Block 15: Discriminant analysis

(Activity solutions can be found at the end of the document.)

We discuss the technique of **discriminant analysis**, initially by examining its relationship to regression analysis. Modelling of discriminant analysis is presented, along with the formulation, estimation, significance, interpretation and validation of results. **Two-group discriminant analysis** and **multiple discriminant analysis** are introduced.

Learning Objectives

- describe the concept of discriminant analysis, its objectives and its applications in market research
- outline the procedures for conducting discriminant analysis, including the formulation of the problem, estimation of the discriminant function coefficients, determination of significance, interpretation and validation
- discuss multiple discriminant analysis and the distinction between two-group and multiple discriminant analysis.

Reading List

Malhotra, N.K., D. Nunan and D.F. Birks. Marketing Research: An Applied Approach. (Pearson, 2017) 5th edition [ISBN 9781292103129] Chapter 23 (up to page 696).

15.1 Discriminant analysis

For each section of *Discriminant analysis*, use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

Overview

Discriminant analysis is a technique for analysing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature. The objectives of discriminant analysis are as follows.

Development of discriminant functions, or linear combinations of the predictor (independent) variables, which will best discriminate between the categories of the criterion (dependent) variable, i.e. the groups.

Examination of whether significant differences exist among the groups, in terms of the predictor variables.

Determination of which predictor variables contribute to most of the intergroup differences.

Classification of cases to one of the groups based on the values of the predictor variables.

Evaluation of the accuracy of classification.

Activity 15.1

What are the objectives of discriminant analysis?

Activity 15.2

Describe four examples of the application of discriminant analysis.

The discriminant analysis model

When the criterion variable has two categories, the technique is known as **two-group discriminant analysis**. When three or more categories are involved, the technique is referred to as **multiple discriminant analysis**. The main distinction is that, in the *two-group case*, it is possible to derive *only one* discriminant function. In multiple discriminant analysis, more than one discriminant function may be computed.

In general, with *G* groups and *k* predictor variables, it is possible to estimate up to the smaller of *G−1* or *k* discriminant functions.

The first discriminant function has the *highest ratio of between-groups to within-groups sum of squares*. The second discriminant function, uncorrelated with the first, has the second highest ratio, and so on. However, not all the discriminant functions may be statistically significant.

Figure 23.1 of the textbook provides a geometric interpretation of two-group discriminant analysis.

The **discriminant analysis model** involves linear combinations of the following form:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_k X_k$$

where:

- D = the discriminant score
- βis = the discriminant coefficients or weights
- Xis = the predictor variables.

The **coefficients** or **weights** (βis), are estimated so that the groups differ as much as possible based on the values of the discriminant function(s). This occurs when the ratio of between-groups sum of squares to within-groups sum of squares for the discriminant scores is at a maximum.

Activity 15.3

What is the main distinction between two-group and multiple discriminant analysis?

Activity 15.4

Describe the relationship between ANOVA, regression and discriminant analysis.

Statistics associated with discriminant analysis

Canonical correlation: Canonical correlation measures the extent of association between the discriminant scores and the groups. It is a measure of association between the discriminant function(s) and the set of dummy variables which define the group membership.

Centroid: The centroid is the mean value for the discriminant scores for a particular group. There are as many centroids as there are groups, as there is one for each group. The means for a group on (all) the discriminant function(s) are the *group centroids*.

Classification matrix: Sometimes also called the confusion or prediction matrix, the classification matrix contains the number of correctly classified and misclassified cases.

Discriminant function coefficients: The (unstandardised) discriminant function coefficients are the multipliers of the predictor variables, when the predictor variables are in the original units of measurement.

Discriminant scores: The unstandardised discriminant function coefficients are multiplied by the values of the predictor variables. These products are summed and added to the constant term to obtain the discriminant scores.

Eigenvalue: For each discriminant function, the eigenvalue is the ratio of between-groups sum of squares to within-groups sum of squares. Large eigenvalues imply superior discriminant functions.

FF **values and their significance**: These are calculated from a one-way ANOVA, with the grouping variable serving as the categorical independent variable. Each predictor variable, in turn, serves as the metric dependent variable in the ANOVA.

Group means and group standard deviations: These are computed for each predictor variable for each group.

Pooled within-groups correlation matrix: The pooled within-groups correlation matrix is computed by averaging the separate covariance matrices for all the groups.

Standardised discriminant function coefficients: The standardised discriminant function coefficients are used as the multipliers when the predictor variables have been standardised to ensure a mean of 0 and a variance of 1.

Structure correlations: Also referred to as *discriminant loadings*, the structure correlations represent the simple correlations between the predictor variables and the discriminant function.

Total correlation matrix: If the cases are treated as if they were from a single random sample and the correlations computed, the total correlation matrix is obtained.

Wilks' λ: Sometimes also called the U statistic, Wilks' λ for each predictor variable is the ratio of the within-groups sum of squares to the total sum of squares. Its value varies between 0 and 1. Large values of λ (near 1) indicate that group means do not seem to be different. Small values of λ (near 0) indicate that the group means seem to be different.

Conducting discriminant analysis

The process to conduct discriminant analysis is as follows:

Formulate the problem

⇓⇓

Estimate the discriminant function coefficients

⇓⇓

Determine the significance of the discriminant function

⇓⇓

Interpret the results

⇓⇓

Assess validity of discriminant analysis

Conducting discriminant analysis

Identify the objectives, the criterion (dependent) variable and the predictor (independent) variables. The criterion variable must consist of two or more *mutually exclusive and collectively exhaustive categories*. The predictor variables should be selected based on a theoretical model, previous research or the experience of the researcher.

One part of the sample, called the **estimation sample** (or *analysis sample*), is used for *estimation* of the discriminant function(s). The other part, called the **validation sample** (or *holdout sample*) is reserved for *validating* the discriminant function(s). Often the distributions of the numbers of cases in the estimation and validation samples follow the distribution in the total sample.

The **direct method** involves estimating the discriminant function(s) so that all the predictor variables are included simultaneously. In **stepwise discriminant analysis**, the predictor variables are entered sequentially, based on their ability to discriminate among groups.

The null hypothesis that, in the population, the means of all discriminant functions in all groups are equal can be statistically tested. In SPSS this test is based on Wilks' λ, such that if several discriminant functions are tested simultaneously (as in the case of multiple discriminant analysis), the Wilks' λ statistic is the product of the univariate Wilks' λs for each discriminant function. The corresponding *p*-value is estimated based on a chi-squared transformation of the Wilks' λλ statistic. *If the null hypothesis is rejected, indicating significant discrimination, one can proceed to interpret the results*.

The interpretation of the discriminant weights, or coefficients, is similar to that in multiple regression analysis. Given the multicollinearity in the predictor variables, there is no unambiguous measure of the relative importance of the predictor variables in discriminating between the groups. With this caveat in mind, we can obtain some idea of the relative importance of the predictor variables by *examining the absolute magnitude of the standardised discriminant function coefficients*.

Some idea of the relative importance of the predictor variables can also be obtained by *examining the structure correlations*, also called *canonical loadings* or *discriminant loadings*. These simple correlations between each predictor variable and the discriminant function(s) represent the variance which the predictor variable shares with the discriminant function(s).

Another aid to interpreting discriminant analysis results is to develop a **characteristic profile** for each group by describing each group in terms of the group means of the predictor variables.

Many computer programs, such as SPSS, offer a *leave-one-out cross-validation option*. The discriminant weights, estimated by using the estimation sample, are multiplied by the values of the predictor variables in the validation sample to generate discriminant scores for the cases in the validation sample. The cases are then assigned to groups based on their discriminant scores and an *appropriate decision rule*.

The **hit ratio**, or the percentage of cases correctly classified, can then be determined by summing the diagonal elements and dividing by the total number of cases.

It is helpful to compare the percentage of cases correctly classified by discriminant analysis to the percentage which would be obtained by chance. Classification accuracy achieved by discriminant analysis should be at least 25% greater than that obtained by chance.

Activity 15.5

What are the steps involved in conducting discriminant analysis?

Activity 15.6

How should the total sample be split for estimation and validation purposes?

Activity 15.7

What is Wilks’ λλ? For what purpose is it used?

Activity 15.8

Define discriminant scores.

Activity 15.9

Explain what is meant by an eigenvalue.

Activity 15.10

What is a classification matrix?

Activity 15.11

Explain the concept of structure correlations.

Activity 15.12

How is the statistical significance of discriminant analysis determined?

Activity 15.13

Describe a common procedure for determining the validity of discriminant analysis.

Activity 15.14

When the groups are of equal size, how is the accuracy of chance classification determined?

Activity 15.15

How does the stepwise discriminant procedure differ from the direct method?

Illustrative examples of discriminant analysis

Suppose a travel agency wanted to determine the salient characteristics of families who have visited a skiing resort during the last two years. Data were obtained from a pre-test of 42 households, split into an estimation sample of 30 households (shown in Table 23.2 of the textbook) and a validation sample of 12 households (shown in Table 23.3 of the textbook). The data can be downloaded from the file *Ski_resort.sav* or viewed in the table below.

Number	Visit	Income	Attitude	Importance	Size	Age	Amount spent	Usage
1	1	50.2	5	8	3	43	2	1
1	1	70.3	6	7	4	61	3	1
1	1	62.9	7	5	6	52	3	1
1	1	48.5	7	5	5	36	1	1
1	1	52.7	6	6	4	55	3	1
1	1	75.0	8	7	5	68	3	1
1	1	46.2	5	3	3	62	2	1
1	1	57.0	2	4	6	51	2	1
1	1	64.1	7	5	4	57	3	1
1	1	68.1	7	6	5	45	3	1
1	1	73.4	6	7	5	44	3	1
1	1	71.9	5	8	4	64	3	1
1	2	37.3	2	7	4	54	1	1
1	2	41.8	5	1	3	56	2	1
1	2	57.0	8	3	2	36	2	1
1	2	33.4	6	8	2	50	1	1
1	2	37.5	3	2	3	48	1	1
1	2	41.3	3	3	2	42	1	1
2	1	50.8	4	7	3	45	2	0
2	1	63.6	7	4	7	55	3	0
2	1	54.0	6	7	4	58	2	0
2	1	45.0	5	4	3	60	2	0
2	1	68.0	6	6	6	46	3	0
2	1	62.1	5	6	3	56	3	0
2	2	35.0	4	3	4	54	1	0
2	2	49.6	5	3	5	39	1	0
2	2	39.4	6	5	3	44	3	0
2	2	37.0	2	6	5	51	1	0
2	2	54.5	7	3	3	37	2	0
2	2	38.2	2	2	3	49	1	0

The households which visited a resort during the last two years were coded 12 1; those which did not, as 2. Data were obtained on:

- Annual family income
- Attitude towards travel
- Importance attached to family skiing holiday
- Household size
- Age of head of household
- Amount spent on family skiing holiday (used for three-group discriminant analysis).

Table 23.4 of the textbook provides the results of two-group discriminant analysis.

Table 23.5 of the textbook provides the results of three-group discriminant analysis.

Discussion forum, activities and discussion points

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Activitiess on the block's topics

1. In investigating the differences between heavy, light and non-users of frozen foods, it was found that the two largest standardised discriminant function coefficients were 0.97 for convenience orientation and 0.61 for income. Is it correct to conclude that convenience orientation is more important than income when each variable is considered by itself?

2. Given the following information, calculate the discriminant score for each participant. The value of the constant is 2.04.

Unstandardised discriminant function coefficients				
Age	0.38			
Income	0.44			
Risk-taking	-0.39			
Optimistic	1.26			
Respondent ID	**Age**	**Income**	**Risk-taking**	**Optimistic**
0246	36	43.7	21	65
1337	44	62.5	28	56
2375	57	33.5	25	40
2454	63	38.7	16	36

3.

4. Replicate the discriminant analysis results from the lecture example using the data file *Ski_resort.sav*. (An Excel version of the dataset is *Ski_resort.xlsx*.)

Video walkthrough of activity 3 (part 1).

Video walkthrough of activity 3 (part 2).

5. In a survey pre-test, data were obtained from 45 participants on Benetton clothes. These data are given in the file *Benetton.sav*, which gives the usage, gender, awareness, attitude, preference, intention and loyalty toward Benetton of a sample of Benetton users. Usage was coded as 1, 2 or 3, representing light, medium or heavy users, respectively. Gender was coded as 1 for females and 2 for males. Awareness, attitude, preference, intention and loyalty were measured on a 7-point Likert-type scale (1 = Very unfavourable, 7 = Very favourable). Note

that five participants have missing values which are denoted by 9. (An Excel version of the dataset is *Benetton.xlsx*.)

Analyse the Benetton data. Do the three usage groups differ in terms of awareness, attitude, preference, intention and loyalty toward Benetton when these predictor variables are considered simultaneously? Treat the whole dataset as the estimation sample.

Video walkthrough of activity 4.

Discussion point

1. ‘Is it meaningful to determine the relative importance of predictor variables in discriminating between the groups? Why or why not?’

Learning outcomes checklist

Use this to assess your own understanding of the chapter. You can always go back and amend the checklist when it comes to revision!

- ○ Describe the concept of discriminant analysis, its objectives and its applications in market research
- ○ Outline the procedures for conducting discriminant analysis, including the formulation of the problem, estimation of the discriminant function coefficients, determination of significance, interpretation and validation
- ○ Discuss multiple discriminant analysis and the distinction between two-group and multiple discriminant analysis.

Block 15: Discriminant analysis

Solution to Exercise 15.1

The main objectives of discriminant analysis are to:

1. develop linear combinations of the predictor variables
2. test the existence of significant differences among the groups in terms of the predictor variables
3. identify the predictor variables which contribute most to the inter-group differences
4. classify cases to one of the groups based on the values of the predictor variables
5. evaluate the accuracy of classification.

Solution to Exercise 15.2

Discriminant analysis can be used to answer questions such as the following.

1. In terms of demographic characteristics, how do customers who exhibit bank loyalty differ from those who do not?
2. Do heavy, medium and light users of bottled beer differ in terms of their consumption of frozen foods?
3. What psychographic characteristics help differentiate between price-sensitive and non-price-sensitive buyers of electronic equipment?
4. What are the distinguishing characteristics of consumers who respond to direct mail solicitations?

Solution to Exercise 15.3

In two-group discriminant analysis, the criterion variable has two categories and it is possible to derive only one discriminant function. However, in multiple discriminant analysis, the criterion variable has three or more categories and more than one discriminant function can be computed.

Solution to Exercise 15.4

Discriminant analysis, being a data analytical technique, is related to both regression and ANOVA since they all involve a single criterion (or dependent) variable and multiple predictor (or independent) variables. However, the dependent variable is metric in regression and ANOVA, but it is categorical in the case of discriminant analysis. The predictor variables are categorical in the case of ANOVA, but metric in the other two procedures.

Two-group discriminant analysis, in which the criterion variable has two categories, is closely related to multiple regression analysis. Here, multiple regression with dummy variables results in partial regression coefficients, which are proportional to the discriminant function coefficients.

Solution to Exercise 15.5

The steps involved in conducting discriminant analysis are the following.

- *Formulation* – defining the discriminant analysis problem by identifying the objectives, criterion variable and predictor variables, and dividing the data into the estimation sample and the validation sample.
- *Estimation* – developing a linear combination of the predictor variables and estimating the discriminant function(s).
- *Determination of statistical significance* – testing the null hypothesis that, in the population, the means of (all) the discriminant function(s) in all groups are equal.
- *Interpretation* – interpreting the discriminant weights or coefficients and determining the relative importance of the predictor variables.
- *Validation* – developing the classification matrix and determining the percentage of cases correctly classified.

Solution to Exercise 15.6

The total sample is divided into two parts as follows.

- *Estimation (or analysis) sample* – used for estimating the discriminant function(s).
- *Validation (or holdout) sample* – used for validating the discriminant function(s).

If the sample is large enough, it is split into two halves. One half serves as an estimation (or analysis) sample, and the other half serves as the validation (or holdout) sample. The role of the halves is then interchanged and the analysis repeated. The distributions of the numbers of cases in the estimation and holdout samples usually follow their distribution in the total sample.

Solution to Exercise 15.7

Wilks’ λλ (also called the UU statistic) is the ratio of the within-groups sum of squares to the total sum of squares for each predictor variable. It varies between 0 and 1. It is used to determine whether the group means are equal. Large values (near 1) of λλ indicate that the group means may be similar. Small values (near 0) indicate that the group means may be different.

Solution to Exercise 15.8

The unstandardised discriminant function coefficients are multiplied by the values of the predictor variables. These products are summed and added to the constant term to obtain the discriminant scores.

Solution to Exercise 15.9

For each discriminant function, the eigenvalue is the ratio of the between-groups sum of squares to the within-groups sum of squares. Larger eigenvalues imply superior discriminant functions.

Solution to Exercise 15.10

A classification matrix contains the number of correctly classified and misclassified cases. The correctly classified cases appear on the diagonal, whereas the off-diagonal elements represent cases which have been incorrectly classified. The sum of the diagonal elements divided by the total number of cases represents the hit ratio.

Solution to Exercise 15.11

Predictor variables are associated with the estimation of the discriminant function(s). Structure correlations (also known as discriminant loadings) measure the simple correlations between each predictor variable and the discriminant function(s). This represents the variance which the predictor variable shares with the discriminant function(s) and gives some idea of the relative importance of the predictor variables.

Solution to Exercise 15.12

The null hypothesis is that in the population, the means of all discriminant functions in all groups are equal. In SPSS, this test is based on Wilks' λλ. If several functions are tested simultaneously (as in the case of multiple discriminant analysis), the Wilks' λλ statistic is the product of the univariate Wilks' λλs for each discriminant function. The significance level is estimated based on a chi-squared transformation of the Wilks' λλstatistic. Rejection of the null hypothesis indicates significant discrimination.

Solution to Exercise 15.13

Validation involves developing the classification matrix. To do this, the data are randomly divided into two classes as follows.

- *Estimation (or analysis) sample* – for estimating the discriminant function(s).
- *Validation (or holdout) sample* – for developing the classification matrix.

The discriminant weights estimated by using the estimation sample are multiplied by the values of the predictor variables in the holdout sample to generate discriminant scores for the cases in the holdout sample. The cases are then assigned to groups based on their discriminant scores and an appropriate decision rule. The hit ratio, or the percentage of cases correctly classified, can then be determined by adding the diagonal elements and dividing by the total number of cases. This is compared to the rate which would be expected by chance classification.

Solution to Exercise 15.14

If the groups are equal in size, the percentage of chance classifications is 1 divided by the number of groups. Although no general guideline is available, it is often suggested that classification accuracy achieved by discriminant analysis should be at least 25% greater than that obtained by chance to indicate satisfactory validity.

Solution to Exercise 15.15

In the stepwise discriminant procedure, the predictors are entered sequentially based on their ability to discriminate between the groups. A univariate ANOVA is conducted with the groups as categorical variables and the predictor as the criterion, hence calculating an FF ratio for each predictor. The predictor with the highest FF ratio is the first to be selected for inclusion in the discriminant function, provided it meets certain significance and tolerance criteria. The second predictor is added based on the highest adjusted or partial FF ratio, taking into account the predictor already selected. Also, each predictor is tested for retention based on its association with the other selected predictors. These procedures are continued until all the predictors are covered.

Solutions to exercises on the block’s topics

1. It cannot be inferred unequivocally that convenience orientation is more important than income because we do not know the correlation between the two predictor variables, and hence multicollinearity may be present. However, given the magnitude of the standardised discriminant function coefficients, the conclusion seems reasonable.

2. We have the following estimated discriminant function:

$D=2.04+0.38\times\text{Age}+0.44\times\text{Income}-0.39\times\text{Risk-taking}+1.26\times\text{Optimistic}.$

Therefore, substituting in the given values for the predictor variables we have:

Respondent	**Score**
0246	108.658
1337	105.900
2375	79.090
2454	82.128

3. Using the ‘Usage’ variable as a dummy variable for selecting the estimation sample, we obtain the following output:

Analysis Case Processing Summary

Unweighted Cases		**N**	**Percent**
Valid		30	71.4
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Unselected	12	28.6
	Total	12	28.6
Total		42	100.0

The above output confirms that 30 of the 42 observations are being used as the estimation sample, as indicated by the dummy variable ‘Usage’.

Group Statistics

Resort visited		**Mean**	**Std. Deviation**	**Valid N (listwise)**	
				Unweighted	**Weighted**
Visited resort during last 2 years	Annual family income (£000)	60.520	9.8307	9.8307	15.000
	Attitude towards travel	5.400	1.9198	15	15.000
	Importance attached to family skiing holiday	5.800	1.8205	15	15.000
	Household size	4.333	1.2344	15	15.000
	Age of head of household	53.733	8.7706	15	15.000
Did not visit resort during last 2 years	Annual family income (£000)	41.913	7.5511	15	15.000
	Attitude towards travel	4.333	1.9518	15	15.000
	Importance attached to family skiing holiday	4.067	2.0517	15	15.000
	Household size	2.800	.9411	15	15.000
	Age of head of household	50.133	8.2710	15	15.000
Total	Annual family income (£000)	51.217	12.7952	30	30.000
	Attitude towards travel	4.867	1.9780	30	30.000

	Importance attached to family skiing holiday	4.933	2.0998	30	30.000
	Household size	3.567	1.3309	30	30.000
	Age of head of household	51.933	8.5740	30	30.000

Group statistics show the mean and standard deviation of each of the predictor variables by resort visited, and also combined (in the ‘Total’ area). Visual inspection of these confirm some differences between predictor variables for each resort (for example, mean income is £60,520 vs. £41,913) and these differences will be used to develop the discriminant function to best discriminate between the groups.

Tests of Equality of Group Means

	Wilks’ Lambda	**F**	**df1**	**df2**	**Sig.**
Annual family income ($000s)	.453	33.796	1	28	.000
Attitude towards travel	.925	2.277	1	28	.143
Importance attached to family skiing holiday	.824	5.990	1	28	.021
Household size	.657	14.636	1	28	.001
Age of head of household	.954	1.338	1	28	.257

For each predictor variable, the Wilks’ λλ is reported (the smaller the better). One-way ANOVA FF statistics are reported using the grouping variable (resort) as the factor and each predictor variable in turn as the criterion variable. The FF statistics have the stated degrees of freedom (‘df1’ = degrees of freedom in the numerator, and ‘df2’ = degrees of freedom in the denominator) along with the pp-value (denoted ‘Sig.’). The larger the FF statistic, the more significant the predictor variable. It appears that annual family income, household size and importance attached to a family skiing holiday are statistically significant predictor variables.

Pooled Within-Groups Matrices

		Annual family income (£000s)	**Attitude towards travel**	**Importance attached to family skiing holiday**	**Household size**	**Age of head of household**
Correlation	**Annual family income (£000s)**	1.000	.197	.091	.089	-.014
	Attitude towards travel	.197	1.000	.084	-.017	-.197
	Importance attached to family skiing holiday	.091	.084	1.000	.070	.017
	Household size	.089	-.017	.070	1.000	-.043
	Age of head of household	-.014	-.197	.017	-.043	1.000

We visually inspect the correlation matrix and observe no 'large' sample correlation coefficients indicating that there is no risk of multicollinearity among the predictor variables.

Eigenvalues

Function	**Eigenvalue**	**% of Variance**	**Cumulative %**	**Canonical Correlation**
1	1.786[a]	100.0	100.0	.801

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	**Wilks' Lambda**	**Chi-square**	**df**	**Sig.**
1	.359	26.130	5	.000

Since there are only two groups (the two different ski resort levels) we can only estimate one discriminant function. The eigenvalue is 1.786 which is the ratio of the between-groups sum of squares to the within-groups sum of squares. This value is ‘large’, which is encouraging. Indeed, the discriminant function can explain 100% of the variation (rounded) and there is a high canonical correlation (strong correlation between the discriminant scores and the groups). Wilks’ λ (0.359) is transfomed into a test statistic with a chi-squared distribution (with 5 degrees of freedom) resulting in a *p*-value of 0.000, which is highly significant and so the discriminant function is highly effective in discriminating between the groups.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Annual family income (£000s)	.743
Attitude toward travel	.096
Important attached to family skiing holiday	.233
Household size	.469
Age of head of household	.209

The standardised discriminant function coefficients are scale-invariant coefficients which allow us to assess the relative importance of the predictor variables. Larger values imply more important predictor variables. Therefore, in decreasing order of importance, we have:

- annual family income
- household size
- importance attached to family skiing holiday
- age of head of household
- attitude towards travel.

Structure Matrix

	Function
	1
Annual family income (£000s)	.822

Household size	.541
Important attached to family skiing holiday	.346
Attitude towards travel	.213
Age of head of household	.164

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

The structure matrix provides the correlations between each predictor variable and the discriminant function(s) (here there is only one discriminant function). Higher correlations indicate better predictor variables. These are shown decreasing in the correlations, with the order similar to that based on the standardised discriminant function coefficients:

- annual family income
- household size
- importance attached to family skiing holiday
- attitude towards travel
- age of head of household.

Canonical Discriminant Function Coefficients

	Function
	1
Annual family income (£000s)	.085
Attitude toward travel	.050
Important attached to family skiing holiday	.120
Household size	.427
Age of head of household	.025
(Constant)	-7.975

Unstandardized coefficients

The canonical discriminant function coefficients are the *unstandardised* discriminant function coefficients which can then be used to determine the discriminant scores. Therefore, the estimated discriminant function is:

$$D=-7.975+0.085\times\text{Income}+0.050\times\text{Attitude}+0.120\times\text{Importance}+0.427\times\text{Size}+0.025\times\text{Age}.$$

Functions at Group Centroids

Resort visited	**Function**
	1
Visited resort during last 2 years	1.291
Did not visit resort during last 2 years	-1.291

The above box provides the discriminant scores for the group centroids (i.e. for ‘average’ or ‘representative’ members of each group). This can then be used to establish a decision rule for classification. Here, an individual with a positive discriminant score would be classified as having visited the resort during the last 2 years, and an individual with a negative discriminant score would be classified as not having visited the resort. The cut-off point being:

$$1.291 + \frac{(-1.291)}{2} = 0$$

Classification Results [b, c, d]

			Resort visited	**Predicted Group Membership**		**Total**
				Visited resort during last 2 years	**Did not visit resort during last 2 years**	
Cases Selected	**Original**	Count	Visited resort during last 2 years	12	3	15
			Did not visit resort during last 2 years	0	15	15
		%	Visited resort during last 2 years	80.0	20.0	100.0

			Did not visit resort during last 2 years	.0	100.0	100.0
	Cross-validated[a]	Count	Visited resort during last 2 years	11	4	15
			Did not visit resort during last 2 years	2	13	15
		%	Visited resort during last 2 years	73.3	26.7	100.0
			Did not visit resort during last 2 years	13.3	86.7	100.0
Cases Not Selected	**Original**	**Count**	Visited resort during last 2 years	4	2	6
			Did not visit resort during last 2 years	0	6	6
		%	Visited resort during last 2 years	66.7	33.3	100.0
			Did not visit resort during last 2 years	.0	100.0	100.0

l. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

m. 90.0% of selected original grouped cases correctly classified.

n. 83.3% of unselected original grouped cases correctly classified.

o. 80.0% of selected cross-validated grouped cases correctly classified.

The classification matrix is used to determine the predictive ability of the discriminant analysis model. Our interest lies in the ‘Cases Not Selected’ predictions, which is the validation sample. Of the 12 cases in the validation sample, there are 10 correct classifications (4 correctly classified as having visited the resort, and 6 correctly classified as not having visited). However, there were 2 incorrect classifications (visited, but predicted not to have visited). Therefore, the hit ratio based on the validation sample is (4+6)/12=0.833, that is 83.3%.

If you look at your data matrix, you will see four extra columns which represent, respectively, predicted group membership, discriminant scores and probabilities of group membership (one column for each group).

Now we repeat the above exercise but use ‘Amount_spent’ as the grouping variable. This takes three values (1, 2 and 3) representing three different expenditure levels (low, medium and high, respectively).

Selected SPSS output follows. Think about how to interpret each box based on the above analysis for a two-group model.

Group Statistics

Amount spent on family skiing holiday		**Mean**	**Std. Deviation**	**Valid N (listwise)**	
				Unweighted	**Weighted**
Low	Annual family income (£000)	38.570	5.2972	10	10.000
	Attitude towards travel	4.500	1.7159	10	10.000
	Importance attached to family skiing holiday	4.700	1.8886	10	10.000
	Household size	3.100	1.1972	10	10.000
	Age of head of household	50.300	8.0973	10	10.000
Medium	Annual family income (£000)	50.110	6.0023	10	10.000
	Attitude towards travel	4.000	2.3570	10	10.000
	Importance attached to family skiing holiday	4.200	2.4855	10	10.000
	Household size	3.400	1.1353	10	10.000
	Age of head of household	49.500	9.2526	10	10.000
High	Annual family income (£000)	64.970	8.6143	10	10.000
	Attitude towards travel	6.100	1.1972	10	10.000
	Importance attached to family skiing holiday	5.900	1.6633	10	10.000
	Household size	4.200	1.1353	10	10.000

	Age of head of household	56.000	7.6012	10	10.000
Total	Annual family income (£000)	51.217	12.7952	30	30.000
	Attitude towards travel	4.867	1.9780	30	30.000
	Importance attached to family skiing holiday	4.933	2.0998	30	30.000
	Household size	3.567	1.3309	30	30.000
	Age of head of household	51.933	8.5740	30	30.000

Tests of Equality of Group Means

	Wilks' Lambda	**F**	**df1**	**df2**	**Sig.**
Annual family income ($000s)	.262	37.997	2	27	.000
Attitude towards travel	.788	3.634	2	27	.040
Importance attached to family skiing holiday	.881	1.830	2	27	.180
Household size	.874	1.944	2	27	.163
Age of head of household	.882	1.804	2	27	.184

Eigenvalues

Function	**Eigenvalue**	**% of Variance**	**Cumulative %**	**Canonical Correlation**
1	3.819[a]	93.9	93.9	.890

2	.247[a]	6.1	100.0	.445

p. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.166	44.831	10	.000
-2	.802	5.517	4	.238

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Annual family income (£000s)	1.047	-.421
Attitude toward travel	.340	.769
Important attached to family skiing holiday	-.142	.534
Household size	-.163	.129
Age of head of household	.495	.524

Structure Matrix

	Function	
	1	2
Annual family income (£000s)	.856*	-.278
Household size	.193*	.077

Attitude towards travel	.219	.588*
Important attached to family skiing holiday	.149	.454*
Age of head of household	.166	.341*

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 and any discriminant function and any discriminant function

Canonical Discriminant Function Coefficients

	Function	
	1	**2**
Annual family income (£000s)	.154	-.062
Attitude towards travel	.187	.422
Important attached to family skiing holiday	-.070	.261
Household size	-.127	.100
Age of head of household	.059	.063
(Constant)	-11.094	-3.792

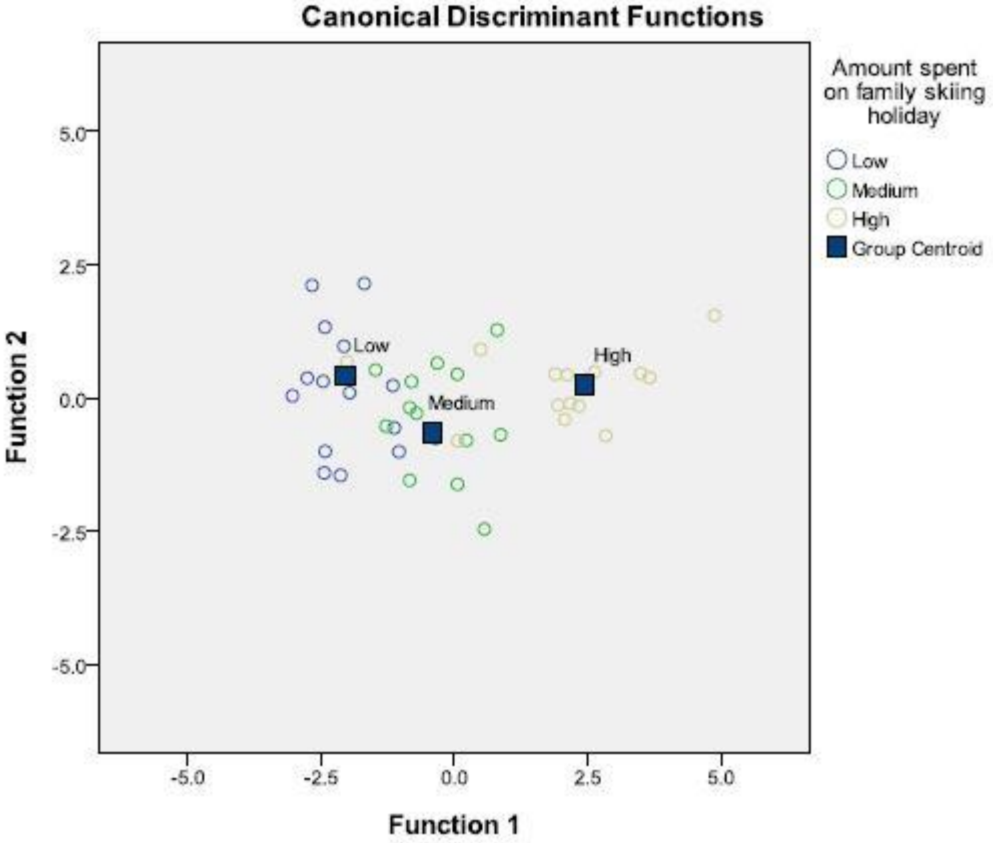
Unstandardized coefficients

Functions at Group Centroids

Amount spent on family skiing holiday	**Function**	
	1	**2**
Low	-2.041	.418
Medium	-.405	-.659

High	2.446	.240

Unstandardized canonical discriminant functions evaluated at group means



The diagram plots each observation using the discriminant scores based on both discriminant functions as the coordinates (discriminant function 1 score on the x-axis, and discriminant function 2 score on the y-axis). The squares are the group centroids - that is, the ‘average’ individual for each group. We see that discriminant function 1 is very effective at discriminating between the groups (based on the discriminant score using the first function), while the second discriminant function is less effective (perhaps not surprising, given it was insignificant).

Classification Results [b, c, d]

			Amount spent on family skiing holiday	**Predicted Group Membership**			**Total**
				Low	**Medium**	**Heavy**	
Cases Selected	**Original**	**Count**	Low	9	1	0	10
			Medium	1	9	0	10
			High	0	2	8	10
		%	Low	90.0	10.0	.0	100.0

			Medium	10.0	90.0	.0	100.0
			High	.0	20.0	80.0	100.0
	Cross-validated[a]	Count	Low	7	3	0	10
			Medium	4	5	1	10
			High	0	2	8	10
		%	Low	70.0	30.0	.0	100.0
			Medium	40.0	50.0	10.0	100.0
			High	.0	20.0	80.0	100.0
Cases Not Selected	Original	Count	Low	3	1	0	4
			Medium	0	3	1	4
			High	1	0	3	4
		%	Low	75.0	25.0	.0	100.0
			Medium	.0	75.0	25.0	100.0
			High	25.0	.0	75.0	100.0

q. Cross validation is done only for those cases in the analysis.In cross validation, each case is classified by the functions derived from all cases other than that case.

r. 86.7% of selected original grouped cases correctly classified.

s. 75.0% of unselected original grouped cases correctly classified.

t. 66.7% of selected cross-validated grouped cases correctly classified.

The hit ratio based on the validation sample is (3+3+3)/12=0.75(3+3+3)/12=0.75, which is 75%. However, note the severity of one of the incorrectly classified cases - one individual is predicted to be a low spender, when they were actually a high spender. This is a ‘worse’ error than had s/he been predicted to be a medium spender. So two discriminant analysis models with the same hit ratio are not necessarily equally good - it depends on the severity of the incorrectly classified cases.

Again, check your data matrix to see the predicted group membership, discriminant scores and probabilities of group membership (one column for each group).

4. Some selected SPSS output follows.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Awareness	.271	49.758	2	37	.000
Attitude	.296	44.070	2	37	.000
Preference	.502	18.350	2	37	.000
Intention	.989	.214	2	37	.808
Loyalty	.988	.220	2	37	.804

Pooled Within-Groups Matrices

		Awareness	Attitude	Preference	Intention	Loyalty
Correlation	**Awareness**	1.000	.308	.135	.107	.194
	Attitude	.308	1.000	.082	.239	.249
	Preference	.135	.082	1.000	.375	.354
	Intention	.107	.239	.375	1.000	.765
	Loyalty	.194	.249	.354	.765	1.000

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	5.277[a]	96.7	96.7	.917
2	.182[a]	3.3	100.0	.393

. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.135	70.151	10	.000

2	.846	5.858	4	.210

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	**2**
Awareness	.547	-.647
Attitude	.588	.124
Preference	.473	.876
Intention	-.165	-.654
Loyalty	-.337	.401

Structure Matrix

	Function	
	1	**2**
Awareness	.708*	-.484
Attitude	.672*	-.060
Preference	.414	.696*
Loyalty	-.042	.117*
Intention	-.046	-.058*

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

Canonical Discriminant Function Coefficients

	Function

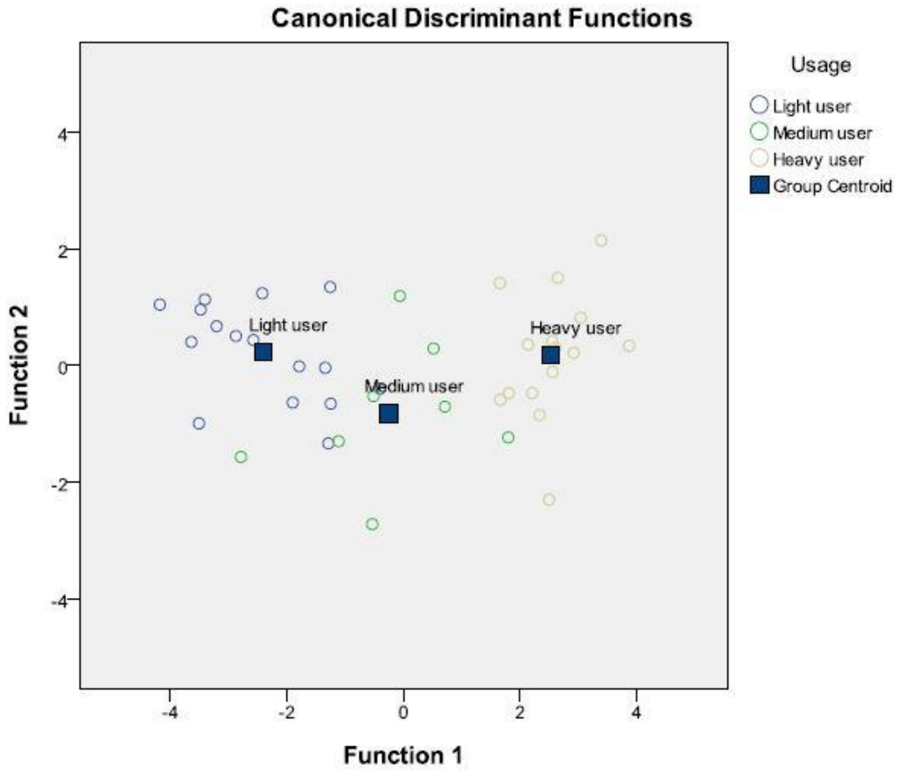
	1	2
Awareness	.540	-.639
Attitude	.548	.115
Preference	.410	.759
Intention	-.097	-.387
Loyalty	-.192	.229
(Constant)	-5.145	-.308

Unstandardized coefficients

Functions at Group Centroids

Usage	**Function**	
	1	**2**
Light user	-2.405	.230
Medium user	-.246	-.820
Heavy user	2.528	.179

Unstandardized canonical discriminant functions evaluated at group means



Classification Results [b, c]

		Usage	Predicted Group Membership			Total
			Light user	**Medium user**	**Heavy user**	
Original	Count	Light user	13	3	0	16
		Medium user	1	6	1	8
		Heavy user	0	0	16	16
	%	Light user	81.3	18.8	.0	100.0
		Medium user	12.5	75.0	12.5	100.0
		Heavy user	.0	.0	100.0	100.0
Cross-validated[a]	Count	Light user	12	4	0	16
		Medium user	2	4	2	8
		Heavy user	0	0	16	16
	%	Light user	75.0	25.0	.0	100.0

		Medium user	25.0	50.0	25.0	100.0
		Heavy user	.0	.0	100.0	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 87.5% of original grouped cases correctly classified.

c. 80.0% of cross-validated grouped cases correctly classified.

The first discriminant function accounts for 96.7% of the explained variance and is significant (the chi-squared test statistic value is 70.151, and the pp-value is 0.000). The second discriminant function accounts for only 3.3% of the explained variance and is not significant. Univariate FF tests show that only awareness, attitude and preference are significantly different across the three groups.

The standardised canonical discriminant function coefficients are:

	Function	
	1	**2**
Awareness	0.547	-0.647
Attitude	0.588	0.124
Preference	0.473	0.876
Intention	-0.165	-0.654
Loyalty	-0.337	0.401

The combined-groups plot shows the variation of the group centroids across the first discriminant function (which was highly significant), and little variation across the second (insignificant) discriminant function.

80.0% of the cases are correctly classified when using the leave-one-out cross-validation approach. This is a reasonable hit ratio.

Issues relevant to this discussion include the following.

Commentary on Discussion point

- With multicollinearity in the predictor variables, there is no unambiguous measure of the relative importance of the predictor variables in discriminating between the groups.

- Some idea of the relative importance of the predictor variables may be obtained by examining the absolute magnitude of the standardised discriminant function coefficients and the structure correlations or discriminant loadings.
- Both the standardised discriminant function coefficients and the structure correlations must be interpreted with caution.

Therefore, it would be meaningful to determine the relative importance of predictor variables. However, no method for reliable interpretation exists.