Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). Applied multivariate research: Design and interpretation. Thousand Oaks, CA: Sage.

CHAPTER 7A

Discriminant Function Analysis

Discriminant and Logistic Analysis Compared

The purpose of a discriminant function analysis is similar to logistic regression in that we use it to develop a weighted linear composite to predict membership in two or more groups (Cooley & Lohnes, 1971). In this chapter, we limit ourselves to a discriminant function analysis design that predicts membership in one of two groups.

The Criterion and Predictor Variables

In both discriminant function analysis and logistic regression, the variable being predicted, the *groups* variable as it is sometimes called in discriminant analysis, is a categorical variable. In logistic regression, the predictor variables may be either continuous, dichotomous (categorical variables can be used if they are first dummy coded), or a combination of both. In discriminant function analysis, the predictors are continuous variables (Duarte Silva & Stam, 1995).

The Model Generated

In logistic regression, the model is nonlinear (i.e., sigmoidal or S shape). This means that the changes in the dependent variable (usually dichotomous) are not as great at the low and high ends of the predictors and that the greater changes occur in the middle ranges of the predictors. In discriminant function analysis, the model is linear; thus, the predictors and the criterion have a constant relationship to each other. The weighted combination of the independent variables (known as the discriminant function) is constructed to maximally separate the groups. These weights are calculated to derive a discriminant score for each case. If these scores are used as the dependent variable in a one-way ANOVA with group as the independent variable, the mean discriminant scores for the groups (called *centroids*) would maximally separate the groups. That is, no other derived centroids could separate the groups more.

Discriminant and Regression Analysis Compared

Comparable to multiple regression, discriminant function analysis likewise assumes proper model specification—that is, the inclusion of all important predictor variables and the exclusion of extraneous variables. As is the case for multiple regression, these predictors are considered to be independent variables. Unlike what we have seen for multiple regression, the criterion or dependent variable is a categorical variable representing the groups being differentiated or distinguished. For those who are familiar with experimental design and ANOVA, this may seem a bit reversed because the groups in such research typically represent the independent rather than the dependent variable. Because we are predicting the group membership of the cases in discriminant analysis, however, this categorical variable plays the role of a dependent variable in discriminant analysis.

Assumptions Underlying Discriminant Function Analysis

Unlike the nonlinear nature of logistic regression, discriminant function analysis conforms to the general linear model. Thus, it makes the same rigorous assumptions as multiple regression, including linearity, normality, independence of predictors, homoscedasticity, absence of multicollinearity, and the influence of outliers; it also presumes the use of near interval data (see Chapters 3A and 3B).

Discriminant function analysis, as is true for some of the other general linear model techniques, is fairly robust to violations of most of these assumptions. However, it is highly sensitive to outliers, and these should be resolved as described in Chapter 3A prior to the analysis. If one group contains extreme outliers, these will bias the mean (draw the mean toward it and away from the bulk of the other values) and increase the variance (the squared difference between the outlier and the mean will result in an exceptionally large value in the calculation of the variance). Because overall significance tests are calculated on pooled variances (i.e., the average variance across all groups), these significance tests are prone to a Type I error—that is, reporting statistical significance erroneously in the presence of outliers (McLachlan, 1992).

Because discriminant function analysis assumes that the continuous independent variables are normally distributed, violation of this assumption suggests that we should probably opt for logistic regression. Recall that logistic regression can accommodate both dichotomous and continuous variables, and the predictors do not have to be normally distributed, linearly related, or have demonstrated equal variance within each group (Tabachnick & Fidell, 2001b). Logistic regression is sometimes preferable to discriminant analysis in studies where the variables violate multivariate normal distributions within the criterion groups (Klecka, 1980). However, if the assumption of normality is not violated, logistic regression is less powerful than discriminant analysis (Lachenbruch, 1975; Press & Wilson, 1978).

Sample Size for Discriminant Analysis

Besides the assumptions presented above, there are issues concerning sample size that need to be taken into account. Discriminant analysis permits the groups to be of different sample sizes, but the sample size of the smallest group should exceed the number of predictor variables (by a lot). The maximum number of independent variables should be taken as $N - 2$, where n is the sample size. But although this minimum sample size may allow the analysis to be conducted, it is certainly not recommended (Huberty, 1994). The recommended sample size for the smallest group should be at least 20 times the number of predictors.

Purposes of Discriminant Function Analysis

Whereas logistic regression is used mainly for predictive purposes (i.e., estimating the probability that an event will occur), discriminant function analysis can serve two distinct purposes: (a) prediction, referred to as predictive discriminant analysis, and (b) explanation, referred to as descriptive discriminant analysis (Huberty, 1994).

If the primary purpose of the researcher is to gain understanding of the nature of the variate derived from the linear combination of the predictors, then descriptive discriminant analysis would be performed. Descriptive discriminant analysis is often used as a follow-up analysis to a significant MANOVA to determine the structure of the linear combination of the dependent variables. In fact, discriminant function analysis is computationally identical to MANOVA. Generally, descriptive discriminant function analysis focuses on "revealing major differences among the groups" (Stevens, 2002, p. 285). If, however, the primary purpose of the researcher is to compute a case's membership in one of the groups from the weighted linear combination of the predictors—that is, to classify the cases into the different groups—then predictive discriminant analysis would be performed.

It is true, as students sometimes note, that we already know the predicted values of the quantitative criterion variable in multiple regression and of the group membership in logistic and in discriminant analyses. Students therefore wonder why we engage in the procedure in the first place with the inevitable result that we have prediction error when we can just look at the data file to see these values directly. The answer is linked to the purposes of these analyses. From a descriptive perspective, we are trying to put together a general model—the discriminant function in the present case—to describe the predictive relationship of the independent variables to the dependent variable; our interest is not restricted exclusively to the cases in the sample. From a predictive perspective, we have hopes of using the model to predict the group membership of future samples; again our interest goes well beyond the particular cases in the present sample.

A Two-Group Discriminant Example

To demonstrate a two-group discriminant function analysis, we will use an example in which researchers are trying to determine if successful clinical psychology interns (i.e., those who passed their internship) could be identified by (a) a pre-self-evaluation questionnaire, (b) a preliminary clinical examination score, and (c) an initial committee evaluation rating. The self-evaluation asks the students to project the degree to which they feel they would be successful in their internship. It uses a 5-point response scale, with higher scores indicating greater confidence in clinical ability. The preliminary exam was administered by the internship sites participating in the study and was graded A to F, with A being awarded 5 points and F being awarded 0 points. The initial committee evaluation rating was based on judgments of the entry-level skills exhibited by the new interns. The items on this rating form were scored on a 7-point Likert-type scale, with greater scores indicating a more positive evaluation.

These three continuous variables are the independent variables (or predictors). One year later, the criterion or outcome measure was obtained. This dependent variable was whether or not the student passed the internship. In this study, the outcome variable was dummy coded as 1, representing successful completion, and 0, representing unsuccessful completion (i.e., failure).

Recognizing the Discriminant Design

When we ask if we can project what someone's score on some measure might be or determine with which subset a certain individual or case might be associated, we are dealing in a general way with prediction. That tells us, within the context of the topics covered in this text, that we are in the domain of multiple regression, logistic regression, or discriminant function analysis. In the present example, because the variable we wish to predict (identify) is a categorical and not a quantitative variable, we know that multiple regression is not the technique we should use. The choice between logistic regression and discriminant analysis is decided by the scale of measurement underlying the predictors. In this case, the predictors are all quantitative variables. We have also determined that these variables meet the assumption of normal within-group distributions. With this information, we can then assert that discriminant analysis becomes the procedure of choice.

It is also worthwhile to note two points in connection with the selection of the statistical procedure we would use to analyze the data in this hypothetical study. First, we could use logistic regression with these quantitative predictors even though it is less powerful here. Second, if we did use standard multiple regression with this dichotomous criterion variable, we would actually obtain a mathematically identical solution to the one yielded by standard discriminant function analysis. The unstandardized and standardized discriminant coefficients would be some constant multiple of the b and beta coefficients from the multiple regression equation. Nonetheless, discriminant is the proper choice of procedure to use with the specified variables.

The Discriminant Function

Recall that in multiple regression, the purpose is to develop a linear function (i.e., the best weighted combination of the independent variables) that predicts some quantitative measure (the criterion variable or dependent variable) from a set of other quantitative variables (predictors or independent variables). This linear equation is the prediction model. This prediction model in discriminant analysis is called the discriminant function. The discriminant function in the present example will weight the three predictor variables (pre-self-evaluation, preliminary clinical examination, and initial committee evaluation) such that the two criterion groups (whether or not the student passed the internship) are maximally differentiated. The students will thus be classified based on their score for the discriminant function into a pass or fail group. One of the ways we will evaluate the quality of the solution is by examining how accurately the students were classified into these groups.

In a multiple regression analysis, a predicted score can be calculated from the weighted combination of the independent variables. Similarly, in a discriminant function analysis, one can calculate a discriminant score (designated as D_i) from the weighted combination of the independent variables. The unstandardized discriminant coefficients are used in the formula for making the classifications, much as b coefficients are used in regression in making predictions. For each case, we multiply the score on each predictor by its discriminant coefficient or weight and add the constant. The result of this computation yields the discriminant score. The equation for the discriminant score is as follows:

$$D_i = \mathbf{a} + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

In this equation, D_i is the predicted score on the criterion variable, the *X*s are the predictor variables in the equation (i.e., pre-self-evaluation, preliminary clinical examination, and initial committee evaluation), and the bs are the weights or coefficients associated with the predictors. Analogous to multiple regression, these b weights reflect the relative contribution of its independent variable when the effects of all the other predictors are statistically controlled. The equation also contains a constant, shown as **a** in the equation (representing the *Y* intercept).

Unlike the multiple regression equation, which employs the ordinary least squares solution, the weights in the discriminant function are derived through the maximum likelihood method (the same strategy used in logistic regression). The maximum likelihood method is an iterative process that starts with an initial arbitrary "guesstimate" of the weights and then determines the direction and magnitude of the b coefficients to minimize the number of classification errors. This maximum likelihood technique ultimately assigns a case to a group from a specific discriminant cutoff score. The cutoff score is the one that results in the fewest classification errors. When group sizes are equal, the cutoff score is selected between the mean scores (the centroids) of the two groups. If the group sizes are unequal, the cutoff score is calculated from the weighted means. Because the mean discriminant score (centroid) of the unsuccessful interns in this example is −.782 and the centroid of the successful interns is +.782, the cutoff score is 0. Scores greater than 0 are predicted as belonging to the successful group, and scores equal to or less than 0 are predicted as belonging to the unsuccessful group.

For example, assume that a case has the following scores for the three-predictor variables: 2 (self-evaluation), 3.12 (exam score), 3 (committee evaluation), and 0 (the student was unsuccessful in the internship).

The discriminant function analysis produced the following weights: .703 (self-evaluation), 1.725 (exam score), and .092 (committee evaluation), with −8.043 as the **a** constant. The equation for this case would then be as follows:

$$\mathbf{2}(.703) + \mathbf{3.12}(1.725) + \mathbf{3}(.092) + (-8.043) = \mathbf{-.98}$$

Because this D_i is less than 0, this case is classified as unsuccessful, an accurate classification. Assume that another case had the following scores: 4 (self-evaluation), 4.09 (exam score), 3 (committee evaluation), and 1 (the student was successful). The equation for this case would then be as follows:

$$\mathbf{4}(.703) + \mathbf{4.09}(1.725) + \mathbf{3}(.092) + (\mathbf{-8.043}) = \mathbf{2.11}$$

Because this D_i is greater than 0, this case is classified as successful. Thus, this student also would be correctly classified on the basis of her score on the discriminant function.

Interpreting the Discriminant Function

The predictor variables have been aggregated (with appropriate weights) to create the discriminant function, which is conceptually identical to the variate in multiple regression. To infer the meaning of the discriminant function, we need to review the structure matrix and the standardized canonical discriminant function coefficients.

The structure matrix reports the discriminant loadings of the variables on the discriminant function. These discriminant loadings are the simple Pearson correlations between the predictor variables and the discriminant function. Usually, any predictor with a loading of .30 or more is considered to be central in defining the discriminant dimension. Researchers gain insight into how to label the function by identifying the largest absolute correlation associated with the discriminant function.

The standardized canonical discriminant function coefficients indicate the partial contribution of each variable to the discriminant function controlling for other predictors in the equation. Although the structure coefficients are used to assign meaningful labels to the discriminant function, the standardized discriminant function coefficients are used to assess each independent variable's unique contribution to the discriminant function. These are akin to beta weights in multiple regression and carry with them the same strengths and liabilities. After reviewing the structure matrix and the standardized canonical discriminant function coefficients, researchers may consider eliminating variables that do not significantly contribute to prediction (variables whose coefficients are not statistically significant).

Researchers may also want or need to assess the relative importance of the discriminating variables. Recall that in multiple regression, the relative importance can be assessed by comparing beta weights. In discriminant function analysis, the standardized canonical discriminant function coefficients are used to assess relative importance. Note that the standardized discriminant coefficients will change if variables are added to or deleted from the equation.

Tests of Significance

The null hypothesis for discriminant function analysis is that the means of the two groups on the discriminant function—the centroids—are equal. Several methods are available to test if the discriminant model as a whole is statistically significant, but we present only Wilks's lambda here. Wilks's lambda varies from 0 to 1 and tells us the variance of the categorical groups variable that is not explained by the discriminant function; thus researchers desire lower values.

Wilks's lambda is also used to test for significant differences between the groups on the individual predictor variables. The *F* test of Wilks's lambda indicates which variables are statistically significant—that is, which variables contribute a significant amount of prediction to help differentiate the groups. Those predictors failing to demonstrate statistically significant differences between the groups could be considered for deletion from the model. If the groups do not differ on the individual predictor variables, then it is unlikely the groups will differ on the discriminant function.

The canonical correlation coefficient is the measure of association between the discriminant function and the outcome variable. The square of the canonical correlation is the percentage of variance explained in the dependent variable.

Accuracy of Classification

The performance of the discriminant function can be evaluated by estimating error rates (probabilities of misclassification). The classification table, also called the prediction matrix, is simply a table in which the rows are the observed categories of the independent variable and the columns are the predicted categories of the independent variable. Such a table is shown in Figure 7a.1.

When prediction is perfect (which is not going to happen in empirical research), all cases will lie on the upper left to lower right diagonal. The

		Predicted Category	
		Group 1	Group 2
Observed Category	Group 1	**Hit**	**Miss**
	Group 2	**Miss**	**Hit**

Figure 7a.1 Classification Table

percentage of cases on the diagonal is the percentage of correct classifications, called the *hit ratio.* This hit ratio is not compared with zero but with the percentage that would have been correctly classified by chance. For our two-group discriminant analysis, the expected hit ratio is 50% (for *n*-way groups with equal splits, the expected percent is $1/n$). Mahalanobis D^2 Rao's *V,* Hotelling's trace, Pillai's trace, and Roy's gcr (see Chapters 10A and 10B) are other indexes of the validity of the discriminant function. Each of these indexes has its own significance tests. Some of these tests are used in stepwise discriminant analysis to determine if adding an independent variable to the model will significantly improve classification of the dependent variable.

Mahalanobis distances are also used in evaluating cases in discriminant analysis. A researcher might wish to analyze a new, unknown set of cases in comparison with an existing set of known cases. Mahalanobis distance is the distance between a case and the centroid for each group. A case will have 1 Mahalanobis distance for each group, and it will be assigned to the group with the smallest Mahalanobis distance. Because Mahalanobis distance is measured in terms of standard deviations from the centroid, a case that is more than 1.96 Mahalanobis distance units from the centroid has less than a .05 chance of belonging to the group represented by the centroid, and a case with 3 distance units would have less than a .01 chance of belonging to that group.

Different Discriminant Function Methods

Just as was true for multiple regression, there are several different methods by which researchers can build the variate or linear function. One important way that these methods differ is the order in which they call for the predictor variables to be entered into the discriminant function. In general, we start with a set of potential predictors of some criterion variable. We finish with either some or all these potential predictors in the discriminant function. The methods differ in how we get from the starting state to the finished state. Here are some of the available strategies:

- We can enter all the independent variables into the equation at once.
- We can enter the independent variables one or more at a time.
- We can sequentially enter only those independent variables that add predictive value to those variables already in the equation.
- We can first put all the independent variables in the equation and then remove those not contributing to prediction one at a time.

Similar to multiple regression, these different discriminant methods can be organized into two groups: the *statistical* methods (standard and stepwise) and the researcher-controlled methods known as *sequential, hierarchical,* or *covariance* methods. The statistical methods rely exclusively on statistical decision-making criteria built into the computer programs to decide during the development of the equation which predictors are to be entered and which predictors will be removed. In using the statistical methods, researchers allow the computer program to autonomously carry out the analyses.

The standard discriminant method, also known as the simultaneous or the direct method, is what most authors refer to if they leave the method unspecified. Under this method, all the predictors are entered into the equation in a single "step" (stage in the analysis). The standard method provides a full model solution in that all the predictors are part of it. The idea that these variables are entered into the equation simultaneously is true only in the sense that the variables are entered in a single step. But that single "step" is not at all simple and unitary; when we look inside this step, we will find that process of determining the weights for the independent variables is governed by a complex strategy similar to what we described for multiple regression.

The stepwise method of developing the discriminant function is another popular method. The stepwise method begins with an empty equation and builds it one step at a time. Once a third independent variable is in the equation, the stepwise method invokes the right to remove an independent variable if that predictor is not earning its keep. Predictors are allowed to be included in the equation if they significantly add to the predictive function. With correlated predictor variables, as we have seen, the predictors in the equation admitted under a probability level of .05 can still overlap with each other.

With t
architects
control ov
Exactly th
regression

Efron, B. (1
nant ar
Huberty, C.
Psycho
Huberty, C
Measu
Huberty, C.
in disc
Joachimsth
the cl
Behavi
Konishi, S.,
in disc
Compu
McLaughlin
Multiv
New Yo
Panel on d
analysi
Press, S. J.,
nant ar
Ragsdale, C
statisti
Spicer, J. (2
Sage.

With the hierarchical or covariance method, the researchers are the architects of the discriminant function equation. Here, researchers are in full control over the order of entry of variables in the discriminant function. Exactly the same sort of options are available as described for multiple regression, and readers are referred to Chapter 5A for that coverage.

Recommended Readings

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association, 70,* 892–898.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin, 95,* 156–171.

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

Huberty, C. J., Wisenbaker, J. M., & Smith, J. C. (1987). Assessing predictive accuracy in discriminant analysis. *Multivariate Behavioral Research, 22,* 307–329.

Joachimsthaler, E. A., & Stam, A. (1990). Mathematical programming approaches for the classification problem in two-group discriminant analysis. *Multivariate Behavioral Research, 25,* 427–454.

Konishi, S., & Honda, M. (1990). Comparison procedures for estimation of error rates in discriminant analysis under non-normal populations. *Journal of Statistical Computing and Simulation, 36,* 105–115.

McLaughlin, M. L. (1980). Discriminant analysis. In P. Monge & J. Cappella (Eds.), *Multivariate techniques in human communication research* (pp. 175–204). New York: Academic Press.

Panel on discriminant analysis, classification and clustering. (1989). Discriminant analysis and clustering. *Statistical Science, 4,* 34–69.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73,* 699–705.

Ragsdale, C. T., & Stam, A. (1992). Introducing discriminant analysis to the business statistics curriculum. *Decision Sciences, 23,* 724–745.

Spicer, J. (2005). *Making sense of multivariate data analysis.* Thousand Oaks, CA: Sage.