Discriminant Analysis

MA4128 Advanced Data Modelling

April 23, 2013

Contents

1 Discriminant Analysis

The major purpose of discriminant analysis is to predict membership in two or more mutually exclusive groups from a set of predictors, when there is no natural ordering on the groups. So we may ask whether we can predict whether people vote Labour or Conservative from a knowledge of their age, their class, attitudes, values etc etc.

1.1 The purposes of discriminant analysis (DA)

Discriminant Function Analysis (DA) undertakes the same task as multiple linear regression by predicting an outcome. However, multiple linear regression is limited to cases where the dependent variable on the Y axis is an interval variable so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values for given values of weighted combinations of X values. But many interesting variables are categorical, such as political party voting intention, migrant/non-migrant status, making a profit or not, holding a particular credit card, owning, renting or paying a mortgage for a house, employed/unemployed, satisfied versus dissatisfied employees, which customers are likely to buy a product or not buy, what distinguishes O'Briens customers from Starbucks clients, whether a person is a credit risk or not, etc.

1.2 Discriminant Analysis and Other Types of Analysis

Discriminant analysis is just the inverse of a one-way MANOVA, the multivariate analysis of variance. The levels of the independent variable (or factor) for MANOVA become the categories of the dependent variable for discriminant analysis, and the dependent variables of the MANOVA become the predictors for discriminant analysis. In MANOVA we ask whether group membership produces reliable differences on a combination of dependent variables. If the answer to that question is 'yes' then clearly that combination of variables can be used to predict group membership. Mathematically, MANOVA and discriminant analysis are the same; indeed, the SPSS MANOVA command can be used to print out the discriminant functions that are at the heart of discriminant analysis, though this is not usually the easiest way of obtaining them.

These discriminant functions are the linear combinations of the standardized independent variables which yield the biggest mean differences between the groups. If the dependent variable is a dichotomy, there is one discriminant function; if there are k levels of the dependent variable, up to k-1 discriminant functions can be extracted, and we can test how many it is worth extracting. Successive discriminant functions are orthogonal to one another, like principal components, but they are not the same as the principal components you would obtain if you just did a principal components analysis on the independent variables, because they are constructed to maximise the differences between the values of the dependent variable.

The commonest use of discriminant analysis is where there are just two categories in the dependent variable; but as we have seen, it can be used for multi-way categories (just as MANOVA can be used to test the significance of differences between several groups, not just two). This is an advantage over logistic regression, which is always described for the problem of a dichotomous dependent variable.

You will encounter discriminant analysis fairly often in journals. But it is now being replaced with multinomial logistic regression, as this approach requires fewer assumptions in theory, is more statistically robust in practice, and is easier to use and understand than discriminant analysis.

- Discriminant function analysis is multivariate analysis of variance (MANOVA) reversed.
- In MANOVA, the independent variables are the groups and the dependent variables are the predictors.
- In Discriminant Analysis , the independent variables are the predictors and the dependent variables are the groups.

As previously mentioned, Discriminant Analysis is usually used to predict membership in naturally occurring groups. It answers the question: can a combination of variables be used to predict group membership? Usually, several variables are included in a study to see which ones contribute to the discrimination between groups.

1.3 Assumptions of Discriminant Analysis

The major underlying assumptions of DA are:

- the observations are a random sample;
- each predictor variable is normally distributed;
- each of the allocations for the dependent categories in the initial classification are correctly classified;

- there must be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group);

- each group or category must be well defined, clearly differentiated from any other group(s) and natural. Putting a median split on an attitude scale is not a natural way to form groups. Partitioning quantitative variables is only justifiable if there are easily identifiable gaps at the points of division; for instance, three groups taking three available levels of amounts of housing loan; the groups or categories should be defined before collecting the data;

- the attribute(s) used to separate the groups should discriminate quite clearly between the groups so that group or category overlap is clearly non-existent or minimal;

- group sizes of the dependent should not be grossly different and should be at least five times the number of independent variables.

1.4 Steps in Discriminant Analysis

Discriminant function analysis is broken into a 2-step process:

(1) testing significance of a set of discriminant functions,

- (2) classification

. The first step is computationally identical to MANOVA. There is a matrix of total variances and covariances; likewise, there is a matrix of pooled within-group variances and covariances. The two matrices are compared via multivariate $F - tests$ in order to determine whether or not there are any significant differences (with regard to all variables) between groups. One first performs the multivariate test, and, if statistically significant, proceeds to see which of the variables have significantly different means across the groups.

1.5 Discriminant Function

Discriminant Analysis involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is:

$$D = v_1 X_1 + v_2 X_2 + \ldots + a$$

Where D = discriminate function
v = the discriminant coefficient or weight for that variable
X = respondents score for that variable
a = a constant
i = the number of predictor variables

This function is similar to a regression equation . The **v**s are unstandardized discriminant coefficients analogous to the **b**s in the regression equation. These vs maximize the distance between the means of the criterion (dependent) variable. Standardized discriminant coefficients can also be used like beta weight in regression.

Good predictors tend to have large weights. What you want this function to do is maximize the distance between the categories, i.e. come up with an equation that has strong discriminatory power between groups.

After using an existing set of data to calculate the discriminant function and classify cases, any new cases can then be classified. The number of discriminant functions is one less the number of groups. There is only one function for the basic two group discriminant analysis.

The coefficients for the first discriminant function are derived so as to maximize the differences between the group means. The coefficients for the second discriminant function are derived to maximize the difference between the group means, subject to the constraint that the values on the second discriminant function are not correlated with the values on the first discriminant function, and so on.

In other words, the second discriminant function is **orthogonal** to the first, and the third discriminant function is orthogonal to the second, and so on. The maximum number of unique functions that can be derived is equal to the number of groups minus one or equal to the number of discriminating variables, whichever is less.

The discriminant functions are generated from a sample of individuals (or cases), for which group membership is known. The functions can then be applied to new cases with measurements on the same set of variables, but unknown group membership.

- A latent variable of a linear combination of independent variables
- One discriminant function for 2-group discriminant analysis
- For higher order discriminant analysis, the number of discriminant function is equal to $n - 1$ (where n is the number of categories of dependent/grouping variable).

- The first function maximizes the difference between the values of the dependent variable.
- The second function maximizes the difference between the values of the dependent variable while controlling the first function.etc etc
- The first function will be the most powerful differentiating dimension.
- The second and later functions may also represent additional significant dimensions of differentiation

1.6 Assumptions of discriminant analysis

The major underlying assumptions of DA are:

- the observations are a random sample;
- each predictor variable is normally distributed;
- each of the allocations for the dependent categories in the initial classification are correctly classified;
- there must be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group);
- each group or category must be well defined, clearly differentiated from any other group(s) and natural. Putting a median split on an attitude scale is not a natural way to form groups. Partitioning quantitative variables is only justifiable if there are easily identifiable gaps at the points of division;
- for instance, three groups taking three available levels of amounts of housing loan; the groups or categories should be defined before collecting the data;
- the attribute(s) used to separate the groups should discriminate quite clearly between the groups so that group or category overlap is clearly non-existent or minimal;
- group sizes of the dependent should not be grossly different and should be at least five times the number of independent variables.

There are several purposes of DA:

- To investigate differences between groups on the basis of the attributes of the cases, indicating which attributes contribute most to group separation. The descriptive technique successively identies the linear combination of attributes known as canonical discriminant functions (equations) which contribute maximally to group separation.
- Predictive DA addresses the question of how to assign new cases to groups. The DA function uses a persons scores on the predictor variables to predict the category to which the individual belongs.
- To determine the most parsimonious way to distinguish between groups.
- To classify cases into groups. Statistical significance tests using chi square enable you to see how well the function separates the groups.
- To test theory whether cases are classified as predicted.

1.7 Discriminant Score

The aim of the statistical analysis in DA is to combine (weight) the variable scores in some way so that a single new composite variable, the discriminant score, is produced. A discriminant score is a weighted linear combination (sum) of the discriminating variables.

1.8 Discriminant Analysis : Comparison to Logistic Regression

Discriminant function analysis is very similar to logistic regression, and both can be used to answer the same research questions. Logistic regression does not have as many assumptions and restrictions as discriminant analysis. However, when discriminant analysis assumptions are met, it is more powerful than logistic regression. Unlike logistic regression, discriminant analysis can be used with small sample sizes. It has been shown that when sample sizes are equal, and homogeneity of variance/covariance holds, discriminant analysis is more accurate.With all this being considered, logistic regression is the common choice nowadays, since the assumptions of discriminant analysis are rarely met.

Discriminant analysis uses a collection of interval variables to predict a categorical variable that may be a dichotomy or have more than two values. The technique involves finding a linear combination of predictors variables the discriminant function that creates the maximum difference between group membership in the categorical dependent variable.