DONALD G. MORRISON

With the availability of "canned" computer programs, it is extremely easy to run complex multivariate statistical analyses. However, it is not as easy to interpret the output of these programs. This article offers some comments about the well-known technique of linear discriminant analysis; potential pitfalls are also mentioned.

On the Interpretation of Discriminant Analysis

BACKGROUND

Many theoretical- and application-oriented articles have been written on the multivariate statistical technique of linear discriminant analysis. However, on a practical level little has been written on how to evaluate results of a discriminant analysis—at least in managerial, as opposed to statistical, terminology. This article looks at the problem of evaluation from various viewpoints and thus highlights some features pertaining to other statistical techniques.

Overview of Discriminant Analysis

The objective of a discriminant analysis is to classify objects, by a set of independent variables, into one of two or more mutually exclusive and exhaustive categories. For example, on the basis of an applicant's age, income, length of time at present home, etc., a credit manager wishes to classify this person as either a good or poor credit risk. For expository purposes we will limit this discussion to two classifications; later we will comment on n-group discriminant analysis.

For notation, let

\[ X_{ji} \] be the \( i \)th individual's value of the \( j \)th independent variable

\[ b_j \] be the discriminant coefficient for the \( j \)th variable

\[ Z_i \] be the \( i \)th individual's discriminant score

\[ z_{crit.} \] be the critical value for the discriminant score.

Linear Classification Procedure

Let each individual's discriminant score \( Z_i \) be a linear function of the independent variables. That is,

\[ Z_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_n X_{in}. \]

The classification procedure follows:

- if \( Z_i > z_{crit.} \), classify Individual \( i \) as belonging to Group 1;
- if \( Z_i < z_{crit.} \), classify Individual \( i \) as belonging to Group 2.

The classification boundary will then be the locus of points,

\[ b_0 + b_1 X_{i1} + \cdots + b_n X_{in} = z_{crit}. \]

When \( n \) (the number of independent variables) = 2, the classification boundary is a straight line. Every individual on one side of the line is classified as Group 1; on the other side, as Group 2. When \( n = 3 \), the classification boundary is a two-dimensional plane in 3-dimensional space; the classification boundary is generally an \( n - 1 \) dimensional hyperplane in \( n \) space.

Advantages of a Linear Classification Procedure

The particularly simple form of (1) allows a clear interpretation of the effect of each of the independent variables. Suppose the independent variable \( X_1 \) is income, and the classification procedure is if \( Z_i > z_{crit.} \), classify the individual as being a good credit risk, i.e., the higher the value of \( Z_i \), the more likely the individual is a good credit risk. If the sign of \( b_1 \) is positive, then higher income implies a better credit risk, and the larger the size of \( b_1 \), the more important variable \( X_1 \) is in discriminating between Group 1 and Group 2 individuals. Clearly, if \( b_1 = 0 \), then \( X_1 \) has no effect.

If we had a more complex discriminant function, we could not isolate the effect of each variable so easily. Suppose we had a nonlinear discriminant function, say

\[ Z_i' = a + b X_i + c X_i^2 + d Y_i + e Y_i^2 + f X_i Y_i. \]
ON THE INTERPRETATION OF DISCRIMINANT ANALYSIS

The effect on $Z_i$ of increasing $X_i$ by one unit depends on the value of $X_i$, $b$, $c$, $f$, and even $Y$.

Hence, for interpretation, a linear discriminant function is highly desirable. This raises the following question.

**When is a Linear Classification Procedure Valid?**

The technical details of this section are in the appendix. However, the essence of these details can be easily expressed. A linear classification procedure is optimal if the spreads (variance) of the independent variables (the $X_i$'s) in Group 1 are the same as the spreads in Group 2 and if the interrelations (correlations) among the independent variables in Group 1 are the same as the interrelations in Group 2. Really we are saying that the covariance matrices of Group 1 and Group 2 are equal.

The appendix also gives a brief example of the kind of nonlinear classification region that can arise when the assumption of equal covariance matrices is not true.

Next is the discussion of evaluating the results after a discriminant analysis has been run.

**STATISTICAL SIGNIFICANCE**

**Distance between Groups**

One of the standard quantities that appear on the output of a discriminant analysis is a distance measure, the Mahalanobis $D^2$ statistic, between the two groups. After a transformation this $D^2$ statistic becomes an $F$ statistic, which is then used to see if the two groups are statistically different from each other. In fact this test is simply the multidimensional analog of the familiar $t$ test for the statistical significance of the difference between one sample mean $\bar{x}_1$ and another sample mean $\bar{x}_2$. The $D^2$ (or transformed $F$) statistic tests the difference between the $n$-dimensional mean vector $\bar{x}_i$ for Group 1 and the corresponding $n$-dimensional mean vector $\bar{x}_j$ for Group 2. However, the statistical significance per se of the $D^2$ statistic means very little.

Suppose the two groups are significantly different at the .01 level. With large enough sample sizes, $\bar{x}_1$ could be virtually identical to $\bar{x}_2$, and we would still have statistical significance. In short, the $D^2$ statistic (or any of its transformed statistics) suffers the same drawbacks of all classical tests of hypotheses. The statistical significance of the $D^2$ statistic is a very poor indicator of the efficacy with which the independent variables can discriminate between Group 1 individuals and those in Group 2.

**PERCENTAGE CORRECTLY CLASSIFIED**

**A Bias Exists in Many “Canned” Programs**

One common source of misinterpretation of discriminant analysis results comes from the way in which most of the “canned” computer programs construct the classification table (sometimes called the confusion matrix). The computer will print out the following table:

```
<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>n_{11}</td>
<td>n_{12}</td>
</tr>
<tr>
<td>Group 2</td>
<td>n_{21}</td>
<td>n_{22}</td>
</tr>
</tbody>
</table>
```

The entry $n_{ij}$ is the number of individuals who are actually in Group $i$, but were classified under Group $j$. Then $(n_{11} + n_{22})/n$ is the proportion of individuals correctly classified. However, the typical canned program uses all $n$ observations to calculate the discriminant function and then classifies these same $n$ individuals with this function. Frank, Massy, and Morrison [3] discuss in detail the upward biases that can occur in classification tables constructed in this way. One method of avoiding this bias is to fit a discriminant function to part of the data and then use this function to classify the remaining individuals. It is the classification table for these last individuals that we will discuss now.

**Percent Correctly Classified by Chance**

Suppose a researcher is interested in determining the socioeconomic variables that distinguish adopters of a new product from nonadopters. His “fresh” second half of the split sample contains 30 adopters and 70 nonadopters. He applies his discriminant function obtained in the first half of the split sample to this second
half and gets 70 percent correct classifications. He then says, "By chance I could get 70 percent correct classifications; therefore, my discriminant function is not effective in separating adopters from nonadopters." Notice that the chance model has not been explicitly stated. The remainder of this section will develop a more appropriate chance model that will show that in a statistical sense this hypothetical researcher is being overly pessimistic.

Assume that there exists a population with only two types of individuals, Type I and Type II. Let \( p \) be the proportion of the population that is Type I, and \( 1 - p \) the proportion that is Type II. If the variables (age, income, etc.) actually have no effect on discriminating I's from II's, we can expect to get a proportion \( p \) correctly classified if we classify everyone as Type I. Hence, if \( p > \frac{1}{2} \), we would classify everyone as Type I; if \( p < \frac{1}{2} \), would classify everyone as Type II.

Our hypothetical researcher wishes to identify adopters. However, since his sample has only 30 percent adopters and 70 percent nonadopters, he defies the pure chance odds if he classifies an individual as an adopter. This is true because any individual has an a priori \(.7\) probability of being a nonadopter and only a \(.3\) probability of being an adopter. But what if the researcher says, "I want to try to identify the adopters. I believe my discriminant function has some merit; therefore, I am going to classify 30 percent of the indi-

For our researcher, \( p = \alpha = .3 \). Hence the chance proportion correctly classified is \((.3)^3 + (.7)^3 = .58\).

Note that when \( p = .5 \), i.e., two groups of equal size, \( P(\text{Correct}) = .50 \) regardless of the value of \( \alpha \).

More formally this proportional chance criterion is

\[
P(\text{Correct}) = \alpha^2 + (1 - \alpha)^2.
\]

where

\[
\alpha = \frac{\text{the proportion of individuals in Group } 1}{1 - \alpha} = \frac{\text{the proportion of individuals in Group } 2}.
\]

The researcher who said, "By chance I could get 70 percent correct," was using the maximum chance criterion,

\[
C_{\text{max}} = \max(\alpha, 1 - \alpha).
\]

where max(\(\alpha, 1 - \alpha\)) is read, "the larger of \(\alpha\) or \(1 - \alpha\)."

For example, \(\max(.3, .7) = .7\).

**Situations Where \(C_{\text{pro}}\) and \(C_{\text{max}}\) Should Be Used**

If the sole objective of the discriminant analysis is to maximize the percentage correctly classified, then clearly \(C_{\text{max}}\) is the appropriate chance criterion. If the discriminant function cannot do better than \(C_{\text{max}}\), you are wiser to disregard it and merely classify everyone as belonging to the larger of the two groups. Obviously, this is rarely true for a marketing research study. Usually a discriminant analysis is run because someone wishes to correctly identify members of both groups. As indicated, the discriminant function defies the odds by classifying an individual in the smaller group. The chance criterion should take this into account. Therefore, in most situations \(C_{\text{pro}}\) should be used. Recall that our discussion on chance models applies to individuals not used in calculating the discriminant function. If the individuals were used in calculating it, then some upward adjustment must be made on \(C_{\text{pro}}\) or \(C_{\text{max}}\). Frank, Massy, and Morrison [3] give methods for estimating these biases.

**Analogy with Regression**

Perhaps an analogy with regression will clarify these concepts. We have all read articles in which the author has found "significant" relations; however, he has "explained" only four percent of the variance, i.e., \(R^2 = 0.04\). But since the sample size is large, this sample \(R^2\) is statistically significantly different from zero. In discriminant analysis, the percentage correctly classified is somewhat analogous to \(R^2\). One tells how well we classified the individual; the other tells how much variance we explained. Statistical significance of the \(R^2\) is analogous to the statistical significance of the \(D^2\) statistic. Clearly, with a large enough sample size in discriminant analysis we could classify 52 percent correctly (when chance was 50 percent) and yet have a statistically significant difference (distance) between the two groups.

**Evaluation Criteria for Discriminant Analysis**

When results of a discriminant analysis are obtained, there are three basic questions to ask: (1) Which independent variables are good discriminators? (2)
How well do these independent variables discriminate among the two groups? (3) What decision rule should be used for classifying individuals? We have already discussed the first two questions; the third one obviously involves economic considerations. More complete answers to these questions require a synopsis of the theoretical derivation of the discriminant function.\footnote{Mathematical details on this derivation are in [1, Chapter 6].}

**Deriving the Discriminant Function**

Let us look at Individual $i$ and observe his values of the $n$ independent variables. That is, we see

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{in}).$$

Let

$$P(I)$$

be the unconditional (prior) probability that an individual belongs to Group 1

$$P(I | x_i)$$

be the conditional (posterior) probability that an individual belongs to Group 1, given we have observed $x_i$

$$l(x_i | I)$$

be the likelihood that an individual has the vector of values $x_i$, given that he belongs to Group 1.

Analogous definitions hold for Group 2. From Bayes Theorem we have

$$P(I | x_i) = \frac{l(x_i | I)}{l(x_i | II)} \cdot \frac{P(I)}{P(II)}.$$  \hspace{1cm} (5)

Or,

$$\text{Posterior Odds} = \text{Likelihood Ratio} \times \text{Prior Odds}.$$  \hspace{1cm} (5)$^*$

The classification procedure will then be as follows. If the odds are strongly enough in favor of Group 1, classify the individual as belonging to Group 1, (If the odds were 3 to 1 in favor of Group 1, this would mean a probability greater than 0.5) is equivalent to the logarithm of the odds being greater than zero. We may write (5) as

$$\log(\text{posterior odds}) = \log(\text{likelihood ratio}) + \log(\text{prior odds}).$$  \hspace{1cm} (6)

If the assumptions of normality and equal covariance matrices discussed earlier are true, the logarithm of the likelihood ratio is of the form,

$$\log(\text{likelihood ratio}) = b_0 + b_1 X_{i1} + \cdots + b_n X_{in}.$$  \hspace{1cm} (7)

This is the discriminant function. When the two groups are of equal size, each group's prior probabilities are equal. The prior odds are then one, and the posterior odds are merely the likelihood ratio. When the prior odds are different from one, then

$$\log(\text{posterior odds}) = b_0 + b_1 X_{i1} + \cdots + b_n X_{in} + \log(\text{prior odds}).$$

However, since the prior odds contain none of the independent variables, this quantity is a constant, and the discriminant function is

$$b_0' + b_1 X_{i1} + \cdots + b_n X_{in},$$

where

$$b_0' = b_0 + \log(\text{prior odds}).$$

An understanding of the foregoing nonmathematical material is sufficient to answer the three basic questions.

**Determining the Effect of Independent Variables**

The sign and size of the $b_j$'s determine the effect of the independent variables $X_j$. The size of the coefficient $b_j$ in the discriminant function,

$$Z_i = b_0 + b_1 X_{i1} + \cdots + b_j X_{ij} + \cdots + b_n X_{in},$$

will clearly be influenced by the scale that we use for $X_j$. Suppose $X_j$ is family income. A change of $X_j$ from $6,000 to $7,000 will have the same effect on $Z_i$ whether or not $X_j$ is scaled in dollars or thousands of dollars. Therefore if $X_j$ is measured in thousands of dollars, $b_j$ will be one thousand times larger than if the units of $X_j$ are in dollars. However, if we normalized (divided) each variable by its standard deviation, the original units become irrelevant. As units are scaled by a factor $k$, the standard deviation is also scaled by the same factor $k$. That is, if the standard deviation of $X_j$ is $\sigma_j$, then the standard deviation of $kX_j$ is $k\sigma_j$. Then since $X_j/\sigma_j = kX_j/k\sigma_j$, we need not worry about the scale of $X_j$.

Let $b_j^*$ be the discriminant coefficient that results when the standardized variables $X_j^* = X_j/\sigma_j$ are used.\footnote{We assume that the variables form at least an interval scale. That is, any variable $X$ can be transformed to a new variable $Y = a + bX$, where $a$ and $b$ are arbitrary constants, without affecting the analysis. The standardized coefficients $b^*$ will remain unaffected by these linear transformations of the data. Some of the other multivariate methods, e.g., some cluster analysis techniques, do not require such strict assumptions about the scale of the independent variables.} Suppose $|b_j^*| > |b_k^*|$. Then variable $X_j$ is a better discriminator between Group 1 and Group 2 than variable $X_k$. A unit change in $X_j^*$ has more effect on $Z_i$ than a unit change in $X_k^*$. The more a variable affects $Z_i$, the better it discriminates. We are justified in normalizing our variables by their standard deviations, since we are discriminating on the basis of statistical distance between the two groups and statistical distances are measured in units of standard deviations.\footnote{See [3] for a detailed discussion of distance concepts.}

If the discriminant analysis is run with nonstandardized variables, it is extremely easy to obtain $b_j^*$
from $b_j$. We have seen that

$$b_j X_j = b_j^* X_j = b_j^* \frac{X_j}{\alpha_j}.$$  

Hence,

$$(9) \quad b_j^* = b_j \alpha_j.$$  

Recall that the sign of $b_j^*$, which is the same as that of $b_j$, determines the direction of the effect of $X_j$. If $b_j^*$ is positive, as $X_j$ increases, $Z_i$ increases; the larger $Z_i$, the more likely that Individual $i$ belongs to Group 1.

We want to obtain the best possible estimates $b_j$. As in all statistical estimation, the larger the sample size (assuming it is a representative sample), the better the estimates. Suppose we have 900 individuals in Group 1 and 100 individuals in Group 2. If we use only 100 of Group 1 individuals in calculating the discriminant function, the prior probability of an individual belonging to Group 1 is 0.5. But if we use all 900 members of Group 1, this prior probability drops to 0.1. Does this affect any of the $b_j$'s of interest? No. Recall from (8) that the prior probabilities affect only $b_0$ and have no effect on $b_1, b_2, \ldots, b_n$. Therefore, in determining which variables are the best discriminators, we should use all the data. (By this we mean all individuals and not necessarily all available independent variables. As in any multivariate technique, if $X_j$ and $X_k$ are highly correlated, they are measuring almost the same thing. The coefficients $b_j$ and $b_k$ will be unstable and hard to interpret.)

The advisability of using all data in calculating the $b_j$'s is not surprising; in fact it is intuitively obvious. However, in assessing the discriminant function's performance, we may not want to use all the data.

**How Well Do the Variables Discriminate?**

To answer this question we need to use the classification table and an appropriate chance criterion. Throughout this discussion we will assume that we either have fresh data or that we have adjusted for “fitting-the-discriminant-function-to-data” bias. The question of how to use the data arises when the two groups are of greatly unequal size.

We saw from (6) that when the two groups are of equal size, the likelihood ratio (which contains all sample information) completely determines the discriminant function. However, when the prior probabilities are unequal, this influences the classification procedure. If the groups are greatly unequal, the term $\log(\text{prior odds})$ can completely dominate the term $\log(\text{likelihood ratio})$. Here we cannot determine how well the independent variables discriminate. We would obtain the clearest picture if the prior odds were equal and, hence, did not affect the classification.

Assume that we were attempting to discriminate adopters of a new product from nonadopters. If we had a sample of 1,000 people, a result of 50 adopters and 950 nonadopters would not be unusual. If we attempted to classify all 1,000 individuals, we might get a classification table like the following:

<table>
<thead>
<tr>
<th>Classified</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>937</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>980</td>
</tr>
</tbody>
</table>

Here we classified 944 (or 94.4 percent) individuals correctly. The proportional chance criterion is (see (2))

$$C_{pro.} = (.05)^2 + (.95)^2 = .907.$$  

However, given that we classified 98 percent as Group 2, the outcome should have been

$$(.98)(.95) + (.02)(.05) = .932,$$  

or 93.2 percent correctly classified.

The maximum chance criterion is

$$C_{max} = .95.$$  

Therefore our 94.4 percent correct classification is not too impressive. However, of the 20 individuals classified as Group 1, seven were correct. This is 35 percent compared with a chance percentage of 5. This last result is fairly impressive.

Now let us change the hypothetical classification slightly.

<table>
<thead>
<tr>
<th>Classified</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>941</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>990</td>
</tr>
</tbody>
</table>

Here, we still classified slightly over 94 percent correctly; however, only one in ten was correct for Group 1 classifications.

In summary, when one group is much larger than the other, almost all individuals are classified as the larger group. This means several will automatically be correctly classified. When we allow the posterior odds to classify the individuals—see (5)—we usually get even fewer classified in the smaller group than actually belong in it. There is often more interest in the smaller group, and classification tables like the preceding two
are not the best way to assess the discrimination power of the independent variables.

One possibility is to rank all 1,000 individuals by their Z values and put the 50 highest in Group 1. This assures that a sufficient number will be classified as Group 1. We can now see how well we classified these individuals.

Another method would be to randomly divide the 950 Group 2 individuals into 19 groups, each with 50 members. We could construct 19 classification tables; the same 50 Group 1 members and the 19 different Group 2's. Then we could see on the average how well we did. This procedure has the advantage that the chance model is unambiguously 50 percent. Working with 50 percent chance models also makes interpretation easier. It is clear that correctly classifying 75 percent when chance is 50 percent is a good classification. (Of the 50 percent by which we could improve chance, we got half or 25 percent.) When the sample sizes were 50 and 950, the proportional chance criterion was 90 percent. Suppose we again obtain half of that remaining after chance and classify 95 percent correctly. This could occur by doing well with Group 1 individuals or by merely classifying everyone as Group 2. However, the interpretation is not as clear.

In summary we can say that (a) when the groups are of greatly unequal size, it may be difficult to interpret the classification table, and (b) regardless of the total sample size, the effective sample size (for determining ability to discriminate) is governed by the smaller of the two groups.

This last point is particularly relevant in the planning stages of a research project. A large total sample size is of little comfort without a sufficient number of individuals in each group.

Classification Decision

The last two sections dealt with which variables are good discriminators and their ability to discriminate. However, if the discriminant function is used to classify individuals, then clearly the misclassification costs must enter the decision.

As before, let

\[ P(I | x_i) \] be the posterior probability that an individual belongs to Group 1, given that we observed his vector of independent variables \( x_i \).

\[ P(II | x_i) \] be analogous definition for Group 2

\( C_{12} \) be the opportunity cost of classifying an individual in Group 2 when he actually belongs in Group 1

\( C_{21} \) be the opportunity cost of classifying an individual in Group 1 when he actually belongs in Group 2.

Any rational cost structure would have \( C_{12} = C_{21} = 0 \). If we classify Individual \( i \) as Group 1, the expected opportunity cost is

\[ K_i(I) = P(II | x_i)C_{21}. \]

Similarly, if we classify him as Group 2, the expected opportunity cost is

\[ K_i(II) = P(I | x_i)C_{12}. \]

The classification procedure becomes: if \( K_i(I) < K_i(II) \), we classify Individual \( i \) as belonging in Group 1 and vice versa.

By the same reasoning used to examine the effect of prior probabilities on the discriminant function, it is clear that \( C_{12} \) and \( C_{21} \) affect only the \( b_0 \) term of the discriminant function (or equivalently it simply changes the \( z_{crit} \) value).

Let the logarithm of the likelihood ratio—see (6)—be

\[ \log(\text{likelihood ratio}) = b_0 + b_1X_{11} + b_2X_{21} + \cdots + b_nX_{ni}. \]

The classification rule is then:

classify Individual \( i \) as Group 1 if

(10a) \( b_0 + b_1X_{11} + b_2X_{21} + \cdots + b_nX_{ni} > \log k. \)

Classify Individual \( i \) as Group 2 if

(10b) \( b_0 + b_1X_{11} + b_2X_{21} + \cdots + b_nX_{ni} < \log k, \)

where

\[ k = \frac{P(II)C_{21}}{P(I)C_{12}}. \]

In a real application, the difficult problem will be obtaining good estimates for the opportunity costs \( C_{12} \) and \( C_{21} \).

IMPLEMENTATION OF THE RESULTS

One of the first successful business applications of discriminant analysis was in credit selection. Good credit risks were separated from poor credit risks on the basis of demographic and socioeconomic variables. Since on the credit application the individual fills in information on these same demographic and socioeconomic variables, the discriminant function can be applied directly to his application. The classification procedure (10a) and (10b) is then used to determine whether the applicant is to be given credit.

A main problem with this kind of project is obtaining representative past data. Chances are that the company only has data on individuals accepted as good credit risks. Of these previously screened individuals, some were actually good credit risks, others were not. However, this sample is not representative of the applicants applying for credit. In other words, the discriminant function for past data may not be the best for discriminating among current applicants. Of course, there is always the problem that the past discriminant function is outdated. Time or the competitive situation has changed the environment enough to make old results inapplicable. But at least for credit selection the variables used to discriminate were operational; the
independent variables were used in decision making. This is not always true.

Suppose a researcher were able to discriminate adopters of a new product from nonadopters on the basis of demographic characteristics. If the product is sold through supermarkets and advertised in the mass media, it may be difficult to direct in-store displays and ads specifically at the likely adopters. However, if cents-off coupons are mailed, it may be relatively easy to direct this mailing to the more likely adopters. If a discriminant analysis is to be considered a decision-making aid (as opposed to a strictly research-oriented study), management needs a clear idea how the results will be implemented before the project is undertaken.

When the independent variables are obtained by personal interviews, there is a whole new set of problems. It may be particularly hard to get comparability across interviewers. High degrees of collinearity (high correlations) among the independent variables should be avoided. The resulting discriminant coefficients will be unstable, and it will be more difficult to interpret the contribution of each independent variable. Hence, if two independent variables are highly correlated, e.g., \( r = 0.95 \), only one of these variables should be included in the analysis. Otherwise the variances of the \( b \)'s (the discriminant coefficients) will be unnecessarily large.

**DISCUSSION**

In summary, some considerations follow:

1. A linear discriminant function is appropriate only when the groups' covariance matrices are equal (or nearly equal).
2. The \( D^2 \) statistic (which may be transformed to an \( F \) statistic) only tests the statistical significance of the difference between groups. Recall the effect of the sample size on statistical significance.
3. Beware of the upward bias that results from classifying the same individuals used to calculate the discriminant function.
4. Beware of the different chance models that can result when groups have different sizes. Remember that greatly unequal-sized groups make interpretation of the classification table difficult.
5. The effective sample size is really governed by the smaller group.
6. Have the discriminant coefficients been normalized by the standard deviations of the independent variables?
7. In forming the classification decision, be sure that prior probabilities and opportunity costs of misclassification have been considered.
8. Will the independent variables used for discrimination be operational?

All of these apply to multiple discriminant analysis, i.e., when we are classifying individuals into more than two groups. The only main difference is that it is not as easy to assess the effect of the independent variables in discriminating among the groups. For example, variable \( j \) might be the best discriminator for Group 1 and Group 2, but variable \( k \) is best between Group 2 and Group 3. Strictly speaking, all eight points also apply to discriminant analysis for more than two groups.

**APPENDIX**

*Conditions Required for Optimal Linear Classification*

Let

\[ \mu_{ji} \] be the mean of the \( j \)th variable for individuals who belong to Group 1

\[ \sigma_{jk} \] be the covariance between variables \( j \) and \( k \) for individuals who belong to Group 1.

The mean vector \( \mu_1 \) is formed as follows:

\[
\mu_1 = (\mu_{11}, \mu_{21}, \ldots, \mu_{n1}).
\]

The covariance matrix \( V_1 \) is

\[
V_1 = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\
\sigma_{12} & \sigma_{22} & \cdots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{nn}
\end{bmatrix}
\]

\( \sigma_{11} \) is merely the variance of \( X_1 \). The covariance between \( X_j \) and \( X_k \) is equal to the covariance between \( X_k \) and \( X_j \); hence, the matrix is symmetrical. Finally, the covariance is related to the simple correlation between two variables. Letting \( r_{jk} \) be the correlation between variables \( X_j \) and \( X_k \), we have

\[
r_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}.\]

Analogous definitions hold for the mean vector \( \mu_2 \) and covariance matrix \( V_2 \). With these preliminaries we can now state the conditions for optimality of a linear classification procedure.

A linear classification procedure is optimal if: (a) the independent variables in Groups 1 and 2 are multivariate normal with mean vectors \( \mu_1 \) and \( \mu_2 \) and covariance matrices \( V_1 \) and \( V_2 \), respectively, and (b) \( V_1 = V_2 \).

We will now illustrate how unequal covariance matrices can lead to nonlinear classification boundaries. Suppose we are classifying two groups on the basis of two variables \( X_1 \) and \( X_2 \). Assume that the mean vectors \( \mu_1 \) and \( \mu_2 \) are equal, but that the covariance matrices are of the form,

\[
V_1 = \begin{bmatrix}
\sigma_{11} & 0 \\
0 & \sigma_{22}
\end{bmatrix},
\]

\[
V_2 = \begin{bmatrix}
\alpha\sigma_{11} & 0 \\
0 & \alpha\sigma_{22}
\end{bmatrix},
\]

where \( \alpha \neq 1 \).
ON THE INTERPRETATION OF DISCRIMINANT ANALYSIS

where

\[ \alpha > 1. \]

Then intuitively, the farther an individual's \( x_i \) is from the common mean vector \( \mu \), the more likely it is that he belongs to Group 2. Mathematically we would calculate the distance from \( \mu \) at which the likelihood functions for both groups were equal. Because the covariance matrices are symmetrical, the locus of such points will be a circle with \( \mu \) as the center. The classification boundary will be this circle. If the prior probabilities favor Group 1, the radius of the circle will increase and vice versa.

Interpretation of the variables is very difficult (or at least not simple). An increase in \( X_1 \) may increase or decrease the likelihood of an individual belonging to Group 1; it depends on the previous values of \( X_1 \) and \( X_2 \).

REFERENCES
