ED 451 211 TM 032 438

AUTHOR	Hwang, Dae-Yeop
TITLE	Issues in Predictive Discriminant Analysis: Using and Interpreting the Leave-One-Out Jackknife Method and the Improvement-Over-Change "I" Index Effect Size.
PUB DATE	2001-02-00
NOTE	24p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (24th, New Orleans, LA, February 1-3, 2001).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Classification; *Discriminant Analysis; *Effect Size
IDENTIFIERS	*Jackknifing Technique; *Predictive Discriminant Analysis

ABSTRACT

Prediction of group membership is the goal of predictive discriminant analysis (PDA) and the accuracy of group classification is the focus of PDA. The purpose of this paper is to provide an overview of how PDA works and how it can be used to answer a variety of research questions. The paper explains what PDA is and why it is important, and it presents an example of PDA that uses the leave-one-out method and interpretation of the "I" effect size proposed by C. Huberty and L. Lowman (2000). An illustrative data set is used to make the discussion concrete and accessible to applied researchers. (Contains 4 tables and 13 references.) (Author/SLD)

Running head: LOO Method and I Index

Issues in Predictive Discriminant Analysis: Using and Interpreting the Leave-One-Out Jackknife Method and the Improvement-Over-Chance I Index Effect Size

Dae-Yeop Hwang

University of North Texas

Paper presented at the annual meeting of the Southwest Educational Research Association (SERA), New Orleans, February 1-3, 2001

Abstract

Prediction of group membership is the goal of predictive discriminant analysis (PDA) and the accuracy of group classification is the focus of PDA. The purpose of the present paper is to provide an overview of how PDA works and can be used to answer a variety of research questions. Specifically, the paper will: a) explain what PDA is and why it is important, b) present an example of PDA that uses the leave-one-out method and interpretation of the I effect size as proposed by Huberty and Lowman (2000). An illustrative data set will be used to make the discussion concrete and accessible to applied researchers.

Introduction

Multivariate statistical analyses have become very popular in social science research (Emmons, Stallings, & Layne, 1990; Grimm & Yarnold, 1995). In their book devoted to understanding multivariate methods, Grimm and Yarnold (1995) noted that " the use of multivariate statistics has become commonplace. Indeed, it is difficult to find empirically based articles that do not use one or another multivariate analysis" (p. vii).

Many multivariate methods can handle a variety of research situations and different types of data. For example, predictive discriminant analysis (PDA) focuses on whether two or more continuous variables can predict group membership (categorical variable) (Buras, 1996; Huberty, 1994; Lancaster, 1999). Klecka (1980) presented an illustrative example where PDA could be employed to predict the outcome of a hostage situation (e.g., successful/not successful) given several relevant predictors (e.g., number of weapons, number of terrorists). Assuming analytic assumptions are met, PDA can essentially be used whenever multiple variables can be used to assess the probability of some discrete outcome. Of course, PDA is conceptually similar to logistic regression, although the two methods may obtain different results.

Important in understanding how PDA works is the leave-one-out (LOO) jackknife approach of classification (Huberty, 1994; Huberty & Lowman, 2000). The LOO method uses jackknife methodology to make these classifications. Appropriate interpretation of results is dependent on understanding how this methodology works. Furthermore, the accuracy of group classification is the focus of PDA and the classification results should be reported and interpreted. To facilitate interpretation, Huberty and Lowman (2000) have presented a PDA related effect size, called the I index. The I index describes the accuracy of group classification as an improvement-over-chance estimate, and as such, captures the accuracy of group classification in terms of group overlap. Use of effect sizes in result interpretation is consistent with current movement in the field and the recommendations presented by the APA Task Force on statistical Inference (Henson & Smith, 2000; Wilkinson & APA Task Force on Statistical Inference, 1999).

The purpose of the present paper is to provide an overview of how PDA works and can be used to answer a variety of research questions. Specifically, the paper will: a) explain what PDA is and why it is important, and b) present an example of PDA that uses the LOO method and interpretation of the I effect size as proposed by Huberty and Lowman (2000).

Brief Review of DDA and PDA

There are two aspects of discriminant analysis: PDA (predictive discriminant analysis) and DDA (descriptive discriminant analysis). PDA and DDA are different analyses for different purposes; one is for prediction of group membership and the other is for description of MANOVA results of grouping variable effects. PDA is useful in allocating new cases (observations) to previously defined groups. DDA is useful in identifying a structure consisting of certain linear combination of outcome variables. This procedure interprets the linear combination of variables with regard to group difference.

PDA focuses on deriving a rule (i.e., classification functions) that can be used to optimally assign a new case to different groups. DDA focuses on finding dimensions that can be used to separate distinct sets of cases. In one word, PDA is for classification and DDA is for separation. The role of the two variable sets is reversed in PDA and DDA.

In PDA, the p response variables are the predictors (the independent variables in the univariate case like multiple regression) and the grouping variables are criterion variables (the dependent variables in multiple regression). In DDA/ MANOVA, the response variables are the criterion (outcome) variables (the dependent variables in the univariate case like in ANOVA) and the grouping variables are the predictor variables (the independent variables in ANOVA).

General linear model is a unifying conceptual framework in all classical parametric methods. All classical parametric methods compute synthetic variables by multiplying the weights to the observed variables. The number of equations of weights differs across PDA and DDA. In DDA, the maximum number of linear discriminant functions (LDFs) is the number of groups minus one, or the number of discriminant variables, whichever is fewer. In PDA, the number of linear classification functions (LCFs) is the number of groups.

The only similarity between an LDF and an LCF is that they are both linear combinations of the response variables (DDA: criterion variables, PDA: predictor variables) (Huberty, 1984, p.158). The fundamental difference in deriving the two functions is that LDF coefficients are obtained using an eigenanalysis and LCF coefficients are obtained by multiplying a mean vector by the inverse of a pooled covariance matrix.

DDA is conducted following the results of main effect from a MANOVA. MANOVA examines the effects of grouping variables, separately or jointly. The MANOVA omnibus test results of statistical significance will be identical with those in DDA. DDA describes the effects of grouping variables (from MANOVA) in terms of the linear combination of the dependent (outcome) variables. The set of scores on the synthetic variable (the discriminant scores) separate various groups from each other and find dimensions along which groups differ. In DDA, LDFs provide for maximal group separation and correspond to a linear combination of the discriminating variables. There is a different set of coefficients that provide maximal group differences. The first (canonical) discriminant function will account for the greatest group differences (Klecka, 1980).

One goal of PDA is to set up a rule. A rule in PDA, termed classification rule, can take three different forms. A first form is the rule of a composite of the predictor variables; a second form is the rule of a probability of population membership (i.e., prior probability); a third form is the rule of a distance between two points (Huberty, 1994, p.40). There are four possible types of classification rules: (1) Linear/Equal prior probability rules, (2) linear/Unequal prior probability rules, (3) Quadratic/Equal prior probability rules, and (4) Quadratic/Unequal prior probability rules. Further discussions on these rules are available in Marcoulides & Hershberger (1997, p.107). Researchers seek a rule that will yield relatively high prediction accuracy.

Tabachnick & Fidell (1996) described a lot of practical issues such as unequal sample sizes and missing data, multivariate normality, outliers, homogeneity of variance-covariance matrices, linearity, and multicollinearity and singularity. If classification is the primary goal, most of the requirements are relaxed except for outliers and homogeneity of variance and covariance matrices (p. 512). An extensive discussion on practical issues is available in Tabachnick & Fidell.

This study focuses on PDA and its application. Specifically, this paper stresses on the hit rate estimation and its assessment using a graphical representation of the group overlap.

Separation and Classification

The discriminant function differs from the classification function (Hair, Anderson, Tatham, & Black, 1999). The distinction between classification and separation becomes blurred (Johnson and Wichern, 1992, p.521). Klecka (1980) suggested that the classification procedures could use either the discriminating variables by themselves or the canonical discriminant functions stating

> In the first instance, one is not performing a 'discriminant analysis' at all. This activity merely uses the theory of maximum group differences to derive classification functions....When the canonical discriminant functions are derived first and classification is based upon them, we can perform a more thorough analysis (Klecka, 1980, p.42).

Thompson (1995) basically disagrees that PDA hit rate can be a possible aid to evaluating DDA effect (p.344). He (1998) also states that "Normally only LCFs are used for classification purposes, even though SPSS incorrectly uses LDF scores for this purposes" (Huberty & Loman, 1997). Huberty (1984) points out that

> Some behavioral researchers have expressed apparent confusion in their writings when they discuss the use of LDFs in classification. Classification decisions can be based directly on LDF scores only in a two-group situation (p.159).

Huberty and Lowman (2000) provide a graphical representation of the group membership prediction using the group-overlap concept. The concept of overlapping is closely related to the concept of separation in social science (Yitzhaki, 1994). Separation may be an important factor that determines the ability to prediction. Distribution i overlaps distribution j if the observations of group j are encircled by the observations of distribution i. The quantification of the term "overlapping" is the major concept of this section. First, this study discusses the concept of overlapping in the DDA context and, next, in the PDA context.

In DDA context, for each case, the predicted value of the discriminant function is calculated by multiplying each discriminating variable by its weights and adds these products. This predicted value becomes the discriminant Z scores in standardized term. The group mean (or centroid) is calculated by averaging the discriminant Z scores for all the cases (observations) within a particular group.

We can compute the distance between the group centroids by comparing the distributions of the discriminant Z scores for the groups. It becomes the test for the statistical significance of the discriminant function. If the overlap in the distributions of discriminant scores for a function is small, the discriminant function separates the groups well. On the other hand, if the overlap in the distribution is large, the discriminant function does not separate the groups. When two groups exist, there are two centroids. When three groups exist, there are three centroids; and so forth. If there are more than two groups, more than one discriminant function will exist.

When two groups exist, each case will have a discriminant Z score from a discriminant function. When three groups exist, each case will have a discriminant Z score from the two discriminant functions. In this case, each discriminant function represents a dimension and two discriminant functions correspond to two dimensions. This is a major difference between multivariate discriminant analysis (MDA) and univariate analysis like multiple regressions; MDA creates multiple variates.

Fisher (1936) developed the linear classification statistic. His idea was to transform the multivariate observations to univariate observations. He suggested that classification should be based on a linear combination of the discriminant variables and that this linear combination of the multivariate observation provides a univariate observation. He proposed selecting the linear combination of the multivariate observation to achieve maximum separation of the univariate sample means.

The linear combination of discriminating variables (multivariate observations) on a two axis becomes single value on a single axis (the Z axis). A score for each case (or observation) on each group's classification function is calculated and the case is assigned to the group by using the highest score. This differs from the calculation of the discriminant Z score, which is calculated for each discriminant function. Fisher's linear function maximally separates the two groups by maximizing the univariate "between" sample variability relative to the "within" sample variability.

Group difference can be compared with the group centroids, which is the average discriminant Z scores for all observations in a group.

The separation of these two sets of univariate y's is assessed in terms of the difference between $\bar{y}_1$ and $\bar{y}_2$ expressed in standard deviation units.

$$\text{Separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \text{ where } s_y^2 = \frac{\sum_{j=1}^{n_1}\left(y_{1j} - \bar{y}_1\right)^2 + \sum_{j=1}^{n_2}\left(y_{2j} - \bar{y}_2\right)^2}{n_1 + n_2 - 2}$$

The linear combination, $y = \hat{l}'x$, maximize the ratio. The maximum of the ratio is $D^2 = \left(\bar{x}_1 - \bar{x}_2\right)' s_{pooled}^{-1}\left(\bar{x}_1 - \bar{x}_2\right)$. An example of a linear combination is $y = \hat{l}'x = \left(\bar{x}_1 - \bar{x}_2\right)' s_{pooled}^{-1} x = 37.61x_1 - 28.92x_2$. The above Fisher's linear function maximally separates the two populations. The maximal separation is:

$$D^2 = (\bar{x}_1 - \bar{x}_2)' s_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2) \quad = [.2418 \quad -.0652] \begin{bmatrix} 131.58 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} .2418 \\ -.0652 \end{bmatrix}$$

$$= 10.98$$

Fisher's solution to the separation problem can be used to classify new observation.

Classify x_0 to π_1 if

$$y_0 = (\bar{x}_1 - \bar{x}_2)' s_{pooled}^{-1} x$$

$$\geq \hat{m} = \tfrac{1}{2} (\bar{x}_1 - \bar{x}_2)' s_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2) \quad \text{or} \quad y_0 - \hat{m} \geq 0$$

Classify x_0 to π_2 if

$$y_0 < \hat{m} \quad \text{or} \quad y_0 - \hat{m} < 0$$

This classification rule says that if the linear composite score of the new observation is closer to the mean composite score for population 1, classify this observation to group 1; otherwise, classify this observation into group 2. In the PDA context, the classification scores consist of sets of scores on the synthetic variables of the predictor variables. The classification rule assumes both equal prior probabilities (equal proportions) of the two populations and equal cost of misclassification for the two populations. Extensive classification mathematics is available in Johnson and Wichern (1992).

The distinction between separation and classification is not very clear. Good classification may depend upon the separation of the population. However, Johnson and Wichern (1992) commented, "significant separation does not necessarily imply good classification" (p.525). They mentioned that a classification rule could be developed independently of any test of separation. Separation may be a necessary condition but not a sufficient condition for classification. Even if the separation is not significant, it is hard to obtain a good classification rule (Johnson and Wichern, 1988, p.525).

Assessment of Classification Accuracy

In PDA, the "hit rate" is the goal of the analysis, but the weights are generally irrelevant when we interpret the result. In DDA, however, the weights and the structure of the synthetic variables (i.e., linear combination of outcome variables) are very important to interpretation, while the concept of hit rate becomes irrelevant. The basic question of DDA pertains to group separation of group differences with respect to the outcome variables (multiple response variables). The focus of DDA is to interpret the linear combination of outcome variables that are associated with group differences Thompson, 1998).

The predictive accuracy of the classification function is measured by the hit rate, which is obtained from the classification matrix. The hit rate can be considered as an assessment of group overlap, which indicates the proportion of correctly classified cases. Percent distribution overlap may be equivalent to proportions of correctly classified. The hit rate is analogous to R^2 in the multiple regression. The hit ratio shows how well the classification function classified the cases; the R^2 indicates how much variance the regression equation explained.

The highlight in PDA is estimate the hit rate and effectiveness of classification rules. The leave-one-out (LOO) method, which provides unbiased hit-rate estimates, uses jackknife methodology to make classification. An internal analysis can be expected to be positively biased because both the samples classified and the sample considered in rule formulation are the same. However, using an external analysis, the classification rule is determined from one set of samples and then used to classify another set of sample. The bias correction can be achieved through leave-one-out (LOO) method (Huberty, 1994), which is often known as jackknifing". The LOO method follows two steps: (1) One unit is deleted and linear classification functions (LCFs) are determined on the remaining N-1 units and (2) These LCFs are used to classify the deleted unit into one of the k criterion groups. For each classification, it may be considered that a training sample of size N-1 and a test sample of size 1 are being used. This process is carried out N times and the proportions of deleted units correctly classified are used as hit-rate estimators. The classification function is fitted to repeatedly drawn samples of the original sample.

An assessment tool for the accuracy of classification is the I index, an improvement-over-chance estimate. A current study develops the I index as a tool for the accuracy of group classification in terms of group overlap (Huberty and Lowman, 2000). Improvement over chance suggests how much better than chance we can predict group membership and how much better we did by using a classification rule than by relying on chance assignment.

$$I = \frac{H_0 - H_e}{1 - H_e}$$

where H_0 = the observed hit rate and H_e = the hit rate expected by chance.

It can be interpreted, by using a linear classification rule, about a certain percentage of fewer classification errors would be made than if classification were done by chance. The index I depends on (1) the definition of "chance" and (2) the rule whatever to determine observed hit rate.

Thompson (1995) suggests that PDA does not belong to the generalized linear model (GLM), while DDA does. The conclusions based on GLM concepts may not apply to the PDA case. In particular, there is a paradoxical dynamics in PDA versus

DDA. In any GLM analysis, more variables always lead to greater effect sizes. However, in PDA, more response variables can hurt the PDA hit rate. More predictor variables will usually improve the Wilks lambda value (i.e., smaller), but these additional variables may provide misinformation about the particular cases, resulting in their moving across the actual group boundary and become misclassification (Thompson, 1995). Extensive variable selection issue in the PDA context is explained in Huberty (1994, chapter 8).

Example

Most empirical research in higher education involves the study of multiple characteristics of students, faculty, and institutions. Metropolitan universities in the United States consists of various types of institutions that range from a group with a strong research orientation and doctoral granting to a number of small campuses. Administrators and researchers are interested in improving ways to classify institutions of higher education for descriptive, comparative, and analytic purposes.

The Carnegie Commission (1973 & 1976) developed a grouping or typology to classify American higher education institutions for research or administrative purposes. The criteria of the traditional Carnegie classification may not be appropriate for the characteristics of metropolitan universities. The research question of this example is to know how homogeneous groups in metropolitan universities differ each other and how well homogeneous groups can be classified.

The purpose of this example is that it seeks to find a classification structure that can increase the number of dimensions on the basis of which institutions can be grouped together and it tries to explore the possibility of classifying each institution into some subgroups of metropolitan universities.

To do this, this example will employ three multivariate analysis techniques. While this example uses a number of variables, the broader conceptual context is critical to gaining a deeper understanding of metropolitan universities. By using data reduction technique, such as factor analysis, helps guide the development of key measurement concepts for assessing key characteristics of metropolitan universities. Factor analysis can reduce the dimension of the multiple predictor variables. Clustering analysis can identify institutions that resemble one another. Using both factor analysis and cluster analysis can produce several homogeneous groups of institutions from the population of metropolitan universities.

Discriminant analysis can assess the homogeneity of groups derived and the relative contribution of each grouping (predictor) variable. Discriminant analysis also can predict group membership by classifying each institution into subgroups of metropolitan universities.

A descriptive discriminant analysis (DDA) is employed as an interpretive aid to assess the homogeneity of the clusters in the pattern. DDA would indicate the dimensions (factor, or clustering criteria) that maximized the differentiation among groups. Finally, a predictive discriminant analysis (PDA) is performed in which each institution is classified using the linear classification functions. Thus, the DDA and PDA serve as tests of the clearness of the clusters, indicating the degree to which the groups derived from the cluster analysis overlap.

The principal component analysis (PCA) extracted 3 factors from the 9 institutional characteristics, using eigenvalues-greater-than-one and scree tests. The PCA shows the content of the rotated factors, accompanied with the percentage of the total variance explained by each variable. In orthogonal rotation, the factor pattern matrix and the factor structure matrix remain equivalent. The analysis shows that the four-factor model accounted for about 84 percent of the total variance among the 9 variables.

Since total enrollment, full-time enrollment (FTE) students, and financial data have high loadings, this factor has been labeled "Size." The second factor has been named "Proportion" because the percentage of full-time and the percentage of part-time students have high loadings. The strong negative loading for part-time students simply means that the variable is related to the factor in opposite directions. The third factor has been labeled "Growth" because two variables, the growth rate of enrollment during the last one and eight years, have high loadings.

The 32 institutions with scores on each of the three factors were entered into the cluster analysis. Three groups with ranging from 7 to 16 were clustered of the 32 institutions for which factor scores were calculated. Table 1 shows the relative sizes of each of the three groups, as well as each group's three factor scores means and standard deviations.

Table 1. Group Means and Standard Deviations on Three Factors

	Group 1 (n=9)		Group 2 (n=16)		Group 3 (n=7)	
Factor	Mean	SD	Mean	SD	Mean	SD
Size	-.36	.84	-.09	.93	.65	1.16
Proportion	-1.06	.53	.83	.46	-.54	.65
Growth	-.64	.57	-.18	.65	1.24	1.08

The three groups (clusters) were entered into the multiple-group discriminant function analysis. Tabachnick and Fidell (1998) describe various practical issues for discriminant analysis. In fact, classification makes fewer statistical requirements than does inference. In this example, since the sample size is unequal and small, the assumption for the homogeneity of variance-covariance matrices is significant. If there is heterogeneity of the covariance matrices, results of significant testing may be misleading. Bos's M in SPSS DISCRIMINANT indicates the homogeneity of variance-covariance matrices. This example also uses SAS DISCRIM with POOL=TEST for the homogeneity test. Thus, this example employs the pooled covariance matrices and linear discriminant function analysis.

Descriptive discriminant analysis (DDA) can access the degree of separation among the three groups and the relative contribution of each of the three factors to group difference. Table 2 shows the result of DDA.

Table 2. Results of Discriminant Analysis (Structure Matrix)

	Discriminant Function Weights	
Factor	I	II
Size	-.055	.307
Proportion	.955	.295
Growth	-.151	.741

Two discriminant functions were produced, each statistically significant at $p < .001$. With both functions included, the $\chi^2(6)$ of 65.693 indicates a highly reliable relationship between groups and predictors. With the first discriminant function removed, there is still a reliable relationship between groups and predictors as indicated by $\chi^2(2)=26.597$, p=.000. This finding indicates that the second discriminant function is also reliable. Wilks' Lambda shows statistically significant association between groups and predictors, with discriminant functions accounting for 90.4% of the variance.

DDA is useful in identifying a structure underlying group comparison effect. Linear composites identify certain linear combination of variables that produce group difference. The first discriminant function accounts for the greatest group difference. The first discriminant function receives its largest contribution from the Proportion factor, and the second function from the growth factor mostly and from the Size factor next. The differences between group centroids (means in multidimensional space) indicate that each group differs from the others. They do suggest that the cluster analysis yielded three quite distinct groups of metropolitan universities. Figure 1 shows the relative separation of the groups, as well as their relation to one another along the two discriminant functions.

The first discriminant function is represented by the horizontal axis and has been labeled "Proportion" because of the strong contribution, which derives from the Proportion factor (see Table 1). The second discriminant function (vertical) has been named "Growth." Figure 1 also shows the plot of group centroids and variates for the pair of discriminant functions. The group centroids are plotted with respect to their values on the X- and Y-axes.

The plot shows the uses of both discriminant functions in separating the three groups. On the first discriminant function (X axis), the second group is some distance from the other two groups, but the first group and the third group are close together. On the second function (Y axis), the first group is far from the third group, but the second group has almost the same distance from the other two.

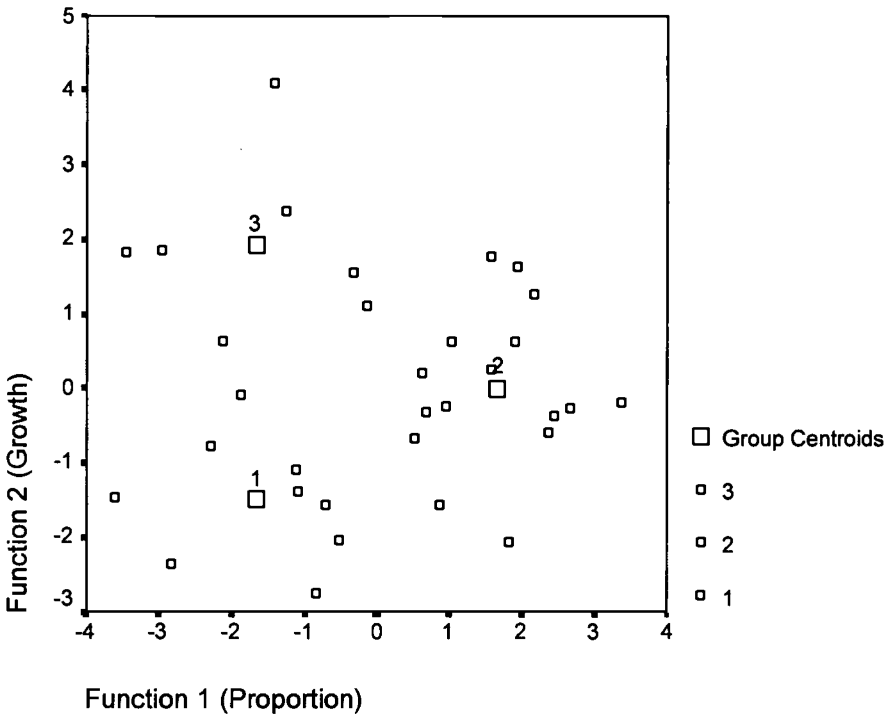

Figure 1. Centroids of three clusters of metropolitan universities on two discriminant functions

Since there is a clear difference between the centroid of one group and the centroid of another along a discriminant function axis, the discriminant function separates the two groups. We can conclude that the three groups are different in terms of three predictors.

These results are corroborated in the predictive discriminant analysis (PDA) (also called the classification analysis). PDA is useful in determining group assignment of experimental units. This classification is performed by calculating a score for each unit on each group's classification function and then assigns the unit to the group with the highest score. Prior probabilities were specified as unequal number of cases based on the number of cases in each group. Table 3 shows the result of the predictive discriminant analysis.

Table 3. Results of Classification Analysis (LOO method)

		Classified Group		
Actual Group	% Correctly classified	1	2	3
1 (n = 9)	100%	9	0	0
2 (n = 16)	100%	0	16	0
3 (n = 7)	71.4%	1	1	5
Overall	93.8%	10	17	5

This example uses the LOO method (an external analysis) so that the classification rule is determined from one set of samples and then uses to classify another set of samples. Overall, 93.8% of the institutions were correctly classified from the leave-one-out (LOO) method. The classification results identify the two cases that "incorrectly" classified and indicate the particular group they most closely resemble. Two cases (institutions) may be outliers in the third cluster and, thus, inappropriately assigned to the first and second groups. Table 4 shows the prediction of group membership that identifies right and wrong classifications.

This example also assesses the degree of prediction effectiveness. How much better than chance can we predict group membership? First, this example computes chance rate using proportional chance criterion. Second, this study employs the index I. The observed hit rate is the sum of the main diagonal elements of the classification matrix. The sample size in each group is not the same. The overall number of correctly classified cases is $o = 9 + 16 + 5 = 30$. The chance frequency of hits for each group: $e_1 = \frac{9}{32} \times 9 = 2.53$, $e_2 = \frac{16}{32} \times 16 = 8$, and $e_3 = \frac{7}{32} \times 7 = 1.53$. Thus, the overall chance hit rate is $e = e_1 + e_2 + e_3$=12.06. The observed hit rate is $H_o = \frac{30}{32} = .938$. The hit rate expected by chance is $H_e = \frac{12.06}{32} = .377$. Therefore, I index $I = \dfrac{.938 - .377}{1 - .377} = \dfrac{.561}{.623} = .90$. I index can be used as an effect size estimate of the hit rate in conjunction with a Z test of the significance between H_o and H_e. The index I in this example is 90 %. Thus, this concludes that by using a linear classification rule, about 90 % fewer classification errors would be made than if classification were done by chance. How much is the hit rate statistically significantly better than "chance"? The value of I Index can be considered as the value of an effect size index. The magnitude of I index is a matter of judgment (Huberty, 1994, p.108).

Summary, Conclusions, and Implications

Prediction of group membership is the goal of PDA and the accuracy of group classification is the focus of PDA. The predictive accuracy of the classification function is measured by hit rate. A hit rate represents the proportion of correctly classified cases. The bias correction can be achieved through leave-one-out (LOO) method (Huberty, 1994). Currently, Huberty and Lowman (2000) provide a graphical representation of the misclassified observation. The plot shows the observations based on the discriminant Z scores and portray the group overlap and the misclassified groups. They consider a classification error rate as an assessment of group overlap. They develop the I index as a tool for the accuracy of group classification in terms of group overlap. The I index suggests how much better than chance we can predict group membership and how much better we did by using a classification rule than by relying on chance assignment. How much is the hit rate statistically significantly better than "chance"? The value of I index can be considered as the value of an effect size index. Use of effect sizes in result interpretation is consistent with current movement in the field and the recommendations presented by the APA Task Force on statistical inference (Henson & Smith, 2000; Wilkinson & APA Task Force on Statistical Inference, 1999).

Table 4. Group Membership (LOO method)

Case	Name	Actual Group	Classified Group
1	U of Alaska	1	1
2	CSU, Fresno	2	2
3	CSU, Sacramento	2	2
4	San Diego SU	2	2
5	San Jose SU	2	2
6	Metropolitan SC, Denver	1	1
7	Florida Atlantic U	3	3
8	Florida International U	3	3
9	U of Central Florida	3	3
10	U of South Florida	3	1*
11	Georgia SU	1	1
12	U of Ill, Chicago	2	2
13	Indiana–Purdue U, Indianapolis	1	1
14	U of Lousville	2	2
15	Towson SU	2	2
16	U of Maryland, College Park	2	2
17	Wayne SU	1	1
18	Southwest Missouri SU	2	2
19	U of Missouri, St. Louis	1	1
20	U of Nevada, Las Vegas	3	3
21	CUNY, Brooklyn	1	1
22	CUNY, Hunter	1	1
23	U of North Carolina,	2	2
24	Cleveland SU	2	2
25	U of Toledo	2	2
26	Wright SU	2	2
27	Portland SU	1	1
28	Southwest Texas SU	2	2
29	UT, San Antonio	3	3
30	U of Houston	3	2*
31	UNT	2	2
32	Virginia Commonwealth U	2	2

Note: * indicates the misclassification of cases.

Lancaster, B.P. (1999). Predictive discriminant analysis: The classification dilimma. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio.

Thompson, B. (1995). Review of applied discriminant analysis by C. J. Huberty. Educational and Psychological Measurement, 55, 340-350.

Thompson, B. (1998). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas, http://acs.tamu.edu/~bbt6147/aeraaddr.htm

Tabachnick, Barbara G. & Linda S. Fidell. (1998). Using multivariate statistics, Allyn & Bacon Pearson Higher Education.

Wilkinson, L. & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanation. American Psychologist, 54, 594-604.

Yitzhaki, Shlomo. (1994). Economic distance and overlapping of distributions. Journal of Econometrics, 61, 147-159.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM032438

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Issues in Predictive Discriminant Analysis: Using and Interpreting the Leave-One-Out Jackknife Method and the Improvement-Over-Chance I Index Effect Size

Author(s): Hwang, Dae-Yeop

Corporate Source: University of North Texas

Publication Date: Feb. 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials.of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:

Printed Name/Position/Title: Dae-Yeop Hwang - Res. Assoc.

Organization/Address: UNT - Dept. of Tech. & Cog.
PO Box 311337 Denton, TX 76203-1337

Telephone: 940-565-2093

FAX: —

E-Mail Address: daeyeop@hotmail.com

Date: 3/12/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com