*IOSR Journal of Mathematics (IOSR-JM)*
*e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 11, Issue 1 Ver. V (Jan - Feb. 2015), PP 12-15*
*www.iosrjournals.org*

# An Overview and Application of Discriminant Analysis in Data Analysis

Alayande, S. Ayinla[1] Bashiru Kehinde Adekunle[2]

*Department of Mathematical Sciences College of Natural Sciences Redeemer's University Mowe, Redemption City Ogun state, Nigeria*

*Department of Mathematical and Physical science College of Science, Engineering &Technology Osun State University*

***Abstract:*** *The paper shows that Discriminant analysis as a general research technique can be very useful in the investigation of various aspects of a multi-variate research problem. It is sometimes preferable than logistic regression especially when the sample size is very small and the assumptions are met. Application of it to the failed industry in Nigeria shows that the derived model appeared outperform previous model build since the model can exhibit true ex ante predictive ability for a period of about 3 years subsequent.*

***Keywords:*** *Factor Analysis, Multiple Discriminant Analysis, Multicollinearity*

## I. Introduction

In different areas of applications the term "discriminant analysis" has come to imply distinct meanings, uses, roles, etc. In the fields of learning, psychology, guidance, and others, it has been used for prediction (e.g., Alexakos, 1966; Chastian, 1969; Stahmann, 1969); in the study of classroom instruction it has been used as a variable reduction technique (e.g., Anderson, Walberg, & Welch, 1969); and in various fields it has been used as an adjunct to MANOVA (e.g., Saupe, 1965; Spain & D'Costa, Note 2). The term is now beginning to be interpreted as a unified approach in the solution of a research problem involving a com-parison of two or more populations characterized by multi-response data. Discriminant analysis as a general research technique can be very useful in the investigation of various aspects of a multi-variate research problem. In the early 1950s Tatsuoka and Tiede-man (1954) emphasized the multiphasic character of discriminant analysis: "(a) the establishment of significant group-differences, (b) the study and 'explanation' of these differences, and finally (c) the utilization of multivariate information from the samples studied in classifying a future individual known to belong to one of the groups represented" (p. 414). Essentially these same three problems related to discriminatory analysis.

Originally developed in 1936 by R.A. Fisher, Discriminant Analysis is a classic method of classification that has stood the test of time. Discriminant analysis often produces models whose accuracy approaches (and occasionally exceeds) more complex modern methods.

Discriminant analysis can be used only for classification (i.e., with a categorical target variable), not for regression. The target variable may have two or more categorical data.

The objective of a discriminant analysis is to classify objects, by a set of independent variables, into one of two or more mutually exclusive and exhaustive categories. For example, on the basis of an applicant's age, income, length of service, time at present home, etc., a teacher want to categorise student as either good or bad. For notation, let $X_{ij}$ be the i<sup>th</sup> individual's value of the j<sup>th</sup> independent variable $b_j$ be the discriminant coefficient for the j<sup>th</sup> variable $K_i$ be the i<sup>th</sup> individual's discriminant score, and $k_{erit}$ be the critical value for the discriminant score. Under the linear classification procedure, let each individual's discriminant score $K_i$ be a linear function of the independent variables. That is, $Ki = \beta_0 + \beta_1 x_{i1} + \beta_2 X_{2i} + ... \beta_i x_{ni}$ (1)

The classification procedure follows: if $K_i > k_{rit.,}$ classify Individual i as belonging to Group 1; if $K_i < k_{rit.}$ , classify Individual i as belonging to Group 2. The classification boundary will then be the locus of points, where $\beta_0 + \beta_1 x_{i1} + .......... + \beta_n x_{ni} = k_{rit}$

When n (the number of independent variables) = 2, the classification boundary is a straight line. Every individual on one side of the line is classified as Group 1; on the other side, as Group 2. When n = 3, the classification boundary is a two-dimensional plane in 3- dimensional space; the classification boundary is generally an n - 1 dimensional hyperplane in n space.

**Different Types of Discriminant Analysis**

Multiple Discriminant Analysis
Linear Discriminant Analysis
K-NNs Discriminant Analysis

There is no best discrimination method. A few remarks concerning the advantages and disadvantages of the methods studied are as follows.

- Analytical simplicity or computational reasons may lead to initial consideration of linear discriminant analysis or the NN-rule.
- Linear discrimination is the most widely used in practice. Often the 2-group method is used repeatedly for the analysis of pairs of multigroup data (yielding $\frac{k(k-1)}{2}$ decision surfaces for k groups).
- To estimate the parameters required in quadratic discrimination more computation and data is required than in the case of linear discrimination. If there is not a great difference in the group covariance matrices, then the latter will perform as well as quadratic discrimination.

The discriminant model has the following assumptions:

- Multivariate Normality
- Data values are from a normal distribution. We can use a normality test to verify this. However, please note that normal assumptions are usually not "fatal". The resultant significance tests may still be reliable.
- Equality of variance-covariance within group
- The covariance matrix within each group should be equal. Equality Test of Covariance Matrices can be used to verify it. When in doubt, try re-running the analyses using the Quadratic method, or by adding more observations or excluding one or two groups.
- Low multicollinearity of the variables

When high multicollinearity among two or more variables is present, the discriminant function coefficients will not reliably predict group membership. We can use the pooled within-groups correlation matrix to detect multicollinearity. If there are correlation coefficients larger than 0.8, exclude some variables or use Principle Component Analysis first.

**Preparing Data for the Analysis**

- Enough sample size
- Independent random sample (no outliers)
- Discriminant analysis requires that the observations are independent of one another, i.e., no repeated measures or matched pairs data
- Selecting Proper Variables Suppressor variables should be excluded. We can judge by observing the Univariate ANOVA table
- Dividing The Sample

**Reasons Why Discriminant Analysis is better than Logistics 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.

**Discriminant Analysis as Part of a System for Classifying Cases in Data Analysis**

Usually discriminant analysis is presented conceptually in an upside down sort of way, where what you would traditionally think of as dependent variables are actually the predictor variables, and group membership rather than being the levels of the IV are groups whose membership is being predicted

When DA is used in this predictive way it is usually followed up by classification procedures to classify new cases based on the obtained discriminant function(s

**Evaluation Criteria for Discriminant Analysis**

When results of a discriminant analysis are obtained, there are three basic questions to ask: (1) Which independent variables are good discriminators? (2) How well do these independent variables discriminate among the two groups? (3) What decision rule should be used for classifying individuals? More complete answers to these questions require a synopsis of the theoretical derivation of the discriminant function. The other steps to look for are;

(i) Deriving the Discriminant Function, and (ii)Determining the Effect of Independent Variables

**Application:** Most of the company/firms in Nigeria that are so viable in the 70's to 80's suddenly disappear, especially from the financial sector. This is major reason of using data from Nigerian economy. The set of variables for the discriminant analysis was chosen using stepwise selection. Variables were chosen to enter or leave the model using the significance level of an F test from an analysis of covariance, where the already selected variables act as covariates and the variable under consideration is the dependent variable.

In our analysis we selected the significance level for adding or retaining variables in the model to be 0.05. All the 31 ratios for every firm were calculated and the stepwise selection was done among these variables 1, 2, 3 and 4 years prior to failure futile. The variables that were selected into the discriminant analysis models are presented in Table 1

**Table 1. Variables selected for discriminant analysis.**

| Year | One Year to failure | Two years to failure | Three years to failure | Four Years to failure |
|---|---|---|---|---|
| | K1<br>K6<br>K7 | K12<br>K21<br>k16<br>K4<br>K15 | K2<br>K5<br>K10<br>K11 | K17<br>K22<br>K25<br>K40<br>K20<br>K13<br>K19 |

To analyse the models we divided the group of 31 original ratios into three very general dimension, namely liquidity (L), solidity (S), and profitability (P) measures and stockburn(ST). It is obvious that some of the ratios are rather measuring some other elements like effectiveness than any of these three, but to make the analysis more simple we did this rough classification. The stepwise model used for discriminant analysis selects two liquidity measures, one profitability measure and one solidity measure, one stockburn measure one year prior to failure. Two years prior to failure it selects one liquidity measure, three solidity measures and two profitability measures and two stockburn measure, and three years prior to failure the corresponding measures are two liquidity, one stckburn and one solid measure.

**Factor analysis**

The first stage in any multivariate analysis of this nature should be a factor analysis of the data set to identify its underlying dimensionality, aid interpretation of the derived models and avoid the inclusion of variables in the computed functions measuring closely related aspects of the firm. Such variable parsimony not only reduces the complexity of a multivariate statistical model, with little if any decrease in its efficiency, but also reduces the likelihood of sample bias being present in the model's construction

To study further if the models really are measuring different economic characteristics of a firm we applied factor analysis using all variables included in original data one, two, and three years prior to failure, separately. This was done to find out if the variables in alternative models are describing different financial dimensions so that the selection of one variable into the model is not only a consequence of extremely small differences in the value of test statistics.

To study further the consequences of different model selection approaches we have applied corresponding statistical method to test the predictive ability of constructed models.

Varimax rotated principal component analysis of the 30 ratio set used in the main analyses of the study was undertaken for the superior17 and Futile13 firms considered both together and separately. The relative importance of the variable is measured as suggested by the approach apparently originally by Mosteller and Wallace (1963, p. 283) for measuring the relative discriminant power of the j$^{th}$ variable between the two groups, subscripted 1 and 2, viz.:

$$r_j = \frac{c_j\left(\bar{x}_{j1} - \bar{x}_{j2}\right)}{\sum_{i=1}^{p} c_i\left(\bar{x}_{i1} - \bar{x}_{i2}\right)}$$

representing the proportion of the Mahalanobis distance accounted for by the j$^{th}$ variable in a p-variable linear discriminant model with coefficients $c_i$ and variables $x_i$ would appear to overcome such problems and make intuitive sense although not generalizable to more than two group.

| | **Standard Coefficient** | **Mostellr and Wallace (5) contribution** | **Conditinal delection (F-value)** |
|---|---|---|---|
| Solidity | 0.62 | 34.2 | 53.4 |
| Liquidity | -0.48 | 33.9 | 70.5 |
| Profitability | 0.53 | 4.2 | 6.4 |
| Stockburn | 0.61 | 6.2 | 11.1 |
| | | | |

Gnanadesikan (1977, Section 6.4) describes a number of approaches for testing for multivariate normality including the use of principal component residuals. Malkovich and Afifi (1973) also discuss various methods including a multivariate generalization of the univariate Shapiro and Wilks test criteria. The classification procedure adopted here explicitly took account of differential prior probability estimates and misclassification costs in determining the appropriate classification criterion in I, where I is the likelihood or probability-cost ratio given by the odds ratio x the loss ratio, viz.:

$$\frac{P_1}{P_2} X \frac{C_1}{C_2}$$

where $P_1$ and $P_2$ represent the prior probability estimates of an insolvent or solvent firm and $C_1$ and $C_2$ the estimated costs of type I and type II errors (Eisenbeis and Avery, 1972; Marriott, 1974).

**Table 2: Classified at risk**

| | 3 | 2 | 1 |
|---|---|---|---|
| At risk | 25 | 14 | 11 |
| Solvent | 32 | 14 | 16 |

An integral part of this type of analysis should be an examination of a random sample of continuing firms to estimate the proportion of all concerns "at risk", i.e. with financial profiles more similar to those of previous bankrupts. An analysis of the 2005 z-score distribution of the 280 quoted industrial companies with accounts held in the Nigerian Dataquest database indicated that 37.8 per cent were at risk. This type of information is essential for any user of such a model to identify the proportion of the population under consideration with an at risk profile and consequently the percentage of enterprises for which further investigation is necessary.

**Table 3: Classification of the Futile 13 firms from the past.**

| | | Futile | Superior | Total |
|---|---|---|---|---|
| Actual group | Futile firms | 11 | 9 | 20 |
| Membership | Firms | 0 | 22 | 22 |

## II. Conclusion

This paper describes discriminant analysis and the case where it is better than logistic regression. The paper also highlight the step to take to form a simple discriminant model. It also developed a simple linear discriminant model for the identification of potential Nigeria bankrupt concerns which uses only accounting statement- based financial ratios as variables. The derived model appeared outperform previous model build concerning failed company in Nigeria. Since the model can exhibit true ex ante predictive ability for a period of about 3 years subsequent

Though the models is well truly predictive in a statistical sense, such an approach is best used in an operational context as a means for identifying a short list of firms which might experience financial distress and which consequently justify further detailed investigation.

## References

[1]. Alexakos, C. E.(1966). Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions. Journal of Educational Psychology, 57, 297-306.

[2]. Anderson, G. J., Walberg, H. J., & Welch, W. W(1969). Curriculum effects on the social climate of learning: A new representation of discriminant functions. American Educational Research Journal, No. 6, 315-328

[3]. Chastian, K.(1969) Prediction of success in audio-lingual and cognitive classes. Lan-guage Learning, Vol. 19, 27-39.

[4]. Cochran, W. G. (1974). On the performance of the linear discriminant function. Technometrics, 6, 179-190.

[5]. Eisenbeis, R. A. and Avery, R. B. (1972). Discriminant Analysis and Classification Procedures: Theory and Applications. Lexington, Mass.: D.C. Heath & Co.

[6]. Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. London: Wiley.

[7]. Joy, 0. M. and Tollefson, J. 0. (1975). On the Financial Applications of Discriminant Analysis. J. Finan. Quantitat. Anal., 10, 723-739.

[8]. Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 23, 639-645

[9]. Malkovich, J. F. and Afifi, A. A. (1973). On Tests for Multivariate Normality. J. Amer. Statist. Ass., 68, 176-179

[10]. Marriott, F. H. C. (1974). The Interpretation of Multiple Observations. London: Academic Press.

[11]. Mosteller, F. and Wallace, D. L. (1963). Inference in the authorship problem. J. Amer. Statist. Ass., 58, 275-309.

[12]. Saupe, J. L.(1965). Factorial-design Multiple Discriminant Analysis: A description and an illustration. American Educational Research Journal, Vol. 2, 175-184.

[13]. Stahmann, R. F.(1969). Predicting graduation major field from freshman entrance data. Journal of Counseling Psychology, Vol. 16, 109-113.

[14]. Tatsuoka, M. M., & Tiedeman, D. V.(1954). Discriminant Analysis. Review of Educational Research, No. 24, 402-420.