SAS/STAT® 14.1 User's Guide
Introduction to Discriminant Procedures

This document is an individual chapter from *SAS/STAT® 14.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2015. *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.1 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

July 2015

Chapter 10

Introduction to Discriminant Procedures

Contents

Overview: Discriminant Procedures

The SAS procedures for discriminant analysis fit data with one classification variable and several quantitative variables. The purpose of discriminant analysis can be to find one or more of the following:

- a mathematical rule, or *discriminant function*, for guessing to which class an observation belongs, based on knowledge of the quantitative variables only
- a set of linear combinations of the quantitative variables that best reveals the differences among the classes
- a subset of the quantitative variables that best reveals the differences among the classes

The SAS discriminant procedures are as follows:

DISCRIM — computes various discriminant functions for classifying observations. Linear or quadratic discriminant functions can be used for data with approximately multivariate normal within-class distributions. Nonparametric methods can be used without making any assumptions about these distributions.

CANDISC — performs a canonical analysis to find linear combinations of the quantitative variables that best summarize the differences among the classes.

STEPDISC — uses forward selection, backward elimination, or stepwise selection to try to find a subset of quantitative variables that best reveals differences among the classes.

Background: Discriminant Procedures

The term *discriminant analysis* (Fisher 1936; Cooley and Lohnes 1971; Tatsuoka 1971; Kshirsagar 1972; Lachenbruch 1975, 1979; Gnanadesikan 1977; Klecka 1980; Hand 1981, 1982; Silverman 1986) refers to several different types of analyses. Classificatory discriminant analysis is used to classify observations into two or more known groups on the basis of one or more quantitative variables. Classification can be done by either a parametric method or a nonparametric method in the DISCRIM procedure. A parametric method is appropriate only for approximately normal within-class distributions. The method generates either a linear discriminant function (the within-class covariance matrices are assumed to be equal) or a quadratic discriminant function (the within-class covariance matrices are assumed to be unequal).

When the distribution within each group is not assumed to have any specific distribution or is assumed to have a distribution different from the multivariate normal distribution, nonparametric methods can be used to derive classification criteria. These methods include the kernel method and nearest-neighbor methods. The kernel method uses uniform, normal, Epanechnikov, biweight, or triweight kernels in estimating the group-specific density at each observation. The within-group covariance matrices or the pooled covariance matrix can be used to scale the data.

The performance of a discriminant function can be evaluated by estimating error rates (probabilities of misclassification). Error count estimates and posterior probability error rate estimates can be evaluated with PROC DISCRIM. When the input data set is an ordinary SAS data set, the error rates can also be estimated by cross validation.

In multivariate statistical applications, the data collected are largely from distributions different from the normal distribution. Various forms of nonnormality can arise, such as qualitative variables or variables with underlying continuous but nonnormal distributions. If the multivariate normality assumption is violated, the use of parametric discriminant analysis might not be appropriate. When a parametric classification criterion (linear or quadratic discriminant function) is derived from a nonnormal population, the resulting error rate estimates might be biased.

If your quantitative variables are not normally distributed, or if you want to classify observations on the basis of categorical variables, you should consider using the CATMOD or LOGISTIC procedure to fit a categorical linear model with the classification variable as the dependent variable. Press and Wilson (1978) compare logistic regression and parametric discriminant analysis and conclude that logistic regression is preferable to parametric discriminant analysis in cases for which the variables do not have multivariate normal distributions within classes. However, if you do have normal within-class distributions, logistic regression is less efficient than parametric discriminant analysis. Efron (1975) shows that with two normal populations having a common covariance matrix, logistic regression is between one-half and two-thirds as effective as the linear discriminant function in achieving asymptotically the same error rate.

Do not confuse discriminant analysis with cluster analysis. All varieties of discriminant analysis require prior knowledge of the classes, usually in the form of a sample from each class. In cluster analysis, the data do not include information about class membership; the purpose is to construct a classification. See Chapter 11, "Introduction to Clustering Procedures."

Canonical discriminant analysis is a dimension-reduction technique related to principal components and canonical correlation, and it can be performed by both the CANDISC and DISCRIM procedures. A discriminant criterion is always derived in PROC DISCRIM. If you want canonical discriminant analysis without the use of a discriminant criterion, you should use PROC CANDISC. Stepwise discriminant analysis is a variable-selection technique implemented by the STEPDISC procedure. After selecting a subset of variables with PROC STEPDISC, use any of the other discriminant procedures to obtain more detailed analyses. PROC CANDISC and PROC STEPDISC perform hypothesis tests that require the within-class distributions to be approximately normal, but these procedures can be used descriptively with nonnormal data.

Another alternative to discriminant analysis is to perform a series of univariate one-way ANOVAs. All three discriminant procedures provide summaries of the univariate ANOVAs. The advantage of the multivariate approach is that two or more classes that overlap considerably when each variable is viewed separately might be more distinct when examined from a multivariate point of view.

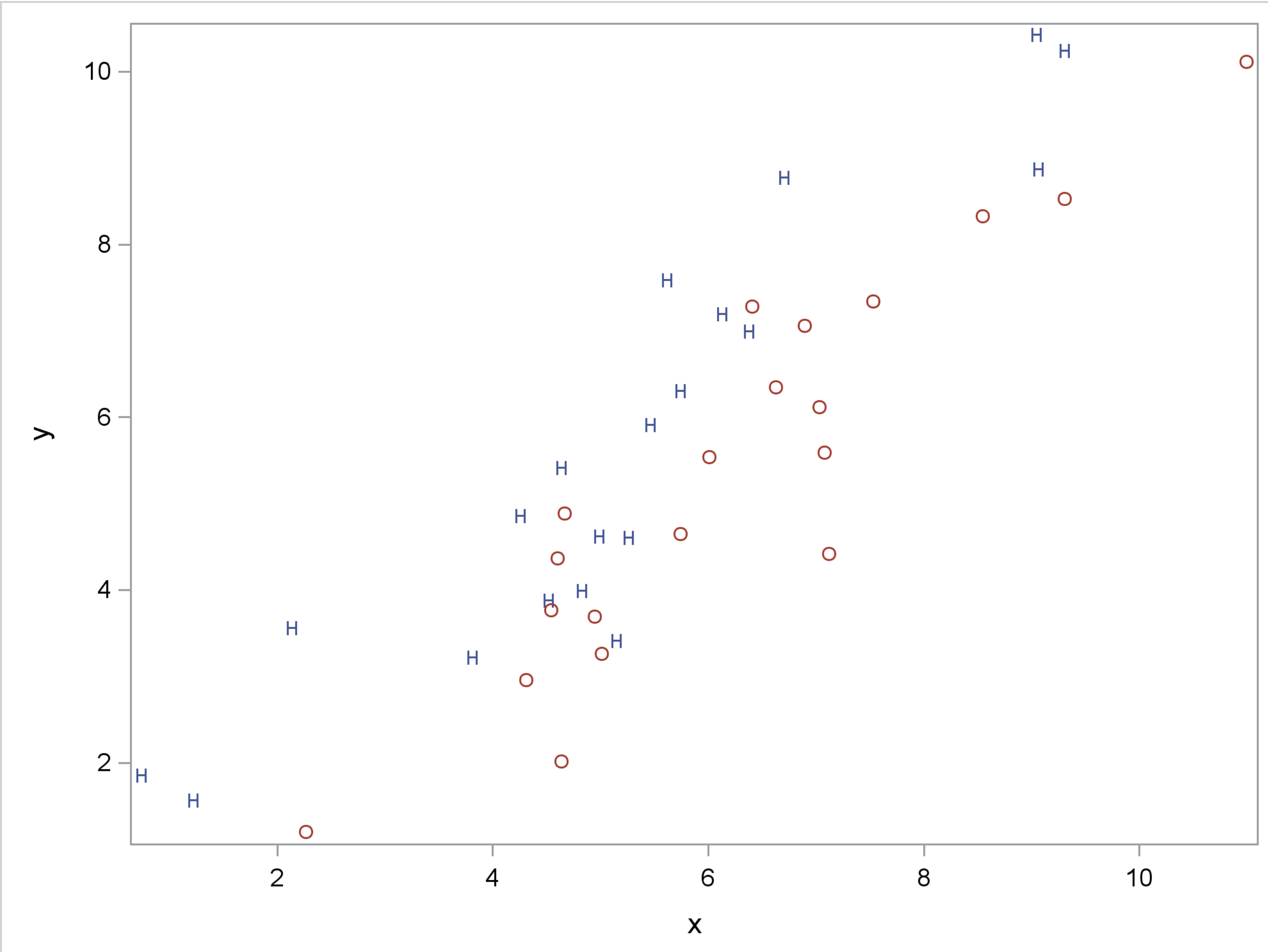
Example: Contrasting Univariate and Multivariate Analyses

Consider an artificial data set with two classes of observations indicated by 'H' and 'O'. The following statements generate and plot the data:

```
data random;
   drop n;

   Group = 'H';
   do n = 1 to 20;
      x = 4.5 + 2 * normal(57391);
      y = x + .5 + normal(57391);
      output;
   end;

   Group = 'O';
   do n = 1 to 20;
      x = 6.25 + 2 * normal(57391);
      y = x - 1 + normal(57391);
      output;
   end;

run;

proc sgplot noautolegend;
   scatter y=y x=x / markerchar=group group=group;
run;
```

The plot is shown in Figure 10.1.

Figure 10.1 Groups for Contrasting Univariate and Multivariate Analyses



The following statements perform a canonical discriminant analysis and display the results in Figure 10.2:

```
proc candisc anova;
   class Group;
   var x y;
run;
```

Figure 10.2 Contrasting Univariate and Multivariate Analyses

The CANDISC Procedure

Total Sample Size	40	**DF Total**	39
Variables	2	**DF Within Classes**	38
Classes	2	**DF Between Classes**	1

Number of Observations Read	40
Number of Observations Used	40

Figure 10.2 *continued*

Class Level Information				
Group	**Variable Name**	**Frequency**	**Weight**	**Proportion**
H	H	20	20.0000	0.500000
O	O	20	20.0000	0.500000

The CANDISC Procedure

Univariate Test Statistics

F Statistics, Num DF=1, Den DF=38

Variable	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
x	2.1776	2.1498	0.6820	0.0503	0.0530	2.01	0.1641
y	2.4215	2.4486	0.2047	0.0037	0.0037	0.14	0.7105

Average R-Square	
Unweighted	0.0269868
Weighted by Variance	0.0245201

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=17.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.64203704	10.31	2	37	0.0003
Pillai's Trace	0.35796296	10.31	2	37	0.0003
Hotelling-Lawley Trace	0.55754252	10.31	2	37	0.0003
Roy's Greatest Root	0.55754252	10.31	2	37	0.0003

The CANDISC Procedure

					Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
	Canonical Correlation	**Adjusted Canonical Correlation**	**Approximate Standard Error**	**Squared Canonical Correlation**	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**
1	0.598300	0.589467	0.102808	0.357963	0.5575		1.0000	1.0000

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	0.64203704	10.31	2	37	0.0003

Note: The F statistic is exact.

The CANDISC Procedure

Total Canonical Structure	
Variable	**Can1**
x	-0.374883
y	0.101206

Figure 10.2 *continued*

Between Canonical Structure	
Variable	**Can1**
x	-1.000000
y	1.000000

Pooled Within Canonical Structure	
Variable	**Can1**
x	-0.308237
y	0.081243

The CANDISC Procedure

Total-Sample Standardized Canonical Coefficients	
Variable	**Can1**
x	-2.625596855
y	2.446680169

Pooled Within-Class Standardized Canonical Coefficients	
Variable	**Can1**
x	-2.592150014
y	2.474116072

Raw Canonical Coefficients	
Variable	**Can1**
x	-1.205756217
y	1.010412967

Class Means on Canonical Variables	
Group	**Can1**
H	0.7277811475
O	-.7277811475

The univariate R squares are very small, 0.0503 for x and 0.0037 for y, and neither variable shows a significant difference between the classes at the 0.10 level.

The multivariate test for differences between the classes is significant at the 0.0003 level. Thus, the multivariate analysis has found a highly significant difference, whereas the univariate analyses failed to achieve even the 0.10 level. The raw canonical coefficients for the first canonical variable, Can1, show that the classes differ most widely on the linear combination -1.205756217 x + 1.010412967 y or approximately y

- 1.2 x. The R square between Can1 and the CLASS variable is 0.357963 as given by the squared canonical correlation, which is much higher than either univariate R square.

In this example, the variables are highly correlated within classes. If the within-class correlation were smaller, there would be greater agreement between the univariate and multivariate analyses.

References

Cooley, W. W., and Lohnes, P. R. (1971). *Multivariate Data Analysis*. New York: John Wiley & Sons.

Efron, B. (1975). "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis." *Journal of the American Statistical Association* 70:892–898.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7:179–188.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: John Wiley & Sons.

Hand, D. J. (1981). *Discrimination and Classification*. New York: John Wiley & Sons.

Hand, D. J. (1982). *Kernel Discriminant Analysis*. New York: Research Studies Press.

Klecka, W. R. (1980). *Discriminant Analysis.* Vol. 07-019 of Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage Publications.

Kshirsagar, A. M. (1972). *Multivariate Analysis*. New York: Marcel Dekker.

Lachenbruch, P. A. (1975). *Discriminant Analysis*. New York: Hafner Publishing.

Lachenbruch, P. A. (1979). "Discriminant Analysis." *Biometrics* 35:69–85.

Press, S. J., and Wilson, S. (1978). "Choosing between Logistic Regression and Discriminant Analysis." *Journal of the American Statistical Association* 73:699–705.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.

Tatsuoka, M. M. (1971). *Multivariate Analysis*. New York: John Wiley & Sons.