Research Design - - Topic 21
Discriminant Function Analysis

General Rationale and Purpose

Basic mathematics

An Example Using SPSS GLM Discriminant

- Eigenvalues
- Calculating and Plotting Centroids
- Interpreting the Discriminant Functions
- Forecasting

Approaches

1

General Rationale and Purpose

The general rationale, mathematics, and assumptions underlying discriminant function analysis are identical to those of single factor multivariate analysis of variance. Only the purpose is different.

Like multivariate analysis of variance:

- It is concerned with identifying weighted aggregates that account for the relative variation among groupings of subjects.
- The number of such aggregates obtained will be the lesser of the number of variables or the number of groups minus 1.

Unlike multivariate analysis of variance:

- It focuses attention on the tests of significance of the individual discriminant functions

2

Unlike single factor multivariate analysis of variance, the purpose underlying discriminant function analysis is to determine whether scores on the measures account for the separation among the groups. That is, in multivariate analysis of variance, the grouping factor is the independent variable, and the measures are the dependent variables, while in discriminant function analysis, the opposite is the case. The measures are the independent variables and the grouping factor is the dependent variable.

Two overall purposes are:

- Evaluative (and Interpretative) Function. Assessed in terms of significance and interpretability of the function(s).
- Forecasting Function. Can the functions be used to predict group membership. Assessed by the percentage of correct classification.

3

Basic Mathematics

The weighted aggregate is:

$$L_{ai} = w_1 X_{1i} + w_2 X_{2i} + \ldots$$

where:

L_{ai} = the aggregate score for an individual in group *a*
w_1, w_2, … = the weights for variables 1, 2, …
X_{1i}, X_{2i}, … = the scores for variables 1,2, etc…in each of the groups

The weights are determined such that the Sum of Squares Between Groups divided by the Sum of Squares Within Groups is as large as possible if one had done a single factor analysis of variance on the L_{ai} scores . The weights are computed using the notion of the Determinantal Equation.

4

The Determinantal Equation is formed on the matrix equivalent of the Sum of Squares Between groups (SSB) divided by the Sum of Squares within groups (SSW). This is a matrix formed by multiplying the inverse of the SSW matrix by the SSB matrix. This will produce eigenvalues that account for the largest amount of relative variation. That is:

$$\left|SSW^{-1}SSB - \lambda I\right| = 0$$

$$[SSW^{-1}SSB - \lambda_p I][W] = 0$$

$$W^T MS_W W = 1$$

This produces a unique solution of the weights such that the MS_{within} for each set of L scores (i.e., each Discriminant Function) = 1.

The eigenvalue is the ratio of 2 Sums of Squares of the aggregates

$$\lambda = \frac{SSB_L}{SSW_L}$$

5

An Example Using SPSS Discriminant

To demonstrate the similarities and differences between single factor multivariate analysis of variance and discriminant function analysis, we will use the same data as in the previous lecture. In this case, however, consider the groups as four groupings of students (e.g., Arts, Social Science, Science, and Engineering). Consider X1 to X4 to be four personality measures.

The purpose is to determine whether the personality measures predict each student's membership in the correct group (forecasting function) , whether the subsets can be interpreted to understand personality differences between classes of students, and how well group membership is predicted (evaluative and interpretative function).

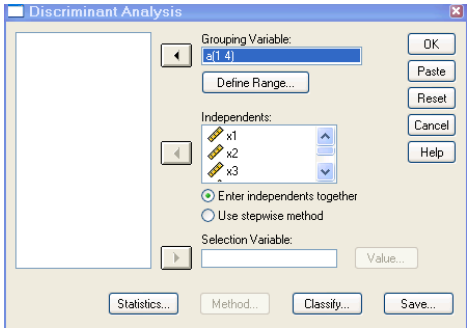
6

The data file is the same as that for multivariate analysis of variance from the previous lecture. In order to run SPSS Discriminant, click on analyze, then Classify, and then Discriminant. The following slide shows the first window with the variables and options that were selected for this example.

The Syntax file from that analysis is as follows:

```
DISCRIMINANT
  /GROUPS=a(1 4)
  /VARIABLES=x1 x2 x3 x4
  /ANALYSIS ALL
  /PRIORS  EQUAL
  /STATISTICS=MEAN STDDEV UNIVF RAW TABLE
  /PLOT=COMBINED
  /CLASSIFY=NONMISSING POOLED .
```

7



8

The tests of the variation of the four groups on each of the variables is presented in the table below. Despite the labelling, this is simply a summary of the F-ratios for each of the variables. The results are identical to those presented in the previous lecture. In fact, for a single dependent variable, Wilk's Lambda (Λ) is simply the SS_{Within}/SS_{Total} and the F-ratio can be shown to be:

$$F = \frac{v_2(1-\Lambda)}{v_1\Lambda} = \frac{MS_{Between}}{MS_{Within}}$$

at v1= (a-1) and v2 = (N-a) degrees of freedom

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
x1	.816	1.199	3	16	.342
x2	.663	2.709	3	16	.080
x3	.523	4.870	3	16	.014
x4	.926	.424	3	16	.739

9

Eigenvalues

The 3 eigenvalues for this analysis are presented below in order of magnitude. The first one is 1.472 and accounts for 50.6 % of the total variation accounted for by the three discriminant functions. The canonical correlation is a measure of the relation of the discriminant function to group membership. Its square is the proportion of the variance of the discriminant function accounted for by group membership in the same way that η^2 is the strength of an effect. For the first function, this value is 59.6%. The formula for the canonical correlation is:

$$R_{CC} = \sqrt{\frac{\lambda}{1+\lambda}} = \sqrt{\frac{1.472}{2.472}} = .772$$

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.472[a]	50.6	50.6	.772
2	1.016[a]	34.9	85.5	.710
3	.422[a]	14.5	100.0	.545

a. First 3 canonical discriminant functions were used in the analysis.

10

The test of significance of the Discriminant functions is a sequential test that makes use of the Chi-square statistic. It initially tests the significance of all 3 functions using the following statistics.

$$\Lambda = \Pi\left(\frac{1}{1+\lambda}\right) = \left(\frac{1}{2.472}\right)\left(\frac{1}{2.016}\right)\left(\frac{1}{1.422}\right) = .141$$

$$\chi^2 = [(N-1) - .5(p+k)]\ln\Lambda \qquad \text{with} \qquad df = p(k-1)$$

The remaining tests eliminate the largest eigenvalue to compute Λ, and reduce both p and k by 1. These results would suggest that the third function is not significant.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.141	29.380	12	.003
2 through 3	.349	15.802	6	.015
3	.703	5.282	2	.071

11

Calculating and Plotting Centroids

In SPSS, a constant is added to each function so that the grand mean for each function is 0. Thus, using the unstandardized coefficients from the table, the first Discriminant Function is:

$$L_{ai} = -10.592 - .382X_1 + .520X_2 + .162X_3 + .292X_4$$

Canonical Discriminant Function Coefficients

	Function		
	1	2	3
x1	-.382	.089	.421
x2	.520	-.157	-.043
x3	.162	.793	.316
x4	.292	.132	.663
(Constant)	-10.592	-11.135	-15.071

Unstandardized coefficients

12

The function can be used to compute a score for each individual, and the means computed for each group. These means are called centroids. Alternatively, the centroids can be computed directly by multiplying the mean on each variable for each group by the unstandardized weights and summing over the variables for that group. These centroids are presented in the following table.

Functions at Group Centroids

	Function		
a	1	2	3
1.00	1.152	1.189	-.213
2.00	-1.785	.475	-.079
3.00	.324	-.495	.939
4.00	.310	-1.169	-.646

Unstandardized canonical discriminant functions evaluated at group means

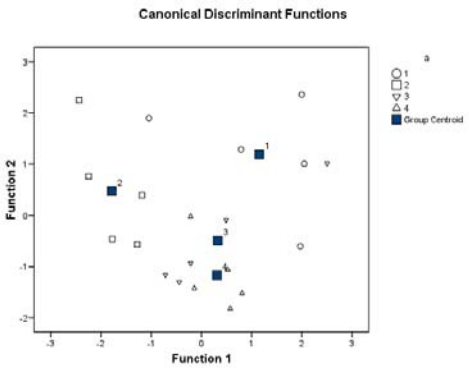
13

Given the two orthogonal and significant Functions, it is possible to understand the relation between group membership and the Functions by plotting them in two dimensional space.

The plot of the means is shown in the next slide. From inspection it can be seen that Function 1 tends to distinguish between groups 1 and 2 (with groups 3 and 4 intermediary but closer to group 1) while Function 2 tends to separate groups 1 and 2 from groups 3 and 4. To the extent that the Functions can be interpreted, it might be possible to understand the reasons for this type of result.

The Functions are determined such that they maximally separate the groups, so it is possible that the functions may not have clear psychological interpretations. Recall the difficulties in interpreting non-rotated factors in factor analysis.

14



15

The standardized coefficients are presented in the table below. These are simply the unstandardized coefficients multiplied by the standard deviations of the variables. Some researchers recommend interpreting the Functions in terms of these coefficients. It is more common, however, to interpret the structure coefficients, which are the within cell correlations of the variables with the Function values (cf., Tabachnick & Fidell, 2006).

Standardized Canonical Discriminant Function Coefficients

	Function		
	1	2	3
x1	-1.083	.252	1.195
x2	1.592	-.482	-.130
x3	.212	1.042	.416
x4	.478	.216	1.084

16

Interpreting the Discriminant Functions

The structure matrix is interpreted much like a factor matrix in factor analysis. The Functions are identified and named, if possible, by the magnitude and the signs of the defining variables. Following are the coefficients.

Structure Matrix

	Function		
	1	2	3
x2	.510*	-.311	.252
x3	.244	.899*	-.096
x1	-.094	-.303	.530*
x4	.072	-.048	.406*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

17

Discriminant Function Analysis requires large sample sizes, and often a number of variables, otherwise the Function and Structure coefficients are unstable. Stevens (1996) suggests a ratio of N/p = 20. This example is for illustrative purposes only.

Function 1 obtains a high loading from variable X2 and a smaller one from X3, suggesting that it is determined by whatever is most characteristic of X2 and shared to some extent with X3.

Function 2 obtains a high positive loading from X3, and lower and negative loadings from X2 and X1, suggesting that it might describe a bipolar variable with X3 defining one end and X2 and X1 defining the other.

Function 3 was not significant, so it would not be meaningful to attempt to identify it.

18

Forecasting

Forecasting concerns the accuracy of predicting group membership. The table below shows the accuracy of prediction in this sample. This will overestimate accuracy to what might be expected if one were to obtain another sample of data and use the Discriminant functions from this analysis to make predictions for the new sample. Cross validation is required for meaningful estimates.

Classification Results[a]

			Predicted Group Membership				
		a	1.00	2.00	3.00	4.00	Total
Original	Count	1.00	3	1	0	1	5
		2.00	0	5	0	0	5
		3.00	1	0	3	1	5
		4.00	0	0	0	5	5
	%	1.00	60.0	20.0	.0	20.0	100.0
		2.00	.0	100.0	.0	.0	100.0
		3.00	20.0	.0	60.0	20.0	100.0
		4.00	.0	.0	.0	100.0	100.0

a. 80.0% of original grouped cases correctly classified.

19

Approaches to Discriminant Function Analysis

Like multiple regression and logistic regression, one could make use of a number of approaches. SPSS provides easy access to two approaches:

- Direct. As demonstrated above.
- Stepwise. Variables are entered (and removed) in order of their contributions to discrimination (based on F to Enter and F to remove).

20

References

Stevens, J. (1996). *Applied multivariate statistics for the behavioral sciences* (Third Edition). Mahwah, NJ: Lawrence Erlbaum.

Tabachnick, B.G. & Fidell, L . S. (2006). *Using multivariate statistics* (fifth edition). Boston, MA: Allyn & Bacon.

21