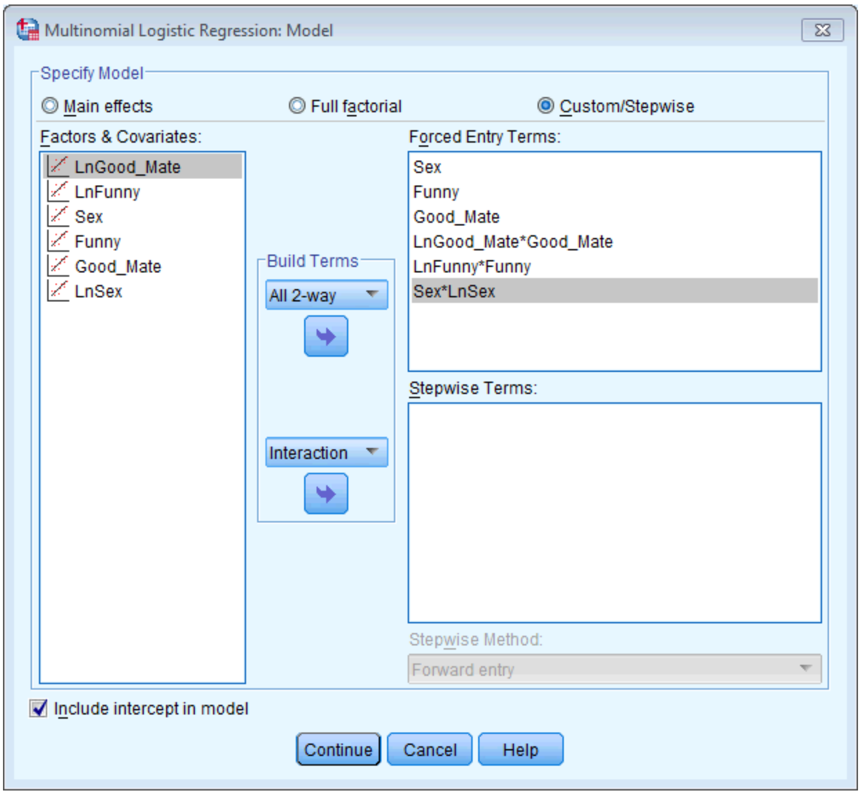
Gender [Gender]

Dependent: Success(First)

Factor(s):

Covariate(s):
- Funniness of Ch...
- Sexual Content of...
- Extent to which C...
- LnFunny
- LnSex

Figure 10



Multinomial Logistic Regression: Model

Specify Model

○ Main effects ○ Full factorial ● Custom/Stepwise

Factors & Covariates:
- LnGood_Mate
- LnFunny
- Sex
- Funny
- Good_Mate
- LnSex

Forced Entry Terms:
- Sex
- Funny
- Good_Mate
- LnGood_Mate*Good_Mate
- LnFunny*Funny
- Sex*LnSex

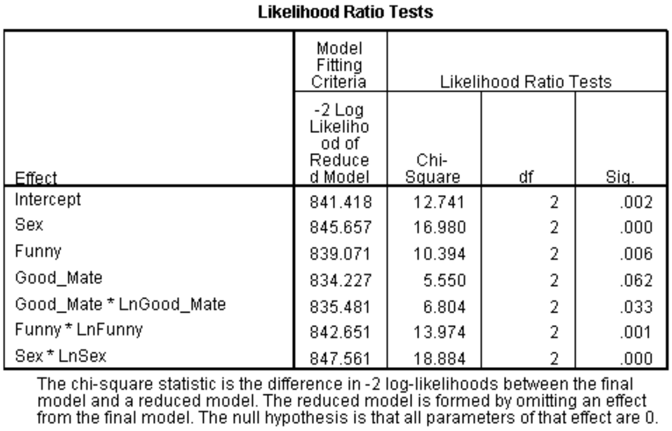
Stepwise Terms:

Stepwise Method: Forward entry

☑ Include intercept in model

Figure 11

Note that (1) we need to enter the log variables in the first screen so that they are listed in the second dialog box, and (2) in the second dialog box we have only included the main effects of **Sex**, **Funny** and **Good_Mate** and their interactions with their log values.



Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	841.418	12.741	2	.002
Sex	845.657	16.980	2	.000
Funny	839.071	10.394	2	.006
Good_Mate	834.227	5.550	2	.062
Good_Mate * LnGood_Mate	835.481	6.804	2	.033
Funny * LnFunny	842.651	13.974	2	.001
Sex * LnSex	847.561	18.884	2	.000

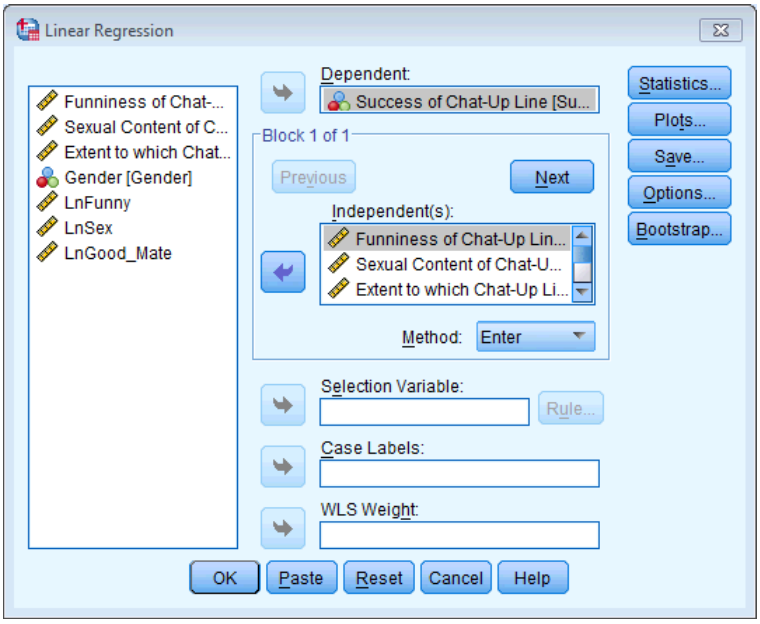
The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Output 14

Output 14 is all that we need to look at because it tells us about whether any of our predictors significantly predict the outcome categories (generally). The assumption of linearity of the logit is tested by the three interaction terms, all of which are significant (p < .05). This means that *all three predictors have violated the assumption*.

Testing for multicollinearity

You can obtain statistics such as the tolerance and VIF by simply running a linear regression analysis using the same outcome and predictors as the logistic regression. It is essential that you click on Statistics... and then select *Collinearity diagnostics* in the dialog box. Once you have selected ☑ Collinearity diagnostics, switch off all of the default options, click on Continue to return you to the *Linear Regression* dialog box, and then click on OK to run the analysis.

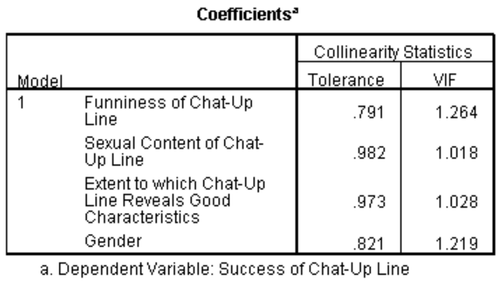


Dependent: Success of Chat-Up Line [Su...

Independent(s):
Funniness of Chat-Up Lin...
Sexual Content of Chat-U...
Extent to which Chat-Up Li...

Method: Enter

Figure 12



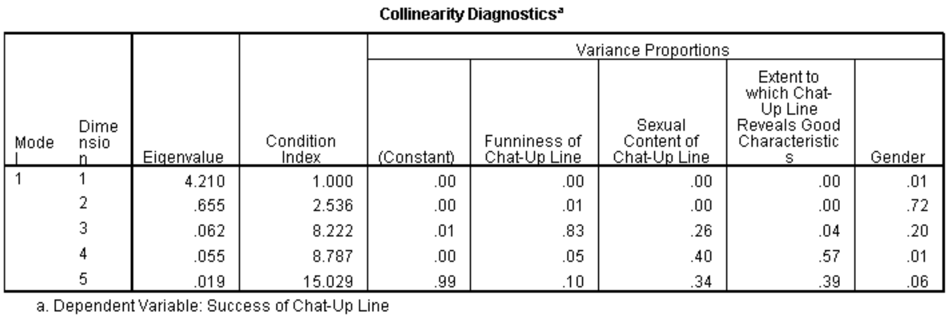
Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Funniness of Chat-Up Line	.791	1.264
	Sexual Content of Chat-Up Line	.982	1.018
	Extent to which Chat-Up Line Reveals Good Characteristics	.973	1.028
	Gender	.821	1.219

a. Dependent Variable: Success of Chat-Up Line

Output 15

Menard (1995) suggests that a tolerance value less than 0.1 almost certainly indicates a serious collinearity problem. Myers (1990) also suggests that a VIF value greater than 10 is cause for concern. In these data all of the VIFs are well below 10 (and tolerances above 0.1); see Output 15. It seems from these values that there is not an issue of collinearity between the predictor variables. We can investigate this issue further by examining the collinearity diagnostics.



Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Funniness of Chat-Up Line	Sexual Content of Chat-Up Line	Extent to which Chat-Up Line Reveals Good Characteristics	Gender
1	1	4.210	1.000	.00	.00	.00	.00	.01
	2	.655	2.536	.00	.01	.00	.00	.72
	3	.062	8.222	.01	.83	.26	.04	.20
	4	.055	8.787	.00	.05	.40	.57	.01
	5	.019	15.029	.99	.10	.34	.39	.06

a. Dependent Variable: Success of Chat-Up Line

Output 16

The table labelled *Collinearity Diagnostics* (Output 16) gives the eigenvalues of the scaled, uncentred cross-products matrix, the condition index and the variance proportions for each predictor. 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 final dimension has a condition index of 15.03, which is nearly twice as large as the previous one. Although there are no hard-and-fast rules about how much larger a condition index needs to be to indicate collinearity problems, this could indicate a problem.

For the variance proportions 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. So we are interested mainly in the bottom few rows of the table (which represent small eigenvalues). In this example, 40–57% of the variance in the regression coefficients of both **Sex** and **Good_Mate** is associated with eigenvalue number 4 and 34–39% with eigenvalue number 5 (the smallest eigenvalue), which indicates *some* dependency between these variables. So, there is some dependency between **Sex** and **Good_Mate**, but given the VIF we can probably assume that this dependency is not problematic.