DOCUMENT RESUME

ED 328 588 TM 016 100

AUTHOR Prosser, Barbara
TITLE Use of the U-Method To Establish the External Validity of Discriminant Analysis Results.
PUB DATE Jan 91
NOTE 29p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 25-27, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; *Discriminant Analysis; Monte Carlo Methods; *Prediction; *Validity
IDENTIFIERS Empirical Methods; *External Validity; Jackknifing Technique; *U Method; Validation Verification and Testing Techniques

ABSTRACT

Accurate classification in discriminant analysis and the value of prediction are discussed, with emphasis on the uses and key aspects of prediction. A brief history of discriminant analysis is provided. C. J. Huberty's discussion of four aspects of discriminant analysis (separation, discrimination, estimation, and classification) is cited. Predictive accuracy results when an investigator understands certain main rules and validation methods. The generalizability of discriminant analysis results, and external versus internal analysis are considered. Four traditional types of external validation methods are discussed: (1) the empirical method; (2) the holdout method; (3) Monte Carlo method; and (4) the random assignment method. Two non-traditional methods, the jackknife and the U-method, are also reviewed. Focus is on the U-method, which is illustrated using a hypothetical data set of 64 cases for which the actual classifications (four groups) are known. Four classification tables are provided, which demonstrate concepts such as hit rates, leave-one-out, and predictor ordering. A summary is presented to improve the interpretation of discriminant analysis results and multivariate procedures in general. A 15-item list of references is included, and selected file commands are outlined. (SLD)

USE OF THE U-METHOD TO ESTABLISH THE EXTERNAL VALIDITY OF DISCRIMINANT ANALYSIS RESULTS

Barbara Prosser

University of New Orleans

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 24, 1991.

ABSTRACT

Accurate classification in discriminant analysis is vitally important. The author discusses the value of prediction, with emphasis on its uses and key aspects, and provides a brief history of discriminant analysis. Predictive accuracy results when an investigator understands certain main rules and validation methods. Four traditional types of external validation methods, as well as two nontraditional ones, receive attention. Of the nontraditional kind, the U-method is the main focus of this paper. A hypothetical data set consisting of 64 cases and for whom the actual classifications (four groups) are known illustrates the U-method. Classification tables show concepts like "hit" rates, "leave-one-out," and predictor ordering. The author presents a summary to improve the interpretation of discriminant analysis results and multivariate procedures in general.

Use of the U-Method to Establish the External Validity of Discriminant Analysis Results

History and Purposes of Discriminant Analysis

When Fisher developed discriminant analysis in 1936, its basic purpose was to provide a way of classifying an item into one of two categories. Rao extended the number of categories to more than two in 1948. However, it was not until the 1960's, partly as a consequence of the development of the electronic computer, that the usefulness and versatility of discriminant analysis increased to a great extent (Huberty, 1975).

Discriminant analysis is a powerful technique for the multivariate study of group differences. It affords a means of examining the extent to which multiple predictor variables relate to group membership (Betz, p. 393). One major problem for novice researchers, however, is the variety of terms used to describe discriminant analysis. "The meaning of discriminant analysis varies somewhat from textbook author to textbook author, from computer programmer to computer programmer, and from statistician to statistician" (Huberty & Barton, 1989, p. 158).

Despite the existence of many meanings, there is common ground. There are two basic characteristics: 1) a number of multiple response variables and 2) multiple

groups of objects or subjects. Therefore, in any context of discriminant analysis, there are two sets of variables. One set consists of a collection of response variables; the other, one or more groupings of nominally scaled variables (Huberty & Barton, 1989). Hoel and Peterson (1949) point out that there are essentially two problems to discriminant analysis: description and prediction. The first problem is to describe population differences, since it would be futile to try a classification if the populations do not differ. The second problem is to find an efficient classification method with which to predict the proper populations for individuals.

Huberty (1975, p. 545) says, "Discriminant analysis as a general research technique can be very useful in the investigation of various aspects of a multivariate research problem." He attempts to counter the confusion surrounding key terms by delineating four aspects of discriminant analysis: 1) separation, 2) discrimination, 3) estimation, and 4) classification.

Separation refers to defining intergroup significant differences of group centroids (mean vectors). Group centroids of each group studied are compared to the discriminant scores to determine

probabilities of group membership. The scores come from discriminant weights. In discriminant analysis, the weights yielded are those that maximally differentiate or separate the groups. Discriminant weights, when multiplied by an individual's standard scores on the variables, yield discriminant scores. When multiplied by the score mean for a group, the discriminant weights yield the group centroid. The centroid to which the individual's score is closest is the group to which he or she is predicted to belong (Betz, 1987). Statistical significance testing is done via the Wilks' Lambda Test.

Discrimination further studies group separation in regard to dimensions and to the discrimination of variable contributions to separation. Some authors equate this aspect with classification, but Huberty (1975) distinguishes between them. This is the stage in which there is interpretation of the linear discriminant function, the equation by which group membership can be predicted (Betz,1987). There is with this aspect a similarity to the procedural "dimensioning" found in factor analysis (factor weights, factor scores).

The third aspect is estimation, that is, obtaining estimates of intergroup differences (distances between

centroids) and the strength (degree) of the relationship between variables and group membership. Estimation is included by Huberty (1975) as an additional aspect for the purpose of underscoring supplementary methods of interpretation of the results of a discriminant analysis.

Classification, the final aspect presented by Huberty (1975), is concerned with developing rules for assigning individuals to groups, which are predetermined and mutually exclusive populations. Its emphasis is prediction rather than description.

Betz (1987), in explaining the aspects and procedures of discriminant analysis, places emphasis on its uniqueness. It enables the investigator to make a prediction of group membership for each individual in the sample. Although it is related to a whole class of methods--including multiple regression and MANOVA--that are based on the general linear model, discriminant analysis addresses distinct research questions, is appropriate for certain types and numbers of variables, and has its own special uses.

Generalizability

Researchers employing discriminant analysis have concern for the validity of the findings in terms of the

general population of interest. There is always the possibility, as with any statistical technique, that results may not be generalizable to a larger population. There is heightened risk in cases in which the sample size is small or when there is a question about the representativeness of the sample (Daniel, 1989). Betz (1987) cautions that if the discriminant function serves for predictive purposes in new populations, the researcher needs to consider the tendency of discriminant analysis to inflate, that is, to overestimate the accuracy of classification. The apparent "hit" rates (correct predictions) are likely to be less than the true "hit" rates. A "hit" results when a case coming from a particular group is assigned to that same group by using the developed prediction rule (Huberty & Barton, 1989).

The researcher's first step is to find the best linear discriminant functions which will discriminate optimally between the groups and maximize the probability of accurate classification. There are two assumptions that meet with wide agreement: 1) each group must come from a population that has a multivariate normal distribution, and 2) the population covariance matrices must be equal. However, the basis for the first

assumption is testing the statistical significance of the resulting discriminant functions (variables) for the purpose of discarding those that do not contribute to group separation. If the researcher uses all functions and variables in the analysis, then no test of statistical significance will be used. Thus, the first assumption need not be met (Jones, 1989). Klecka (1980) suggests a third assumption that no discriminating variable is a linear combination of other discriminating variables or is perfectly correlated with any other discriminating variable.

External Versus Internal Analysis

Problems of generalizability due to unstable results appear, then, to emphasize the need for replication of studies and careful cross-validation of findings (Huberty, 1975). Cross-validation represents external analysis, preferable because in it the classification rule is derived from one set of units and then employed to classify another set of units. This approach, exemplifying the traditional idea of cross-validation, typically gives results in the form of a classification matrix (Huberty, Wisenbaker, & Smith, 1987).

In contrast, internal analysis--classifying units whose own data are used both to derive and to validate

the prediction statistics--causes biased hit rate estimates. The degree of bias in an internal analysis (referred by some as the "empirical" method) is, not surprisingly, a function of the number of variables, the number of units, and the degree of group overlap (Huberty, 1984). In practice, however, it is not uncommon to see in the applied literature results of a PDA (predictive discriminant analysis) based on internal analysis. This means the classification rule is built on the very cases used in obtaining the classification table. Some feel that internal analysis may be acceptable providing the number of cases is large. One rule of thumb for "large" is a data set in which the smallest group size is five times the number of predictor variables (Huberty & Barton, 1989).

Traditional Methods of External Analysis

In addition to the empirical method, Daniel (1989) describes three other traditional approaches for assessing the stability of discriminant function coefficients. There are the "holdout" method, the "Monte Carlo" method, and the "random assignment" method. All these have built-in weaknesses and tend to produce biased results.

The "holdout" method is well known and may be called

by other names: "split half," "cross-validation," or "invariance." "For large samples, the holdout method yields fairly good hit estimates" (Huberty, 1984, p. 165). Using this method, the researcher randomly splits the sample into two equal or approximately equal subsamples. One subsample is then used to develop estimates of the discriminant coefficients, and then these are applied to the other subsample for purposes of classification (Crask & Perrault, 1977, p. 61). The problem with this method is that in small-sample research, dividing the sample into smaller subgroups makes the derived coefficients even less stable.

The "Monte Carlo" method involves the researcher's random generation of synthetic data from which discriminant functions are derived with degrees of freedom equal to the original data. Then, these data can be utilized to validate the predictive discriminant function coefficients derived from the original data set. This method is useful when the predictor variables are independent of one another--i. e., when uncorrelated factor scores are used as predictors (Daniel, 1989; Crask & Perrault, 1977). However, the predictors tend to be correlated in most cases involving multiple predictors. The main "problem" with the Monte Carlo

method is it does require special computer programming to reproduce the variance/covariance structure of the original data using randomly-generated data.

The "random assignment" method is a procedure in which discriminant functions are derived from repeated random assignment of real cases from the original sample to groups. Once the researcher obtains several sets of discriminant functions using the randomly assigned cases, these classification results can be compared to those of the original sample. The advantage to this method is clear. Because it uses actual rather then synthetic data, it holds more appeal for preserving the true interrelationships among the variables. Despite the advantage, though, this method is questionable as an absolute performance assessment because of its reliance on random or chance classification (Daniel, 1989; Crask & Perrault, 1977).

Nontraditional Methods of External Analysis

Both the "jackknife statistic" and the "U-method" represent efforts to remedy the shortcomings of the traditional methods (Daniel, 1989). Traditional methods, as noted earlier, tend to produce biased estimates of the stability of the findings. Assessments of the generalizability of discriminant analysis tend to be

inflated. Crask and Perrault (1977) have demonstrated that the jackknife and the U-method produce more conservative and less biased estimates of true population traits.

The two methods are similar to each other, and some authors like Betz (1987) group the jackknife with this method under the heading of "cross-validation" methods. However, there are differences. Daniel (1989) points out that the jackknife statistic offers a procedure for assessing the stability of discriminant function coefficients while the U-method estimates error rates in the classification of cases. Crask & Perreault (1977) demonstrate that the two methods may be used separately or together simultaneously, depending on the aims of the researcher. However, in regard to the simultaneous use of the two methods, advantages need to be weighed against the large number of computer runs needed, as well as time and expense requirements. The present study illustrates the use of the U-method.

An Overview of the U-Method

Lachenbruch first proposed the U-method, also called the "leave-one-out" (L-O-O) procedure,or the "L-method," in 1967 (Huberty, 1984; Glick, 1978). With this one,

a unit is removed and the classification statistic is formulated on the remaining n-1 units. Then the removed unit is classified. The researcher carries out the steps N times to ascertain a hit rate estimate which is based on classifications of the deleted units (Huberty, 1984). Its almost unbiased estimate of misclassifications in the group can be obtained for each group by taking the total number of the misclassifications in the group and dividing it by the total number of cases in the group (Crask & Perreault, 1977).

The U-method has several advantages. With it, any given observation has no effect on the coefficients of the function used to classify that observation. The analysis is an external one. The U-method lends itself more confidently to smaller sample sizes than does the popular "holdout" method. Although lacking some of the bias-reducing properties of the jackknife, the U-method is similar in that respect in that it, too, involves the efficient partitioning of the sample (Crask & Perreault, 1977). It makes use of all the available data without serious bias in the estimation of error rates (Dillon & Goldstein, 1984). Furthermore, its results are easily obtained, and it offers a fair degree of robustness to distribution violations (Huberty, Wisenbaker, & Smith,

1987; Lachenbruch, 1968).

Description of Data Set

The fictitious data set in the present study consists of two predictor and four criterion variables. The predictor variables are X and Y. The criterion variables are Groups 1, 2, 3, and 4. The full data set, with 64 cases (16 per group), is listed in Table 1.

INSERT TABLE 1 ABOUT HERE

Analysis of the Data

Data were analyzed using three different statistical methods: 1) regular predictive discriminant analysis, 2) U-method, and 3) deletion of predictor variables. The first analysis represents internal classification and utilizes the two predictor variables and all four criterion variables. The second and the third represent external classification. Appendix A shows the SPSSx command file.

The three analyses can be compared to evaluate the predictive accuracy of the different classification methods. A discussion of each analysis follows.

Internal Classification--Regular Predictive Discriminant Analysis

Data were analyzed first using the internal classification method of regular predictive discriminant analysis. The number of hits relative to chance alone, the "prior probability," was first assessed at .25. Further, an examination of the data revealed no problem with outlier scores. Separate hit rates for each group were obtained, as well as an overall percentage of correct classifications. Classification results show these in Table 2.

INSERT TABLE 2 ABOUT HERE

Hits for Group 1 were 9/16 or 56.3%; Group 2, 5/16 or 31.3%; Group 3, 9/16 or 56.3%; and Group 4, 11/16 or 68.8%. Overall predictive accuracy was 53.13%.

External Classification--U-Method

The second analysis of the data employed external classification, the U-method. First, the data set was divided into eight subsets of eight cases each. Then, as previously outlined, the procedure for removing one subset at a time was begun. At the removal of the subset, the classification statistic was formulated on the remaining seven subsets. Then, the deleted subset was classified. These steps were carried out eight

times and a hit rate based on the classifications of the deleted units was obtained. Table 3 provides a summary of this method of analysis, which yielded an overall classification accuracy rate of 50%.

INSERT TABLE 3 ABOUT HERE

External Classification--Predictor Variable Deletion

The final method of analysis used was predictor variable deletion, the assessment of the relative contribution of each variable to estimate classification accuracy. The attractiveness of this method is readily understood. Perhaps an increase in accuracy will develop if one predictor is deleted. Alternatively, one predictor may show a higher hit rate than the other. (In a study with three or more predictors, one may show a higher hit rate than all the others combined.) By using this method, the researcher can determine an order of importance in terms of predictive accuracy. Table 4 gives a summary of the results of this third type of analysis.

INSERT TABLE 4 ABOUT HERE

The predictor-deletion method resulted in an overall

prediction accuracy rate of 36.72%. Of the two predictor variables, the one contributing more to the accuracy rate is Y: 37.5%. X follows with a rate of 35.94%.

Discussion

Discriminant analysis is a versatile, useful research technique for multivariate statistical problems. Its aspect of prediction makes it uniquely important, but the researcher must consider the tendency of discriminant analysis to overestimate the accuracy of classifications. Establishing validity of findings through cross-validation is an important concern if the researcher is to assure generalizability and to give proper emphasis to the importance of replication of studies.

Cross-validation, representing external analysis, is preferable because in terms of generalization, it gives more accurate classification results. Of the methods of cross-validation mentioned in this paper, nontraditional types are favored. Specifically, the nontraditional approach called the U-method offers several advantages and is gaining wider acceptance in applied research.

The present study offers results from three analyses of an artificial data set. Findings, easily obtained, support the bias-reduction properties of the U-method

in particular and external classification in general. Researchers would do well to investigate further applications of this predictive aspect of discriminant analysis.

REFERENCES

Betz, N. E. (1987). Use of discriminant analysis in counseling psychology research. Journal of Counseling Psychology, 34(4), 393-403.

Crask, M. R., & Perreault, Jr., W. D. (1977,February). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Daniel, L. G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX.

Dillon, W., & Goldstein, M. (1984). Multivariate Analysis. New York: John Wiley & Sons.

Glick, N. (1978). Additive estimators for probabilities of correct classification. Pattern Recognition, 10, 211-222.

Hoel, P. G., & Peterson, R. P. (1949). A solution to the problem of optimal classification. Annals of Mathematical Statistics, 20, 433-438.

Huberty, C. J. (1975). Discriminant analysis. Review of Educational Research, 45(4), 543-598.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis.

Psychological Bulletin, 95, 156-171.

Huberty, C. J., Wisenbaker, J. M., Smith, J. D., & Smith, J. C. (1986, October). Using categorical variables in discriminant analysis. Multivariate Behavioral Research, 21(4), 479-496.

Huberty, C. J., Wisenbaker, J. M., Smith, J. D., & Smith, J. C. (1987, July). Assessing predictive accuracy in discriminant analysis. Multivariate Behavioral Research, 22(5), 327-329.

Huberty, C. J., & Barton, R. M. (1989, October). An introduction to discriminant analysis. Measurement & Evaluation in Counseling and Development, 22, 158-168.

Huberty, C. J., & Wisenbaker, J. M. (in press). Discriminant analysis: potential improvement in typical practice. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2). Greenwich, CT: JAI Press.

Jones, G. (1989, January). Some examples of invariance procedures in discriminant analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX.

Klecka, W. R. (1980). Discriminant analysis. Beverly

Hills: Sage Publications.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence levels for the probability of misclassification in discriminant analysis. Biometrics, 23, 639-645.

Table 1

Data Listing

Case	Group	X	Y	Subgroup
1	1	4	2	5
2	1	5	3	8
3	1	4	4	2
4	1	4	5	3
5	1	3	4	4
6	1	6	5	6
7	1	5	6	7
8	1	7	5	2
9	1	6	6	1
10	1	8	6	8
11	1	7	6	1
12	1	9	7	5
13	1	8	7	4
14	1	8	8	3
15	1	9	8	7
16	1	9	9	6
17	2	1	2	8
18	2	3	3	4
19	2	3	5	3
20	2	3	5	6
21	2	2	5	5
22	2	4	6	4
23	2	4	5	2
24	2	5	6	5
25	2	6	6	6
26	2	6	6	1
27	2	6	7	7
28	2	7	7	8
29	2	7	7	2
30	2	8	9	3
31	2	8	9	7
32	2	9	9	1
33	3	4	1	8
34	3	4	2	6
35	3	3	2	3
36	3	2	4	5
37	3	5	3	2
38	3	7	4	1
39	3	4	5	7
40	3	5	4	5
41	3	7	5	8

(continued next page)

Table 1 (continued)

42	3	9	5	6
43	3	6	5	4
44	3	5	6	1
45	3	7	6	7
46	3	9	7	3
47	3	8	6	5
48	3	8	5	2
49	4	1	7	4
50	4	1	2	3
51	4	1	1	2
52	4	2	2	8
53	4	2	2	3
54	4	2	3	1
55	4	3	2	7
56	4	3	3	4
57	4	3	4	7
58	4	4	5	6
59	4	4	4	5
60	4	4	5	4
61	4	4	6	2
62	4	5	6	1
63	4	5	7	8
64	4	5	7	6

Table 2

Internal Analysis

CLASSIFICATION RESULTS

ACTUAL GROUP		NO. OF CASES	PREDICTED GROUP MEMBERSHIP 1	2	3	4
Group	1	16	9	1	3	3
			56.3%	6.3%	18.8%	18.8%
Group	2	16	5	5	0	6
			31.3%	31.3%	0.0%	37.5%
Group	3	16	4	1	9	2
			25.0%	6.3%	56.3%	12.5%
Group	4	16	0	4	1	11
			0.0%	25.0%	6.3%	48.8%

PERCENT OF "GROUPED" CASES CORRECTLY CLASSIFIED: 53.13%

Table 3

External Analysis: The U-Method

Number of Hits with Eight Subsets of Eight Cases Each

Subset Deleted	No. of Hits Group 1	Group 2	Group 3	Group 4	%
1	2	0	1	1	50.00
2	0	0	2	1	37.50
3	1	1	1	2	62.50
4	1	1	0	2	50.00
5	1	1	2	1	62.50
6	1	0	2	1	50.00
7	1	2	0	1	50.00
8	0	0	2	1	37.50

OVERALL AVERAGE OF PERCENTAGES OF CASES CORRECTLY CLASSIFIED: 50.00%

Table 4

Number of Hits Relative to Each Predictor Variable

Predictor Variable Deleted	Group 1	Group 2	Group 3	Group 4	%
X	8	1	1	13	35.94
Y	0	10	5	9	37.50

OVERALL AVERAGE OF PERCENTAGE OF CASES

CORRECTLY CLASSIFIED: 36.72%

Appendix A

File Commands

```
TITLE "DISCRIMINANT ANALYSIS EXAMPLE--BARBARA PROSSER"
FILE HANDLE DISCRIM/NAME='DIS2CRMT.DAT'
DATA LIST FILE=DISCRIM
  /CASE 1-2 GROUP 7 X 12 Y 17 SUBGROUP 20
SORT CASES BY GROUP
LIST VAR=ALL/CASES=900
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /STATISTICS=MEAN STDEV CORR UNIVF RAW TABLE
 /PLOT=ALL
TEMPORARY
IF (SUBGROUP GT 1) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
TEMPORARY
IF (SUBGROUP LT 2 OR SUBGROUP GT 2) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
TEMPORARY
IF (SUBGROUP LT 3 OR SUBGROUP GT 3) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
TEMPORARY
IF (SUBGROUP LT 4 OR SUBGROUP GT 4) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
TEMPORARY
IF (SUBGROUP LT 5 OR SUBGROUP GT 5) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
```

(continued next page)

Appendix A (continued)

```
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
TEMPORARY
IF (SUBGROUP LT 6 OR SUBGROUP GT 6) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
TEMPORARY
IF (SUBGROUP LT 7 OR SUBGROUP GT 7) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
TEMPORARY
IF (SUBGROUP LT 8 OR SUBGROUP GT 8) TEMP=1
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /SELECT=TEMP(1)
 /STATISTICS=TABLE
 /PLOT=CASES
DISCRIM GROUPS=GROUP(1,4)
 /VAR=X Y
 /ANALYSIS=X
 /ANALYSIS=Y
 /STATISTICS=TABLE
 /PLOT=CASES
```