Chapter 4

Classification into Groups: Discriminant Analysis

Section 4.1

Introduction: Canonical Discriminant Analysis

Objectives

- Understand the goals of discriminant analysis.
- Identify similarities between discriminant analysis and multivariate general linear models.
- Explain how to perform canonical discriminant analysis.
- Use the CANDISC procedure to perform canonical discriminant analysis.

The Research Questions

- A credit card company wants to use financial information to decide whether a potential customer is a good risk or a bad risk before offering a credit card.
- A school district wants to use classroom behavior and scores to identify candidate students for a learning intervention program.
- An insurance company wants to understand what demographic and behavioral variables are most characteristic of different types of drivers.

Why Discriminant Analysis?

With discriminant analysis, you can

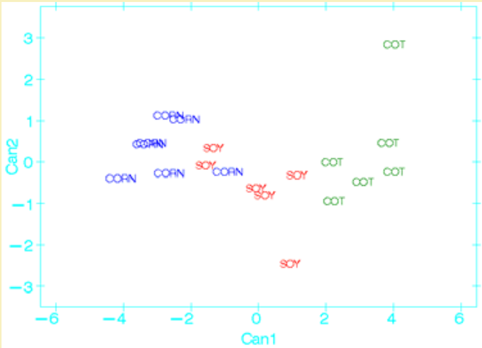
- interpret variables that are most characteristic of group differences
- use a linear combination of variables to predict group membership
- validate the model on a new sample in one step
- easily score new observations into groups.

Supervised Data Analysis

There are numerous types of analysis that are used to classify observations based on a set of variables. However,

- discriminant analysis is not the same as cluster analysis
- to perform discriminant analysis, you must have information about actual group membership in order to estimate discriminant functions
- discriminant analysis finds combinations of predictors that best differentiate the groups, so you can apply those linear combinations in the future to predict groups when group membership is **not** known.

How Does Discriminant Analysis Work?



7

Canonical Discriminant Analysis

If you were to perform canonical analysis on the crops data with all four remote sensing measures as one set of variables and the crops (dummy-coded) as the second set of variables, you would be performing the equivalent of canonical discriminant analysis.

- The number of discriminant functions is the minimum of the number of predictors or the number of groups minus one.
- For the crops example, you would have min (2,4) = 2 discriminant functions.

8

The Multivariate Linear Model

The linear model underlying discriminant analysis is essentially the same as MANOVA: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$

- The assumptions are the same as for MANOVA
- If the data are not multivariate normal, a nonparametric method may be preferred.

9

Two Goals for Discriminant Analysis

1. Interpretation: "How are the groups different?"

   Find and interpret linear combinations of variables that optimally predict group differences

2. Classification: "How accurately can observations be classified into groups?"

   Using functions of variables to predict group membership for a data set and evaluate expected error rates

10

Two Goals for Discriminant Analysis

1. Interpretation: "How are the groups different?"

   Find and interpret linear combinations of variables that optimally predict group differences

2. Classification: "How accurately can observations be classification into groups?"

   Using functions of variables to predict group membership for a data set and evaluate expected error rates

11

The CANDISC Procedure

General form of the CANDISC procedure:

```
PROC CANDISC <options>;
     CLASS variable;
     VAR variables;
RUN;
```

12

The %PLOTIT Macro

General form of the %PLOTIT macro:

```
%plotit (data=data set, plotvars=var1 var2,
         labelvar = varname, symvar=group_var,
         typevar=group_var, symsize = option,
         symlen=option);
```

13

Pathological Gambling Example

14

Canonical Discriminant Analysis

ch4s1d1.sas

This demonstration illustrates the CANDISC procedure for canonical discriminant analysis, and shows the %PLOTIT macro for visualizing discriminant functions.

15

Exercises

This exercise reinforces the concepts discussed previously.

16

Section 4.2

Fisher Linear Discriminant Analysis

Objectives

- Describe the steps involved in Fisher linear discriminant analysis.
- Contrast Fisher linear discriminant analysis with canonical discriminant analysis.
- Perform Fisher linear discriminant analysis using the DISCRIM procedure.

18

What Have You Learned?

- The predictors discriminate between groups.
- Both functions discriminate significantly.
- There are useful interpretations for the two functions.
- Both functions discriminate between different pairs of groups.

19

Two Goals for Discriminant Analysis

Recall the two aspects of discriminant analysis:

✓ Interpretation: "How are the groups different?"
Find and interpret linear combinations of variables that optimally predict group differences

2. Classification: "How accurately can observations be classified into groups?"
Using functions of variables to predict group membership for a data set and evaluate expected error rates

20

Compare and Contrast

Assuming number of groups < number of predictors:

Canonical discriminant analysis

- Number of functions = groups – 1
- Seek functions that maximally separate group centroids.

Fisher linear discriminant analysis

- Number of functions = groups
- Score observations on similarity to group centroids. Scores are converted to probability of membership in each group.

21

The DISCRIM Procedure

PROC DISCRIM can be used for many different types of analysis including

- canonical discriminant analysis
- assessing and confirming the usefulness of the functions (empirical validation and crossvalidation)
- predicting group membership on new data using the functions (scoring)
- linear and quadratic discriminant analysis
- nonparametric discriminant analysis

22

The DISCRIM Procedure

General form of the DISCRIM procedure:

```
PROC DISCRIM <options>;
      CLASS variable;
      PRIORS expression;
      VAR variables;
RUN;
```

23

Prior Probability Estimates

In discriminant analysis, it may be useful to specify prior probabilities, or *priors*.

By using PRIORS statement, you can specify how to estimate the probabilities of group membership in the population.

You will use *proportional*, *equal*, and *user-specified* priors in this chapter.

24

Fisher Linear Discriminant Analysis

ch4s2d1.sas

This demonstration illustrates the DISCRIM procedure for Fisher linear discriminant analysis.

25

Exercises

This exercise reinforces the concepts discussed previously.

26

sas | Education

Section 4.3

Quadratic Discriminant Analysis

Objectives

- Use the DISCRIM procedure to test the assumption of homogeneous covariance matrices
- Perform quadratic discriminant analysis

28

Homogeneity of Covariance Matrices

Recall that one assumption of the multivariate linear model is homogeneity of covariance matrices.

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
12	-1.27952

$$D_t^2(\mathrm{x}) = \underbrace{d_t^2(\mathrm{x})}_{\text{Mahalanobis distance}} + \underbrace{g_2(t)}_{-2(\ln(\text{prior}))}$$

If groups do have heterogeneous covariance structures, classifications based on a pooled covariance matrix can be prone to greater error in classification.

29

Quadratic Discriminant Analysis

Quadratic discriminant analysis uses a separate estimate of the covariance matrix for each group in calculating distances from group centroids:

$$D_t^2(\mathrm{x}) = d_t^2(\mathrm{x}) + \underbrace{g_1(t)}_{\ln|\mathbf{S}|} + g_2(t)$$

Within Covariance Matrix Information

type	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
Binge	12	-5.61076
Control	12	-1.26247
Steady	12	-10.13772
Pooled	12	-1.27952

30

Quadratic Discriminant Analysis

```
PROC DISCRIM … POOL=option SLPOOL=option;
```

OPTION:	ANALYSIS:
POOL=YES	Linear
POOL=NO	Quadratic
POOL=TEST	Depends on test

31

Demonstration

ch4s3d1.sas

This demonstration illustrates a test for homogeneity of covariance matrices and performs quadratic discriminant analysis using the DISCRIM procedure.

32

sas Education

Section 4.4

Discriminant Analysis and Empirical Validation

Objectives

- Use the DISCRIM procedure to validate discriminant analysis.

34

What Have You Learned?

- Using quadratic discriminant analysis with proportional priors, your expected error rate in classification is only 4%.
- Most of the errors in classification are expected to be among Controls.

35

Empirical Validation

Validation is particularly important in discriminant analysis.

- The observations that were classified were the same ones used to develop the equations, resulting in downwardly biased error count estimates.
- Discriminant analysis capitalizes on chance associations in the data.
- Apply the equations to a new set of data to get a better estimate of the expected error rate for the population.

36

PROC DISCRIM for Empirical Validation

General form of the DISCRIM procedure:

```
PROC DISCRIM DATA = old-data TESTDATA = new-data
                    TESTLIST;
      CLASS variable;
      PRIORS priors;
      VAR variables;
RUN;
```

37

Testing Discriminant Functions on a New Data Set

ch4s4d1.sas

This demonstration illustrates the DISCRIM procedure for empirical validation of discriminant functions.

38

What Have You Learned?

- Validating the discriminant functions on a new sample resulted in an estimated error rate of about 19%, which is still quite low.
- Most of the errors in the validation data set were from the Binge and Steady groups.

39

sas Education

Section 4.5

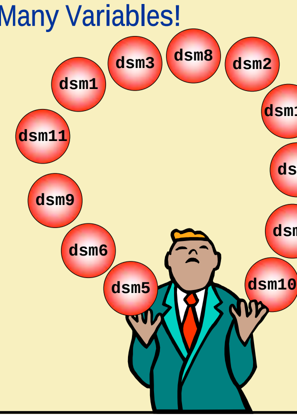
Stepwise Discriminant Analysis

Objectives

- Explain different methods for stepwise discriminant analysis.
- Understand some of the potential problems with stepwise analysis.
- Fit a stepwise model using PROC STEPDISC and interpret the output.
- Validate the results of a stepwise analysis using PROC DISCRIM.

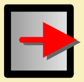
41

Too Many Variables!



42

Stepwise Selection Methods


Forward Selection


Backward Selection


Stepwise Selection

Analyze with Care

Although it is useful and efficient, stepwise methods have limitations:

- Correlated predictors
- Chance relationships in the data

Always validate your findings from a stepwise analysis.

The STEPDISC Procedure

General form of the STEPDISC procedure:

```
PROC STEPDISC DATA = data-set METHOD = method;
      CLASS variable;
      VAR variables;
RUN;
```


Stepwise Discriminant Analysis

ch4s5d1.sas

This demonstration illustrates the STEPDISC procedure for stepwise discriminant analysis and the DISCRIM procedure for empirical validation of the final model.

What Have You Learned?

- The reduced model with four predictors resulted in an estimated error rate of 11% in the calibration data (compared to 4% with the full, 12-predictor model).
- The reduced model resulted in an estimated error rate of about 21% in the validation data, compared with 19% with the full model.
- Reducing the model from 12 predictors to 4 predictors resulted in very little loss of predictive power.

Exercises

This exercise reinforces the concepts discussed previously.