Source: © Getty Images

21 Discriminant analysis

> *Discriminant analysis is used to estimate the relationship between a categorical dependent variable and a set of interval scaled, independent variables.*

Objectives

After reading this chapter, you should be able to:

1 describe the concept of discriminant analysis, its objectives and its applications in marketing research;

2 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;

3 discuss multiple discriminant analysis and the distinction between two-group and multiple discriminant analysis;

4 explain stepwise discriminant analysis and describe the Mahalanobis procedure.

STAGE 1 Problem definition

STAGE 2 Research approach developed

STAGE 3 Research design developed

STAGE 4 Fieldwork or data collection

STAGE 5 Data preparation and analysis

STAGE 6 Report preparation and presentation

Overview

This chapter discusses the technique of discriminant analysis. We begin by examining the relationship of this procedure to regression analysis (Chapter 20) and analysis of variance (Chapter 19). We present a model and describe the general procedure for conducting discriminant analysis, with an emphasis on formulation, estimation, determination of significance, interpretation, and validation of the results. The procedure is illustrated with an example of two-group discriminant analysis, followed by an example of multiple (three-group) discriminant analysis. The stepwise discriminant analysis procedure is also covered.

We begin with an example that illustrates an application of multiple discriminant analysis.

Example

An eye for a bargain[1]

A study of 294 consumers was undertaken to determine the correlates of 'rebate proneness': in other words, the characteristics of consumers who respond favourably to direct mail promotions that offer a discount on the normal purchase price. The predictor variables were four factors related to household shopping attitudes and behaviour and selected demographic characteristics (gender, age and income). The dependent variable was the extent to which respondents were predisposed to take up the offer of a rebate, of which three levels were identified. Respondents who reported no purchases triggered by a rebate during the past 12 months were classified as *non-users*, those who reported one or two such purchases as *light users*, and those with more than two purchases as *frequent users* of discounts. Multiple discriminant analysis was used to analyse the data.

Two primary findings emerged. First, consumers' perception of the effort/value relationship was the most effective variable in discriminating among frequent users, light users and non-users of rebate offers. Clearly, 'rebate-prone' consumers associate less effort with fulfilling the requirements of the rebated purchase, and are willing to accept a relatively smaller refund than other customers. Second, consumers who were aware of the regular prices of products, so that they recognise bargains, are more likely than others to respond to rebate offers.

These findings were used by DIRECTV (**www.directtv.com**) when it added the National Geographic Channel to its programming. As a way to promote its service, DIRECTV offered three levels of rebates for new customers. The company felt that this would encourage the rebate-sensitive new customers to choose DIRECTV, which it did.

In this example, significant intergroup differences were found using multiple predictor variables. An examination of differences across groups lies at the heart of the basic concept of discriminant analysis.

Basic concept

Discriminant analysis
A technique for analysing marketing research data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature.

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.[2] For example, the dependent variable may be the choice of the make of a new car (A, B or C) and the independent variables may be ratings of attributes of PCs on a seven-point Likert scale. The objectives of discriminant analysis are as follows:

1 Development of discriminant functions, or linear combinations of the predictor or independent variables, that best discriminate between the categories of the criterion or dependent variable (groups).
2 Examination of whether significant differences exist among the groups, in terms of the predictor variables.
3 Determination of which predictor variables contribute to most of the intergroup differences.
4 Classification of cases to one of the groups based on the values of the predictor variables.
5 Evaluation of the accuracy of classification.

Discriminant analysis techniques are described by the number of categories possessed by the criterion variable. 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**, but in multiple discriminant analysis more than one function may be computed.[3]

Two-group discriminant analysis
Discriminant analysis technique where the criterion variable has two categories.

Multiple discriminant analysis
Discriminant analysis technique where the criterion variable involves three or more categories.

Discriminant function
The linear combination of independent variables developed by discriminant analysis that will best discriminate between the categories of the dependent variable.

Examples of discriminant analysis abound in marketing research. This technique can be used to answer questions such as the following:[4]

- In terms of demographic characteristics, how do customers who exhibit loyalty to a particular car manufacturer differ from those who do not?
- Do heavy users, medium users and light users of soft drinks differ in terms of their consumption of frozen foods?
- What psychographic characteristics help differentiate between price-sensitive and non-price-sensitive buyers of groceries?
- Do market segments differ in their media consumption habits?
- What are the distinguishing characteristics of consumers who respond to direct mail offers?

Relationship to regression and ANOVA

The relationships between discriminant analysis, analysis of variance (ANOVA) and regression analysis are shown in Table 21.1.

We explain these relationships with an example in which the researcher is attempting to explain the amount of life insurance purchased in terms of age and income. All three procedures involve a single criterion or dependent variable and multiple predictor or independent variables. The nature of these variables differs, however. In ANOVA and regression analysis, the dependent variable is metric or interval scaled (amount of life

Table 21.1 Similarities and differences among ANOVA, regression and discriminant analysis

	ANOVA	*Regression*	*Discriminant analysis*
Similarities			
Number of dependent variables	One	One	One
Number of independent variables	Multiple	Multiple	Multiple
Differences			
Nature of the dependent variable	Metric	Metric	Categorical Binary
Nature of the independent variable	Categorical	Metric	Metric

insurance purchased in euros), whereas in discriminant analysis, it is categorical (amount of life insurance purchased classified as high, medium or low). The independent variables are categorical in the case of ANOVA (age and income are each classified as high, medium or low) but metric in the case of regression and discriminant analysis (age in years and income in euros, i.e. both measured on a ratio scale).

Two-group discriminant analysis, in which the dependent variable has only two categories, is closely related to multiple regression analysis. In this case, multiple regression, in which the dependent variable is coded as a 0 or 1 dummy variable, results in partial regression coefficients that are proportional to discriminant function coefficients (see the following section on the discriminant analysis model).

Discriminant analysis model

Discriminant analysis model
The statistical model on which discriminant analysis is based.

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

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

where D = discriminant score
b's = discriminant coefficients or weights
X = predictor or independent variable

The coefficients or weights (b) are estimated so that the groups differ as much as possible on the values of the discriminant function. This occurs when the ratio of between-group sum of squares to within-group sum of squares for the discriminant scores is at a maximum. Any other linear combination of the predictors will result in a smaller ratio. The technical details of estimation are described in the appendix to this chapter.

Statistics associated with discriminant analysis

The important statistics associated with discriminant analysis include the following:

Canonical correlation. Canonical correlation measures the extent of association between the discriminant scores and the groups. It is a measure of association between the single discriminant function and the set of dummy variables that define the group membership.

Centroid. The centroid is the mean values 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 functions are the *group centroids.*

Classification matrix. Sometimes also called confusion or prediction matrix, the classification matrix contains the number of correctly classified and misclassified cases. The correctly classified cases appear on the diagonal, because the predicted and actual groups are the same. The off-diagonal elements represent cases that have been incorrectly classified. The sum of the diagonal elements divided by the total number of cases represents the *hit ratio.*

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

Discriminant scores. The unstandardised coefficients are multiplied by the values of the 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-group to within-group sums of squares. Large eigenvalues imply superior functions.

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

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

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

Standardised discriminant function coefficients. The standardised discriminant function coefficients are the discriminant function coefficients that are used as the multipliers when the variables have been standardised to 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 predictors and the discriminant function.

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

Wilks' λ. Sometimes also called the U statistic, Wilks' λ for each predictor is the ratio of the within-group 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.

The assumptions in discriminant analysis are that each of the groups is a sample from a multivariate normal population and that all the populations have the same covariance matrix. The role of these assumptions and the statistics just described can be better understood by examining the procedure for conducting discriminant analysis.

Conducting discriminant analysis

The steps involved in conducting discriminant analysis consist of formulation, estimation, determination of significance, interpretation and validation (see Figure 21.1). These steps are discussed and illustrated within the context of two-group discriminant analysis. Discriminant analysis with more than two groups is discussed later in this chapter.

Formulate the problem

The first step in discriminant analysis is to formulate the problem by identifying the objectives, the criterion variable and the independent variables. The criterion variable

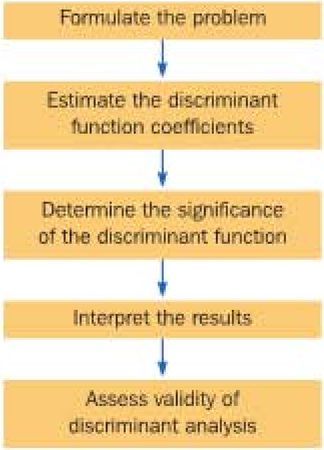

Formulate the problem
Estimate the discriminant function coefficients
Determine the significance of the discriminant function
Interpret the results
Assess validity of discriminant analysis

Figure 21.1 Conducting discriminant analysis

must consist of two or more mutually exclusive and collectively exhaustive categories. When the dependent variable is interval or ratio scaled, it must first be converted into categories. For example, attitude towards the brand, measured on a seven-point scale, could be categorised as unfavourable (1, 2, 3), neutral (4) or favourable (5, 6, 7). Alternatively, one could plot the distribution of the dependent variable and form groups of equal size by determining the appropriate cut-off points for each category. The predictor variables should be selected based on a theoretical model or previous research, or in the case of exploratory research, the experience of the researcher should guide their selection.

Analysis sample
Part of the total sample used to check the results of the discriminant function.

Validation sample
That part of the total sample used to check the results of the estimation sample.

The next step is to divide the sample into two parts. One part of the sample, called the *estimation* or **analysis sample**, is used for estimation of the discriminant function. The other part, called the *holdout* or **validation sample**, is reserved for validating the discriminant function. When the sample is large enough, it can be split in half. One half serves as the analysis sample, and the other is used for validation. The roles of the halves are then interchanged and the analysis is repeated. This is called double cross-validation and is similar to the procedure discussed in regression analysis (Chapter 20).

Often, the distribution of the number of cases in the analysis and validation samples follows the distribution in the total sample. For instance, if the total sample contained 50% loyal and 50% non-loyal consumers, then the analysis and validation samples would each contain 50% loyal and 50% non-loyal consumers. On the other hand, if the sample contained 25% loyal and 75% non-loyal consumers, the analysis and validation samples would be selected to reflect the same distribution (25% vs. 75%).

Finally, it has been suggested that the validation of the discriminant function should be conducted repeatedly. Each time, the sample should be split into different analysis and validation parts. The discriminant function should be estimated and the validation analysis carried out. Thus, the validation assessment is based on a number of trials. More rigorous methods have also been suggested.[5]

To illustrate two-group discriminant analysis better, let us look at an example. Suppose that we want to determine the salient characteristics of families that have visited a skiing resort during the last two years. Data were obtained from a pretest sample of 42 households. Of these, 30 households, shown in Table 21.2, were included in the analysis sample and the remaining 12, shown in Table 21.3, were part of the validation sample. The households that visited a resort during the last two years were coded as 1; those that did not, as 2 (VISIT). Both the analysis and validation samples were balanced in terms of VISIT. As can be seen, the analysis sample contains 15 households in each category, whereas the validation sample had 6 in each category. Data were also obtained on annual family income (INCOME), attitude towards travel (TRAVEL, measured on a nine-point scale), importance attached to a family skiing holiday (HOLIDAY, measured on a nine-point scale), household size (HSIZE) and age of the head of the household (AGE).

Estimate the discriminant function coefficients

Direct method
An approach to discriminant analysis that involves estimating the discriminant function so that all the predictors are included simultaneously.

Stepwise discriminant analysis
Discriminant analysis in which the predictors are entered sequentially based on their ability to discriminate between the groups.

Once the analysis sample has been identified, as in Table 21.2, we can estimate the discriminant function coefficients. Two broad approaches are available. The **direct method** involves estimating the discriminant function so that all the predictors are included simultaneously. In this case, each independent variable is included, regardless of its discriminating power. This method is appropriate when, based on previous research or a theoretical model, the researcher wants the discrimination to be based on all the predictors. An alternative approach is the stepwise method. In **stepwise discriminant analysis**, the predictor variables are entered sequentially, based on their ability to discriminate among groups. This method, described in more detail later in this chapter, is appropriate when the researcher wants to select a subset of the predictors for inclusion in the discriminant function.

The results of running two-group discriminant analysis on the data of Table 21.2 using a popular statistical analysis package are presented in Table 21.4. Some intuitive feel for the results may be obtained by examining the group means and standard deviations. It appears that the two groups are more widely separated in terms of income than other

Table 21.2 Information on resort visits: analysis sample

Number	*Resort visit*	*Annual family income (e000)*	*Attitude towards travel*	*Importance attached to family skiing holiday*	*Household size*	*Age of head of household*	*Amount spent spent on family skiing holiday*
1	1	50.2	5	8	3	43	M (2)
2	1	70.3	6	7	4	61	H (3)
3	1	62.9	7	5	6	52	H (3)
4	1	48.5	7	5	5	36	L (1)
5	1	52.7	6	6	4	55	H (3)
6	1	75.0	8	7	5	68	H (3)
7	1	46.2	5	3	3	62	M (2)
8	1	57.0	2	4	6	51	M (2)
9	1	64.1	7	5	4	57	H (3)
10	1	68.1	7	6	5	45	H (3)
11	1	73.4	6	7	5	44	H (3)
12	1	71.9	5	8	4	64	H (3)
13	1	56.2	1	8	6	54	M (2)
14	1	49.3	4	2	3	56	H (3)
15	1	62.0	5	6	2	58	H (3)
16	2	32.1	5	4	3	58	L (1)
17	2	36.2	4	3	2	55	L (1)
18	2	43.2	2	5	2	57	M (2)
19	2	50.4	5	2	4	37	M (2)
20	2	44.1	6	6	3	42	M (2)
21	2	38.3	6	6	2	45	L (1)
22	2	55.0	1	2	2	57	M (2)
23	2	46.1	3	5	3	51	L (1)
24	2	35.0	6	4	5	64	L (1)
25	2	37.3	2	7	4	54	L (1)
26	2	41.8	5	1	3	56	M (2)
27	2	57.0	8	3	2	36	M (2)
28	2	33.4	6	8	2	50	L (1)
29	2	37.5	3	2	3	48	L (1)
30	2	41.3	3	3	2	42	L (1)

variables, and there appears to be more of a separation on the importance attached to the family skiing holiday than on attitude towards travel. The difference between the two groups on age of the head of the household is small, and the standard deviation of this variable is large.

Table 21.3 Information on resort visits: validation sample

Number	*Resort visit*	*Annual family income (€000)*	*Attitude towards travel*	*Importance attached to family skiing holiday*	*Household size*	*Age of head of household*	*Amount spent on family skiing holiday*
1	1	50.8	4	7	3	45	M (2)
2	1	63.6	7	4	7	55	H (3)
3	1	54.0	6	7	4	58	M (2)
4	1	45.0	5	4	3	60	M (2)
5	1	68.0	6	6	6	46	H (3)
6	1	62.1	5	6	3	56	H (3)
7	2	35.0	4	3	4	54	L (1)
8	2	49.6	5	3	5	39	L (1)
9	2	39.4	6	5	3	44	H (3)
10	2	37.0	2	6	5	51	L (1)
11	2	54.5	7	3	3	37	M (2)
12	2	38.2	2	2	3	49	L (1)

Table 21.4 Results of two-group discriminant analysis

Group means

Visit	*INCOME*	*TRAVEL*	*HOLIDAY*	*HSIZE*	*AGE*
1	60.52000	5.40000	5.80000	4.33333	53.73333
2	41.91333	4.33333	4.06667	2.80000	50.13333
Total	51.21667	4.86667	4.93333	3.56667	51.93333

Group standard deviations

1	9.83065	1.91982	1.82052	1.23443	8.77062
2	7.55115	1.95180	2.05171	0.94112	8.27101
Total	12.79523	1.97804	2.09981	1.33089	8.57395

Pooled within-groups correlation matrix

INCOME	1.00000				
TRAVEL	0.19745	1.00000			
HOLIDAY	0.09148	0.08434	1.00000		
HSIZE	0.08887	−0.01681	0.07046	1.00000	
AGE	−0.01431	−0.19709	0.01742	−0.04301	1.00000

Table 21.4 Continued

Wilks' λ (*U* statistic) and univariate *F* ratio with 1 and 28 degrees of freedom

Variable	*Wilks' λ*	*F*	*Significance*
INCOME	0.45310	33.800	0.0000
TRAVEL	0.92479	2.277	0.1425
HOLIDAY	0.82377	5.990	0.0209
HSIZE	0.65672	14.640	0.0007
AGE	0.95441	1.338	0.2572

Canonical discriminant functions

Function	*Eigenvalue*	*Per cent of variance*	*Cumulative percentage*	*Canonical correlation*	*After function*	*Wilks' λ*	*Chi-square*	*df*	*Sig.*
1*	1.7862	100.00	100.00	0.8007	0	0.3589	26.13	5	0.0001

*Marks the one canonical discriminant function remaining in the analysis.

Standard canonical discriminant function coefficients

	Func 1
INCOME	0.74301
TRAVEL	0.09611
HOLIDAY	0.23329
HSIZE	0.46911
AGE	0.20922

Structure matrix: pooled within-groups correlations between discriminating variables and canonical discriminant functions (variables ordered by size of correlation within function)

	Func 1
INCOME	0.82202
HSIZE	0.54096
HOLIDAY	0.34607
TRAVEL	0.21337
AGE	0.16354

Unstandardised canonical discriminant function coefficients

	Func 1
INCOME	0.8476710E-01
TRAVEL	0.4964455E-01
HOLIDAY	0.1202813
HSIZE	0.4273893
AGE	0.2454380E-01
(Constant)	−7.975476

Table 21.4 Continued

Canonical discriminant functions evaluated at group means (group centroids)

Group	*Func 1*
1	1.29118
2	−1.29118

Classification results

			Predicted group membership		*Total*
		Visit	1	2	
Original	Count	1	12	3	15
		2	0	15	15
	%	1	80.0	20.0	100.0
		2	0.0	100.0	100.0
Cross-validated	Count	1	11	4	15
		2	2	13	15
	%	1	73.3	26.7	100.0
		2	13.3	86.7	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] 90.0% of original grouped cases correctly classified.

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

Classification results for cases not selected for use in analysis (holdout sample)

			Predicted group membership	
	Actual group	*No. of cases*	1	2
Group	1	6	4	2
			66.7%	33.3%
Group	2	6	0	6
			0.0%	100%

Percentage of grouped cases correctly classified: 83.33%

The pooled within-groups correlation matrix indicates low correlations between the predictors. Multicollinearity is unlikely to be a problem. The significance of the univariate *F* ratios indicates that, when the predictors are considered individually, only income, importance of holiday and household size significantly differentiate between those who visited a resort and those who did not.

Because there are two groups, only one discriminant function is estimated. The eigenvalue associated with this function is 1.7862, and it accounts for 100% of the explained variance. The canonical correlation associated with this function is 0.8007. The square of this correlation, $(0.8007)^2 = 0.64$, indicates that 64% of the variance in the dependent variable (VISIT) is explained or accounted for by this model. The next step is determination of significance.

Determine the significance of the discriminant function

It would not be meaningful to interpret the analysis if the discriminant functions estimated were not statistically significant. 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. If several functions are tested simultaneously (as in the case of multiple discriminant analysis), the Wilks' λ statistic is the product of the univariate λ for each function. The significance level is estimated based on a chi-square transformation of the statistic. In testing for significance in the holiday resort example (see Table 21.4), it may be noted that the Wilks' λ associated with the function is 0.3589, which transforms to a chi-square of 26.13 with 5 degrees of freedom. This is significant beyond the 0.05 level. In SAS, an approximate *F* statistic, based on an approximation to the distribution of the likelihood ratio, is calculated. If the null hypothesis is rejected, indicating significant discrimination, one can proceed to interpret the results.[6]

Interpret the results

The interpretation of the discriminant weights, or coefficients, is similar to that in multiple regression analysis. The value of the coefficient for a particular predictor depends on the other predictors included in the discriminant function. The signs of the coefficients are arbitrary, but they indicate which variable values result in large and small function values and associate them with particular groups.

Given the multicollinearity in the predictor variables, there is no unambiguous measure of the relative importance of the predictors in discriminating between the groups.[7] With this caveat in mind, we can obtain some idea of the relative importance of the variables by examining the absolute magnitude of the standardised discriminant function coefficients. Generally, predictors with relatively large standardised coefficients contribute more to the discriminating power of the function, as compared with predictors with smaller coefficients, and are therefore more important.

Some idea of the relative importance of the predictors can also be obtained by examining the structure correlations, also called *canonical loadings* or *discriminant loadings*. These simple correlations between each predictor and the discriminant function represent the variance that the predictor shares with the function. Like the standardised coefficients, these correlations must also be interpreted with caution.

An examination of the standardised discriminant function coefficients for the holiday resort example is instructive. Given the low intercorrelations between the predictors, one might cautiously use the magnitudes of the standardised coefficients to suggest that income is the most important predictor in discriminating between the groups, followed by household size and importance attached to the family skiing holiday. The same observation is obtained from examination of the structure correlations. These simple correlations between the predictors and the discriminant function are listed in order of magnitude.

The unstandardised discriminant function coefficients are also given. These can be applied to the raw values of the variables in the holdout set for classification purposes. The group centroids, giving the value of the discriminant function evaluated at the group means, are also shown. Group 1, those who have visited a resort, has a positive value, whereas Group 2 has an equal negative value. The signs of the coefficients associated with all the predictors are positive, which suggests that higher family income, household size, importance attached to family skiing holiday, attitude towards travel and age are more likely to result in the family visiting the resort. It would be reasonable to develop a profile of the two groups in terms of the three predictors that seem to be the most important: income, household size and importance of holiday. The values of these three variables for the two groups are given at the beginning of Table 21.4.

The determination of relative importance of the predictors is further illustrated by the following example.

Example

Satisfied salespeople stay[8]

A survey asked business people about the climate of hiring and maintaining employees in harsh economic conditions. It was reported that 85% of respondents were concerned about recruiting employees and 81% said they were concerned about retaining employees. When the economy is down, turnover is rapid. Generally speaking, if an organisation wants to retain its employees, it must learn why people leave their jobs and why others stay and are satisfied with their jobs. Discriminant analysis was used to determine what factors explained the differences between salespeople who left a large computer manufacturing company and those who stayed. The independent variables were company rating, job security, seven job satisfaction dimensions, four role-conflict dimensions four role-ambiguity dimensions and nine measures of sales performance. The dependent variable was the dichotomy between those who stayed and those who left. The canonical correlation, an

Discriminant analysis results

Variable		Coefficients	Standardised coefficients	Structure correlations
1	Work[a]	0.0903	0.3910	0.5446
2	Promotion[a]	0.0288	0.1515	0.5044
3	Job security	0.1567	0.1384	0.4958
4	Customer relations[b]	0.0086	0.1751	0.4906
5	Company rating	0.4059	0.3240	0.4824
6	Working with others[b]	0.0018	0.0365	0.4651
7	Overall performance[b]	–0.0148	–0.3252	0.4518
8	Time-territory management[b]	0.0126	0.2899	0.4496
9	Sales produced[b]	0.0059	0.1404	0.4484
10	Presentation skill[b]	0.0118	0.2526	0.4387
11	Technical information[b]	0.0003	0.0065	0.4173
12	Pay benefits[a]	0.0600	0.1843	0.3788
13	Quota achieved[b]	0.0035	0.2915	0.3780
14	Management[a]	0.0014	0.0138	0.3571
15	Information collection[b]	–0.0146	–0.3327	0.3326
16	Family[c]	–0.0684	–0.3408	–0.3221
17	Sales manager[a]	–0.0121	–0.1102	0.2909
18	Coworker[a]	0.0225	0.0893	0.2671
19	Customer[c]	–0.0625	–0.2797	–0.2602
20	Family[d]	0.0473	0.1970	0.2180
21	Job[d]	0.1378	0.5312	0.2119
22	Job[c]	0.0410	0.5475	–0.1029
23	Customer[d]	–0.0060	–0.0255	0.1004
24	Sales manager[c]	–0.0365	–0.2406	–0.0499
25	Sales manager[d]	–0.0606	–0.3333	0.0467
26	Customer[a]	–0.0338	–0.1488	0.0192

Note: Rank order of importance is based on the magnitude of the canonical loadings:
[a] Satisfaction
[b] Performance
[c] Ambiguity
[d] Conflict.

index of discrimination ($R = 0.4572$), was significant (Wilks' $\lambda = 0.7909$, $F(26,173) = 1.7588$, $p = 0.0180$). This result indicated that the variables discriminated between those who left and those who stayed.

The results from simultaneously entering all variables in discriminant analysis are presented in the table opposite. The rank order of importance, as determined by the relative magnitude of the canonical loadings, is presented in the first column. Satisfaction with the job and promotional opportunities were the two most important discriminators, followed by job security. Those who stayed in the company found the job to be more exciting, satisfying, challenging and interesting than those who left.

In this example, promotion was identified as the second most important variable based on the canonical loadings. However, it is not the second most important variable based on the absolute magnitude of the standardised discriminant function coefficients. This anomaly results from multicollinearity.

Characteristic profile
An aid to interpreting discriminant analysis results by describing each group in terms of the group means for the predictor variables.

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 for the predictor variables. If the important predictors have been identified, then a comparison of the group means on these variables can assist in understanding the intergroup differences. Before any findings can be interpreted with confidence, however, it is necessary to validate the results.

Assess the validity of discriminant analysis

Many computer programs, such as SPSS, offer a leave-one-out cross-validation option. In this option, the discriminant model is re-estimated as many times as there are respondents in the sample. Each re-estimated model leaves out one respondent and the model is used to predict for that respondent. When a large holdout sample is not possible, this gives a sense of the robustness of the estimate using each respondent in turn, as a holdout.

As explained earlier, the data are randomly divided into two subsamples. One, the analysis sample, is used for estimating the discriminant function, and the validation sample is used for developing the classification matrix. The discriminant weights, estimated by using the analysis 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. For example, in two-group discriminant analysis, a case will be assigned to the group whose centroid is the closest. 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.[9]

Hit ratio
The percentage of cases correctly classified by discriminant analysis.

It is helpful to compare the percentage of cases correctly classified by discriminant analysis with the percentage that would be obtained by chance. When the groups are equal in size, the percentage of chance classification is one divided by the number of groups. How much improvement should be expected over chance? No general guidelines are available, although some authors have suggested that classification accuracy achieved by discriminant analysis should be at least 25% greater than that obtained by chance.[10]

Most discriminant analysis programs also estimate a classification matrix based on the analysis sample. Because they capitalise on chance variation in the data, such results are invariably better than the classification obtained on the holdout sample.[11]

Table 21.4, for the holiday resort example, also shows the classification results based on the analysis sample. The hit ratio, or the percentage of cases correctly classified, is $(12 + 15)/30 = 0.90$, or 90%. One might suspect that this hit ratio is artificially inflated, as the data used for estimation were also used for validation. Leave-one-out cross-validation correctly classifies only $(11 + 13)/30 = 0.80$ or 80% of the cases. Conducting classification analysis on an independent holdout set of data results in the classification matrix with a slightly lower hit ratio of (4 + 6)/12 = 0.833, or 83.3% (see Table 21.4). Given two groups of equal size, by chance one would expect a hit ratio of 1/2 = 0.50, or 50%. Hence, the improvement over chance is more than 25%, and the validity of the discriminant analysis is judged as satisfactory.

Another application of two-group discriminant analysis is provided by the following example.

Example

Home bodies and couch potatoes[12]

Two-group discriminant analysis was used to assess the strength of each of five dimensions used in classifying individuals as TV users or non-users. The discriminant analysis procedure was appropriate for this use because of the nature of the predefined categorical groups (users and non-users) and the interval scales used to generate individual factor scores.

Source: © Getty Images

Two equal groups of 185 elderly consumers, users and non-users (total n = 370), were created. The discriminant equation for the analysis was estimated by using a subsample of 142 respondents from the sample of 370. Of the remaining respondents, 198 were used as a validation subsample in a cross-validation of the equation. Thirty respondents were excluded from the analysis because of missing discriminant values.

The canonical correlation for the discriminant function was 0.4291, significant at the $p < 0.0001$ level. The eigenvalue was 0.2257. The table opposite summarises the standardised canonical discriminant coefficients. A substantial portion of the variance is explained by the discriminant function. In addition, as the table shows, the home-orientation dimension made a fairly strong contribution to classifying individuals as users or non-users of TV. Morale, security and health, and respect also contributed significantly. The social factor appeared to make little contribution.

The cross-validation procedure using the discriminant function from the analysis sample gave support to the contention that the dimensions aided researchers in discriminating between users and non-users of TV. As the table shows, the discriminant function was successful in classifying 75.76% of the cases. This suggests that consideration of the identified dimensions will help marketers understand the elderly market.

Summary of discriminant analysis

Standard canonical discriminant function coefficients

Morale	0.27798
Security and health	0.39850
Home orientation	0.77496
Respect	0.32069
Social	–0.01996

Classification results for cases selected for use in the analysis

		Predicted group membership	
Actual group	*No. of cases*	*Non-users*	*Users*
TV non-users	77	56	21
		72.7%	27.3%
TV users	65	24	41
		36.9%	63.1%
Per cent of grouped cases correctly classified: 68.31%			

Classification results for cases selected for cross-validation

		Predicted group membership	
Actual group	*No. of cases*	*Non-users*	*Users*
TV non-users	108	85	23
		78.7%	21.3%
TV users	90	25	65
		27.8%	72.2%
Per cent of grouped cases correctly classified: 75.76%			

The extension from two-group discriminant analysis to multiple discriminant analysis involves similar steps and is illustrated with an application.

Conducting multiple discriminant analysis

Formulate the problem

The data presented in Tables 21.2 and 21.3 can also be used to illustrate three-group discriminant analysis. In the last column of these tables, the households are classified into three categories, based on the amount spent on their family skiing holiday (high, medium or low). Ten households fall in each category. The question of interest is whether the households that spend high, medium or low amounts on their holidays (AMOUNT) can be differentiated in terms of annual family income (INCOME), attitude towards travel (TRAVEL), importance attached to family skiing holiday (HOLIDAY), household size (HSIZE) and age of the head of household (AGE).[13]

Estimate the discriminant function coefficients

Table 21.5 presents the results of estimating three-group discriminant analysis. An examination of group means indicates that income appears to separate the groups more widely than any other variable. There is some separation on travel and holiday. Groups 1 and 2 are very close in terms of household size and age. Age has a large standard deviation relative to the separation between the groups. The pooled within-groups correlation matrix indicates some correlation of holiday and household size with income. Age has some negative correlation with travel. Yet these correlations are on the lower side, indicating that, although multicollinearity may be of some concern, it is not likely to be a serious problem. The significance attached to the univariate *F* ratios indicates that, when the predictors are considered individually, only income and travel are significant in differentiating between the two groups.

In multiple discriminant analysis, if there are *G* groups, *G* – 1 discriminant functions can be estimated if the number of predictors is larger than this quantity. In general, with *G* groups and *k* predictors, it is possible to estimate up to the smaller of *G* – 1, or *k*, discriminant functions. The first function has the highest ratio of between-groups to within-groups sum of squares. The second function, uncorrelated with the first, has the second highest ratio, and so on. Not all the functions may be statistically significant, however.

Because there are three groups, a maximum of two functions can be extracted. The eigenvalue associated with the first function is 3.8190, and this function accounts for 93.93% of variance in the data. The eigenvalue is large, so the first function is likely to be superior. The second function has a small eigenvalue of 0.2469 and accounts for only 6.07% of the variance.

Table 21.5 Results of three-group discriminant analysis

Group means

Visit	INCOME	TRAVEL	HOLIDAY	HSIZE	AGE
1	38.57000	4.50000	4.70000	3.10000	50.30000
2	50.1100	4.00000	4.20000	3.40000	49.50000
3	64.97000	6.10000	5.90000	4.20000	56.00000
Total	51.21667	4.86667	4.93333	3.56667	51.93333

Group standard deviations

1	5.29718	1.71594	1.88856	1.19722	8.09732
2	6.00231	2.35702	2.48551	1.50555	9.25263
3	8.61434	1.19722	1.66333	1.13529	7.60117
Total	12.79523	1.97804	2.09981	1.33089	8.57395

Pooled within-groups correlation matrix

INCOME	1.00000				
TRAVEL	0.05120	1.00000			
HOLIDAY	0.30681	0.03588	1.00000		
HSIZE	0.38050	0.00474	0.22080	1.00000	
AGE	–0.20939	–0.34022	–0.01326	–0.02512	1.00000

Table 21.5 Continued

Wilks' λ (*U* statistic) and univariate *F* ratio with 2 and 27 degrees of freedom

Variable	*Wilks' λ*	*F*	*Significance*
INCOME	0.26215	38.000	0.0000
TRAVEL	0.78790	3.634	0.0400
HOLIDAY	0.88060	1.830	0.1797
HSIZE	0.87411	1.944	0.1626
AGE	0.88214	1.804	0.1840

Canonical discriminant functions

Function	*Eigenvalue*	*Per cent of variance*	*Cumulative percentage*	*Canonical correlation*	*After function*	*Wilks' λ*	*Chi-square*	*df*	*Sig.*
					0	0.1664	44.831	10	0.00
1*	3.8190	93.93	93.93	0.8902	1	0.8020	5.517	4	0.24
2*	0.2469	6.07	100.00	0.4450					

* Marks the two canonical discriminant functions remaining in the analysis.

Standardised canonical discriminant function coefficients

	Func 1	*Func 2*
INCOME	1.04740	–0.42076
TRAVEL	0.33991	0.76851
HOLIDAY	–0.14198	0.53354
HSIZE	–0.16317	0.12932
AGE	0.49474	0.52447

Structure matrix: pooled within-groups correlations between discriminating variables and canonical discriminant functions (variables ordered by size of correlation within function)

	Func 1	*Func 2*
INCOME	0.85556*	–0.27833
HSIZE	0.19319*	0.07749
HOLIDAY	0.21935	0.58829*
TRAVEL	0.14899	0.45362*
AGE	0.16576	0.34079*

Unstandardised canonical discriminant function coefficients

	Func 1	*Func 2*
INCOME	0.1542658	–0.6197148E-01
TRAVEL	0.1867977	0.4223430
HOLIDAY	–0.6952264E-01	0.2612652
HSIZE	–0.1265334	0.1002796
AGE	0.5928055E-01	0.6284206E-01
(Constant)	–11.09442	–3.791600

Table 21.5 Continued

Canonical discriminant functions evaluated at group means (group centroids)

Group	*Func 1*	*Func 2*
1	–2.04100	0.41847
2	–0.40479	–0.65867
3	2.44578	0.24020

Classification results

			Predicted group membership			
		Amount	1	2	3	*Total*
Original	Count	1	9	1	0	10
		2	1	9	0	10
		3	0	2	8	10
	%	1	90.0	10.0	0.0	100.0
		2	10.0	90.0	0.0	100.0
		3	0.0	20.0	80.0	100.0
Cross-validated	Count	1	7	3	0	10
		2	4	5	1	10
		3	0	2	8	10
	%	1	70.0	30.0	0.0	100.0
		2	40.0	50.0	10.0	100.0
		3	0.0	20.0	80.0	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] 86.7% of original grouped cases correctly classified.

[c] 66.7% of cross-validated grouped cases correctly classified.

Classification results for cases not selected for use in analysis

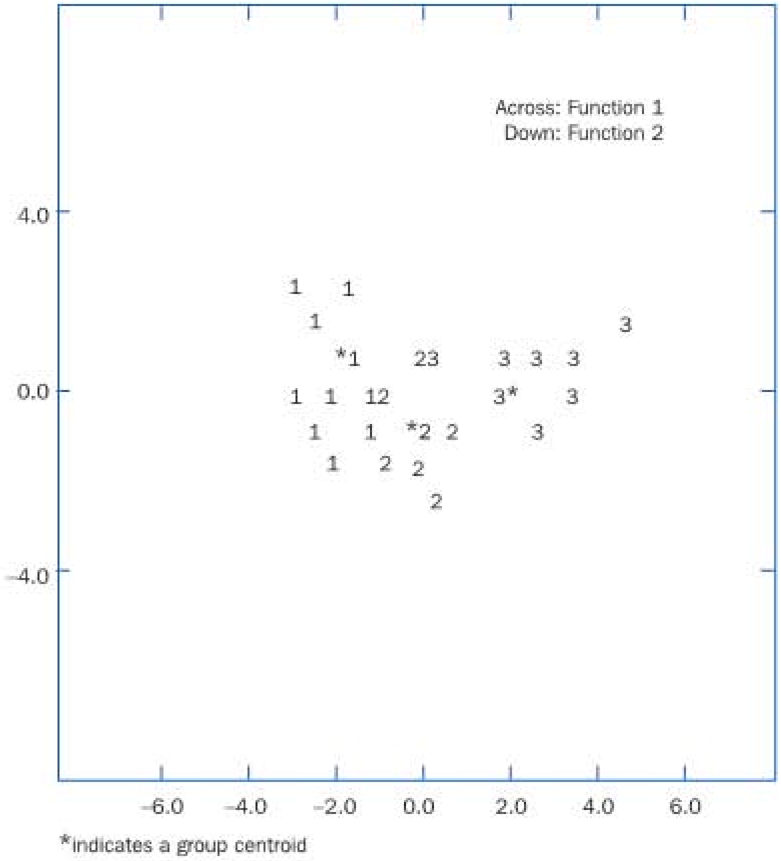
			Predicted group membership		
	Actual group	*No. of cases*	*1*	*2*	*3*
Group	1	4	3	1	0
			75.0%	25.0%	0.0%
Group	2	4	0	3	1
			0.0%	75.0%	25.0%
Group	3	4	1	0	3
			25.0%	0.0%	75.0%
Percentage of grouped cases correctly classified: 75.00%					

Determine the significance of the discriminant function

To test the null hypothesis of equal group centroids, both the functions must be considered simultaneously. It is possible to test the means of the functions successively by first testing all means simultaneously. Then one function is excluded at a time, and the means of the remaining functions are tested at each step. In Table 21.5, the 0 below the 'After function' heading indicates that no functions have been removed. The value of Wilks' λ is 0.1644. This transforms to a chi-square of 44.831, with 10 degrees of freedom, which is significant beyond the 0.05 level. Thus, the two functions together significantly discriminate among the three groups. When the first function is removed, however, the Wilks' λ associated with the second function is 0.8020, which is not significant at the 0.05 level. Therefore, the second function does not contribute significantly to group differences.

Interpret the results

The interpretation of the results is aided by an examination of the standardised discriminant function coefficients, the structure correlations and certain plots. The standardised coefficients indicate a large coefficient for income on function 1, whereas function 2 has relatively larger coefficients for travel, holiday and age. A similar conclusion is reached by an examination of the structure matrix (see Table 21.5). To help interpret the functions, variables with large coefficients for a particular function are grouped together. These groupings are shown with asterisks. Thus income and household size have asterisks for function 1 because these variables have coefficients which are larger for function 1 than for function 2. These variables are associated primarily with function 1. On the other hand, travel, holiday and age are predominantly associated with function 2, as indicated by the asterisks.



*indicates a group centroid

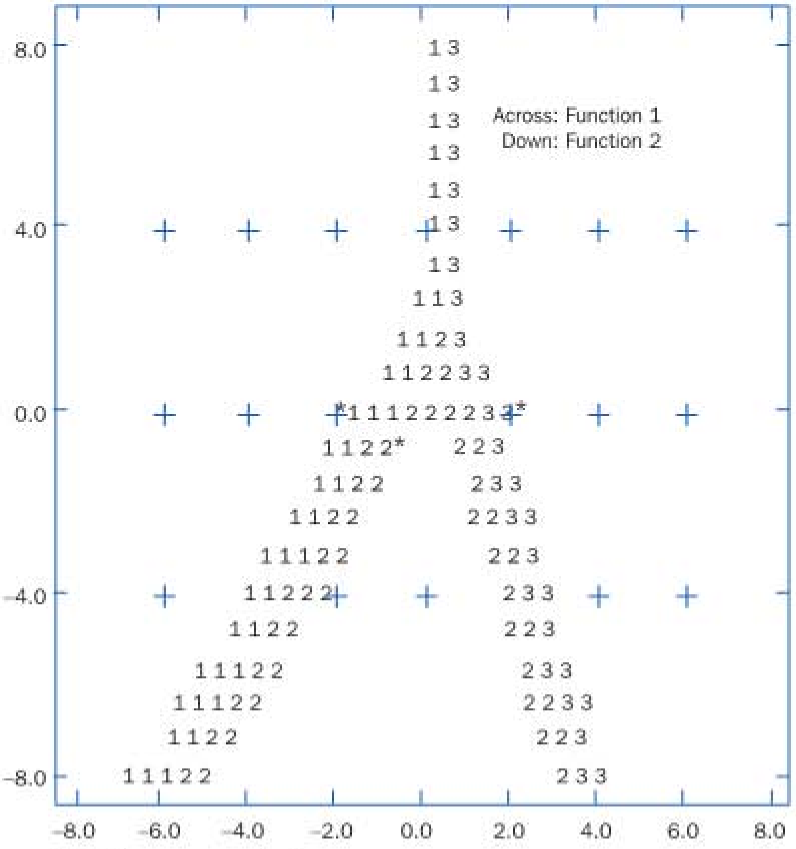
Figure 21.2
All-groups scattergram

Figure 21.2 is a scattergram plot of all the groups on function 1 and function 2. It can be seen that group 3 has the highest value on function 1, and group 1 the lowest. Because function 1 is primarily associated with income and household size, one would expect the three groups to be ordered on these two variables. Those with higher incomes and higher household size are likely to spend large amounts of money on holidays. Conversely, those with low incomes and smaller household size are likely to spend small amounts on holidays. This interpretation is further strengthened by an examination of group means on income and household size.

Figure 21.2 further indicates that function 2 tends to separate group 1 (highest value) and group 2 (lowest value). This function is primarily associated with travel, holiday and age. Given the positive correlations of these variables with function 2 in the structure matrix, we expect to find group 1 to be higher than group 2 in terms of travel, holiday and age. This is indeed true for travel and holiday, as indicated by the group means of these variables. If families in group 1 have more favourable attitudes towards travel and attach more importance to a family skiing holiday than group 2, why do they spend less? Perhaps they would like to spend more on holidays but cannot afford it because they have low incomes.

Territorial map
A tool for assessing discriminant analysis results by plotting the group membership of each case on a graph.

A similar interpretation is obtained by examining a **territorial map**, as shown in Figure 21.3. In a territorial map, each group centroid is indicated by an asterisk. The group boundaries are shown by numbers corresponding to the groups. Thus, group 1 centroid is bounded by ls, group 2 centroid by 2s and group 3 centroid by 3s.



*indicates a group centroid

Figure 21.3
Territorial map

Assess the validity of discriminant analysis

The classification results based on the analysis sample indicate that (9 + 9 + 8)/30 = 86.67% of the cases are correctly classified. Leave-one-out cross-validation correctly classifies only (7 + 5 + 8)/30 = 0.667 or 66.7% of the cases. When the classification analysis is conducted on the independent holdout sample of Table 21.3, a hit ratio of (3 + 3 + 3)/12 = 75% is obtained. Given three groups of equal size, by chance alone one would expect a hit ratio of 1/3 = 0.333 or 33.3%. The improvement over chance is more than 25%, indicating at least satisfactory validity.[14]

Further illustration of multiple group discriminant analysis is provided by the following example.

Example

The home is where the patient's heart is[15]

In the majority of developed nations, the largest sector in the economy is the health care industry. Over the next decade, it is expected that spending on health care services will grow significantly faster than their economies. Contributing to this growth are demographic forces, with shrinking birth rates, ageing populations, longer life expectancies and with rapidly growing numbers of elderly patients seeking health support of some kind.

Consumers were surveyed to determine their attitudes towards four systems of health care delivery (home health care, hospitals, nursing homes and outpatient clinics) along 10 attributes. A total of 102 responses were obtained, and the results were analysed using multiple discriminant analysis (Table 1). Three discriminant functions were identified. Chi-square tests performed on the results indicated that all three discriminant functions were significant at the 0.01 level. The first function accounted for 63% of the total discriminative power, and the remaining two functions contributed 29.4% and 7.6%, respectively.

Table 1 gives the standardised discriminant function coefficients of the 10 variables in the discriminant equations. Coefficients ranged in value from –1 to +1. In determining the ability of each attribute to classify the delivery system, absolute values were used. In the first discriminant function, the two variables with the largest coefficients were comfort (0.53) and privacy (0.40). Because both related to personal attention and care, the first dimension was labelled 'personalised care'. In the second function, the two variables with the largest coefficients were quality of medical care (0.67) and likelihood of faster recovery (0.32). Hence, this dimension was labelled 'quality of medical care'. In the third discriminant function, the most significant attributes were sanitation (–0.70) and expense (0.52). Because these two attributes represent value and price, the third discriminant function was labelled 'value'.

The four group centroids are shown in Table 2. This table shows that home health care was evaluated most favourably along the dimension of personalised care, and hospitals were evaluated least favourably. Along the dimension of quality of medical care, there was a substantial separation between nursing homes and the other three systems. Also, home health care received higher evaluations on the quality of medical care than did outpatient clinics. Outpatient clinics, on the other hand, were judged to offer the best value.

Classification analysis of the 102 responses, reported in Table 3, showed correct classifications ranging from 86% for hospitals to 68% for outpatient clinics. The misclassifications for hospitals were 6% each to nursing homes and outpatient clinics and 2% to home health care. Nursing homes showed misclassifications of 9% to hospitals, 10% to outpatient clinics and 3% to home health care. For outpatient clinics, 9% misclassifications were made to hospitals, 13% to nursing homes and 10% to home health care. For home health care, the misclassifications were 5% to hospitals, 4% to nursing homes and 13% to outpatient clinics. The results demonstrated that the discriminant functions were fairly accurate in predicting group membership.

→

Table 1 Standardised discriminant function coefficients

Variable system	*Discriminant function*		
	1	*2*	*3*
Safe	−0.20	−0.04	0.15
Convenient	0.08	0.08	0.07
Chance of medical complications[a]	−0.27	0.10	0.16
Expensive[a]	0.30	−0.28	0.52
Comfortable	0.53	0.27	−0.19
Sanitary	−0.27	−0.14	−0.70
Best medical care	−0.25	0.67	−0.10
Privacy	0.40	0.08	0.49
Faster recovery	0.30	0.32	−0.15
Staffed with best medical personnel	−0.17	−0.03	0.18
Percentage of variance explained	63.0	29.4	7.6
Chi-square	663.3[b]	289.2[b]	70.1[b]

[a] These two items were worded negatively on the questionnaire. They were reverse coded for purposes of data analysis.

[b] $p < 0.01$.

Table 2 Centroids of health care systems in discriminant space

System	*Discriminant function*		
	1	*2*	*3*
Hospital	−1.66	0.97	−0.08
Nursing home	−0.60	−1.36	−0.27
Outpatient clinic	0.54	−0.13	0.77
Home health care	1.77	0.50	−0.39

Table 3 Classification table

System	*Classification (%)*			
	Hospital	*Nursing home*	*Outpatient clinic*	*Home health care*
Hospital	86	6	6	2
Nursing home	9	78	10	3
Outpatient clinic	9	13	68	10
Home health care	5	4	13	78

Stepwise discriminant analysis

Stepwise discriminant analysis is analogous to stepwise multiple regression (see Chapter 20) in that the predictors are entered sequentially based on their ability to discriminate between the groups. An *F* ratio is calculated for each predictor by conducting a univariate ANOVA in which the groups are treated as the categorical variable and the predictor as the criterion variable. The predictor with the highest *F* ratio is the first to be selected for inclusion in the discriminant function, if it meets certain significance and tolerance criteria. A second predictor is added based on the highest adjusted or partial *F* ratio, taking into account the predictor already selected.

Each predictor selected is tested for retention based on its association with other predictors selected. The process of selection and retention is continued until all predictors meeting the significance criteria for inclusion and retention have been entered in the discriminant function. Several statistics are computed at each stage. In addition, at the conclusion, a summary of the predictors entered or removed is provided. The standard output associated with the direct method is also available from the stepwise procedure.

Mahalanobis procedure
A stepwise procedure used in discriminant analysis to maximise a generalised measure of the distance between the two closest group.

The selection of the stepwise procedure is based on the optimising criterion adopted. The **Mahalanobis procedure** is based on maximising a generalised measure of the distance between the two closest groups. This procedure allows marketing researchers to make maximal use of the available information.[16]

The Mahalanobis method was used to conduct a two-group stepwise discriminant analysis on the data pertaining to the visit variable in Tables 21.2 and 21.3. The first predictor variable to be selected was income, followed by household size and then holiday. The order in which the variables were selected also indicates their importance in discriminating between the groups. This was further corroborated by an examination of the standardised discriminant function coefficients and the structure correlation coefficients. Note that the findings of the stepwise analysis agree with the conclusions reported earlier by the direct method.

The next example illustrates an application of discriminant analysis in marketing research.

Example

Satisfactory results of satisfaction programmes in Europe[17]

In their marketing strategies, computer companies are emphasising the quality of their customer service programmes rather than focusing upon computer features and capabilities. Hewlett-Packard learned this lesson in Europe. Research conducted in the European market revealed that there was a difference in emphasis on service requirements across age segments. Focus groups revealed that customers above 40 years of age had a hard time with the technical aspects of the computer and greatly required the customer service programmes. On the other hand, younger customers appreciated the technical aspects of the product that added to their satisfaction. To uncover the factors leading to differences in the two segments, further research in the form of a large single cross-sectional survey was done. A two-group discriminant analysis was conducted with satisfied and dissatisfied customers as the two groups, with several independent variables such as technical information, ease of operation, variety and scope of customer service programmes, etc. Results confirmed the fact that the variety and scope of customer satisfaction programmes was indeed a strong differentiating factor. This was a crucial finding because Hewlett-Packard could better handle dissatisfied customers by focusing more on customer services than on technical details. Consequently, Hewlett-Packard successfully started three programmes on customer satisfaction: customer feedback, customer satisfaction surveys and total quality control. This effort resulted in increased customer satisfaction.

After seeing the successful results of this approach in Europe, HP Malaysia decide to launch a customer service programme called HP Cares!. This programme was developed in Malaysia as part of HP's Total Customer Experience (TCE) initiative to be rolled out worldwide. HP Malaysia's marketing and sales manager belives this programme will allow HP to build brand value and to differentiate the company from its competitors. HP realises that no matter how great its products may be, brand loyalty can be achieved only through continuous improvements in customer service.

Internet and computer applications

SPSS

The DISCRIMINANT procedure is used for conducting discriminant analysis. This is a general program that can be used for two-group or multiple discriminant analysis. Furthermore, the direct or the stepwise method can be adopted.

SAS

The DISCRIM procedure can be used for performing two-group or multiple discriminant analysis. If the assumption of a multivariate normal distribution cannot be met, the NEIGHBOR procedure can be used. In this procedure, a non-parametric nearest neighbour rule is used for classifying the observations. CANDISC performs canonical discriminant analysis and is related to principal components analysis and canonical correlation. The STEPDISC procedure can be used for performing stepwise discriminant analysis.

Minitab

Discriminant analysis can be conducted using the STATS>MULTIVARIATE>DISCRIMINANT ANALYSIS function. It computes both linear and quadratic discriminant analysis in the classification of observations into two or more groups.

Excel

At the time of writing, discriminant analysis was not available.

Summary

Discriminant analysis is useful for analysing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval scaled. When the criterion variable has two categories, the technique is known as two-group discriminant analysis. Multiple discriminant analysis refers to the case when three or more categories are involved.

Conducting discriminant analysis is a five-step procedure. First, formulating the discriminant problem requires identification of the objectives and the criterion and predictor variables. The sample is divided into two parts. One part, the analysis sample, is used to estimate the discriminant function. The other part, the holdout sample, is reserved for validation. Estimation, the second step, involves developing a linear combination of the predictors, called discriminant functions, so that the groups differ as much as possible on the predictor values.

Determination of statistical significance is the third step. It involves testing the null hypothesis that, in the population, the means of all discriminant functions in all groups are equal. If the null hypothesis is rejected, it is meaningful to interpret the results.

The fourth step, the interpretation of 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 predictors in discriminating between the groups. Some idea of the relative importance of the variables, however, may be obtained by examining the absolute magnitude of the standardised discriminant function coefficients and by examining the structure correlations or discriminant loadings. These simple correlations between each predictor and the discriminant function represent the variance that the predictor shares with the function. Another aid to interpreting discriminant analysis results is to develop a characteristic profile for each group, based on the group means for the predictor variables.

Validation, the fifth step, involves developing the classification matrix. The discriminant weights estimated by using the analysis 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 percentage of cases correctly classified is determined and compared with the rate that would be expected by chance classification.

Two broad approaches are available for estimating the coefficients. The direct method involves estimating the discriminant function so that all the predictors are included simultaneously. An alternative is the stepwise method in which the predictor variables are entered sequentially, based on their ability to discriminate among groups.

In multiple discriminant analysis, if there are G groups and k predictors, it is possible to estimate up to the smaller of $G - 1$, or k, discriminant functions. The first function has the highest ratio of between-group to within-group sums of squares; the second function, uncorrelated with the first, has the second highest ratio; and so on.

Questions

1 What are the objectives of discriminant analysis?

2 Describe four examples of the application of discriminant analysis.

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

4 Describe the relationship of discriminant analysis to regression and ANOVA.

5 What are the steps involved in conducting discriminant analysis?

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

7 What is Wilks' λ? For what purpose is it used?

8 Define discriminant scores.

9 Explain what is meant by an eigenvalue.

10 What is a classification matrix?

11 Explain the concept of structure correlations.

12 How is the statistical significance of discriminant analysis determined?

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

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

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

Exercises

1 In investigating the differences between heavy and light or 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 respondent. 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 Analyse the Benetton data (taken from Exercise 4, Chapter 18). Do the three usage groups differ in terms of awareness, attitude, preference, intention and loyalty towards Benetton when these variables are considered simultaneously?

4 Conduct a two-group discriminant analysis on the data given in Tables 21.2 and 21.3 using different statistical analysis packages, e.g. SPSS, SAS and Minitab. Compare the output from all the packages. Discuss the similarities and differences.

5 In a small group discuss the following issue: 'Is it meaningful to determine the relative importance of predictors in discriminating between the groups? Why or why not?'

Appendix: Estimation of discriminant function coefficients

Suppose that there are G groups, $i = 1, 2, 3, \ldots, G$, each containing n_i observations on K independent variables, $X_1, X_2, \ldots, X_k$. The following notations are used:

N = total sample size = $\sum_{i=1}^{G} n_i$

W_i = matrix of mean corrected sum of squares and cross-products for the ith group

W = matrix of pooled within-groups mean corrected sum of squares and cross-products

B = matrix of between-groups mean corrected sum of squares and cross-products

T = matrix of total mean corrected sum of squares and cross-products for all the N observations = $W + B$

$\bar{X}_i$ = vector of means of observations in the ith group

$\bar{X}$ = vector of grand means for all the N observations

λ = ratio of between-groups to within-group sums of squares

b = vector of discriminant coefficients or weights

Then,

$$T = \sum_{i=1}^{G} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})'$$

$$W_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'$$

$$W = W_1 + W_2 + W_3 + \ldots + W_G$$

$$B = T - W$$

Define the linear composite $D = b'_1X$. Then, with reference to D, the between-groups and within-groups sums of squares are b'_1Bb and b'_1Wb, respectively. To discriminate the groups maximally, the discriminant functions are estimated to maximise the between-group variability. The coefficients b are calculated to maximise λ, by solving

$$\text{Max } \lambda = \frac{b'Bb}{b'Wb}$$

Taking the partial derivative with respect to λ and setting it equal to zero, with some simplification, yields

$$(B - \lambda W)b = 0$$

To solve for b, it is more convenient to premultiply by W^{-1} and solve the following characteristic equation:

$$(W^{-1}B - \lambda I)b = 0$$

The maximum value of λ is the largest eigenvalue of the matrix $W^{-1}B$, and b is the associated eigenvector. The elements of b are the discriminant coefficients, or weights, associated with the first discriminant function. In general, it is possible to estimate up to the smaller of $G - 1$ or k discriminant functions, each with its associated eigenvalue. The discriminant functions are estimated sequentially. In other words, the first discriminant function exhausts most of the between-group variability, the second function maximises the between-group variation that was not explained by the first one, and so on.

Notes

1 Anon., 'DirecTv adds National Geographic Channel', *Satellite News* (15 January 2001), 1; Lichtenstein, D.R., Burton, S. and Netemeyer, R.G., 'An examination of deal proneness across sales promotion types: a consumer segmentation perspective', *Journal of Retailing* 73 (2) (Summer 1997), 283–297; Jolson, M.A., Wiener, J.L. and Rosecky, R.B., 'Correlates of rebate proneness', *Journal of Advertising Research* (February–March 1987), 33–43.

2 A detailed discussion of discriminant analysis may be found in Kemsley, E.K., *Discriminant Analysis and Class Modeling of Spectroscopic Data* (New York: Wiley, 1998); Tacq, J., *Multivariate Analysis Techniques in Social Science Research* (Thousand Oaks, CA: Sage, 1996); Lachenbruch, P.A., *Discriminant Analysis* (New York: Hafner Press, 1975). For an application, see Deal, K., 'Determining success criteria for financial products: a comparative analysis of CART, logit and factor/discriminant analysis', *Service Industries Journal* 17 (3) (July 1997), 489–506.

3 See Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th edn (Paramus, NJ: Prentice Hall, 2001); Klecka, W.A., *Discriminant Analysis* (Beverly Hills, CA: Sage, 1980). See also Sinclair, S.A. and Stalling, E.C., 'How to identify differences between market segments with attribute analysis', *Industrial Marketing Management* 19 (February 1990), 31–40.

4 For an application, see Khan, Z., Chawla, S.K. and Cianciolo, T.A., 'Multiple discriminant analysis: tool for effective marketing of computer information systems to small business clients', *Journal of Professional Systems Marketing* 12 (2) (1995), 153–162; Sager, J.K. and Menon, A., 'The role of behavioural intentions in turnover of salespeople', *Journal of Business Research* 29 (March 1994), 179–188; Kijewski, V., Yoon, E. and Young, G., 'How exhibitors select trade shows', *Industrial Marketing Management* 22 (November 1993), 287–298.

5 Franses, P.H., 'A test for the hit rate in binary response models', *International Journal of Market Research* 42 (2) (Spring 2000), 239–245; Mitchell, V.-W., 'How to identify psychographic segments: Part 2', *Marketing Intelligence and Planning* 12 (7) (1994), 11–16; Crask, M.R. and Perrault, W.D. Jr, 'Validation of discriminant analysis in marketing research', *Journal of Marketing Research* 14 (February 1977), 60–68.

6 Strictly speaking, before testing for the equality of group means, the equality of group covariance matrices should be tested. Box's M test can be used for this purpose. If the equality of group covariance means is rejected, the results of discriminant analysis should be interpreted with caution. In this case, the power of the test for the equality of group means decreases.

7 See Hanna, N., 'Brain dominance and the interpretation of advertising messages', *International Journal of Commerce & Management* 9 (3/4) (1999), 19–32; Fok, L., Angelidis, J.P., Ibrahim, N.A. and Fok, W.M., 'The utilization and interpretation of multivariate statistical techniques in strategic management', *International Journal of Management* 12 (4) (December 1995), 468–481; Morrison, D.G., 'On the interpretation of discriminant analysis', *Journal of Marketing Research* 6 (May 1969), 156–163. For the use of other techniques in conjunction with discriminant analysis to aid interpretation, see Dant, R.P., Lumpkin, J.R. and Bush, R.P., 'Private physicians or walk-in clinics: do the patients differ?', *Journal of Health Care Marketing* 10 (June 1990), 23–35.

8 Sahl, R.J., 'Retention reigns as economy suffers drought', *Workspan* 44 (11) (November 2001), 6–8; Hawes, J.M., Rao, C.P. and Baker, T.L., 'Retail salesperson attributes and the role of dependability in the selection of durable goods', *Journal of Personal Selling and Sales Management* 13 (4) (Fall 1993), 61–71; Fern, E.E., Avila, R.A. and Grewal, D., 'Salesforce turnover: those who left and those who stayed', *Industrial Marketing Management* (1989), 1–9.

9 For the validation of discriminant analysis, see Reinartz, W.J. and Kumar, V., 'On the profitability of long life customers in a non-contractual setting: an empirical investigation and implications for marketing', *Journal of Marketing* 64 (4) (October 2000), 17–35.

10 Hair, J.E. Jr, Anderson, R.E., Tatham, R.L. and Black, W.C., *Multivariate Data Analysis with Readings*, 5th edn (Upper Saddle River, NJ: Prentice Hall, 1999). See also Glen, J.J., 'Classification accuracy in discriminant analysis: a mixed integer programming approach', *Journal of the Operational Research Society* 52 (3) (March 2001), 328.

11 Mitchell, V.W., 'How to identify psychographic segments: Part 2', *Marketing Intelligence and Planning* 12 (7) (1994), 11–16; Albaum, G. and Baker, K., 'The sampling problem in validation of multiple discriminant analysis', *Journal of the Market Research Society* 18 (July 1976).

12 Anon., 'Interactive TV growth to erupt over the next five years', *Satellite News* 24 (2) (8 January, 2001), 1; Rahtz, D.R., Sirgy, M.J. and Kosenko, R., 'Using demographics and psychographic dimensions to discriminate between mature heavy and light television users: an exploratory analysis', in Bahn, K.D. (ed.), *Developments in Marketing Science*, Vol. 11 (Blacksburg, VA: Academy of Marketing Science, 1988), 2–7.

13 For advanced discussion of multiple discriminant analysis, see Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th edn (Upper Saddle River, NJ: Prentice Hall, 2002). For an application, see Reinartz, W.J. and Kumar, V., 'On the profitability of long life customers in a non-contractual setting: an empirical investigation and implications for marketing', *Journal of Marketing* 64 (4) (October 2000), 17–35.

14 Loucopoulos, C. and Pavur, R.M., 'Computational characteristics of a new mathematical programming model for the three-group discriminant problem', *Computers and Operations Research* 24 (2) (February 1997), 179–191. For an application of multiple discriminant analysis, see O'Connor, S.J., Shewchuk, R.M. and Camey, L.W., 'The great gap', *Journal of Health Care Marketing* 14 (Summer 1994), 32–39.

15 Tudor, J., 'Valuation of the health services industry', *Weekly Corporate Growth Report* (1133) (26 March, 2001), 11237–11238; Dansky, K.H. and Brannon, D., 'Discriminant analysis: a technique for adding value to patient satisfaction surveys', *Hospital and Health Services Administration* 41 (4) (Winter 1996), 503–513; Lim, J.S. and Zallocco, R., 'Determinant attributes in formulation of attitudes toward four health care systems', *Journal of Health Care Marketing* 8 (June 1988), 25–30.

16 Johnson, R.A. and Wichern, D.A., *Applied Multivariate Statistical Analysis*, 5th edn (Upper Saddle River, NJ: Prentice Hall, 2002); Hair, J.E. Jr, Anderson, R.E., Tatham, R.L. and Black, W.C., *Multivariate Data Analysis with Readings*, 5th edn (Englewood Cliffs, NJ: Prentice Hall, 1999), 178–255.

17 Raman, P., 'Taking customer service a step ahead', *Computimes* (22 October 2001), 1; Whitelock, J., Roberts, C. and Blakeley, J., 'The reality of the eurobrand: an empirical analysis', *Journal of International Marketing* 3 (3) (1995), 77–95.

Visit the *Marketing Research* Companion Website at www.pearsoned.co.uk/malhotra_euro for additional learning resources including annotated weblinks, an online glossary and a suite of downloadable video cases.