Assignment 3: Using Discriminant Analysis

Overview

For the third assignment we are going to work on developing a discriminant analysis. I have provided a couple of examples for you to work with for this assignment for classifying individual cases into groups. Discriminant analysis approaches the problem as a number of underlying dimensions that can separate the categories in multivariate space. We have \( C - 1 \) possible distinct functions (i.e., each function is a weighted set of variables uncorrelated with the other functions) for a variable with \( C \) categories. The first function always accounts for the most separation among groups (just like the first factor in factor analysis); the second accounts for the second most separation, and so forth. Note that functions are evaluated for statistical significance together and then again after removing the first function. You should probably let the program use the prior proportions in each category from the data in this instance to facilitate classification, since the data are not equally distributed across categories.

No special problems are posed by unequal sample sizes; however, the sample size of the smallest group should exceed the number of predictors. Tabachnick and Fidell (1996) suggest, for multivariate normality, as a conservative estimate, if there are at least 20 cases in the smallest group and only a few predictors (say 5 or so), results should be robust against violations of multivariate normality. We can examine Box’s M, but that is generally too sensitive to violations. Similar to MANOVA, discriminant analysis is robust to failures of multivariate normality if the violation is caused by skewness (rather than outliers). A sample size that would produce 20 \( df \) for error in the univariate ANOVA case should ensure robustness with respect to multivariate normality, as long as sample sizes are relatively equal and two-tailed tests are used. In cases where homogeneity of variance-covariance matrices may be a concern (e.g., small sample sizes), one can transform predictors or use separate covariance matrices during classification (this is an option in defining the model). As Tabachnick and Fidell note, since this can lead to over-fitting the model, it should be used only when the sample is large enough to permit cross-validation.

Example 1

In the first example, a researcher is interested in investigating what combination of variables best classifies employees into their job categories of managers (coded 3), custodians (coded 2) and clerical (coded 1). In this first case, the research hypothesis is that variables like education, experience, and salary data should be more useful in determining employees’ job categories. A secondary goal is to determine whether gender and race/ethnicity contribute in classifying employees by job category. The hypothesis here is that they should not differentiate employees by job category.

In this case, we can use the following information about employees: beginning and current salary, education (years), previous experience (months), female (0 = male, 1 = female), race/ethnicity (coded 0 = white, 1 = minority by race/ethnicity).
Steps

1. Examine the descriptive statistics and correlations between predictors.
2. Develop a predictive model. Examine the output to determine how well the model aids in classifying employees into their job categories.
3. Develop a plot (i.e., plot the combined groups using within-group covariance matrix) of individuals and their group centroids. Name the underlying functions on the X and Y axes.
4. Report on your findings and offer any recommendations.

Example 2

Example 2 focuses on classifying cars according to the regions in which they were made. A researcher might be interested in knowing whether there are any common characteristics among cars manufactured in particular regions of the world. In this study, the three regions are the United States (coded 1), Asia (coded 2), and Europe (coded 3). Variables include sales, price, engine size, fuel efficiency, horsepower, and curb weight. I transformed several of the predictors by taking their natural logs (i.e., sales, price, engine size, fuel efficiency, and horse power) in order to reduce the skewness and kurtosis. I left both the logged and raw metric measures in the data set so you could evaluate the difference.

Steps

1. Examine the descriptive statistics and correlations between predictors.
2. Develop a predictive model. Examine the output to determine how well the model aids in classifying cars according to where they were made.
3. Develop a plot (i.e., plot the combined groups using within-group covariance matrix) of individual cars and their group centroids. Name the underlying functions on the X and Y axes.
4. Report on your findings and offer any recommendations.

Reference