

## Logistic regression

### *An alternative example with medical data*

A study is carried out to find out how well people with traumatic brain injury (TBI) can be classified shortly after the injury is sustained either as having made a sufficiently good recovery to be back at work at 6 months (WORK = 1) or as not having recovered sufficiently to be back at work at 6 months (WORK = 2). The study is carried out on 54 people with TBI who have Glasgow coma scores below 12. Data are obtained on the following covariates: (1) an EEG-derived score (EEG), (2) the coma score (COMA) and bivariate (yes/no) pupil reactivity (REACT), the first two being the same variables that were used in the preceding discriminant analysis. The data are analyzed using logistic regression. The first few and last few rows of the amended data appear in Table 9.6 (the full dataset is on the book website as med.logistic.sav).

Table 9.6

*The first few and last few rows of amended brain injury data on 54 patients (the full dataset can be found as med.logistic.sav on the book website)*

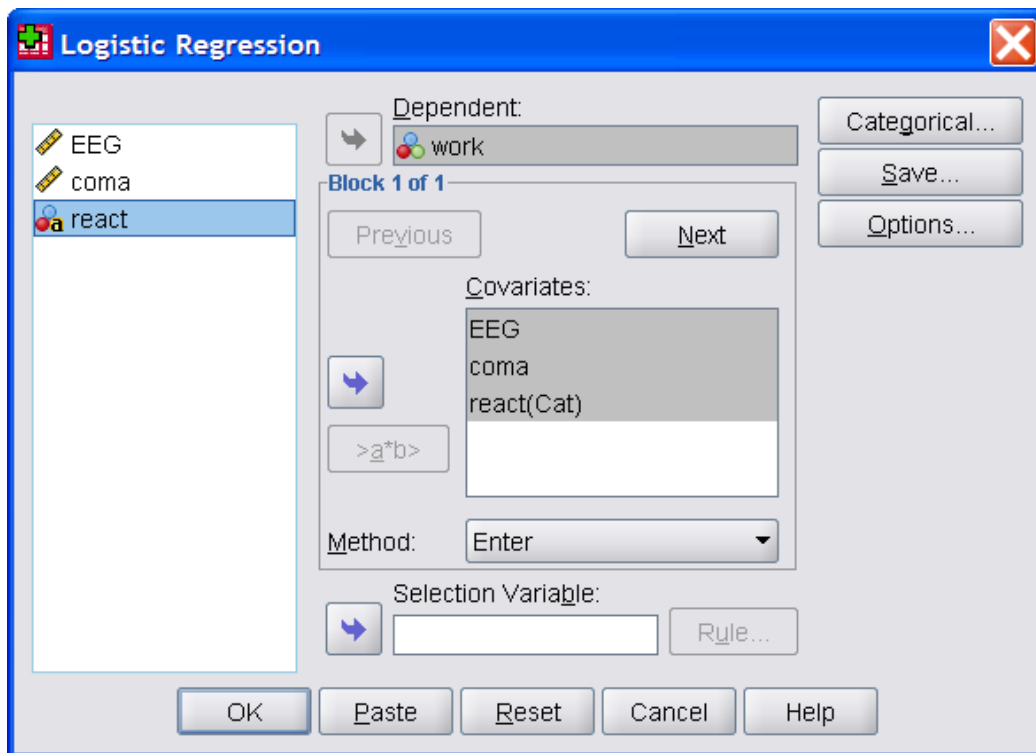
<b>work</b>	<b>EEG</b>	<b>coma</b>	<b>react</b>
1	8	11	yes
1	4	8	yes
1	4	9	no
1	7	10	yes
2	4	5	no
2	5	4	no
2	5	6	yes
2	7	6	no

Now we see how to set up a logistic regression in SPSS.

### *Logistic regression: requesting the analysis in SPSS*

With the data arranged in the SPSS data window as in Table 9.6, we proceed as follows. We choose **Analyze** from the menu bar, then **Regression**, then **Binary**

**Logistic**, and get SPSS Dialog Box 9.4. The grouping variable is the **Dependent** (WORK in our example). Only two levels are possible so we don't have to enter a range as we did in discriminant analysis. The IVs EEG, COMA and REACT are the **Covariates**. If any categorical variables are coded numerically, we would need to press the **Categorical** button to get a dialog box that allows us to say which these variables are. In our case, since we have used the words 'yes' and 'no' as levels of REACT, SPSS recognises this as a categorical variable. If we were to press the **Categorical** button, we would find SPSS has already entered REACT for us. The default **Method** enters all the variables from the **Covariates** list in the model. There are stepwise options that we will consider later. Press the **Save** button to keep the predicted group membership in an extra column of the data sheet, just as we did in discriminant analysis. In the **Save** dialog box, click **Group membership** in the **Predicted Values** group. Click **Continue** and **OK** to get our logistic regression.



*SPSS Dialog Box 9.4. Starting a logistic regression*

*Logistic regression: understanding the output*

The first table in the output, called Case Processing Summary but not reproduced here, shows that all our 54 cases were included in the analysis and none were missing. The next two tables (also not reproduced here) show the codings given within the logistic regression to the DV and our categorical IV, REACT. The DV is always given the codings 0 and 1 within the logistic regression, so REACT is given codes 0 for 'yes' and 1 for 'no' (you can change this by clicking the **Categorical** button and changing the **Reference Category** from **Last** to **First**).

Next comes a heading, Block 0: Beginning Block. This section is not very interesting because here SPSS starts by assigning all cases to the largest group. We don't reproduce this here. The next section is headed Block 1: Method = Enter. This means all our covariates are added, since we did not specify a stepwise method. The first two tables in this section show the overall significance of the model and we consider this later. Then we have a summary of the classification as shown in SPSS Output 9.6. Here we see that we got just over 94% correctly assigned, and underneath the table we are told that the cut value is 0.5, in other words if the probability of belonging to the group with WORK coded as 2 is greater than 0.5, then the case is classified as belonging to this group.

**Classification Table<sup>a</sup>**

		Predicted		
		work		Percentage Correct
Observed		1	2	
Step 1	work 1	25	1	96.2
	2	2	26	92.9
	Overall Percentage			94.4

a. The cut value is .500

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	EEG	1.236	.629	3.857	1	.050	3.441
	coma	-2.691	.971	7.674	1	.006	.068
	react(1)	3.109	1.589	3.831	1	.050	22.402
	Constant	11.471	6.492	3.122	1	.077	95901.429

*SPSS Output 9.6. Classification results using logistic regression*

The second table in SPSS Output 9.6 gives the parameter values in column B. These can be used to classify new cases. The mathematics is a bit off-putting and we will not show how it is done by hand, since you can also use the **Selection Variable** box in SPSS Dialog Box 9.4 to get SPSS to do this for you, as explained in the next section.

If we look in the data window we find a new column called PGR\_1 has been added. This gives the group membership code that was assigned by the logistic regression. If you do this yourself and look down the new column you will see that case 9 is incorrectly assigned to group 2, and cases 31 and 42 are incorrectly assigned to group 1. These are the three misclassified cases shown in SPSS Output 9.6.

*Logistic regression: assigning new cases to WORK= 1 or 2*

For new cases, you can if you wish calculate the probabilities of belonging to WORK = 2 using the parameters from SPSS Output 9.6 and a calculator. However, as indicated above, we can persuade SPSS to do this job for us. We enter the ten new cases in Table 9.7 at the bottom of the data sheet. These new cases are the same ones we used to try out our discriminant functions, but with PUPIL replaced by the REACT variable. The variable WORK will be missing in the data sheet for the new cases.

Table 9.7

*New cases to be classified by logistic regression (med.logistic.newcases.sav)*

EEG	coma	react
-----	------	-------

7	10	yes
4	8	yes
7	7	no
7	8	yes
6	9	no
5	6	yes
6	10	no
5	5	yes
4	6	no
6	8	yes

Now we also need a variable to allow us to select the original cases that formed our training set to be included in the logistic regression, and then use the parameters from the regression to assign all cases, including the new ones, to the two groups. We can call the selecting variable CHOOSE, and give it the value 1 for the 54 training set cases and 2 for the ten new cases. The complete dataset, with the original and new cases, and the CHOOSE variable added in a new column, can be found as med.logistic.predict.sav on the book website. Then in SPSS Dialog Box 9.4, we put the variable CHOOSE in the **Selection Variable** box with the arrow. Click the **Rule** button and put **equals 1** in the box and click **Continue**. The dialog box now has CHOOSE = 1 in the **Selection Variable** box at the bottom.

We remember to **Save** the predicted group membership as before. When we press **OK** we get the usual output, with the additional information that 54 of 64 cases were selected for the analysis. The parameters and classification tables appear as before, but in the data window the new column called PGR\_1 assigns all of the cases, including the 10 new ones, to one of the two groups. In fact they are assigned to groups 1, 1, 2, 1, 1, 2, 1, 2, 2, and 1 respectively.

*Logistic regression: estimating misclassification probabilities for new cases*

SPSS does not offer a cross validation option for logistic regression, so the only ways to get a realistic estimate of misclassification probabilities on new cases are either to

await final correct assignment on a new data set, if that will eventually be available, or else divide the training set at random into two, using half to calculate the parameters, then using these parameters to assign the other half to get the misclassification rate.

Random assignment to the two halves is best, but to make it easy for you to reproduce our results, we used alternate cases, starting with the first, from our dataset in Table 9.6, as a training set to perform a logistic regression. The other 27 cases will be our validation set. Remove the ten new cases if you still have them at the bottom of the dataset. Then we can use a variable called, for example, XVAL, and set it as 1 or 2 in alternate rows, and use the selection procedure described in the previous section. The modified dataset with the original 54 cases and XVAL added as an extra column is available as med.logistic.crossvalidation.sav on the book website.

When we did this cross-validation, two of the 27 cases in the training set were misclassified (we haven't shown the output here). The parameter estimates corresponding to column B in SPSS Output 9.6 were 0.497, -2.322, -3.672, and 17.184, different from those in SPSS Output 9.6 because only the 27 cases with XVAL = 1 were used to calculate the regression. SPSS used these to calculate the predicted group membership for the remaining cases. This time we saved the probabilities as well, and these are shown in Table 9.8, along with the predicted and actual group membership for the 27 cases with XVAL = 2 (the validation set).

Table 9.8  
*Probabilities of belonging to WORK = 2 for validation set*

<b>probability</b>	<b>predicted group</b>	<b>actual group</b>
.04426	1	1

.00198	1	1
.01980	1	1
.22721	1	1
.00325	1	1
.11130	1	1
.07077	1	1
.44291	1	1
.00019	1	1
.00120	1	1
.00007	1	1
.43713	1	1
.01213	1	1
.93022	2	2
.99928	2	2
.77546	2	2
.99555	2	2
.98805	2	2
.95638	2	2
.99270	2	2
.25299	1	2
.99805	2	2
.99948	2	2
.99948	2	2
.99805	2	2
.99997	2	2
.99882	2	2

For these 27 cases we find we have only one misclassification, or 96% correctly classified. This is one less misclassified than we got from the 27 cases we used as the training set, so we could be quite optimistic that in future we should be able to get about the same proportion correctly classified. However, if we reverse the roles of our training and validation sets, by setting  $XVAL = 2$  in the **Selection Variable** box, we get all cases correct in the training set, but 4 misclassified (15%) in the validation set. This should be a warning not to exaggerate our claims for our results. We may prefer to use the parameters calculated from the full training set of 54 in classifying new cases in future, and in that case we expect to get at least  $23/27 * 100 = 85\%$  classified correctly (our worst result from the cross-validation exercises described in this paragraph).

*Logistic regression: more about the model parameters and their significance*

Return to SPSS Output 9.6 and look again at the parameters in the B column of the second table. At the right of the table there is a column labelled 'Sig' which tells you whether each parameter is significant in the model: if the probability in this column is less than 0.05, you would reject at the 5% level the hypothesis that the parameter is zero. The probabilities for our three parameters are 0.050 for EEG, 0.006 for COMA and 0.050 for REACT, so COMA makes a significant contribution to predicting the probability of belonging to WORK = 2, and EEG and REACT are right on the borderline of significance. We could try removing each of these, one at a time, to see the effect on our correct classification rate.

If we remove REACT (use the arrow to remove it from the **Covariates** box in SPSS Dialog Box 9.4), we get four instead of three misclassifications in the summary table similar to that shown in SPSS Output 9.6. The 'Sig' column in the table of variables now gives 0.009 and 0.001 for EEG and COMA as shown in SPSS Output 9.7. If we remove EEG and put REACT back in, we get five misclassified and the 'Sig' column for COMA and REACT reads 0.004 and 0.010 (we don't show the output this time). There is a difficult decision to make here, since we can always get better estimates if we have fewer parameters for any given size of dataset. On the other hand, excluding either of the variables at the borderline of significance increases the number of misclassifications in the training set. But if we remember that the 5% significance level is a boundary imposed on a continuum that runs from 'extremely unlikely to be a chance effect' through to 'very likely to be a chance effect' we may decide that on balance it is better to include both EEG and REACT as predictors with COMA. We consider this further when we look at stepwise procedures.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	EEG	1.442	.555	6.745	1	.009	4.227
	coma	-2.565	.751	11.679	1	.001	.077
	Constant	10.512	4.011	6.869	1	.009	36754.312

*SPSS Output 9.7. Significance of EEG and COMA when these are the only IVs in the logistic regression*

Look at the signs of the parameters in the second column in SPSS Output 9.6 or 9.7, labelled 'B'. EEG has a positive sign for B, so an increase in the EEG score increases the odds of being in group 2 (WORK = 2). COMA is negative, so an increase in the COMA score reduces the odds of belonging to the group WORK = 2. If you look at the data in Table 9.6 you see that the COMA scores for WORK = 1 are generally higher than they are for WORK = 2.

*Logistic regression: checking the fit of the model*

The overall fit of the model can be checked by looking for the  $-2\text{Log}(\text{likelihood})$  in the Model Summary table, which appears at the beginning of the output, and is shown in SPSS Output 9.8. If the model fits well, then  $-2\text{Log}(\text{likelihood})$  is approximately  $\chi^2$  with degrees of freedom equal to the number of cases minus the number of estimated parameters, including the constant. This is  $54-4 = 50$  when we include EEG, COMA and REACT in our model. If the value does not exceed that for significance at the 5% level, we would not reject at the 5% level the hypothesis that the model is a good fit. Our value of 15.533 is way below the value of 67.5, which is the critical value for the 5% significance level with 50 degrees of freedom, so we conclude that our model fits well.

Another way to check the fit is to click the **Options** button and request **Casewise listing of residuals** (accept the default of **Outliers outside 2 sd**). Then if any cases are not well fitted by the model, you get a table listing them like the second one in SPSS Output 9.8. Here we see that our case 9, belonging to WORK = 1, is more typical of WORK = 2 and is not well fit by the model. It has a very large residual,  $Z = -4.343$ . Sometimes examining a case that is an outlier like this one will help us to understand our problem better. But one outlier in 54 cases when the model fit is good overall is probably not enough to make us lose confidence in our ability to classify new cases. Of course, we should always monitor the performance of a classification tool that is being used in a practical situation.

**Model Summary**

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	15.533 <sup>a</sup>	.666	.889

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

**Casewise List<sup>b</sup>**

Case	Selected Status <sup>a</sup>	Observed	Predicted	Predicted Group	Temporary Variable	
		work			Resid	ZResid
9	S	1**	.950	2	-.950	-4.343

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

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

*SPSS Output 9.8. Fit of the model and cases that are not fit well*

### *Logistic regression: using a stepwise method*

In the previous section we saw that all the IVs in our logistic regression had parameters that were significantly different from zero or were on the borderline, so all make a significant contribution in estimating the probability of belonging to WORK = 2. Just as in discriminant analysis, we can choose to enter the variables by a stepwise method. Clicking the arrow to the right of the **Method** box in SPSS Dialog Box 9.4 produces a list of alternative stepwise methods. The first three are *forward* selection methods: start with no variables and add them one at a time if they pass a criterion.

The last three are *backward* elimination methods: start with all the variables and remove them one at a time if they fail a criterion. We prefer forward selection, since parameter estimates are more reliable if we have fewer parameters. The different selection criteria will usually give similar results, so we just take the first one and redo the regression. We find that all our variables are added in three steps, which supports our argument for retaining all three variables in the above section entitled 'More about model parameters and their significance'. SPSS does not offer a procedure for combining forward selection and backward elimination as it does for discriminant analysis, but you can see from the table shown in SPSS Output 9.9 that it does provide the significance of the change if you do remove any of the entered variables. At every step, removing any of the entered variables would be a significant change. We should keep them in.

**Model if Term Removed<sup>a</sup>**

Variable	Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1 coma	-37.401	41.043	1	.000
Step 2 coma	-30.837	39.484	1	.000
react	-18.574	14.959	1	.000
Step 3 EEG	-11.749	7.964	1	.005
coma	-31.111	46.688	1	.000
react	-11.775	8.017	1	.005

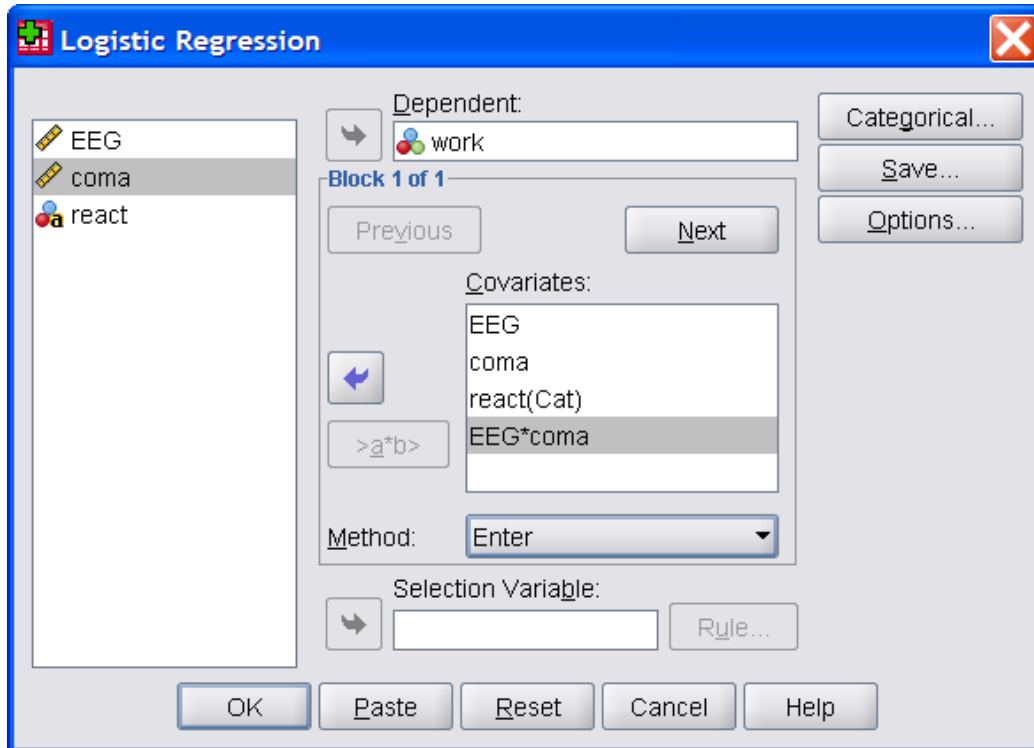
a. Based on conditional parameter estimates

*SPSS Output 9.9. The significance of removing variables from a logistic regression*

### *Logistic regression: interaction terms*

Logistic regression allows *interaction terms* to be included in just the same way as does covariance analysis. However, the main effects for any variables in interaction terms must always be included. Adding them one at a time yourself, and using the **Enter** method rather than a stepwise procedure is the best way to ensure that this is done. We consider this next. First put WORK in the **Dependent** box and the three IVs EEG, COMA and REACT in the **Covariates** box as before. To add the EEG\*COMA

interaction term to the model, we select these two variables, which are still in the box at the left, and the **>a\*b>** button becomes available. Click it and the interaction term EEG\*COMA appears in the **Covariates** box, as shown in SPSS Dialog Box 9.5.



SPSS Dialog Box 9.5. Adding an interaction term to a logistic regression

Click **OK** and we get the parameter estimates shown in SPSS Output 9.10. Now none of our variables is significant. We did still get only 3 misclassified, but the case that was not very well fit by the model is still not well fit. The model summary (not reproduced here) shows that  $-2\text{Log}(\text{likelihood})$  is 15.274. With just the main effects it was 15.533. The difference is only 0.259. This is approximately  $\chi^2$  with 1 degree of freedom (the degrees of freedom is the number of extra parameters in the more complex model), so certainly not significant. We do better without the EEG\*COMA interaction.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	EEG	-1.271	5.268	.058	1	.809	.280
	coma	-4.787	4.741	1.020	1	.313	.008
	react(1)	2.973	1.568	3.594	1	.058	19.544
	EEG by coma	.328	.691	.224	1	.636	1.388
	Constant	27.355	35.403	.597	1	.440	7.588E11

*SPSS Output 9.10. The effect of adding an interaction term*

Similar results are obtained if we add either of the other second order interactions. We cannot add the third order interaction since none of the second order ones is significant.