International Journal of Physical Sciences Vol. 4 (1), pp. 016-021, January, 2009
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 © 2009 Academic Journals

Full Length Research Paper

Application of discriminant analysis to predict the class of degree for graduating students in a university system

Erimafa J.T., Iduseri A. and Edokpa I.W.*

Department of Mathematics and Statistics, Ambrose Alli University, Ekpoma, Nigeria.

Accepted December 22, 2008

In this Paper, discriminant analysis was used to predict the class of degree obtainable in a University system. The conditions for predictive discriminant analysis were obtained, and the analysis yielded a linear discriminant function which successfully classified or predicted 87.5 percent of the graduating students' class of degrees. The function had a hit ratio of 88.2 percent when generalized, as a valid tool to classify fresh students of unknown group membership. It was also discovered that success in classifying or predicting fresh students of unknown group into classes of degree, was essentially similar to that of the historical sample.

Key words: Discrimination, prediction, linear discriminant function, confusion matrix.

INTRODUCTION

The challenge of designing an educational intervention of any kind in higher education has been of great interest to many a researcher and/or educator, over the years. Usoro (2006) carried out a study on classification of students into various departments on the basis of their cumulative results for a one year Foundation Programme otherwise known as Pre-National Diploma (PRE-ND) in Polytechnics system. Charles and June (1970) carried out a study to determine if a differentiation or separation among students graduating, withdrawing or failing could be identified. Adebayo and Jolayemi (1998, 1999), applied the τ-statistic Jolayemi (1990) to investigate how predictable the final-year result would be using the first year result or Grade Point Average (GPA) of some selected University graduates. In the past 25 years, research in academic prediction has centered on graduation, withdrawal, failure and selection of student's on the basis of either their collegiate success or cumulative results of Remedial or PREND; and literature to date suggests no loss of interest.

While the sheer volume of studies may be impressive, we do perceive a skewed interest. Most of the early research and much of the current ones, involved academic prediction in terms of class or college placement with the criterion of collegiate success being first term or year grade point average. In this study, our major task is to identify students who might be termed "at risk" (AR) and "Not at risk" (NAR). The first group are the students who are in danger of graduating with poor class of degree, PCD (that is, Third Class, Pass and Fail); and the second group are those that will graduate with better class of degree, BCD, (that is, First Class, Second Class Upper and Lower Division), within their first two years of study. Secondly, we determine what we call the grade point average booster course. This is a course the understanding of its concepts has a booster effect on sectional Grade Point Average (GPA). This student-identification task, performed by the discriminant analysis seems more appropriate than commonly used educational measures such as correlations, regression weights, e.t.c., because the variable being predicted is categorical. Research has shown that predictive discriminant analysis performs quite well with categorical data (Gilbert, 1968; Moore, 1973). Also violations of the assumptions underlying regression modeling can have serious repercussions (Cook and Weisberg, 1982).

A technique that could identify the factors that are pre-

*Corresponding author. E-mail: wazirip@yahoo.com.

dictive of performance as well as predict graduating students' class of degree would be of great benefit to the design, implementation, and evaluation of any educational programme/policy. This paper will show how discriminant function can be used to help determine what variables have relationship with performance, and an illustration of using the discriminant function to predict graduating students' class of degree in a university system.

Therefore the objective of this paper is to develop a discriminant function that will discriminate among classes of degree, and be generalized as a valid tool for classifying an individual student to one of the classes of degree to which he/she may belong on the basis of the individual profiles of scores on a set of predictor variables in the future.

Discriminant predictive analysis

The concern for the predictive ability of the linear discriminant function has obscured and even confused the fact that two sets of techniques based on the purpose of analysis exist, i.e., predictive discriminant analysis (PDA) and descriptive discriminant analysis (DDA). Stevens (1996) described the distinction between PDA and DDA in the following way; "in the predictive discriminant analysis, the focus is on classifying subjects into one of several groups (or to predicate group membership), whereas in descriptive discriminant analysis, the focus is on revealing major differences among the groups" (Stevens, 1996).

Also, Huberty and Barton (1989) aptly stated, the purposes of the two analyses are different; the techniques in the two analyses are different. There is, perhaps, some feasibility of the "mixing of DDA and PDA for purpose of corroboration of results. Generally research questions are of the descriptive type or of the predictive type; only seldom would both types of questions be addressed in a given research situation.

The two types of discriminant analyses i.e., PDA and DDA have different histories of development. According to Hurberty (1994), "discriminant analysis for the first three or four decades focused on the prediction of group membership", PDA, whereas DDA usage did not appear until the 1960s and "its use has been very limited to apply research settings over the past two decades".

Hence, PDA is appropriate when the researcher is interested in assigning units (individuals) to groups based on composite scores on several predictor variables. The accuracy of such prediction can be assessed by examining "hit rates" as against chance; for example, the most basic question answered by PDA is "given the individual scores on several predictor variables, which group represents their true membership group?" Again, the focus of PDA is predication and the accuracy of hit rates. As Huberty and Barton (1989) noted with respect to PDA, "one is basically interested in determining a classification rule and assessing its accuracy".

MATERIALS AND METHODS

Data collection

The data for this study were from student's academic records for 100 level and 200 levels, in the Department of Statistics, from 2004 to 2007 academic session in a University system as shown in Appendix A. In the first stage of data collection, two groups of student's in terms of their graduating class of degree were formed, and nine possible predictor variables, including the following: overall GPA. For 100 level grades in all the Statistics and Mathematics core courses thought to be predictive of performance. However, using the method of Stepwise Discriminant Analysis (forward stepwise analysis), we found that only two of these variables made significant independent and combined contributions. These were the students overall GPA and grade in statistics course dealing with probability distribution (STA 202). Since Predictive Discriminant Analysis (PDA) is concerned with hit rates and accuracy of classification, and reasonable PDA stepwise procedures must focus on maximizing hit rates. In order to confirm the GPA and STA 202 as the best subsets of the predictor variables using the forward stepwise analysis, we then used an "all- possible-subsets" approach which gave the same result (Huberty, 1989; Thompson, 1995).

Data analysis

The problem is to set up a procedure based on the student's grade, which enables us to predict the student's correct groups when we do not know which of the two groups the student will likely belong.

Basically, the two groups Fisher Linear Discriminant Function (Fisher, 1936) will be adopted in this study since it will discriminate between the two groups better than any other linear function (Ander-son, 1958).

Using an arbitrary linear discriminant function given by:

$$Z = U1X1 + U2X2 \qquad (1)$$

The variance-covariance matrices for the groups are given as:

$$S^1 = \begin{bmatrix} 0.364 & 2.339 \\ 2.339 & 95.475 \end{bmatrix}$$

$$S^2 = \begin{bmatrix} 0.263 & 1.946 \\ 1.946 & 108.118 \end{bmatrix}$$

To determine the vector of discriminant weight, U in equation (1), we compute:

(a) Pooled Sum of Squares and Cross Product Matrix, W

$$W = (N_1 - 1)S^1 + (N_2 - 1)S^2 = \begin{bmatrix} 37.00 & 252.82 \\ 252.82 & 12011.99 \end{bmatrix}$$

(b) Inverse of Matrix, W

$$W^{-1} = \frac{1}{|W|}C$$

$$= \begin{bmatrix} 0.031566831 & -0.000664397 \\ -0.000664397 & 0.000097234 \end{bmatrix}$$

(c) Mean Vectors, d

The deviation of mean vectors of Group 2 from Group 1 gives:

$$d = \begin{bmatrix} X_{1G1} - X_{1G2} \\ X_{2G1} - X_{2G2} \end{bmatrix}$$

$$= \begin{bmatrix} 1.12 \\ 16.02 \end{bmatrix}$$

So that:

$$U = W^{-1}d = \begin{bmatrix} 0.02471122 \\ 0.00081356 \end{bmatrix}$$

Thus, substituting these values of the discriminant weights, U in equation (1), we get:

$$Z = 0.02471122(GPA) + 0.00081356(STA202) \quad (2)$$

Multivariate test of significance

Problems arising in multivariate populations are direct generalization from the Univariate case. Thus, we decided to test for equality of Group Means and equality of Variance-Covariance matrices.

Equality of group means

The hypothesis of interest is:

$$H_0 : \mu^{(1)} = \mu^{(2)} \text{ Vs } H_1 : \mu^{(1)} \neq \mu^{(2)}$$

Using F-transformation of Hotelling's T^2 ,as our test statistic.

$$F = \frac{N_1 + N_2 - P - 1}{(N_1 + N_2 - 2)P}.T^2$$

Where $T^2 = \frac{N_1 N_2}{N_1 + N_2} D_M^2$ and $D_M^2 =$ Mahalanobis distance.

H_0 is rejected if $F_{CAL} > F_{P,N1+N2-P-1;1-\alpha}$

At 5% level of significance, we rejected the hypothesis of equality of group means. This implies that there exist significant differences between the group means.

Equality of covariance matrices

The hypothesis of interest is:

$$H_0 : V^1 = V^2 \quad \text{Vs } H_1 : V^1 \neq V^2$$

As our test statistic, we use Box's M test

$$M = (N-g)Log / S / -\sum_{i=1}^{g} V_i Log / S_i /$$

A reasonable approximation, when each $N_i > 20$ and N_i is large relative to P < 6 and g < 6, is obtained by using the Chi-Square approximation.

$$X_B^2 = (1-c)M$$

Where:

$$C = \frac{2P^2 + 3P - 1}{6(P+1)(g-1)} \left[\sum_{i=1}^{g} \frac{1}{V_i} - \frac{1}{N-g} \right]$$

$$M = (N-g)Log / S / -\sum_{i=1}^{g} V_i Log / S_i /$$

H_0 is rejected if $X_B^2 = (1-c)M > X_{\alpha(v)}^2$; where

$$V = \frac{P(P+1)(g-1)}{2}$$

At 5% level of significance, we accepted the hypothesis of equality of Variance-Covariance matrices. This analysis shows that the equality of variance assumption required when using Hotelling's T^2 statistics is tenable.

Classification rule

We define the cut off as:

$$C = \frac{\overline{Z}_1 + \overline{Z}_2}{2}, \qquad \overline{Z_1} \geq \overline{Z_2}$$

We first of all compute $\overline{Z_1}$ and $\overline{Z_2}$ which denote the functions at Group Centriods, by substituting the means of GPA and STA 202 for each of the two groups. In the linear discriminant function, we obtain $\overline{Z_1}$ = 0.122092085 and $\overline{Z_2}$ = 0.081308367 on calculation. Thus, the discriminating procedure is as follows:

Assign an individual to group 1 if $Z_{DS} > 0.102$ and group 2 if $Z_{DS} \leq 0.102$.

RESULTS

The data were processed on a microcomputer using SPSS statistical software package and confusion matrices for the analysis sample and hold-out sample are shown in Tables 1-3 below.

In Table 1, the rows totals are the observed categories for the class of degree and the columns totals are the

Table 1. Confusion matrix for actual and predicted categories of class of degree.

Class of Degree	Predicted Class of Degree		Total
	BCD	PCD	
2^1	12	0	12
2^2	38	10	48
3^{rd}	5	41	46
Pass	0	14	14
Total	**55**	**65**	**120**

Table 2. Confusion matrix for actual and predicated class of degree with percentage.

Group		Predicted Group Membership		Total
		1	2	
Original Count	1	50	10	60
	2	5	55	60
%	1	83.3	16.7	100
	2	8.3	91.7	100

Total Hit Ratio = 87.5%. Probability of Misclassification = 0.125.

Table 3. Confusion matrix for validated data.

Group		Predicted Group Membership		Total
		1	2	
Original Count	1	19	3	22
	2	1.0	11	12
%	1	86.4	13.6	100
	2	8.3	91.7	100

Total Hit Ratio = 88.2%. Probability of Misclassification = 0.118

predicted categories for the class of degree. It was observed that 50 out of 55 individuals predicted to graduate with Second Class Upper (2^1) or Second Class Lower division (2^2) did so. This represents Hit Ratio of 90.9%. Also, of 65 individuals predicted to graduate with Third Class or Pass, some 55 did so. This also represents a Hit Ratio of 84.6%.

In Table 2, success in identifying students that will graduate with better classes of degree (BCD) was 83.3%, essentially similar to that of the hold-out sample (Table 3). Also, in Table1, success in identifying students that will graduate with Poor Class of Degree (PCD) was 91.7%, again essentially the same as the hold-out sample (Table 3). Looking at the Total Hit-Ratio for both the historical sample (Table 2) and the hold-out sample (Table 3), the results are essentially similar. Hence this shows that, the classification result of the historical sample (analysis sample) was not biased upward.

Conclusion

The consistent high hit rates for both the analysis sample and hold-out sample, i.e., the overall percentage of correct classifications which is 87.5 and 88.2%, as seen in the confusion matrices (Tables 2 and 3), for this study, which is a measure of predictive ability shows that discriminant analysis can be used to predict students' graduating class of degree from knowledge of variable(s) that have relationship with performance. This study tends to illustrate the logicality and wisdom in examining related statistical technique useful for the purpose of prediction.

The use of discriminant analysis in this manner that is, conducting discriminant analysis for predictive purpose enables us to identify the students who might be termed at risk; these are students that will graduate with Poor Class of Degree, PCD. It also identifies STA 202 as having a booster effect on final graduating Cumulative Gra-

APPENDIX A

HISTORICAL DATA												VALIDATED DATA					
VALUES OF GPA AND STA 202 FOR TWO GROUPS												VALUES OF GPA AND STA 202 FOR TWO GROUPS					
GROUP 1 (N_1 = 60)						GROUP 2 (N_2 = 60)						GROUP 1 (N_1 =22)			GROUP 2 (N_2 =12)		
NO	G.P.A	STA 202	NO	G.P.A	STA 202	NO	G.P.A	STA 202	NO	G.P.A	STA 202	NO	G.P.A	STA 202	NO	G.P.A	STA 202
1	2.11	34	31	3.30	56	1	1.57	34	31	2.19	28	1	2.89	80	1	2.11	58
2	3.41	57	32	2.02	65	2	1.54	50	32	1.86	47	2	3.62	76	2	2.19	61
3	2.44	46	33	2.86	64	3	2.27	40	33	2.32	44	3	3.46	78	3	1.49	60
4	2.65	42	34	3.30	67	4	1.08	15	34	3.08	40	4	3.38	75	4	2.57	40
5	2.35	50	35	3.38	32	5	1.11	21	35	1.22	27	5	2.57	57	5	1.86	40
6	3.08	70	36	2.72	55	6	1.56	40	36	1.33	24	6	3.46	63	6	1.97	43
7	3.35	70	37	2.37	54	7	2.19	47	37	2.78	46	7	3.00	63	7	2.08	49
8	3.14	56	38	2.03	71	8	1.76	64	38	1.81	52	8	1.78	71	8	1.86	58
9	4.00	67	39	2.14	53	9	2.57	40	39	1.81	52	9	3.30	68	9	2.63	50
10	2.89	40	40	3.78	62	10	2.35	46	40	2.97	61	10	2.32	67	10	2.21	41
11	2.70	53	41	2.05	58	11	2.11	40	41	2.14	40	11	2.46	42	11	1.41	24
12	3.05	57	42	3.59	53	12	1.76	64	42	2.20	46	12	2.92	66	12	1.89	61
13	2.38	43	43	3.35	64	13	1.06	21	43	1.89	43	13	3.41	55			
14	3.46	73	44	2.13	56	14	1.97	44	44	2.56	41	14	2.11	67			
15	3.92	69	45	2.81	63	15	2.78	45	45	2.81	36	15	1.68	69			
16	2.57	53	46	2.32	67	16	1.14	46	46	1.59	37	16	2.41	62			
17	3.95	79	47	4.11	66	17	2.43	50	47	1.92	40	17	1.59	77			
18	3.73	73	48	4.08	63	18	2.51	48	48	1.97	43	18	3.37	53			
19	3.68	65	49	3.27	60	19	2.00	33	49	2.05	40	19	2.89	65			
20	3.11	40	50	3.78	53	20	2.16	41	50	1.87	40	20	1.70	56			
21	3.19	45	51	2.51	56	21	1.33	40	51	1.64	28	21	2.43	79			
22	3.08	59	52	3.41	62	22	2.27	24	52	1.97	50	22	2.00	59			
23	2.81	53	53	3.49	60	23	1.22	40	53	2.81	53						
24	2.86	60	54	2.35	48	24	1.49	41	54	1.68	62						
25	3.51	60	55	2.00	49	25	1.95	52	55	1.81	48						
26	3.59	64	56	3.22	52	26	1.43	42	56	2.03	41						
27	3.59	57	57	3.32	46	27	1.78	41	57	1.89	35						
28	3.59	62	58	3.19	71	28	2.08	56	58	2.49	45						
29	2.54	51	59	3.16	58	29	1.24	29	59	1.19	40						
30	2.43	56	60	3.86	62	30	1.76	25	60	1.12	41						

de Point Average (CGPA), as well as brought to light the difficulty in understanding its concept. Therefore there is need for an instructional intervention.

In conclusion, this study shows that discriminant analysis provides results that are both more interpretable and statistically sound, in addition to being a statistically correct procedure for prediction purpose than traditional measures.

ACKNOWLEDGEMENT

I thank the Editor and the reviewers for their insightful comments that enhanced the paper.

REFERENCES

Adebayo SB, Jolayemi ET (1998b). On the Effect of Rare Outcome on some Agreement/Concordance Indices. Nig. J. Pure and Appl. Sci. 13: 718-723.

Adebayo SB, Jolayemi ET (1999). Effect of Rare Outcome on the Measure of Agreement Index. J. Nig. Stat. Assoc. 13:1-10.

Anderson TW (1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley

Charles BK, June EH (1970). Predicting Graduation Withdrawal and Failure in College by Multiple Discriminant Analysis. J. Edu. Measur. 7(2): 91-95.

Huberty CJ (1989). Problems with stepwise methods: Better alternatives. In: B Thompson (ED.), Advances in social science methodology. Greenwich, CT: JAI Press. Vol. 1: 43-70.

Cook RD, Weisberg S (1982). Residuals and Influence in Regression”, London: Chapman and Hall.

Fisher RA (1936). The use of Multiple Measurements in Taxonomic problems. Ann. Eug. 7: 179 – 188.

Gilbert ES (1968). On Discrimination using Qualitative Variable. J. Amr. Stat. Soc. pp. 399 – 404.

Huberty CJ (1994). Applied Discriminant Analysis. New York: Wiley and sons.

Huberty CJ, Barton RM (1989). An Introduction to Discriminant Analysis. Measurement and Evaluation in Counseling and Development, 22: 158 – 168.

Jolayemi ET (1999a). On the Measure of Agreement between two raters", Biomed. J. 32: 87-93.

Moore DH (1973). Evaluation of five Discrimination procedures for Binary Variables. J. Amer. Stat. Soc. 13: 399 – 404.

Stevens J (1996). Applied Multivariate Statistics for the Social Sciences.In: Mahwah NJ, (3rd Ed). Lawrence Erlbaum Associate, Inc.

Thompson B (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Edu. Psychol. Measur. 55(4): 525-534.

Usoro AE (2006). Discriminant Analysis and its Application to the Classification of Students on the Basis of their Academic Performances. J. Res. Phy. Sci. 2(3): 53 – 55.