THE DISCRIM PROCEDURE IN SAS: A COMPARATIVE ANALYSIS

Emma L. Frazier, University of South Carolina

I. INTRODUCTION

Using the Statistical Analysis System (SAS) as the primary package for statistical analyses, there are several options that are available to perform discriminant analysis. Three of the options are to utilize 1) the DISCRIM procedure in SAS79; 2) the BMDP7M program through the SAS-BMDP interface of the 1976 version of SAS; and 3) the MULDIS procedure, which is developed from a FORTRAN program written by Robert B. Avery and Robert A. Eisenbeis at the Federal Deposit Insurance Company (FDIC).

The primary objective of this paper is to compare current multiple discriminant analyses and classification routines in the DISCRIM and MULDIS procedures of SAS and the BMDP7M program through the SAS-BMDP interface. Comparisons are made based on the analysis of the three iris groups of the classical Fisher Iris data. In addition to a comparison of the three options, a review of the purpose and basic assumptions of discriminant analysis is presented.

II. PURPOSE OF DISCRIMINANT ANALYSIS

A basic purpose of multiple discriminant analysis is to assign unknown observations or subjects to one of two or more distinct and identifiable groups. While the assignment of an observation or subject to one of the groups is generally based on some multivariate observation, other factors as the a priori probabilities of belonging to one of the groups and the costs of misclassification must be considered.

Multiple discriminant analysis deals with several statistically related problems involving two or more groups when each observation is characterized by measurements on several variables. Four questions are of concern:

1) Are there any significant differences among the groups?
2) What is a "good" allocation rule for assigning each observation to one of the groups, while minimizing the costs of misclassification?
3) How well does the proposed discriminant function perform on existing, as well as new sample observations? This involves the estimation of error rates of the discriminant function.
4) Will a subset of the discriminating variables be sufficient for classification and discrimination among the groups? In general one may be interested in reducing the number of variables in the discriminant function, while minimizing the loss of information.

III. UNDERLYING ASSUMPTIONS OF DISCRIMINANT ANALYSIS

The underlying assumptions of multiple discriminant analysis are:

1) the groups under investigation must be k (two or more) discrete and identifiable groups;
2) the variables used to examine group differences and to classify observations have populations with multivariate normal distributions; and
3) the within-groups dispersion or variance-covariance matrices for the variables are equal for each group investigation.

IV. DESCRIPTION OF DATA

To compare the three options for performing discriminant analysis through the SAS package, the classical Fisher Iris data are examined. Three species of the iris plants, the iris setosa, iris versicolor and iris virginica are compared. Measurements of each plant's sepal length, sepal width, petal length and petal width were recorded and used as predictor variables.

V. RESULTS

Each of the three programs calculates and prints simple statistics as the means and standard deviations for each group of the iris plants. Group means and standard deviations for all groups are not presented in the DISCRIM procedure of the SAS package. Tests of the equality of group means are performed in the MULDIS and BMDP7M programs. This test is not performed in the DISCRIM procecure in SAS, though differences in group means may be calculated using the GLM procedure. Results as reported in Table 1 indicate that the group means are significantly different among the three species of plants.

Both the DISCRIM and MULDIS procedures compute the individual and pooled variance-covariance matrices and a test of the equality of the dispersion matrices. The DISCRIM procedure reports the results of Box's M, while MULDIS presents a F-statistic. Both statistics suggest that homogeneity among the variances of the three iris species does not hold; hence indicating that quadratic classification rules should be examined. Results of the tests of equality of dispersion matrices are presented in Table 2.

The test for the equality of dispersion matrices is not presented in the BMDP7M program. In the BMDP3D program, a F-statistic is computed to determine if the three dispersion matrices are equal. Since this test is not included in the BMDP7M program, the research is restricted to use of the linear classification rule to classify observations.

A summary table, resulting from the use of linear classification rules is given by each program. Results of quadratic classification rules are given by the MULDIS and DISCRIM procedures. Summary tables for the linear and quadratic classification rules are given in Table 3.

Three methods to estimate error rates of the discriminant function are the resubstitution, holdout and jackknife methods. The first two methods are available in the DISCRIM procedure in SAS. The MULDIS and BMDP7M programs utilize all three methods to estimate error rates. Identical results obtained for the resubstitution and jackknife methods are reported in Table 4.

The fourth area of concern is whether a subset of the discriminating variables is sufficient to classify observations into one of two or more groups. The DISCRIM procedures does not make any provisions for reducing the original number of discriminating variables. Univariate F-ratios, in addition to the stepwise forward and backward methods are available in the MULDIS and BMDP7M programs. Results of the univariate F-ratio tests and stepwise methods are shown in Table 5.

The complete selection method chooses the "best" set of discriminating variables independent of all other different size subsets. Results of the complete selection methods revealed the "best" subset of three variables and are presented in Table 5.

Data processing was performed on an AMDAHL 470 V6 system. The DISCRIM program required the least amount of CPU time, 1.02 sec; BMDP7M used 2.46 sec; and the MULDIS program required 3.26 sec of CPU time. CPU time includes the time for stepping routines in the BMDP7M and MULDIS programs.

VI. DISCUSSION

The program with the most available options is the MULDIS program written by R.A. Eisenbeis and R. Avery. One potential limitation of MULDIS is that the keypunching of information requires special care since eavh value is column specific. It contains most routines discussed in this paper, but it does not allow for transformations of variables. When there are more than two groups, the mean values (see Table 1) computed do not compare precisely with mean values calculated from the DISCRIM and BMDP7M programs.

In the BMDP package, the greatest drawback of BMDP7M is the inability to examine quadratic classification rules in addition to linear classification rules. The advantages of BMDP7M include the stepwise procedures, tests of equality of group means and the assessment of error rates by the jackknife method.

Potential limitations in the discriminant analysis procedure, DISCRIM of the SAS package include: a) omission of options to examine a reduced space using one of the stepwise procedures; and b) failure to test for the equality of group means. Advantages of the DISCRIM procedure in SAS include the option to formulate both the linear and the quadratic classification rules and the examination of classification results using the entire test space.

Choice of the statistical program will be contingent upon the researcher's access to each of the programs and the basic questions to be answered from the discriminant analysis. Perhaps a combination of the three programs may be best suited to answer the basic question.

REFERENCES

[1] Barr, Anthony J., Goodnight, James H., Sall, John P., and Helwig, Jane T. (1979), A User's Guide to SAS79, (Raleigh, N.C.: SAS Institute Inc.)

[2] Cooley, William W. and Lohnes, Paul R. (1971), Multivariate Data Analysis. (New York: John Wiley & Sons, Inc.)

[3] Dixon, W.J., ed. (1975), BMDP: Biomedical Computer Programs (Berkeley and Los Angeles, Calif: University of California Press).

[4] Eisenbeis, R.A. (1969), Discriminant Analysis and Classification Procedures, (Washington, D.C.: Federal Deposit Insurance Corporation) Working Paper 69-12

[5] Eisenbeis, Robert A. and Avery, Robert B. (1972), Discriminant Analysis and Classification Procedures, (Lexington, Mass: D.C.

[6] Fisher, R.A. (1936). "The Use of Multiple Measurement in Taxonomic Problems", Annals of Eugenics, 7, 179-188.

[7] Lachenbruch, Peter A. and Mickey, M. Ray. (1968), "Estimation of Error Rates in Discriminant Analysis", Technometrics, Vol 10, No. 1, 1-11.

[8] Lachenbruch, Peter A., (1975), Discriminant Analysis, (New York: Hafner Press).

[9] Lachenbruch, P.A. and Goldstein, M. (1979), "Discriminant Analysis", Biometrics 35, 69-85.

Table 1

Comparisons of Group Means, Standard Deviations, And Standard Errors of the Mean

Variable	Computer Program	Iris Setosa (n=50)		Iris Versicolor (n=50)		Iris Virginica (n=50)		All Groups (n=150)	
		$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S
SEPAL Length	SAS	5.0060	0.3525	5.9360	0.5162	6.6080	0.6262	N.R.	N.R.
	BMDP	5.0060	0.3525	5.9360	0.5162	6.6080	0.6262	5.8500	0.5108
	MULDIS	5.0060	0.3525	5.9360	0.5162	6.5879	0.6359	5.8433	0.8281
SEPAL Width	SAS	3.4280	0.3791	2.7700	0.3138	2.9680	0.3223	N.R.	N.R.
	BMDP	3.4280	0.3791	2.7700	0.3138	2.9680	0.3223	3.0553	0.3396
	MULDIS	3.4280	0.3791	2.7699	0.3138	2.9739	0.3255	3.0573	0.4359
PETAL Length	SAS	1.4620	0.1737	4.2600	0.4699	5.5600	0.5481	N.R.	N.R.
	BMDP	1.4620	0.1737	4.2600	0.4699	5.5600	0.5481	3.7607	0.4287
	MULDIS	1.4620	0.1737	4.2599	0.4699	5.5519	0.5519	3.7580	1.7653
PETAL Width	SAS	0.2460	0.1054	1.3260	0.1978	2.0120	0.2670	N.R.	N.R.
	BMDP	0.2460	0.1054	1.3260	0.1978	2.0120	0.2670	1.1947	0.2012
	MULDIS	0.2460	0.1054	1.3259	0.1978	2.0259	0.2747	1.1995	0.7622

Results of the tests of equality of group means:

Computer Program	Wilk's λ	F-statistic
SAS	N.R.	N.R.
BMDP	0.0237	198.083**
MULDIS	0.0234	199.139**

**Significant at the.01 level of significance.

Table 2

Pooled and Individual Variance-Covariance Matrices

		Computer Package							
		SAS				MULDIS			
Group	Variable	SL	SW	PL	PW	SL	SW	PL	PW
S_1 Iris Setosa	SL	0.1242				0.1242			
	SW	0.0992	0.1437			0.0992	0.1437		
	PL	0.0164	0.1170	0.0302		0.0164	0.0117	0.0302	
	PW	0.0103	0.0093	0.0061	0.0111	0.0103	0.0093	0.0061	0.0111
S_2 Iris Versicolor	SL	0.2664				0.2664			
	SW	0.0852	0.0985			0.0852	0.0985		
	PL	0.1829	0.0827	0.2208		0.1829	0.0827	0.2208	
	PW	0.0558	0.0412	0.0731	0.0391	0.0558	0.0412	0.0731	0.0391
S_3 Iris Virginia	SL	0.3922				0.4044			
	SW	0.0931	0.1039			0.0938	0.1040		
	PL	0.2956	0.0724	0.3004		0.3033	0.0714	0.3046	
	PW	0.0572	0.0484	0.0544	0.0713	0.0491	0.0476	0.0488	0.0754
S_p Pooled Over-all Groups	SL	0.2610				0.2650			
	SW	0.0925	0.1153			0.0927	0.1154		
	PL	0.1650	0.0556	0.1838		0.1675	0.0552	0.1852	
	PW	0.0411	0.0330	0.0445	0.0405	0.0384	0.0327	0.0427	0.0419

Results of the tests of the equality of the dispersion matrices

Computer Package	Box's M	F-statistic
SAS	137.345**	N.R.
MULDIS	N.R.	7.045**

**Significant at the .01 level of significance.

BMDP07M does not perform this test.

TABLE 3

CLASSIFICATION RESULTS
LINEAR AND QUADRATIC RULES

Actual Groups	n	Linear Classification Results from: MULDIS, DISCRIM and BMDP7M			Quadratic Classification Results from: MULDIS and DISCRIM		
		Iris Setosa	Iris Versicolor	Iris Virginica	Iris Setosa	Iris Versicolor	Iris Virginica
Iris Setosa	50	50	0	0	50	0	0
Iris Versicolor	50	0	48	2	0	48	2
Iris Virginica	50	0	1	49	0	1	49
Total Sample	150	50	49	51	50	49	51

The columns indicate the predicted group membership; the number of cases per group, n, denotes the actual group membership.

TABLE 4

PERCENT OF CASES ASSIGNED TO EACH GROUP
LINEAR AND QUADRATIC CLASSIFICATION RULES

Actual Groups	n	Linear Classification Rules from: MULDIS, DISCRIM and BMDP7M			Quadratic Classification Rules from: MULDIS and DISCRIM		
		Iris Setosa	Iris Versicolor	Iris Virginica	Iris Setosa	Iris Versicolor	Iris Virginica
Iris Setosa	50	100.00	0.00	0.00	100.0	0.00	0.00
Iris Versicolor	50	0.00	96.00	4.00	0.00	96.00	4.00
Iris Virginica	50	0.00	2.00	98.00	0.00	2.00	98.00
Total Sample	150	33.33	32.67	34.00	33.33	32.67	34.00

98% of all groups were correctly classified.

The columns indicate the percent of cases classified or assigned to each group.

TABLE 5

Rankings of the Discriminating Variables

Ranking Method	Results from BMDP7M and MULDIS			
	1	2	3	4
Univariate F-ratio	Petal Length	Petal Width	Sepal Length	Sepal Width
Stepwise Methods	Petal Length	Sepal Width	Petal Width	Sepal Length
*Complete Selection (Best 3-Subset)	Sepal Width	Petal Length	Petal Width	--

1 → most significant F-value or most important predictor

4 → least significant F-value or lease important predictor

* The complete selection method is only available in the MULDIS program.

Table 6

Summary of Three Options for Discriminant Analysis

Specific Areas	DISCRIM	BMDP7M	MULDIS
Presentation of Simple Statistics	X	X	X
Tests of Equality of Means	-	X	X
Tests of Equality of Dispersion matrices	X	-	X
Prior Probabilities	X	X	X
Linear Classification Rules	X	X	X
Quadratic Classification Rules	X	-	X
Resubstitution Method	X	X	X
Holdout Method	X	X	X
Jackknife Method	-	X	X
Univariate F-ratio	-	X	X
Stepwise Method	-	X	X
Complete Selection	-	-	X
CPU time used in seconds	1.02	2.46	3.26

X indicates that the option is available for the program denoted in the heading.

- indicates that the option is not available for the program denoted in the heading.