Available online

JOURNAL OF CONTEMPORARY RESEARCH (JOCRES)

RESEARCH ARTICLE VOL. 1 (1) ISSN:2814-2241

www.unicrossjournals.com

Date Accepted: 30th June, 2022

Pages 42 - 48

APPLICATION AND UTILITY OF LINEAR DISCRIMINANT ANALYSIS IN EDUCATIONAL STATISTICS

Dr. Henry Ojating
E-Mail: Henryojating@Gmail.Com
Dept. of Educational Foundations and Childhood Education
University of Cross River State, Nigeria.

Abstract

Linear discriminant analysis is a technique applied in predicting group membership of individuals or cases based on measured characteristics of independent variables. It is analogous to linear multiple regression analysis, except that, linear discriminant analysis, has a categorical dependent variable. Problems of classification of cases (observations) based on measured characteristics are common place in virtually every sphere of life today. Researchers and pundits in various fields of education have attempted to critically address problems associated with delineating groups of learners on the basis of certain features to ensure appropriateness in placement. Others have examined characteristics that best discriminate groups of job seekers for proper job placement. This paper explored, in some detail, the application and utility of linear discriminant analysis in educational research.

Key words: Linear discriminant analysis, educational statistics, explanatory variables.

1. Introduction

Discriminant analysis is a statistical technique deployed in addressing research problems that are multivariate in nature. It is currently in widespread application in the global research community, and with bearing across various fields of learning. Discriminant analysis is concerned with the assignment of observations or objects into known or natural categories based on measured characteristics referred to as independent or explanatory variables. As noted by Dash (2021), Linear Discriminant Analysis (LDA) was first developed by Ronald Fisher in 1936 who examined linear combinations of variables that discriminate between only two categories or classes of observations. But later in 1948, C. R. Rao, had to formulate the LDA for multiple groups/categories. The groups or categories into which these observations or objects are assigned represent the criterion or dependent variable. Unlike what obtains in the Linear Regression Analysis, in Discriminant Analysis, the dependent variable is not metric but categorical. This is essentially where the difference between the two statistical techniques lie.

Statisticians, both applied and basic, are usually faced with problems of delineating groups and their members in the most parsimonious and scientific manner. For instance, how can we ascertain that there is a difference between Primary six pupils who are admittable and those not admittable into Secondary School? What should be the basis for assigning pupils into these two groups? Or, what should we practically do to predict membership into the groups? These questions and more would often come to mind when faced with problems of this nature. In assigning membership into groups of "admittable" and "not admittable" pupils, independent or predictor variables, such as, the pupils' Mock and Common Entrance Examination scores can be used. The dependent variable here represents the groups or categories. This is a case of a Two-group Linear Discriminant Analysis.

Similarly, businessmen and women may often express concern about knowing their customers who are highly, moderately and poorly satisfied with their services or products. Classification in terms of highly, moderately and poorly satisfied with services and products, could be based on values of sales and/or some measured characteristics of customers. This context describes what is called Multiple Linear Discriminant Analysis because it is a case where the criterion variable has more than two groups (categories). According to Pandya (2018), a case of Multiple Linear Discriminant Analysis is one in which there are more than two groups (categories) of dependent variable.

2. Key steps in the application of linear discriminant analysis

Linear Discriminant Analysis is carried out based on these key steps:

(i) **Problem identification:** This involves the formulation of a research question in a curious effort to generate linear combinations of explanatory variables that best discriminate categories or groups of observations. A research question could be in the form, "do students' academic performance, academic motivation and academic self-concept permit their classification by gender"?

(ii) **Data reduction process:** The discriminating or explanatory variables originally identified usually undergo a reduction process through the discriminant analysis technique.

(iii) **Defining the discriminant function:** The discriminant function is analogous to the linear regression equation. When in a two-group case, the groups are coded dichotomously (0 or 1) as dependent variable, the results of the analysis obtained are similar to those obtained through the multiple linear regression analysis. A discriminant function which is meant to maximize the difference between groups of male and female students on the basis of their academic motivation (x_1) and academic self-efficacy (x_2) can be stated as follows:

$$Y = a + b_1X_1 + b_2X_2$$

Where $Y =$ the predicted (categorical) variable, representing male or female students.

$a =$ the y-intercept, that is, the value of y when the other parameters or input variables tend to zero.

$b_1 =$ regression coefficient of academic motivation (x_1)

$b_2 =$ regression coefficient of academic self-efficacy (x_2).

(iv) **Estimate regression coefficients:** Obtaining coefficients for each of the explanatory variables is critical in determining the magnitude of contribution of each in predicting group membership. Standardized coefficients are used to show such magnitudes. The process of obtaining beta weights has been made a lot easier with the availability of numerous statistical packages like Excel, SPSS, and so on, especially where large volumes of data are involved.

(v) **Assess the fit of the regression equation to the data:** This has to do with examining how much of the variation in the dependent variable is accounted for by the independent or explanatory variables. That is, it defines the extent to which observations have been classified between groups based on the linear combinations of the independent variables in the model. The proportion of fitness of the equation to the data is referred to as, 'Coefficient of Determination' or 'Eigenvalue' and is defined by the symbol, R^2.

(vi) **Assess the relative contribution of the explanatory variables to the prediction of the dependent variable:** This is done using the individual beta weights or standardized coefficients obtained through the regression analysis procedure. High beta weight signifies high contribution to the prediction of group membership.

(vii) **Assess the ability of the discriminant function to correctly classify observations:** Discriminant Analysis indicates how accurately observations are classified by the explanatory variables. This is estimated in percentages. The measure of accuracy of classification is referred to as the hit ratio.

3. Assumptions of linear discriminant analysis

Discriminant analysis, like other multivariate techniques, has underlying assumptions, though, uniquely, some may not be robust enough to invalidate outcomes of analysis. The assumptions, as highlighted by Kottari and Garg (2014); and Holland (2019), are summarized as follows.

(i) The groups must be mutually exclusive. That is, no two observations must belong to one group.

(ii) The number of cases or observations for each group must not be greatly different.

(iii) The cases must be independent. Individual measures (scores) of students' performance, for instance, must be independent of each other.

(iv) Discriminant function performs better with larger samples. A good guideline is that there should be at least four times as many samples as there are independent variables. The minimum sample size should be the number of independent variables plus 2.

(v) Discriminant function analysis is highly sensitive to outliers. Each group should have the same variance for any independent variable (that is, be homoscedastic), although the variances can differ among the independent variables. For many types of data, a log transformation will make the data more homoscedastic (that is, have equal variances).

(vi) The independent variables should be multivariate normal; in other words, when all other independent variables are held constant, the independent variable being examined should have a normal distribution.

(vii) Group memberships must be collectively exhaustive, that is, all cases are members of a group.

4. Application of discriminant analysis in educational statistics

Educational statistics is concerned with data gathering, organization, interpretation and analysis for decision making and knowledge building to advance education. Discriminant analysis, as a classification technique, has over the years proved a veritable tool in proffering answers to a myriad of education related problems that are multivariate in nature. For instance, senior secondary school students seeking admission into tertiary institutions can be properly assigned courses of study on the basis of certain measured criteria, including their individual performances in the School Certificate Examinations (SSCE). That is, a student's course of study in the University or other tertiary institution can be predicted using discriminant function analysis. A procedure which guarantees that career choice is based on competence and not intuition or other considerations. Bakari, Isa and Zannah (2016) used discriminant analysis in modelling students' placement in Colleges of Education. Their interest was to appropriately place incoming students with Pre-Nigerian Certificate of Education (Pre-NCE) programmes into various courses of study based on the students' performance at the pre-NCE level. The results obtained indicated that different subjects in different courses were the strongest contributors to the placement of students into the subject combinations.

Discriminant analysis was deployed in the study by Lacruz, Americo, and Carniel (2019) to examine explanatory variables that best differentiate the performance obtained by students from the final grade of primary education in State schools of Espiorito Santo in Prova Bazil. The results showed that the age-series distortion, the teacher regularity index and the abandonment rate formed optimal set of variables to discriminate the schools with "better" and "worse" school performance. Bhalchandra, Muley, Joshi, Khamitkar, Fadewar and Wasnik (2018), attempted to discover the best strategy for the classification of values related to a categorical dependent variable using discriminant function analysis. To perform the analysis, they created a personal dataset of students with social, economic and academic variables. The discriminant function analysis was meant to examine the behavior patterns of students against their performance. It was found that some of the variables played a very important role in discriminating between categories of performance. Based on the data, the accuracy of classification or its ratio represented 67.4%.

Discriminant analysis was applied by Razzah, Ali, and Ali (2015) in predicting the annual performance of various institutes affiliated with Sargodha Board in Pakistan. The researchers used the annual performance of 2nd year (12th year education) results of 2014. The study was meant to discriminate between two categories (group 1: institutes having annual results below board average result and group 2: institutes having annual results above board average result) based on disciplines of institutes, gender, status of institutes and geographical area of four districts in Sargodha division. Successful discrimination was made between institutes with results below or above board average. Clustering annual results of institutes under two categories was statistically significant on the basis of discriminant analysis.

Let's examine a hypothetical case where one is interested in finding out if group membership in terms of students' gender can be predicted based on academic motivation and academic self-efficacy. The students are rated on the variables using the scale of 1-10. The outcome is classified in table 1:

Table 1: Observed discriminant data for groups of male and female students

GROUP 1	MALE	ACADEMIC MOTIVATION (X_1)	ACADEMIC SELF-EFFICACY (X_1)
	1	7	5

	2	5	9
	3	8	6
	4	6	7
	5	4	8
GROUP MEAN		**4**	**7**
GROUP 2	**FEMALE**		
	6	9	4
	7	3	4
	8	2	3
	9	2	2
	10	4	3
GROUP MEAN		**4**	**3.2**

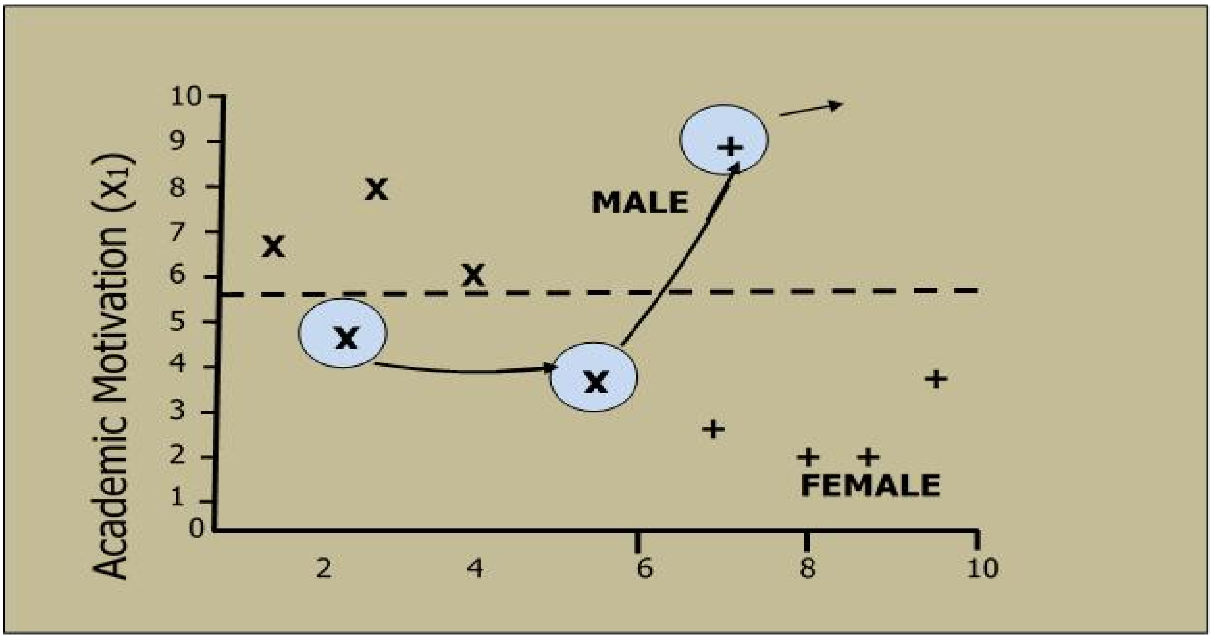
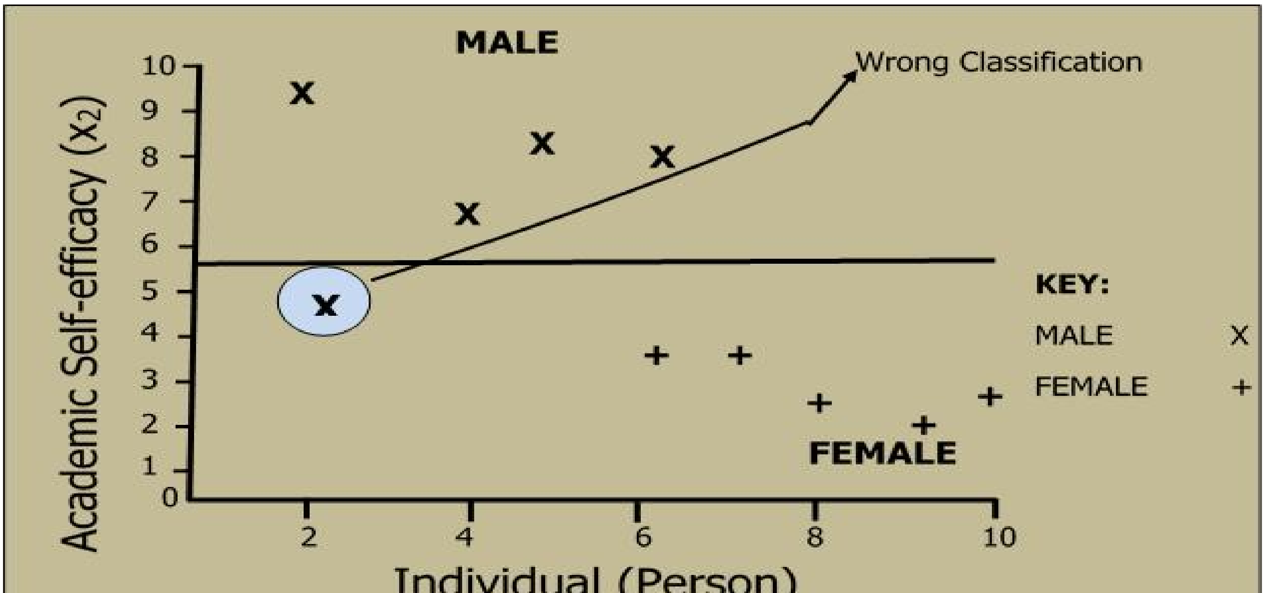


Fig. 1

Table 1 shows observed measures of students' academic motivation (x_1) and academic self-efficacy (x_2) for male and female students, respectively. We can visualize the classification graphically for each of the explanatory variables as in FIG. 1 and 2.



Notice that, from the observed data, academic motivation (x_1) wrongly classified the 2nd and 5th male (group1) and the 1st female (group 2). Academic self-efficacy wrongly classified the 1st male (group 1). Discrimination between the two groups based on the discriminant variables is further illustrated in figure. 3.

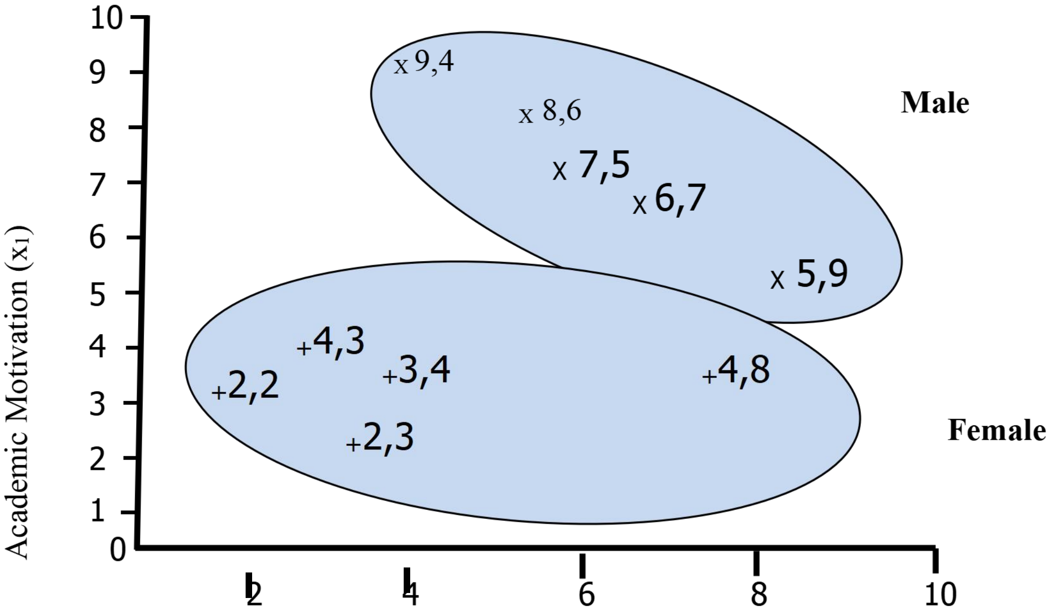


Fig. 3 Academic self-efficacy (x2)

Note that in figure 3, the measures (scores) of academic motivation and academic self-efficacy determined the classification of individual students on their gender. Such that, with a cut-off point of 11, students having ($x_1 + x_2 > 11$) are grouped as males while those with scores ($x_1 + x_2$) < 11 are classified as females. The process of discriminating between groups based on measured characteristics of students as so far examined is essentially based on observed data. To find the linear combination of predictor variables that best discriminate between groups of male and female students will require the fitting of a regression line to the data, along with detailed computations, which, among other things, will show how valid the model is and the relative contribution of each explanatory variable to the prediction of group membership, which are outside the scope of this paper.

5. Utility of discriminant analysis in educational statistics

As a multivariate technique, discriminant analysis has gained extensive recognition and application across various fields of learning and the research community. Apart from its utility in addressing classification problems in the Business world, Medicine, Humanities and so on, discriminant analysis is useful in tackling most societal and educational problems that are multivariate in nature. Educational statistics deploys linear discriminant analysis because of its numerous benefits, which may apply universally. Some of which are outlined here:

(i) Prediction of group membership based on measured explanatory variables that are education related.

(ii) Able to statistically examine difference between observed and expected discriminant educational data.

(iii) Defines the model that best fits the data.

(iv) It is deployed in the computation of the accuracy of group classification or the hit ratio.

(v) It allows for the classification of more than two groups based on measured explanatory variables, which is a case of multiple discriminant analysis.

(vi) It is a data reduction technique which serves the purpose of screening out variables which are very weak in discriminating the groups or categories.

6. Conclusion

Linear discriminant analysis is a versatile multivariate technique which has relevance for problem solving in practically every sphere of life. Its application and utility in educational statistics is clearly evident, considering the myriad of education related cases cited earlier in this paper.

7. Suggestions

(i) In view of its importance, discriminant analysis may be adopted and frequently deployed as a tool in our nation's educational system for career placement.

(ii) Statistical findings from discriminant analysis can be used by school counsellors to guide learners who are classified as low achievers to improve their learning experiences.

(iii) Admissions into educational programmes at the tertiary level should be based strictly on outcomes of discriminant analysis, which would show what course of study an applicant is capable of coping with successfully.

(iv) Job placement across various sectors of the nation should be based on statistical outputs of discriminant analysis so as to assure quality in service delivery in the nation's workforce. With discriminant analysis, an employee is placed on a given job based on competitive competence and not on other considerations.

References

Bakari,H. R., Isa, A. M. & Kannah, U. (2016). Application of discriminant analysis in modelling students‘ placement in Colleges of Education. *Discovery Journals,* 52(249), 1702-1707.

Bhalchandra, P., Muley, A, Joshi, M. Khamitkar, S., Fadewar, H. & Wasnik, P. (2018). Classification through discriminant analysis over educational dataset. *Information and decision sciences,* 701, 99-106.

Dash, S. K. (2021). A brief introduction to Linear Discriminant Analysis. Retrieved from: analyticsvidhya.com. Date: 04/05/22.

Holland, S. (2019). Data analysis in the Geosciences - Discriminant function analysis. Retrieved from: strata.uga.edu. Date: 07/05/22.

Kothari, C. R. & Garg, C. (2014). *Research Methodology: Methods and Techniques* (3rd. ed.). New Delhi: New age International (P) Publishers.

Lacruz, A.J., Americo, B. L. & Carniel, F. (2019). Quality indicators in education: discriminant analysis of the performance in Prova Brasil. *Rev. Bras. Edu.* 24.

Pandya, H. (2018). Multiple Discriminant Analysis. Retrieved from htttp://www.grandacademicportal.eduction.

Razzak, H., Ali, M. Ali, M. (2015). Application of discriminant analysis to predict the institute’s annual performance in Sangodha Board. *World applied sciences journal,* 33(2), 213-219).