International Journal of Scientific Research in Education, September 2024, Vol. 17(3), 297-305.

International Journal of Scientific Research in Education (IJSRE)
ISSN: 1117-3259

Comparison of Multiple Regression and Discriminant Analysis as Tools for Predicting Academic Achievement

Chidirim Esther Nwogwugwu
Department of Educational Foundations
University of Calabar, Nigeria
chidirimesther66@yahoo.com

Sylvia Victor Ovat
Department of Educational Foundations
University of Calabar, Nigeria
sylviaovat@gmail.com

Abstract

The study compared the results obtained from multiple regression and discriminant analysis in the prediction of academic achievement, in order to determine their effectiveness and discrepancies as predictive tools. Admission indices from the universities in Nigeria were used in predicting students' graduation indices. A sample of 1662 university students were drawn using purposive, stratified and simple random sampling techniques. Data collection was done based on the records from Admission, Examinations and Records units of the universities. The findings of the study revealed that, multiple regression was as efficient as discriminant analysis as a predictive tool, in predicting over-all academic achievement. The discrepancies in the predictive power of the two statistical tools were in the magnitude of their probability levels. From the magnitude of their probability levels, discriminant analysis appeared to be more sensitive in prediction than multiple regression. Again, discriminant analysis explains a higher relative variance than multiple regression. Based on the findings, it was recommended that, researchers can use either discriminant analysis or multiple regression as a statistical tool when making predictions in situations where the categorical variables involved in the discriminant analysis has underlying linearity, knowing however that discriminant analysis is more sensitive.

Keywords: Multiple Regression, Discriminant Analysis, Comparison, Efficiency, Discrepancies.

Reference to this paper should be made as follows:

Nwogwugwu, C. E., & Ovat, S. V. (2024). Comparison of multiple regression and discriminant analysis as tools for predicting academic achievement. *International Journal of Scientific Research in Education*, *17*(3), 297-305.

INTRODUCTION

In prediction, the intention is to forecast a condition that will occur at a future time Y (criterion) from a condition at present or past time X (predictor) Y = X + a (Gall et al. in Nwogwugwu, 2016).

According to Anastasi and Urbina (2006), prediction can be used in a broader sense to refer to "prediction from any set rest to any criterion situation, or in a more limited sense as "prediction over a time interval". If the predicted criterion evolves over a time interval, then it is expressed as "predictive validity". Examples of that sort of decisions requiring a knowledge of the predictive validity of tests abound in the society. Some of such examples are: selection and classification of personnel, selecting students for admission into tertiary institutions or professional schools, hiring job applicants, etc.

Whenever the purpose of analysis is to find predictors for score variable, the suggested statistical procedures are: simple regression, stepwise multiple regression, hierarchical multiple regression and simultaneous multiple regression. Also, if the objective is to find predictors for a category variable, then, the suggested statistical procedures are: multinomial logistical regression and binomial logistic regression (Howitt & Cramer, 2011). Furthermore, when the scores on score variables are to be discriminate between groups, we use discriminant function analysis.

Multiple regressions have been explained by many people in different forms. Jing-Jing and Kong-Lai (2010) stated that multiple regression analysis seeks the equation representing the effect of two or more independent variables. According to Anastasi and Urbina (2006), multiple regression equation yields a predicted criterion score for each individual on the basis of his or her scores on all the tests in the battery. Also that "regression equations are computed on the basis of a regression line, the geometric line representing, with least amount of discrepancy, the equal location for points in a scattergram". Furthermore, that in multiple regression, " there is a single criterion variable (Y), and K predictor variables (K ≥2). That these predictor variables are combined into an equation which can be used to predict scores on the criterion variable (Y) from scores on the predictor variables (X's)". The general form of this equation is:

$Y=b_1x_1 + b_2x_{2+}\ldots\ldots\ldots b_kx_k + a$

Where:

b's- are the regression coefficients of the respective predictor variables.

a-is the regression constant

Y- is the criterion variable

X- is the predictor variable

Abrams (2009) opined that, multiple regression which gives a quantitative prediction is used when the dependent and variables independent variables are interval data. Regression analysis relies upon certain assumptions about the variables used in the analysis. These assumptions are: number of cases, out lies, multi collinearity and singularity.

Discriminant analysis involves the determination of a linear equation like regression that will predict which group the case belongs (Matinez & Kak, 2001). Discriminant analysis is based on the idea of finding suitable linear combinations of the original variables with the aim of seeing whether it is possible to separate different groups on the basis of available interval measurements (Bryan, 2005). The form of the linear combinations or equation is:

$D= V_1X_1 + V_2X_2 + V_3X_{3+}\ldots\ldots.. V_iX_i + a$

Where:

D- is the discriminant function
V- is the discriminant coefficient or weight for that variable
X- is the respondent score for that variable
a-is the constant
i- is the number of predictor variable

The V's are unstandardized discriminant coefficients. The V's maximize the distance between the means of the criterion (dependent) variable. Good predictors tend to have large weight. This function maximizes the distance between the categories, (come up with an equation that has strong discriminant power between groups). If an existing set of data is used to calculate the discriminant function and classify cases, then, any new case can be classified (Everitt & Dun, 2000). The number of discriminant function is one less than the number of groups. For instance, there is only one function for any basic two groups' discriminant analysis.

Discriminant analysis has been described in different ways by various researchers. Discriminant analysis discriminates between groups on the basis of some explanatory variables (Kothari & Garg, 2014). The usefulness of discriminant analysis as listed by Nwogwuwu et al. (2015) are:

- To investigate the differences between groups on the basis of the attributes of the cases, indicating which attribute contribute most to group separation.
- Determine the most parsimonious way to assign new cases into groups.
- Classify cases into groups.
- Test theory whether cases are classified as predicted.
- To determine the amount of variance in the dependent variable that is explained by the independent variable.

Furthermore, discriminant analysis is a regression equation that has dependent variable which represents group membership (Udofia, 2011). That is, a regression equation that predicts a categorical variable. The two main objectives of discriminant analysis is prediction and classification. Discriminant analysis as a predictive technique involves the derivation of the linear discriminant functions and not the classification of the objects. Also, this technique uses a set of independent variables with associated weights that are derived using the best-fit fashion to predict the scores on the dependent variables. Then, these discriminant scores predict groups membership that can be used for group classification purposes. Discriminant analysis as a classification technique involves using the derived predictive functions to classify the sets of data derived from the predictive functions that have previously been validated. The purpose of such classification is simply to see how well the derived discriminant functions predict group membership using the data from which it was derived (Manly, 2005).

Studies on comparison of multiple regression and discriminant analysis as statistical technique in prediction have been carried out in different fields of learning in different times. For example, in the University of the Republic of Uruguay, prediction of students' academic achievement by linear and logistic models that aimed at comparing results yielded by both statistical procedures, to identify the most suitable approach in terms of goodness of fit and predictive power (Ayan & Garcia, 2008). Their findings revealed that grades were positively

related to progress; also, logic regression explained less than 20% of the variance in some cases and was shown to be a better technique than linear regression, yielding, more stable estimates with regard to the presence of ill-fitting patterns.

Also, Kong-Lai and Jing-Jing (2010) conducted their own study in China on the topic discriminant analysis and logistic regression model application in credit risk for china's listed companies. The study aimed at determining the model that is superior in establishing credit evaluation models. The result of their empirical research on the credit risk analysis for listed companies was that logistic regression model is superior to the discriminant analysis model.

Furthermore, Daniel (2001) carried out a study in California using regression to predict student's success in a course and discriminant analysis in placing students in courses, aimed at exploring the limitations of using multiples regression for placement, and the use of discriminant function as an alternative. His findings revealed that regression models could only be used to predict success in the course. Without knowing into which course the student was to enroll, these models were useless. He concluded by saying that "discriminant function analysis can provide the necessary classification into the courses".

In prediction, there are recognized methods used by researchers. Some of these are prediction equations written by the multiple regression analysis, logistic regression and classification equations determined by the use of discriminant analysis. Some researchers have described discriminant analysis and multiple regression as being similar to each other. Some researchers argue that discriminant analysis often times is used where regression analysis is more appropriate and powerful technique. This study therefore compared the results obtained from multiple regression and discriminant analysis in the prediction of academic achievement in order to determined their efficiencies and discrepancies as predictive tools. Consequently, the following research questions guided this investigation:

- What is the relative level of efficiency of multiple regression and discriminant analysis as predictive models?
- What are the discrepancies in the regression and discriminant analysis models in their predictive power?
- What is the relative variance explained by discriminant analysis and multiple regression?

METHODS

The study adopted an ex-post-facto research design. The data used in this study were collected from federal Universities in the five states that make up the South Eastern Nigeria. A sample of 1,662 students was drawn using multi-stage sampling, comprising purposive, proportionate stratified and simple random sampling techniques. The data for this study were collected from secondary source.

A results template which contained students' age on admission, unified Tertiary Matriculation Examination (UTME) scores, post-UTME scores, and cumulative grades point Average (CGPA), class of degree, year of admission, faculty, department, matriculation number and year of graduation; was used in collecting the data. The result template was to collect admission indices (age on admission, post-UTME scores and) of students admitted into Faculties of Education and Social Sciences, from 2015/2016 to 2017/2018; was obtained from the Admission Office. The computed cumulative grade point averages and classes of degree of those who successfully completed the program form 2018/2019 to 2020/2021, were also collected

from the Examination and Records Units of these Universities. The information collected constituted the data used for this study.

The data collected was analyzed using multiple regression and discriminant analysis as statistical tools used in prediction. When multiple regressions was used as a predictive tools, the criterion was CGPA while the age, UTME and post-UTME scores were the predictors. Also, in the case of discrepant analysis as the predictive tool, class of degree (first class, second class, etc) was the criterion while the predictors were age, UTME and post UTME scores.

RESULTS

Research Question 1: what is the relative level of efficiency of multiple regression and discriminant analysis as predictive models?

Table 1: Comparison of probability values associated with predictors variables in multiple regression and discriminant analysis

Predictor variables	**Discriminant analysis p-level**	**Multiple regression p-level**	**Remark**
UTME	.000	.000	Significant
Post UTME	.001	.003	Significant
Age	.935	.081	Not significant

Table 1 shows the comparison of the probability level of the predictors variables (UTME, post-UTME scores and age), in multiple regression and discriminant analysis from the SPSS analysis out-put in predicting academic achievement. The result presented in table 1 indicates that, for discriminant analysis, UTME and post-UTME predictors variables were statistically significant in predicting the criterion variable at 5% probability level, while in multiple regression analysis, UTME and post-UTME out of three predictor variables were significant at 5%. However, one of the predictor variable (age) was not statistically significant at 5% probability level in the two multivariate statistical procedures.

Research Question 2: what are the discrepancies in the multiple regression and discriminant analysis models in their predictive power?

Table 2: Discrepancies in the predictive power of multiple regression and discriminant analysis in predicting academic achievement

Predictors	**Discriminant analysis**	**Multiple regression**	**Differences**
UTME	.000	.000	.000 (0%)
Post UTME	.001	.003	.002 (.2%)
Age	..935	.081	.844 (84%)

Table 2 shows the discrepancies in the magnitude associated with significant levels of multiple regression and discriminant analysis models in predicting academic achievement. The result shows that, for the predictor variable "UTME", the levels of significance for discriminant analysis and multiple regression were .000 and .000 respectively. In the case of post-UTME, the probability level associated with disciminant analysis was .001, while that of multiple regression was .003. Also, significance associated with discriminant analysis was .935, while significance associated with multiple regressions was .081.

The results therefore reveal that, for the predictor variable UTME, the level of significance for the both statistical procedure is 0% level (P< .01). In the case of post-UTME, the probability level associated with discriminant analysis is .1% (p< .01), while that of multiple regression is .3% (P< .03). Furthermore, for age, the probability associated with discriminant analysis is 93.5% (P> .05), while for multiple regression it is 8.1% (P> .05).

Research Question 3: what is the relative variance explained by discriminant analysis and multiple regression?

Table 3: Comparison of relative variance explained by the predictors in both multiple regression and discriminant analysis

Predictors	**Discriminant analysis**			**Multiple regression**		**Regression**
	Eigenvalue	variance	p-value	Betaweights	variance	p- value
UTME	.028	67.6%	.000	.121	1.45%	.000
Post UTME	.013	32.2%	.001	.073	.53%	.003
Age	.000	.2%	.935	.043	.18%	.081

In table 3, the relative variance explained by each of the predictors in both multiple regressions and discriminant analysis were compared. In discriminant analysis, the eigenvalues show the percentage of variance explained by each function, while Wilks lambda tests the significance of each function. In multiple regression, the beta weights show the relative contribution of the predictor variables in the explanation of the criterion, while the significance of each of the predictor variable is measured by the t column. From table 3, in multiple regression, the predictor variable UTME has a beta weight of .121, variance of 1.45% that was significant at .000, while in discriminant analysis it had an eigenvalue of .028 that accounted for 67.6% of the variance that is significant at .000. Also, predictor variable post-UTME, in multiple regression had a beta weight of .073, variance of .53% that is significant at .003; while in discriminant analysis, it had an eigenvalue of .000 that accounted for .2% of the variance which is significant at .013.

Furthermore, the predictor variable age, in multiple regression had a beta weigh of .043, variance of .18% that is not significant at .081; while in discriminant analysis age had an eigenvalue of .000, variance of .2% that is not significant at .935. The result therefore is that, for the predictor variable UTME, the percentage of variance explained by multiple regression is .18%, while in discriminant analysis it is 67.6%. In the case of post-UTME, the percentage of variance explained by multiple regression is .53% while in discriminant analysis it is 32.2%. Also, for the predictor variable age, the percentage of variance explained by multiple regression is .18%, while in discriminant analysis it is .2%.

DISCUSSION OF RESULTS

The findings on the relative level of efficiency of multiple regression and discriminant analysis as predictive tools is that, two of the predictors variables (UTME and post-UTME) in the two statistical tools had a significant level of less than .05. The predictor variable, age was not significant in the two statistical tools. This indicates that, discriminant analysis is as efficient as multiple regression as a predictive tool. This finding is in agreement with the opinion of Howitt and Cramer (2011), who reported that it is possible to use discriminant function analysis in any circumstance in which you wish to use a set of score variables to predict membership of groups of participants. They added that, discriminant analysis can thus be regarded as a form of regression in which score variables are used to predict membership of the groups.

The result of this study is however not in agreement with the findings of kong-Lai and Jing-Jing (2010), who studied the "comparison of discriminant analysis and logistic regression model application in credit risk of china's listed companies." Their results showed that, all the predictor variables in discriminant analysis had a significant level of less than .50. They concluded that, discriminant analysis has a stronger predictive power than logistic regression.

The discrepancies in discriminant analysis and multiple regression as predictive tools were revealed to be in the magnitude of the probability levels of their predictors. This finding is in agreement with the opinion of Udofia (2011), who noted that the differences between discriminant analysis and multiple regression, is in the method of obtaining their b-values. He said that, in regression analysis, the b-value is obtained using the least square method whereas in linear discriminant analysis it is obtained using the eigenvectors of the matrix.

This finding disagrees with that of Whitaker (2005), who after conducting his study concluded that, the relative magnitudes of predictors in multiple regression and discriminant analysis are the same in his comparison of descriptive discriminant analysis and regression results. The discrepancies in the magnitude of the probability levels of the predictors in the two statistical techniques may be due to their sensitivity and robust nature. It may also be due to sample size, as multiple regression performs better for extreme small sample sizes and when theory or experience indicates an underlying relationship between dependent and predictor variables.

The relative variance in academic achievement of students explained by discriminant analysis technique was higher than in multiple regression. This finding is similar to that of Ayan and Garcia (2008) who studied prediction of University students' academic achievement by linear and logistic models. They found out that, linear models explain less variances in some cases when achievement scores are used in the standardized University entrance tests, as the predictor variables.

Moreso, this finding is in agreement with the opinion of Udofia (2011), who noted that the differences between discriminant analysis and multiple regression is in the method of obtaining their b-values. He opined that "in regression analysis, the b-value is obtained using the least square method whereas in linear discriminant analysis, it is obtained using the eigenvalues of the matrix.

The high relative variance in academic achievement of students explained by discriminant analysis may be as a result of the fact that, in discriminant analysis, discriminant functions are formed using the number of the correlations of each independent variable. This discriminant function is a means of a particular combination of variables. To decode the language or dimension of the discriminant function, the largest absolute correlation between each

variable and any discriminant function is used. The predictors in discriminant analysis were identified based on being the biggest contributors to any of the functions, while other predictors equally contributed. In the case of multiple regression, the predictors' correlation values are self-evident.

CONCLUSION

Multiple regression as an efficient as discriminant analysis in their predictive power in predicting students' academic achievement. Therefore, in prediction, discriminant analysis can be used in place of multiple regression as a statistical tool, when the conditions needed by the variables in the study for both statistical tools are met.

The discrepancies in the use of discriminant analysis and multiple regression for prediction, lie in the percentage of variations and relative variance explained; and also, in the magnitude of their probability levels. In these discrepancies, discriminant analysis is more sensitive than multiple regression.

REFERENCES

Abrams, D. R. (2009). *Introduction to regression*. Princeton University Press.

Anastasi, A., & Urbina, S. (2006). *Psychological testing (7th edition)*. Dorling kinderly Publishers.

Ayan, M. N. O., & Garcia, M. I. C. (2008). Prediction of University students' academic achievement by linear and logistic models. *The Spanish Journal of Psychology*, 11(1), 275-288.

Bryan, F. J. M. (2005). *Multivariate statistical methods. A Primer (3rd edition)*. Chapman & Hall.

Daniel, M. (2001). Predicting students' outcome using discriminant function analysis research and planning group. Educational Resources Information Centre. http://www.eric.gov/pdf/ed462116.pdf.

Everitt, B. S., & Dun, G. (2000). Applied multivariate data analysis (2nd edition). London: Arnold Publication.

Howitt, D., & Cramer, D. (2011). Introduction to SPSS statistics in psychology for version 19 and earlier (5th edition). Rotolito Lambarda Publishers.

Kong-Lai, Z., & Jing-Jing, L. (2010). Discriminant analysis and regression model application in credit risk for China's listed companies. *The Journal of Science and Engineering*, *4*(4), 24-32.

Kothari, C. R., & Garg, G. (2014). *Research methodology, methods and techniques (3rd edition)*. New Age International.

Manly, B. F. J. (2005). *Multivariate statistical method: A primer*. Chapman & Hall.

Matinez, A. M., & Kak, A. C. (2001). Principal component analysis and linear discriminant analysis. *Journal of the American Statistician Association, 84*(405), 165-175.

Nwogwugwu, C. E., Omole, D. O. K., & Ikiroma, B. (2015). Execution of multivariate discriminant analysis with SPSS. *ASSEREN Journal of Educational Research and Development, 1*, 74-83.

Nwogwugwu, C. E. (2016). Comparison of two multivariate statistics in predicting academic achievement in Federal universities in South East Nigeria. Unpublished Ph.D. Thesis Faculty of Education, University of Calabar, Calabar, Cross River State.

Udofia, E. P. (2011). *Applied statistics with multivariate methods*. Immaculate Publishers.
Whitaker, S. J. (2005). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 20*(4), 401-417.
