

EXERCISE 23

Predicting category membership: Discriminant analysis and binary logistic regression

Before you start

Before proceeding with this practical, please read Chapter 14.

Prediction of reading success at the school-leaving stage

Just before they leave school, students in the most senior class of a school are regularly tested on their comprehension of a difficult reading passage. Typically, only 50% of students can perform the task. We shall also suppose that, for a substantial number of past pupils, we have available data not only on their performance on the comprehension passage but also on the very same variables that were investigated in the exercise on multiple regression, namely, the reading-related measures that we have referred to in Table 1 below as *Logo*, *Syntax* and *Vocal*, all of which were taken in the very earliest stages of the children's education.

The full data set is given in the appendix to this Exercise. As with the multiple regression example, we hope that the data have already been stored for you in a file with a name such as **discrim**, the contents of which you can access by using the **Open** procedure. The data (**Ex23 Reading data for discriminant analysis**) are also available from WWW in the website:

<http://www.psypress.com/spss-made-simple/datasets.asp>

Table 1 shows the first and the last two lines of the data set.

Logo	Syntax	Vocal	Comprehension
10	20	64	1
28	28	58	1
...
82	69	60	2
51	48	52	2

The rightmost variable is a coding variable whose values, 1 and 2, denote, respectively, *failure* and *success* on the comprehension task.

Exploring the data set

Before moving on to the main analysis, a preliminary exploration of the data will bring out at least some of their important features. For example, if a particular variable is going to be useful in assigning individuals to categories, one might expect that, if its scores are subdivided by category membership, there should be a substantial difference between the group means. If there is no difference, the variable will probably play a minimal role in the final discriminant

function. To investigate these differences, **one-way ANOVAs** can be used to compare the group means on the various independent variables. These tests, however, are requested by options in the **Discriminant** procedure. We shall therefore return to the descriptive statistics when we come to prepare the dialog box.

Since discriminant analysis assumes that the distribution of the independent variables is multivariate normal, we shall also need to look at their distributions. Here we suggest you check for extreme scores and outliers by using the **Explore** command (see Section 14.2.2) to examine the distributions of the variables.

In the **Explore** dialog box, transfer the variable names of all the predictors *Logo*, *Syntax* & *Vocal* into the **Dependent List** box. Transfer the variable name *Comprehension* into the **Factor List** box, and the variable name *Case Number* into the **Label Cases by** box. Click the **Plots** radio button in the **Display** options. Click **Plots...** and click off the check box for **Stem-and-leaf**. Click **Continue** and then **OK** to plot three sets of boxplots.

- **Do the boxplots reveal any extreme cases or outliers? Are the distributions relatively normal?**

DISCRIMINANT ANALYSIS

Procedure for discriminant analysis

Run the discriminant analysis as described in Section 14.2.3. There, however, we recommended the **Stepwise** method of minimisation of **Wilks' Lambda**. In the present example, because of its simplicity, it is better to use the default method known as **Enter**, in which all the variables are entered simultaneously. Since **Enter** is the default method, there is no need to specify it. In the **Discriminant Analysis** dialog box, click **Statistics** to open the **Discriminant Analysis: Statistics** dialog box. Select **Univariate ANOVAs** and click **Continue**. In the **Discriminant Analysis** dialog box, click **Classify** to open the **Discriminant Analysis: Classification** dialog box and (in **Display**) select **Summary table**. Click **Continue**, then **OK**.

Output for discriminant analysis

The main features of the output for a discriminant analysis are explained in Section 14.2.4, which you should review. In the present example, the table labelled Group Statistics shows the number of cases in each of the categories of the variable *Comp*. The next table, headed Tests of Equality of Group Means lists **Wilks' Lambda** and **F-ratios** (with their associated p-values in the column **Sig.**) for the comparisons between the groups on each of the three independent variables.

- **Which variables have significant F ratios and which do not?**

There now follows the first of the tables labelled **Eigenvalues**, which show the output of the discriminant analysis proper. Because there are only two groups, there is only one function. The next table, **Wilks' Lambda**, tabulates the statistic **lambda**, its **chi-square value** and the associated p-value (**Sig.**). You will notice immediately that the value of lambda is smaller than the value for any of the three IVs considered separately. That is well and good: the discriminant function *D*, which uses the information in all the IVs should do a better job than any one IV alone. Here there is an obvious parallel with multiple regression, in which the predictive ability of the multiple regression equation cannot (provided there is no multicollinearity) be less than the simple regressions of the target variable on any one predictor alone. In the case of the variable *Vocal*, however, the improvement is negligible. Since, however, two of the IVs can each discriminate reliably between the groups, the result of

the chi-square test of lambda in the discriminant analysis table is a foregone conclusion. As expected, the p-value is very small. The discriminant function D can indeed discriminate reliably between the two groups on the basis of performance on the independent variables.

Ignore the table labelled Standardized Canonical Discriminant Function Coefficients. A more useful table is the next one, labelled **Structure Matrix**, which lists the pooled-within-groups correlations between discriminating variables and the standardized canonical discriminant function.

- **Are the correlations as you expected?**

Ignore the table Functions at Group Centroids.

The next set of tables relate to the classification of cases. We have shown that the discriminant function D discriminates between the two groups; but how effectively does it do this? This is shown under the heading: 'Classification Results'.

- **Write down the percentage of grouped cases correctly classified, the percentage of correct group 1 (failure) predictions and the percentage of correct group 2 (success) predictions.**

Now try out the discriminant function on some fresh data by adding them at the end of the data file (e.g. enter in the columns for *Logo*, *Syntax*, *Vocal*, the values 50, 50, 50; 10, 10, 10; 80, 80, 80 and any others you wish). Leave the column blank for *Compreh*. Then re-run the analysis after selecting **Save** in the **Discriminant Analysis** dialog box, clicking the radio button for **Predicted group membership**, and then clicking **Continue** and **OK**. The predicted memberships will appear in the variable called **dis_1** in the **Data Editor**.

- **Would someone with Logo, Syntax and Vocal scores of 50, 50, 50 respectively be expected to pass or fail the comprehension test?**

Conclusion

This Exercise is intended to be an introduction to the use of a complex and sophisticated statistical technique. Accordingly, we chose an example of the simplest possible application, in which the dependent variable comprises only two categories. The simplicity of our interpretation of a number of statistics such as **Wilks' lambda** breaks down when there are more than two categories in the dependent variable. For a treatment of such cases, see Tabachnick & Fidell (2007).

BINARY LOGISTIC REGRESSION

Procedure for binary logistic regression

We shall use the same data set for binary logistic regression as we used for discriminant analysis at the start of this Exercise. Use the procedure described in Section 14.3.3.

Output for binary logistic regression

The main features of the output for binary logistic regression are explained in the same Section, which you should review.

Examine the tables in Block 1.

- **What is the value of R^2 as calculated by the Nagelkerke formula? What is the meaning of this value?**

- What is the value of chi-square for the Hosmer and Lemeshow test and is it significant? What do you conclude about the fit of the model?
- What is the overall percentage of correct predictions? How does this compare with the success rate of the discriminant analysis?

Conclusion

In this example, the results of the **discriminant analysis** and **binary logistic regression** are similar but where there are several binary predictors, logistic regression would be the preferred analysis.

Appendix to Exercise 23 - Reading data

L is Logo; S is Syntax; V is Vocal; C is Comprehension

L	S	V	C	L	S	V	C	L	S	V	C	L	S	V	C
10	20	64	1	49	59	46	1	41	55	41	2	49	72	72	2
28	28	58	1	39	42	31	1	30	54	20	2	66	61	40	2
55	25	42	1	26	56	78	1	29	67	18	2	84	50	46	2
30	20	30	1	40	31	51	1	28	68	72	2	70	54	51	2
32	27	42	1	34	60	45	1	46	67	80	2	65	64	23	2
25	49	81	1	31	66	50	1	56	44	52	2	69	60	57	2
40	38	43	1	18	61	22	1	69	46	59	2	66	79	50	2
71	22	79	1	43	50	31	1	53	57	52	2	58	82	13	2
19	59	71	1	48	45	44	1	75	48	34	2	45	90	59	2
55	32	75	1	14	77	53	1	71	52	30	2	82	58	65	2
45	45	29	1	64	32	55	1	50	68	75	2	82	69	60	2
62	30	26	1	55	48	9	1	81	54	41	2	51	48	52	2
20	69	78	1					51	62	14	2				