10

# DISCRIMINANT ANALYSIS: A METHOD FOR DETERMINING RELATIVE IMPORTANCE OF PREDICTOR VARIABLES

P.S. Swathi Lekshmi

### Introduction

Discriminant analysis is a technique designed to characterize the relationship between a set of variables often called the response or predictor variables and a grouping variable with a relatively small number of categories. To do so, discriminant creates a linear combination of the predictors that best characterize the differences among the groups. The technique is related to both regression and multivariate analysis of variance and as such it is another general linear modeling technique. Another way to think of discriminant analysis is as a method to study difference between two or more groups of cases on several variables simultaneously. This technique was developed by Sir Ronald Fischer in 1936. Discriminant function analysis is useful in determining whether a set of variables is effective in predicting category membership. (Green *et al*., 2008)

## The Elements of a Discriminant Analysis

The general procedure of doing a predictive discriminant analysis (PDA) is outlined as follows.

1. A grouping variable must be defined whose categories are exhaustive and mutually exclusive.

2. A set of potential predictors must be selected. This is one of the most important steps, although in many real world applications, set of predictors will be limited by what is available in existing dataset.

3. Once the above two steps are accomplished, as with any multivariate technique the next job is to study the data to see if it meets the assumption of doing a discriminant analysis. It is also important to look for outliers and unusual patterns in the data and to look for variables that might not be good predictors. Univariate ANOVAs and correlations can be used to identify such variables.

4. The goal of a PDA is to correctly classify cases into the appropriate group. Given this, as with any multivariate technique parsimony is an important sub goal. This means, using the fewest predictors needed for accurate classification, although not necessarily the smallest set of classification functions. Fewer predictors will mean lower cost of data collection and easier interpretation.

5. The discriminant analysis must be specified and run using statistical software such as the SPSS. A method of model selection must be chosen and prior probabilities for group membership should be considered. A significant test is available to see whether the difference in group means on each function is due to chance or not. The relative importance (in terms of explained variance) of each function is also calculated.

6. Use the classification result to see how well cases have been placed in their known groups.

7. At least two statistics are available to examine the effect of individual predictors on the discriminant functions and in particular to decide whether a particular variable adds little to the classification ability of the model.

8. Look for outliers in the data and examine cases that have been misclassified to check for problems and to see if and how the model can be re-specified.

9. Finally, it is of the utmost importance that the model be validated by some procedure.

## The Discriminant Model: Methodological Framework

Discriminant function analysis is a statistical technique which allows for the study of the differences between two or more groups with respect to several variables simultaneously and provides a means of classifying any object / individual into the group with which it is most closely associated and for assessing the relative importance of each variable used to discriminate between different groups. A linear combination of predictor variables, weighted in such a way that it will best discriminate among groups with the least error is called a linear discriminant function and is given by:

$$D = L_1X_1 + L_2X_2 + \cdots + L_KX_K$$

Where, $X_1$, $X_2$ .... $X_k$ are predictor variables, $L_1$, $L_2$ .... $L_K$ represent the discriminant coefficient, and D is the value of the discriminant function of a particular individual, such that if this value is greater than a certain critical value D, the individual is classified in group I (e.g. a high adopter group), and otherwise the individual would be classified in group II. (e.g. a low adopter group). In the foregoing example, the respondents were classified into two groups, namely low adoption group and high adoption group, based on the mean adoption score. The predictor variables used for the study were the attributes of shrimp culture technologies, perception of cost of technologies, and perception of policies affecting shrimp culture.

## Discriminant Analysis: An example from fisheries sector

Discriminant function analysis in relation to 12 attributes, cost and policy between the high and low adoption categories of 60 shrimp farmers of Nellore, Andhra Pradesh was studied. (Lekshmi *et al*., 2007)

The Mahalanobis $D^2$ value and discriminant function coefficient were computed, to find out the difference between the attributes, cost and policy perceptions of high and low adoption categories of shrimp farmers of Nellore when all the fourteen variables (twelve attributes, perception of cost, and policy) were considered together. The results are presented in Table 1.

**Table 1: Discriminant function analysis in relation to the relative importance of variables in discriminating between the groups** (n=60)

| Sl. No. | Variables | Discriminant function coefficient l (i) | Relative importance (%) |
|---|---|---|---|
| | Efficiency ($X_1$) | 1.0584 | 100.78 |
| | Feasibility ($X_2$) | 0.4455 | -0.788 |
| | Immediacy of returns ($X_3$) | 0.0194 | 0 |
| | Physical compatibility ($X_4$) | -0.0433 | 0 |
| | Observability ($X_5$) | -0.1857 | 0 |
| | Profitability ($X_6$) | -0.4232 | 0 |
| | Perceived risk ($X_7$) | 0.5651 | 0 |
| | Input availability ($X_8$) | -0.4461 | 0 |
| | Cost ($X_9$) | 0.2485 | 0 |
| | Total | | 100 |

Note: $D^2$ = 0.3505 High group ($n_1$) = 31 Low group ($n_2$) = 29 f = 20.56**

As could be seen from Table 1, the $D^2$ value was found to be 0.3505 and the f value was found to be highly significant at one per cent level of significance. Therefore, it could be concluded that the fourteen variables (consisting of perception of twelve attributes, perception of cost and perception of policy) were significantly discriminating between the high and low adoption categories of shrimp farmers.

Thus the null hypothesis, that there will be no difference between the perception of attributes, cost and policy by high and low adoption categories of shrimp farmers is rejected.

Table 1 reveals that out of the fourteen variables studied, 8 variables had shown significant positive influence in differentiating the high from the low adoption categories of shrimp farmers. The 8 variables in the descending order of their importance were efficiency (1.0584), perceived risk (0.5651), feasibility (0.4455), policies (0.3330), cost (0.2485), complexity (0.04313), immediacy of returns (0.0194), and multiple advantages (0.0055).

This indicated that the increased differential scores in these variables would increase the difference between the high and low adoption categories. It suggested that the respondents who scored high in these variables (individuals having higher perception of efficiency, perceived risk, feasibility, policies, cost, complexity, immediacy of returns, and multiple advantages, might have differentiated more significantly between the high and low adoption categories, among the shrimp farmers.

The analysis also revealed that the remaining 6 variables viz., input availability (-0.4461), profitability (-0.4232), trialability (-0.3247), observability (-0.1857), physical compatibility (-0.0433) and cost of technologies (-0.0163) had shown significant negative discriminant function coefficients in the descending order of their importance. The analysis also revealed that these variables had shown significant negative influence in differentiating the high adoption category and low adoption categories. This suggested that the respondents who scored high in these variables (respondents with high perception of input availability, profitability, Trialability observability, physical compatibility and cost of technologies) might have differentiated less between the high and low adoption categories of shrimp farmers.

Further observation of Table 1, shows the relative importance of the variables in discriminating between the high and low adoption categories. It could be seen from the table that the variables having substantial importance in the classification of shrimp farmers in to the high adoption category (first group) and low adoption category (second group) were efficiency and feasibility with a relative importance of 100.78 and– 0.788 percent respectively.

The Discriminant function fitted was, $D = L_1X_1 + L_2X_2 + \cdots + L_KX_K$ , where D is the value of the discriminant function of an individual shrimp farmer, $X_i$'s are the predictor variables and Li's represents the discriminant coefficients. The estimated function takes the form following form:

$$D = 1.0584\ X_1 + 0.4455\ X_2 + 0.0194\ X_3 - 0.0433\ X_4 - 0.1857\ X_5 - 0.4232\ X_6 + 0.5651\ X_7 - 0.4461\ X_8 + 0.2485\ X_9 + 0.04313\ X_{10} - 0.3247\ X_{11} + 0.0055\ X_{12} - 0.0163\ X_{13} + 0.3330\ X_{14}$$

The significance of the function was tested using the following analysis of variance presented in Table 2.

**Table 2: Analysis of variance for discriminant function**

| Source | Degrees of freedom | Sum of Squares | NS | F-Value |
|---|---|---|---|---|
| Between population | 14 | 2.73 | 0.19 | $20.56^{**}$ |
| Within population | 15 | 0.42 | 9.48 | |

## Discriminant scores for categories I and II were

$D_1 = 5.3149$ $D_2 = 5.0055$

$D^* = \frac{5.314+5.005}{2} = 5.16$, where D* is the critical value

If the Discriminant score, D is greater than the critical value (D*) then the individual is assigned to the first category i.e., high adoption category, otherwise the individual is assigned to the second category i.e. low level of adoption. The classification of the shrimp farmers into high and low adoption categories is presented in Table 3.

**Table 3: Classification of respondents in to high and low adopter categories based on discriminant function** (n=60)

| Adopter category | Assigned locations using discriminatory function | | Total |
|---|---|---|---|
| | High | Low | |
| High | 30 | 1 | 31 |
| Low | 28 | 1 | 29 |
| Total | 58 | 2 | 60 |

From Table 3, it is observed that, out of the 60 farmers in Nellore district, 31 farmers were correctly classified. Hence the percentage of correct classification is 51.66 per cent. The significance of the F value as well as the per cent of correct classification of shrimp farmers, using the observed values, clearly indicates the overall significance and adequacy of the model.

## Conclusion:

The discriminant analysis helps us in finding out the independent variables which best differentiate between two given categories of individuals or cases. It also helps to classify or assign individuals to a particular category to which they belong. It helps researchers and technology developers in analyzing the important attributes of a particular technology which would help in increasing its adoption among end users.

## Suggested Readings:

- Cohen, J., Cohen, P., West, S.G. and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences*, 3rd Edition. Taylor & Francis Group.

- Green, S.B. Salkind, N. J. & Akey, T. M. (2008). Using SPSS for Windows and Macintosh: Analyzing and understanding data. New Jersey: Prentice Hall.

- Swathi Lekshmi, P.S, Balasubramani, N, Deboral Vimala, D and Chandrakandan, K (2007) Factors responsible for discriminating between high and low adopter categories of shrimp farmers. *Fishery Technology*, 44 (1). pp. 113-116.