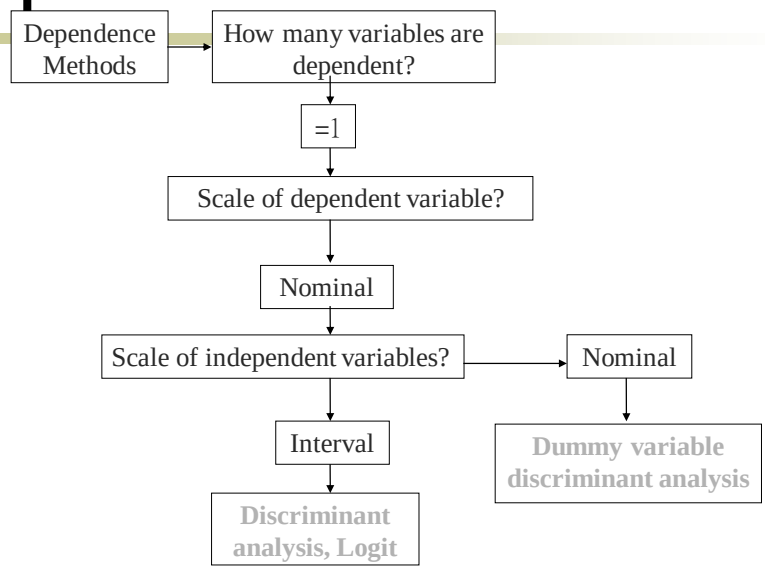
Chapter Eighteen

Discriminant and Logit Analysis

Review: Multivariate Analytics



Chapter Outline

1) Overview
2) Basic Concept
3) Relation to Regression and ANOVA
4) Discriminant Analysis Model
5) Statistics Associated with Discriminant Analysis
6) Multiple Discriminant Analysis
7) Stepwise Discriminant Analysis
8) Logit Regression

Similarities and Differences between ANOVA, Regression, and Discriminant Analysis

	ANOVA	REGRESSION	DISCRIMINANT (Logit) ANALYSIS
Similarities			
Number of dependent variables	**One**	**One**	**One**
Number of independent variables	**Multiple**	**Multiple**	**Multiple**
Differences			
Nature of the dependent variables	**Metric**	**Metric**	**Categorical**
Nature of the independent variables	**Categorical**	**Metric**	**Metric**

Discriminant Analysis

Discriminant analysis is a technique for analyzing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature.

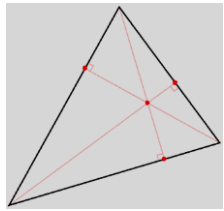
The objectives of discriminant analysis are as follows:

- Development of **discriminant functions**, or linear combinations of the predictor or independent variables, which will best discriminate between the categories of the criterion or dependent variable (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.

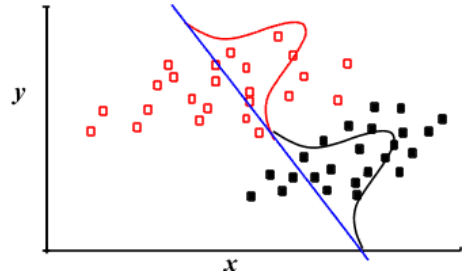
Discriminant Analysis

- 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 (**e.g., L, M, H**)**.
- 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. However, not all the functions may be statistically significant.

Discriminant Analysis

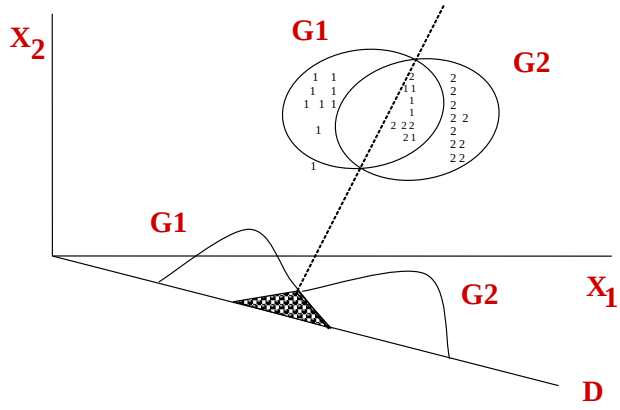


Centroid



Geometric Interpretation

Fig. 18.1



Discriminant Analysis Model

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

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

where

D = discriminant score (predicted/estimated)

b 's= discriminant coefficient or weight

X 's= 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.

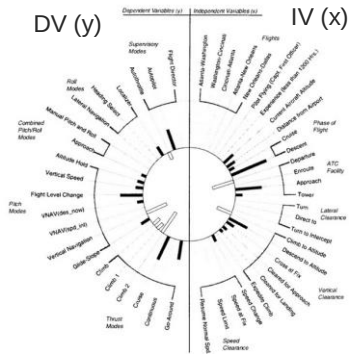
Statistics Associated with Discriminant Analysis

- **Canonical correlation**. A **canonical correlation** is the correlation of two canonical (latent) variables, one representing a set of independent variables (predictors), the other a set of dependent variables (discriminant groups or segments).
- **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.

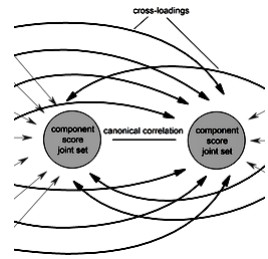
Canonical Correlation Analysis
典型相關分析

DV (y)

IV (x)



$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \left(\begin{bmatrix} \mu \\ \nu \end{bmatrix}, \begin{bmatrix} \sum_{XX} & \sum_{XY} \\ \sum_{YX} & \sum_{YY} \end{bmatrix} \right)$$



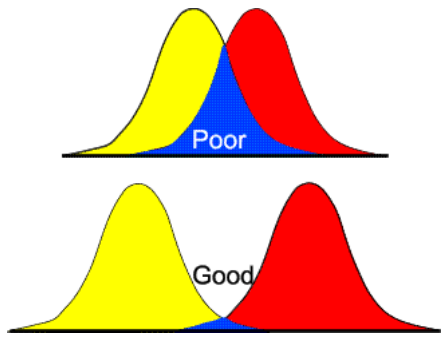
Statistics Associated with Discriminant Analysis

- **Discriminant function coefficients**. The discriminant function coefficients (unstandardized) are the multipliers of variables, when the variables are in the original units of measurement.
- **Discriminant scores.** The unstandardized 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. (e.g., if the ratio of two eigenvalues is 1.4, then the first discriminant function accounts for 40% more between-group variance in the dependent categories than does the second discriminant function.)

Statistics Associated with Discriminant Analysis

- ***F* values and their significance**. These 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**. These 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.

Statistics Associated with Discriminant Analysis



Statistics Associated with Discriminant Analysis

- **Standardized discriminant function coefficients**. The standardized discriminant function coefficients when the variables have been standardized 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 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.

Conducting Discriminant Analysis

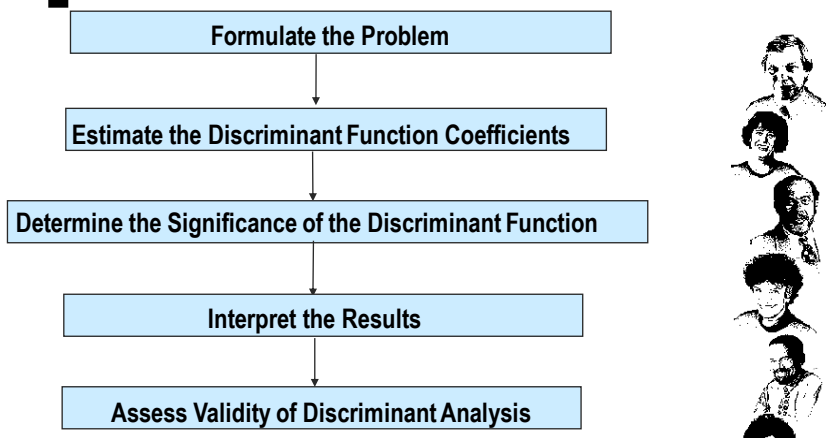


Fig. 18.1

Conducting Discriminant Analysis Formulate the Problem

- Identify the objectives, the criterion variable, and the independent variables (from your theoretical construct)
- 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 or previous research, or the experience of the researcher (theory).
- 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.
- Often the distribution of the number of cases in the analysis and **validation samples** follows the distribution in the total sample.

Information on Resort Visits: Analysis Sample

Table 18.2

No.	Resort Visit	Annual Family Income ($000)	Attitude Toward Travel	Importance Attached to Family Vacation	Household Size	Age of Head of Household	Amount Spent on Family Vacation
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)

Information on Resort Visits: Holdout Sample

Table 18.3

No.	Resort Visit	Annual Family Income ($000)	Attitude Toward Travel	Importance Attached to Family Vacation	Household Size	Age of Head of Household	Amount Spent on Family Vacation
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)

Conducting Discriminant Analysis Estimate the Discriminant Function Coefficients

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



Results of Two-Group Discriminant Analysis

Table 18.4

GROUP MEANS

VISIT	INCOME	TRAVEL	VACATION	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	4.9333	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	.94112	8.27101
Total	12.79523	1.97804	2.09981	1.33089	8.57395

Pooled Within-Groups Correlation Matrix

	INCOME	TRAVEL	VACATION	HSIZE	AGE
INCOME	1.00000				
TRAVEL	0.19745	1.00000			
VACATION	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

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

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

Contd.

Results of Two-Group Discriminant Analysis

Table 18.4 cont.

CANONICAL DISCRIMINANT FUNCTIONS

Function	Eigenvalue	% of Variance	Cum %	Canonical Correlation	After Function	Wilks' λ	Chi-square	df	Significance
1*	1.7862	100.00	100.00	0.8007	: 0 :	0 .3589	26.130	5	0.0001

* marks the 1 canonical discriminant functions remaining in the analysis.

Standard Canonical Discriminant Function Coefficients

	FUNC 1
INCOME	0.74301
TRAVEL	0.09611
VACATION	0.23329
HSIZE	0.46911
AGE	0.20922

Structure Matrix:

Pooled within-groups correlations between discriminating variables & canonical discriminant functions (variables ordered by size of correlation within function)

	FUNC 1
INCOME	0.82202
HSIZE	0.54096
VACATION	0.34607
TRAVEL	0.21337
AGE	0.16354

Results of Two-Group Discriminant Analysis

Table 18.4 cont.

Unstandardized Canonical Discriminant Function Coefficients

	FUNC 1
INCOME	0.8476710E-01
TRAVEL	0.4964455E-01
VACATION	0.1202813
HSIZE	0.4273893
AGE	0.2454380E-01
(constant)	-7.975476

Canonical discriminant functions evaluated at group means (group centroids)

Group	FUNC 1
1	1.29118
2	-1.29118

Classification results for cases selected for use in analysis

	Actual Group	Predicted No. of Cases	Group Membership 1	2
Group	1	15	12 80.0%	3 20.0%
Group	2	15	0 0.0%	15 100.0%

Percent of grouped cases correctly classified: 90.00%

Contd.

Results of Two-Group Discriminant Analysis

Table 18.4 cont.

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

	Actual Group	Predicted No. of Cases	Group Membership 1	2
Group	1	6	4 66.7%	2 33.3%
Group	2	6	0 0.0%	6 100.0%

Percent of grouped cases correctly classified: 83.33%.

Conducting Discriminant Analysis
Determine the Significance of Discriminant Function

- The **null hypothesis** that, in the population, the means of all discriminant functions in all groups are equal can be statistically tested.
- In SPSS/SAS 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 for each function. The significance level is estimated based on a chi-square transformation of the statistic.
- If the null hypothesis is rejected, indicating significant discrimination, one can proceed to interpret the results.

Conducting Discriminant Analysis
Interpret the Results

- The interpretation of the discriminant weights, or coefficients, is similar to that in **multiple regression** analysis.
- 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 (similar to the *partial correlation* concept).
- 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.

Conducting Discriminant Analysis Access Validity of Discriminant Analysis

- Many computer programs, such as SPSS and SAS, offer a leave-one-out cross-validation option.
- 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 **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.
- Classification accuracy achieved by discriminant analysis should be **at least 25% greater than that obtained by chance** (i.e., two groups, 75%; three groups, 60%)

Results of Three-Group Discriminant Analysis

Table 18.5

Group Means

AMOUNT	INCOME	TRAVEL	VACATION	HSIZE	AGE
1	38.57000	4.50000	4.70000	3.10000	50.30000
2	50.11000	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	TRAVEL	VACATION	HSIZE	AGE
INCOME	1.00000				
TRAVEL	0.05120	1.00000			
VACATION	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

Contd.

Results of Three-Group Discriminant Analysis

Table 18.5 cont.

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

Variable	Wilks' Lambda	*F*	Significance
INCOME	0.26215	38.00	0.0000
TRAVEL	0.78790	3.634	0.0400
VACATION	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	% of Variance	Cum %	Canonical Correlation	After Function	Wilks' λ	Chi-square	df	Significance
					: 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.

Standardized Canonical Discriminant Function Coefficients

	FUNC 1	FUNC 2
INCOME	1.04740	-0.42076
TRAVEL	0.33991	0.76851
VACATION	-0.14198	0.53354
HSIZE	-0.16317	0.12932
AGE	0.49474	0.52447

Contd.

Results of Three-Group Discriminant Analysis

Table 18.5 cont.

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
VACATION	0.21935	0.58829*
TRAVEL	0.14899	0.45362*
AGE	0.16576	0.34079*

Unstandardized canonical discriminant function coefficients

	FUNC 1	FUNC 2
INCOME	0.1542658	-0.6197148E-01
TRAVEL	0.1867977	0.4223430
VACATION	-0.6952264E-01	0.2612652
HSIZE	-0.1265334	0.1002796
AGE	0.5928055E-01	0.6284206E-01
(constant)	-11.09442	-3.791600

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

Contd.

Results of Three-Group Discriminant Analysis

Table 18.5 cont.

Classification Results:

	Actual Group	No. of Cases	Predicted Group Membership 1	2	3
Group	1	10	9 90.0%	1 10.0%	0 0.0%
Group	2	10	1 10.0%	9 90.0%	0 0.0%
Group	3	10	0 0.0%	2 20.0%	8 80.0%

Percent of grouped cases correctly classified: 86.67%

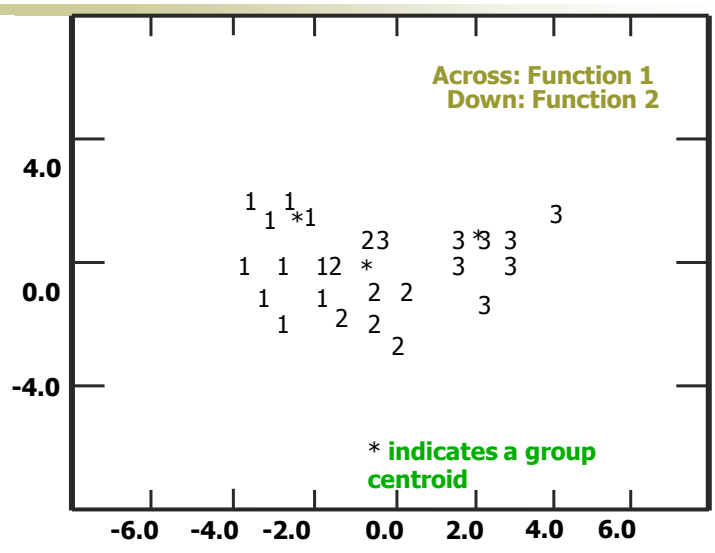
Classification results for hold out sample

	Actual Group	No. of Cases	Predicted Group Membership 1	2	3
Group	1	4	3 75.0%	1 25.0%	0 0.0%
Group	2	4	0 0.0%	3 75.0%	1 25.0%
Group	3	4	1 25.0%	0 0.0%	3 75.0%

Percent of grouped cases correctly classified: 75.00%

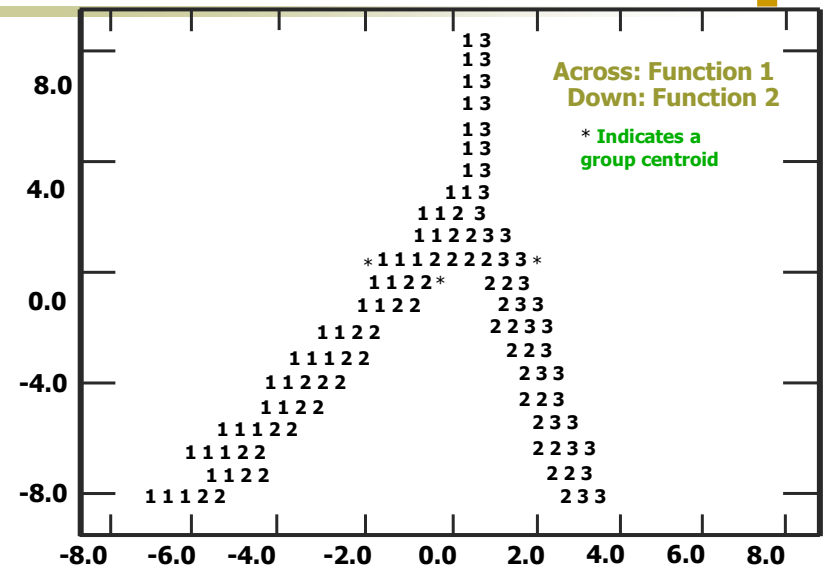
All-Groups Scattergram

Fig. 18.2



Territorial Map

Fig. 18.3



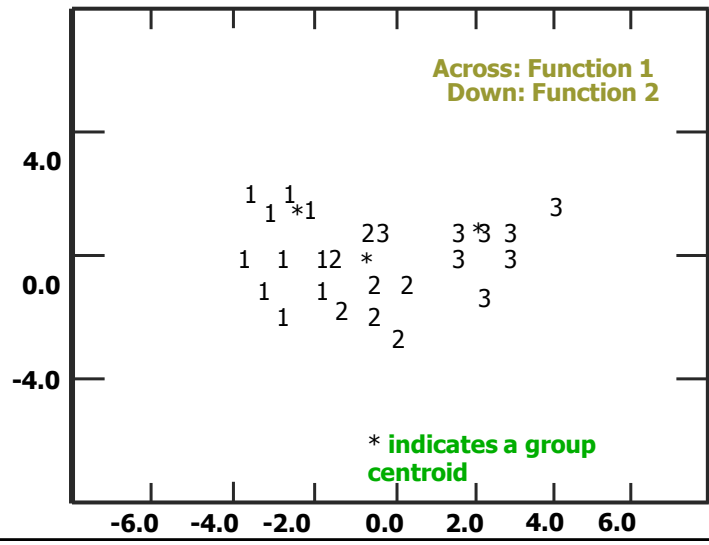
Stepwise Discriminant Analysis

- Stepwise discriminant analysis is analogous to stepwise multiple regression (see Chapter 17) 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 analysis of variance 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.

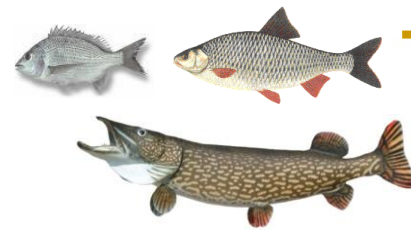
Stepwise Discriminant Analysis

- A second predictor is added based on the highest adjusted or partial *F* ratio, taking into account the predictor already selected.
- The selection of the stepwise procedure is based on the optimizing criterion adopted. The **Mahalanobis procedure** is based on **maximizing** a generalized measure of the distance between the two closest groups.
- The order in which the variables were selected also indicates their importance in discriminating between the groups.

Mahalanobis procedure



SAS Exercise – Discriminant Analysis



```
proc format;
   value specfmt
      1='Bream'
      2='Roach'
      3='Whitefish'
      4='Parkki'
      5='Perch'
      6='Pike'
      7='Smelt';
data fish (drop=HtPct WidthPct);
   title 'Fish Measurement Data';
   input Species Weight Length1 Length2 Length3 HtPct
         WidthPct @@;
   Height=HtPct*Length3/100;
   Width=WidthPct*Length3/100;
   format Species specfmt.;
   symbol = put(Species, specfmt2.);
   datalines;
1  242.0 23.2 25.4 30.0 38.4 13.4
1  290.0 24.0 26.3 31.2 40.0 13.8
1  340.0 23.9 26.5 31.1 39.8 15.1
1  363.0 26.3 29.0 33.5 38.0 13.3
 ...[155 more records]
```

A

```
proc candisc data=fish ncan=3 out=outcan;
   class Species;
   var Weight Length1 Length2 Length3 Height Width;
run;
```

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion
Bream	Bream	34	34.0000	0.215190
Parkki	Parkki	11	11.0000	0.069620
Perch	Perch	56	56.0000	0.354430
Pike	Pike	17	17.0000	0.107595
Roach	Roach	20	20.0000	0.126582
Smelt	Smelt	14	14.0000	0.088608
Whitefish	Whitefish	6	6.0000	0.037975

Multivariate Statistics and F Approximations

S=6 M=-0.5 N=72

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00036325	90.71	36	643.89	<.0001
Pillai's Trace	3.10465132	26.99	36	906	<.0001
Hotelling-Lawley Trace	52.05799676	209.24	36	413.64	<.0001
Roy's Greatest Root	39.13499776	984.90	6	151	<.0001

```
proc candisc data=fish ncan=3 out=outcan;
   class Species;
   var Weight Length1 Length2 Length3 Height Width;
run;
```

The CANDISC Procedure

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.987463	0.986671	0.001989	0.975084
2	0.952349	0.950095	0.007425	0.906969
3	0.838637	0.832518	0.023678	0.703313
4	0.633094	0.623649	0.047821	0.400809
5	0.344157	0.334170	0.070356	0.118444
6	0.005701	.	0.079806	0.000033

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	39.1350	29.3859	0.7518	0.7518
2	9.7491	7.3786	0.1873	0.9390
3	2.3706	1.7016	0.0455	0.9846
4	0.6689	0.5346	0.0128	0.9974
5	0.1344	0.1343	0.0026	1.0000
6	0.0000		0.0000	1.0000


```
proc candisc data=fish ncan=3 out=outcan;
   class Species;
   var Weight Length1 Length2 Length3 Height Width;
run;
```


Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.00036325	90.71	36	643.89	<.0001
2	0.01457896	46.46	25	547.58	<.0001
3	0.15671134	23.61	16	452.79	<.0001
4	0.52820347	12.09	9	362.78	<.0001
5	0.88152702	4.88	4	300	0.0008
6	0.99996749	0.00	1	151	0.9442

Figure 21.4. Likelihood Ratio Test

The first canonical variable, Can1, shows that the linear combination of the centered variables Can1= $-0.0006\times$Weight $- 0.33\times$Length1 $- 2.49\times$Length2 $+ 2.60\times$Length3 $+ 1.12\times$Height $- 1.45\times$Width separates the species most effectively (see Figure 21.5).

Raw Canonical Coefficients

Variable	Can1	Can2	Can3
Weight	-0.000648508	-0.005231659	-0.005596192
Length1	-0.329435762	-0.626598051	-2.934324102
Length2	-2.486133674	-0.690253987	4.045038893
Length3	2.595648437	1.803175454	-1.139264914
Height	1.121983854	-0.714749340	0.283202557
Width	-1.446386704	-0.907025481	0.741486686

PROC CANDISC computes the means of the canonical variables for each class. The first canonical variable is the linear combination of the variables Weight, Length1, Length2, Length3, Height, and Width that provides the greatest difference (in terms of a univariate F-test) between the class means. The second canonical variable provides the greatest difference between class means while being uncorrelated with the first canonical variable.

Fish Measurement Data

The CANDISC Procedure

Class Means on Canonical Variables

Species	Can1	Can2	Can3
Bream	10.94142464	0.52078394	0.23496708
Parkki	2.58903743	-2.54722416	-0.49326158
Perch	-4.47181389	-1.70822715	1.29281314
Pike	-4.89689441	8.22140791	-0.16469132
Roach	-0.35837149	0.08733611	-1.10056438
Smelt	-4.09136653	-2.35805841	-4.03836098
Whitefish	-0.39541755	-0.42071778	1.06459242

A plot of the first two canonical variables (Figure 21.7) shows that Can1 discriminates between three groups: 1) bream; 2) whitefish, roach, and parkki; and 3) smelt, pike, and perch. Can2 best discriminates between pike and the other species.

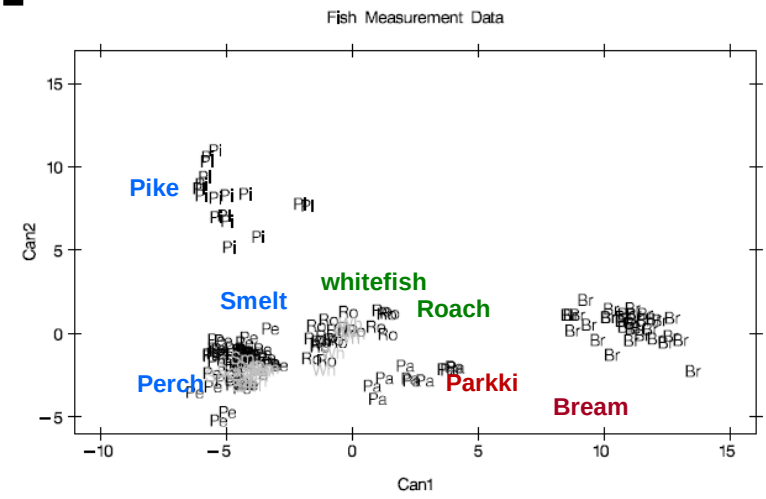
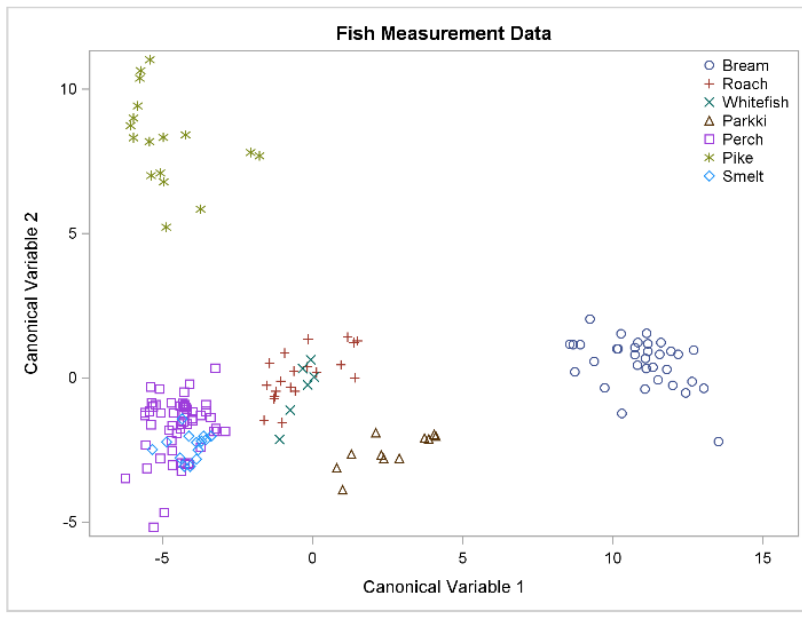


Figure 27.6 Plot of First Two Canonical Variables



B

```
proc discrim data=fish;
   class Species;
run;
```

The coefficients of the linear discriminant function are displayed (in Figure 31.4) with the default options METHOD=NORMAL and POOL=YES.

Figure 31.4 Linear Discriminant Function

Linear Discriminant Function for Species

Variable	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish
Constant	-185.91682	-64.92517	-48.68009	-148.06402	-62.65963	-19.70401	-67.44603
Weight	-0.10912	-0.09031	-0.09418	-0.13805	-0.09901	-0.05778	-0.09948
Length1	-23.02273	-13.64180	-19.45368	-20.92442	-14.63635	-4.09257	-22.57117
Length2	-26.70692	-5.38195	17.33061	6.19887	-7.47195	-3.63996	3.83450
Length3	50.55780	20.89531	5.25993	22.94989	25.00702	10.60171	21.12638
Height	13.91638	8.44567	-1.42833	-8.99687	-0.26083	-1.84569	0.64957
Width	-23.71895	-13.38592	1.32749	-9.13410	-3.74542	-3.43630	-2.52442

A summary of how the discriminant function classifies the data used to develop the function is displayed last. In Figure 31.5, you see that only three of the observations are misclassified. The error-count estimates give the proportion of misclassified observations in each group. Since you

Number of Observations and Percent Classified into Species

From Species	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Bream	34	0	0	0	0	0	0	34
	100.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
Parkki	0	11	0	0	0	0	0	11
	0.00	100.00	0.00	0.00	0.00	0.00	0.00	100.00
Perch	0	0	53	0	0	3	0	56
	0.00	0.00	94.64	0.00	0.00	5.36	0.00	100.00
Pike	0	0	0	17	0	0	0	17
	0.00	0.00	0.00	100.00	0.00	0.00	0.00	100.00
Roach	0	0	0	0	20	0	0	20
	0.00	0.00	0.00	0.00	100.00	0.00	0.00	100.00
Smelt	0	0	0	0	0	14	0	14
	0.00	0.00	0.00	0.00	0.00	100.00	0.00	100.00
Whitefish	0	0	0	0	0	0	6	6
	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
Total	34	11	53	17	20	17	6	158
	21.52	6.96	33.54	10.76	12.66	10.76	3.80	100.00
Priors	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	

C

```
proc stepdisc data=fish;
   class Species;
run;
```

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Height		0.7553	77.69	<.0001	0.24466983	<.0001	0.12588836	<.0001
2	2	Length2		0.9229	299.31	<.0001	0.01886065	<.0001	0.25905822	<.0001
3	3	Length3		0.8826	186.77	<.0001	0.00221342	<.0001	0.38427100	<.0001
4	4	Width		0.5775	33.72	<.0001	0.00093510	<.0001	0.45200732	<.0001
5	5	Weight		0.4461	19.73	<.0001	0.00051794	<.0001	0.49488458	<.0001
6	6	Length1		0.2987	10.36	<.0001	0.00036325	<.0001	0.51744189	<.0001

Figure 67.5. Step Summary

All the variables in the data set are found to have potential discriminatory power.

Edition. The Multivariate>Discriminant Analysis task offers both two-group and multiple discriminant analysis.

Both two-group and multiple discriminant analysis can be performed using the Discriminant Analysis task within the SAS Learning Edition. To select this task click:

The steps for running three-group discriminant analysis are similar to these steps.

To run logit analysis or logistic regression using the SAS Learning Edition, click:

Analyze>Regression>Logistic . . .

The following are the detailed steps for running logit analysis with brand loyalty as the dependent variable and attitude toward the brand, attitude toward the product category, and attitude toward shopping as the independent variables using the data of Table 18.6.

1. Select ANALYZE from the SAS Learning Edition menu bar.
2. Click REGRESSION and then LOGISTIC.
3. Move LOYALTY to the Dependent variable task role.
4. Move BRAND, PRODUCT, and SHOPPING to the Quantitative variables task role.
5. Select MODEL EFFECTS.
6. Choose BRAND, PRODUCT, and SHOPPING as Main Effects.
7. Select MODEL OPTIONS.
8. Check SHOW CLASSIFICATION TABLE and enter 0.5 as the critical probability value.
9. Click RUN.

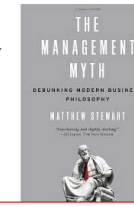
Strategic Management Journal
Strat. Mgmt. J., **31**: 39–57 (2010)
Published online EarlyView in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/smj.799
Received 26 December 2007; Final revision received 4 July 2009

DOES FIRM PERFORMANCE REVEAL ITS OWN CAUSES? THE ROLE OF BAYESIAN INFERENCE

YING-CHAN TANG[1] and FEN-MAY LIOU[2]*
[1] *Institute of Business and Management, National Chiao Tung University, Taipei, Taiwan, ROC*
[2] *Graduate Institute of Business and Management, Yuanpei University, Hsin Chu, Taiwan, ROC*

What Killed Michael Porter's Monitor Group? The One Force That Really Matters

What killed the Monitor Group, the consulting firm co-founded by the legendary business guru, Michael Porter? In November 2012, Monitor was unable to pay its bills and was forced to file for bankruptcy protection. Why didn't the highly paid consultants of Monitor use Porter's famous five-force analysis to save themselves?



What went wrong?

Was Monitor's demise something that happened

Table 3. Discriminant analysis on advantaged and disadvantaged firms

	Standardized canonical coefficients	F value	Prob > F
ART	−0.0517	1.08	0.3014
CGS/S	−0.2782	0.15	0.6958
APT	−0.1720	0.66	0.4184
INVT	0.1710	3.74	0.0552*
R&D/S	−0.2515	23.85	<0.0001***
SG&A/S	−0.2504	37.89	<0.0001***
Dep/S	−0.3525	2.01	0.1587
FAT	−0.1317	0.14	0.7136
Tax/S	0.2282	19.35	<0.0001***

*$p < 0.10$; **$p < 0.05$; *** $p < 0.01$.

Eigenvalue	Canonical correlation	Likelihood ratio	F value	Prob > F
1.1121	0.725628	0.47346	15.82	<0.0001

Classification results used for cross-validation[a]

Groups	Competitive advantage	Competitive disadvantage	Total
Competitive advantage	64	13	77
Competitive disadvantage	15	46	61

[a] Cross-validation is done by recalculating the discriminant function for all firms other than the validated firm.
[b] 88.4% ((70 + 52)/138) of firms are correctly classified.
[c] 79.7% ((64 + 46)/138) of the cross-validated firms remain correctly classified.

The Logit Model

- DV is binary, and IX are metric
- Binary Logit model, or, Logistic Regression

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p).$$

$$= a0 + a1X1 + a2X2 + a3X3 + \ldots + akXk = \sum aiXi$$

- Where p = probability of success; Xi are independent variables; ai are parameters to be estimated
- If *p* is a probability of success then *p*/(1 − *p*) is the corresponding odds, and the logit of the probability is the logarithm of the odds; similarly the difference between the logits of two probabilities is the logarithm of the **odds-ratio**.

The Logit Model

Further features of the LOGISTIC procedure enable you to do the following:

- control the ordering of the response categories
- compute a generalized R^2 measure for the fitted model
- reclassify binary response observations according to their predicted response probabilities
- test linear hypotheses about the regression parameters
- create a data set for producing a receiver operating characteristic curve for each fitted model
- specify contrasts to compare several receiver operating characteristic curves
- create a data set containing the estimated response probabilities, residuals, and influence diagnostics
- score a data set by using a previously fitted model

Logistic Regression with SAS

- Suppose the response variable Y is 0 or 1 binary (loyal, non-loyal), and X1 and X2 are two regressors of interest.
- SAS PROC LOGISTIC models the probability of Y=0 by default. In other words, SAS chooses the smaller value to estimate its probability. One way to change the default setting in order to model the probability of Y=1 in SAS is to specify the DESCENDING option on the PROC LOGISTIC statement. That is, use:
- proc logistic descending;
- model y=x1 x2;
- run;

Health Care Manage Sci (2008) 11:353–358
DOI 10.1007/s10729-008-9054-y

Detecting hospital fraud and claim abuse through diabetic outpatient services

Fen-May Liou · Ying-Chan Tang · Jean-Yi Chen

Received: 3 July 2007 / Accepted: 8 January 2008 / Published online: 19 January 2008
© Springer Science + Business Media, LLC 2008

Abstract Hospitals and health care providers tend to get involved in exaggerated and fraudulent medical claims initiated by national insurance schemes. The present study applies data mining techniques to detect fraudulent or

1 Introduction

Healthcare fraud and abuse are of major concern in many countries, in some cases costing public and private financial

Table 1 Descriptive statistics for normal and fraudulent hospitals

Variable (per case)	Normal Hospital		Fraudulent Hospital	
	Mean	SD	Mean	SD
Average days of drug dispense	7.72	5.60	7.39	1.50
Average drug cost	221.63	274.06	208.25	88.13
Average consultation and treatment fees	358.71	176.88	259.58	113.69
Average diagnosis fees	265.42	42.93	265.00	43.46
Average dispensing service fees	24.48	8.13	30.01	11.21
Average medical expenditure	548.04	408.33	584.75	145.92
Average amount claimed	487.99	394.69	511.81	131.97
Average drug cost per day	28.81	27.79	33.82	10.07
Average medical expenditure per day	134.37	92.29	173.13	73.56

claim was associated with the value "0" if regular, and "1" if irregular.

Stepwise logistic regression was performed on each variable individually to identify the most effective factors

weighted sum of the input variables, and transform that sum to an output signal using some kind of threshold function (typically a step function or sigmoid) [30, 31]. The output layer (often a single node) receives a weighted sum of the

THE WALL STREET JOURNAL.
ASIA
PRINT ◦ ONLINE ◦ MOBILE

2013/11/19 07:43:43

"僵屍論文"中埋葬的美麗年華

正文 評論(1)

投稿 打印 轉發 0

字體

董潔林

世界上有些人具備獨特的洞察力，可以從糞朽中發現神奇，物理學家普萊斯（Derek John de Solla Price）就是其中一位。20世紀60年代初，普萊斯的腦角累積了一堆老舊論文集，他在隨便翻閱中發現了一系列有關論文的數字規律，例如在所發表的論文中，約35%的論文從來未被引用，49%的論文被引用一次，9%被引用兩次，3%被引用三次，2%被引用四次，而被引用次數大於等於五次的論文，共佔2%。他進而對這個現象進行了系統地分析並發表了裡程碑性質的論文：科學的網絡。他發現的規律後來被很多人用不同的論文集合驗証，數字雖然略有不同，但規律不會變，因為這反映了自組織社會的通性，被現在的學者們稱之為"無尺度網絡"。

無獨有偶，美國結構語言學博士尤金•加菲爾德（Eugene Garfield）在普萊斯之前也發現論文的引文關系可以提供很多重要的信息，例如論文的被引數可以反映論文的質量。因此，他成立了一家公司，專門收集和發布學術論文索引的統計數據，也即現在學術界流行的SCI和SSCI文獻索引數據庫，如今這兩個統計系統已經成為國際上各類學術期刊、論文和學者排行榜的重要參照，論文被引數也被認可為較好地反映了論文的原創性和權威性。