Israel Nolazco

israelnolazco@lewis.edu

A Summary of Multiple Discriminant Analysis

August 9, 2020

Table of Contents

1. What is it?
2. How are the scores created?
3. How do we know it works?
4. How are the categories chosen?
5. Research the design
6. Verifying the Assumptions
7. Estimation of the Discriminant Model and Assessing Overall Fit

**What is it?**

Discriminant Analysis is a type of model that looks at categorical dependent variables and compares them to the independent variables (metric). This comparison gives the independent variables weight in order to categorize them and determine when a variate falls within a group. As an example, good credit vs bad credit based on the customer’s metric variables (income, debt, etc. ).

**How are the scores created?**

Discriminant Analysis involves deriving a variate through discriminant function. This function has the appearance of a linear function, which can be seen in the following:

Zjk= a + W1K1k+ W2K2k+ … + WnKnk

Where:

Zjk = Discriminant Z score of discriminant function j for object k

a = intercept

Wi = discriminant weight for independent variable i

Xik = independent variable i for object k

**How do we know it works?**

Having created a linear model that looks into several independent variables and determine two or more categories those variables fall within, how can we tell those categories are significantly different from one another. A clever way to do so is by measuring their centroids. A centroid is measuring the group mean of each category and then sees how far from each centroid is from one another. If the overlap in the distribution is small, the discriminant function separates the groups well, if the overlap is large, then the discriminant function is poor.

**How are the categories chosen?**

We first get started by determining if there are statistically significant differences that exist between the two (or more) a priori defined groups. Then identify the relative importance of each independent variable in predicting group membership (centroids). Establish the number and composition of the dimensions of discrimination between groups formed from the set of independent variables. The number of significant functions determines the "dimensions“ / discriminant functions and what they represent in distinguishing the groups. Finally, we develop procedures for classifying objects (individuals, firms, products, etc.) into groups, and then examining the predictive accuracy (hit ratio) of the discriminant function to see if it is acceptable (> 25% increase).

**Research the design**

After examining the categories, we must then look at the dependent variables. The groups being derive from those dependent variables should be mutually exclusive and exhaustive. Additionally, let us not forget of instances where there are no non-metric variables for our dependent variables, i.e. the dataset is express with metric values rather than no-metric. In that case, we must look into converting a metric scale to a non-metric scale by comparing extremes of each case or develop subcategories in that scalar metric. For example, use a question asking the typical number of soft drinks consumed per day and develop a three-category variable of 0 drinks for non-users, 1 – 5 for light users, and 5 or more for heavy users.

**Verifying the Assumptions**

It is without a doubt recognize that this type of model analysis requires some sort of inherent assumptions. The dependent variable and independent variables fall within a group and mathematically speaking it could make sense; however, in real scientific world causation does not equal correlation. Therefore, it is important to identify any violations of the assumptions. The most common way to do so is by assessing the equality of covariance matrices with the Box’s M test. In this test, the researcher is looking for a nonsignificant probability level which would indicate that there were no differences between the group covariance matrices. A rule of thumb is to apply a conservative significance level of .01 and become even more conservative as the analysis becomes more complex with a larger number of groups and/or independent variables.

**Estimation of the Discriminant Model and Assessing Overall Fit**

After verifying the assumptions and clearing any doubts regarding the analysis. We must select an estimation model. A Simultaneous Estimation involves computing the discriminant function so that all of the independent variables are considered concurrently. Stepwise Estimation is an alternative to the simultaneous approach. It involves entering the independent variables into the discriminant function one at a time on the basis of their discriminating power. Although stepwise estimation may seem “optimal” by selecting the most parsimonious set of maximally discriminating variables, beware of the impact of multicollinearity on the assessment of each variable’s discriminatory power. Overall model fit assesses the statistical significance between groups on the discriminant Z score(s) but does not assess predictive accuracy. With more than two groups, do not confine your analysis to only the statistically significant discriminant function(s), but consider if nonsignificant functions (with significance levels of up to .3) add explanatory power. Then we define the prior probabilities based either on the relative sample sizes of the observed groups or specified by the researcher (either assumed to be equal or with values set by the researcher), and calculate the optimum cutting score value as a weighted average based on the assumed sizes of the groups (derived from the sample sizes).

**Interpretation of the Results**

There are three methods to interpret the results: Standardized discriminant weights, Discriminant loadings (structure correlations), and Partial F values. The standardized discriminant weights examine the sign and magnitude of the standardized discriminant weight. Discriminant loadings referred to sometimes as structure correlations and are increasingly used as a basis for interpretation because of the deficiencies in utilizing weights. As for partial F values is accomplished by examining the absolute sizes of the significant F values and ranking them. Large F values indicate greater discriminatory power.