Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.5, No.6, 2014

Improving Admissions into Technical and Vocational Institutions Through a Statistical Classification Technique

Joseph Dadzie[1*], Martin Owusu Amoamah[1], Ben Apau- Dadson[1], Prince Koranteng Kumi[2]
1. Department of Mathematics and Statistics, Accra Polytechnic. P. O. Box GP561, Accra.
2. Department of Accountancy, Accra Polytechnic, Ghana.
*Corresponding Author's Email: kobenagyesi@yahoo.com

Abstract
Polytechnic education at the Tertiary level is part of the the crowning achievement or the capstone of the traditional educational structure. It is the hub of human development worldwide. Therefore the system of admissions into these institutions anywhere is an area of research interest. Many factors are considered before students are admitted into programmes at the tertiary level of education and this differs from country to country. However, a growing conflict is the entry requirements into the traditional universities on one hand and the polytechnic and the vocational institutions on the other hand. According to Stubbs (Stubbs, 1998), at least half of the differences in students' academic performance could be attributed to the students' social background and prior attainment rather than the school they attended. Using Ghana's example as a case study, the multivariate classification tool of linear Discriminant Analysis was conducted on students' academic performance with entry grades as the predictors and the classes they obtain at the end of their study at the Polytechnic level as the response variable. Exploratory data analytical tools of normality, multicolinearity, homocedasticity amongst others were employed on the data. Further analysis on the transformed data using linear discriminant analysis revealed that the elective subjects as predictors have far stronger discriminating powers than the core subjects. This suggests that the current entry requirement policy into the polytechnics that focuses on the core subjects must be looked at again. The current entry regime which is more like that of the universities tends to turn away many otherwise very good technical and vocational education materials because of fails in one or two core courses.
Keywords: Discriminant Analysis, Multicollinearity, Homocedasticity, Centroids, Eigen value, Normality

1. Introduction

The polytechnics as a hub of technical and vocational education in Ghana were upgraded to tertiary status in 1993 with the promulgation of the Polytechnic Law, PNDC Law 321, 1992. The polytechnics were charged with responsibility for producing middle level manpower in manufacturing, applied science and technology, and business and commerce for national development. They are also to encourage the study of technical subjects at the tertiary level and provide opportunity for research and publication of research findings (Polytechnic Law:1992).Ghana has 10 regionally based polytechnics located in Accra, Kumasi, Takoradi, Ho, Tamale, Sunyani, Cape Coast, Koforidua, Wa and Bolgantanga that offer various programmes leading to Higher National Diploma (HND). In addition, they also offer non-tertiary education. About 1 per cent of the total population of post-secondary age goes to the Polytechnics. At present, polytechnics in Ghana have the requisite legal mandate to award degrees subject to such conditions that the Polytechnic Council of that polytechnic may determine. In preparing of these polytechnic documents, contributions were received from various stakeholders all with the view of meeting educational as well as technological demands of the current global chances. The objective of the polytechnics is to produce hands-on skilled manpower and professionals for employment in industry and other sectors of the economy. The Higher National Diploma (HND) programs at the various polytechnic levels have been with us for some time now since they were first introduced in 1993. Majority of students have graduated from these programs and are contributing in many areas of the countries developmental progress. Some of these graduates could be found in different sectors of the economy, playing various roles in industry, business and commerce. Currently, some of the polytechnics are offering Bachelor of Technology (B. Tech) degree programs in some engineering and other fields, while other polytechnics are awaiting approval to commence their degree programs.

A growing concern is the entry requirements into the polytechnics. All students who gain admission into the country's tertiary institutions are required to have requisite passes in core subjects of English language, core mathematics, social studies and integrated science as well as their elective subjects. In recent years many students fail to gain admission into polytechnics in Ghana because they fail to obtain the requisite passes in their core subjects most especially. There is the need to investigate whether these core and elective subjects have any significant effect on the performance of the students at the polytechnic level. It is worth finding out whether irrespective of the course being pursued the students should be made to have passes in all these core subjects.

2. Method

The research used a statistical classification technique of multiple linear discriminant method with entry grades of students as the predictor variables and their classes at the end of the programme as the response variable. Linear discriminant analysis is statistical technique that is more useful in education intervention programmes than the commonly used correlation and regression technique since the response variables are categorical. It allows the researcher to study the difference between two or more groups of objects with respect to several variables simultaneously. Ideas associated with Discriminant analysis can be traced back to the 1920's through the works completed by English statistician Karl Pearson (1857-1936), who translated multivariate inter group distance into linear combination of variables to aid in inter group discrimination (Anderson, 1996). Since then Discriminant analysis has been applied to a wide variety of disciplines. As a statistical classification technique, it generates functions from a sample of cases for which group membership is known; the functions can then be applied to new cases with measurement for the predictor variables but unknown group membership. When there are two groups only one discriminant function is generated. When there are more than two groups, several functions will be generated. Thus in general, if there are G groups, then we have (G-I) discriminant functions. Usually, only the first three of these functions will be useful (Sharma, 1996). The central research objectives by which discriminant analysis is most often evaluated are to maximize either the discriminating power of the predictive function or the overall correct classification within the confusion matrix.

The discriminant analysis model involves linear combinations of the following form:

$$D = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots\ldots + b_k X_k$$

Where $D =$ discriminant score

$b_s =$ discriminant coefficients or weights $s = 0, 1, 2, 3, \ldots,$

$X_s =$ predictors or independent variables $s = 1, 2, 3, \ldots,$

The coefficients, or weights (b_s), are estimated so that the groups differ as much as possible on the values of the discriminant function. The assumptions in discriminant analysis are that each of the groups is a sample from a multivariate normal population and all of the populations have the same covariance matrix. Eigen values are significant in a discriminant function analysis process in that they reflect the ratio of importance of the dimensions which classify cases of the dependent variable. . They are determined by solving the equation $\left|W^{-1}A - \gamma I\right| = 0$

Where γ is the eigenvalue, W is the within group sums of squares and cross products matrix. Suppose $x_{i1}, x_{i2}, \ldots\ldots, x_{in_i}$ is a random sample of size n, from a population with mean μ_i, $i = 1,2,\ldots\ldots,g$ where g is the number of populations and each x_{ij} is an observation vector of p components (number of variables). The within group sums of squares and cross products matrix can be expressed as

$$\underline{W} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)' \quad \text{where } i = 1,2,\ldots\ldots g \text{ and } j = 1,2,\ldots.n_i$$

Also the between groups sums of squares and cross products (SSCP) matrix is given by the expression.

$$\underline{A} = \sum_{i=1}^{g} n_i x(\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'$$

. An eigen analysis also produces an eigenvector that is associated with each eigenvalue. Eigen vectors are determined by solving the equation $\left(W^{-1}A - \gamma_i I\right)\hat{v}_i = 0$. Where γ_i is the eigen value corresponding to the i^{th} canonical function, and $\hat{v}_i$ is the normalized eigenvector associated with the i^{th} eigen value, $\underline{W}$ is the within-group SSCP matrix and $\underline{A}$ is the between-groups SSCP matrix. Each solution yields its own γ and the set of $\hat{v}_i$'s corresponds to one discriminant function. Another significant statistic is the (Model) Wilks' Lambda, λ. It is used to test the significance of the discriminant function as a whole. A significant lambda means one can reject the null hypothesis that the groups have the same mean discriminant function scores and conclude the model is significantly discriminating. The corresponding Wilks' Lambda, λ for variable is used to test which independent variable contribute significantly in the discriminant function. The smaller the variable Wilk's lambda for an independent variable, the more that variable contributes to the discriminant function. Lambda varies from 0 to 1, with 0 meaning group means differ. (Thus the variable differentiates the groups), and 1 meaning all group means are the same. The F test of Wilks's lambda shows which variables' contributions are significant. Wilk's Lambda follows the equation $\lambda = \frac{|W|}{|A+W|}$. The smaller the value of λ the greater the probability that the null hypothesis will be rejected and vice versa. To assess the statistical significance of the Wilks' λ, it can be converted into an F-ratio using the transformation

$$F = \left(\frac{1-\lambda}{\lambda}\right) x \left(\frac{n_1 + n_2 - p - 1}{p}\right)$$

Where p is the number of variables for which the statistic is computed. Given that the null hypothesis is true, the F-ratio follows an F-distribution, with p and $(n_1 + n_2 - p - 1)$ degrees of freedom.

2 Research data

Accra Polytechnic was used for the study with emphasis on Statistics which is a mathematics related program and Secretaryship and Management Studies, a language related programme. For the purpose of the research Statistics program would be referred to as Stats and Secretary and Management Studies be referred to as Sec. The data is a secondary data collected from the academic affairs of Accra Polytechnic consisting of the entry results and exit performances of the 2007/2008 year group. The students entered with passes in four core subjects and three elective subjects. At the entry point, the students were selected into their various courses based on their performances in core subjects of English language, core mathematics, integrated science and three electives. The grades of the subjects were weighted for the research as shown in the appendix table 1

When the students complete their courses of study, depending on their cumulative grade point, they are graded into First, Second Upper, Second Lower and Pass classes. Again for the purpose of this research, First and Second Upper classes become the Group 1, Second lower class the Group 2 and Pass class, the Group3. Those who were not able to complete their course of study as at time of doing the study were excluded from the analysis. A data size of 86 cases was obtained for the Stats cohort. 22 of this representing about 25% of the data size was used as the Holdup or Validation Sample whilst the rest of 64 cases was used as the Estimation or Analysis sample. The core subjects used were English language (EG), Social Studies (SS), Core Mathematics (CM), Integrated Science (IS). Depending on the percentage of students who offered it the following elective subjects were also identified; Elective Mathematics (EM), Physics (PH), Chemistry (CH), Biology (BI), Economics (EC) and Geography (GO). For all these subjects, Groups 1, 2 and3 have sizes of 30, 29 and 27 respectively. For all the groups, there are no obvious outliers. Missing observations were replaced by the mean value of that group. A data set of 130 cases was used for the Sec cohort. 30 of this representing 23.1% was used as the Validation or Holdout sample and the rest used for the estimation of the discriminant function coefficients. In addition to the core courses of English (EG), Core mathematics (CM), Integrated Science (IS), Social Studies (SS), the following electives: Typing (TY), Economics (EC), Business Management (BM), History (HI), Accounting (AC) and Christian Religious Studies.(CR) were also selected based on the percentage of students who offered them. The missing observations were replaced with the mean value for the groups. For all the subjects, Groups 1, 2 and 3 have sizes of 39, 37 and 24 respectively. In the test for equality of means for the Stats cohort, the wilks' lambda for all the predictors with exception of EG and CM are less than 0.5. This indicates that for these subjects there appear to be differences in the group means. Though by using Shapiro-Wilk test of Normality all the predictor variables exhibit significant departures from normality at 5%, examination of Boxplots and Q-Q plots show that for the Stats, some of the predictors such as IS, EM, SS appear to follow the normal distribution. EM has the largest Shapiro-Wilk Statistic whilst that of PH is the least. In case of Sec, examination of the relevant boxplot shows that the predictors IS, BM, HI and CR appear to exhibit normality behavior. The Levene test of homogeneity of variances show that at 5% there is no significant differences between the variances of CM, EG, EM, PH, BI and EC. The Levene test of homogeneity of variances for Sec group shows that at 5%, there are no differences in the variances of all the predictors with the exception of CR. The one way test of equality of means show that for both Stat and Sec cohort, significant differences exist in the means of the predictors. This means stable discrimination could be done between the groups with all the predictors.

3. Analysis

The first two canonical discriminant functions were used in the analysis. To test the null hypothesis of equal group centroids, both functions must be considered simultaneously. From the table 2 in the appendix, the value of Wilk's lambda for function 1 is 0.167 for the Stats cohort. This transforms to a chi-square of 101.092, with 20 degrees of freedom, which is significant beyond 0.05 level. Thus, the two functions together significantly discriminate among the three groups. Similarly for the Sec cohort the result in appendix table 3 shows that, the value of Wilk's lambda for function 1 through 2 is 0.068. This transforms to a chi-square of 248.616, with 20 degrees of freedom, which is significant beyond 0.05 level. Again the two functions together significantly discriminate among the three groups.

Since there are three groups, a maximum of two functions can be extracted. It could be seen from the table 4 at the appendix, that the eigen value associated with the first function for the Stat cohort is 2.819, and this function accounts for 83.3% of the explained variance. Since the eigen value is large, the first function is more superior than the second function which accounts for only 16.7% of the variation with an eigen value of 0.567. The standardized canonical function coefficients indicate a large coefficients for EM, IS, GO and SS. Apart from BI and IS, all the standardized canonical discriminant function coefficients are all positive. From table 5, the same trend could be seen in the Sec cohort. A maximum of two functions can be extracted. The eigenvalue associated with the first function is 11.322 and this accounts for 98.3% of the explained. Because the eigenvalue is large, the first function is superior. The second function accounts for only 1.7% of the explained variation. Examining the standardized discriminant function coefficients for function 1, as shown in the table 6 at the appendix indicate that for the Stat cohort, Elective Mathematics has the largest discriminating power with that of Core mathematics being the least. Apart from Integrated Science and Biology, all the other predictors have positive standardized function coefficients values. Similarly in the case of the Sec cohort in table 7, for function 1 Typing (TY) has the largest discriminating value followed by Economics (EC) with English (EN) having the least. For the function 1 of the Sec cohort, all the discriminant function coefficients are positive.

The table 8 and table 9 in the appendix show the classification matrix or the prediction matrix for the Stat and Sec respectively. They contain the number of correctly classified and misclassified cases. The correctly classified cases appear on the leading diagonal elements. The classification results based on the analysis sample indicate that for the Stats cohort, 91.8% of the Group 1 cases are correctly classified. When the classification analysis is conducted on the independent holdout sample; the result as indicated by table 8 in the appendix shows that 85.7% of unselected original grouped cases are correctly classified.

Similarly, the appendix table 9 shows the classification results based on the analysis sample and the validation sample for the Sec cohort. It contains the number of correctly classified and misclassified cases. The correctly classified cases appear on the diagonal. The classification results based on the analysis sample indicate that 97.4% of selected original group 1 cases are correctly classified and 90% of unselected original group 1 cases are correctly classified. In the case of the validation sample, Group 1 has the largest percentage of correct classification of 90% with Group 3 having a least value of 72.7%.

4. Discriminant analysis model

From the further analysis above and the unstandardized coefficients in the appendix, the discriminant analysis model for stats becomes;

$$D_{sta} = -17.005 + 3.228SS + 0.516CM + 1.459EG - 3.523IS + \\ + 4.603EM + 2.874PH + 3.605CH - 3.243BI + 4.217GO$$

And for the sec

$$D_{\sec} = -20.171 + 0.375EN + 0.856CM + 0.768IS + 1.191SS + 2.539TY + 2.204EC \\ + 1.085BM + 1.769HI + 1.731AC + 1.746CR$$

Considering a hypothetical results of a student applying to do Stat;

$$SS = B2,\ CM = B3,\ EG = C4,\ IS = B3,\ EM = D7,\ PH = C4,\ CH = C6$$

The discriminant score for such a case is:

$$D = -17.005 + 3.228\times1.6094 + 0.516\times1.3863 + 1.459\times1.0986 - 3.523\times1.3863 + 4.603\times0.9631 \\ + 2.874\times1.0984 + 3.605\times1.0986 + 3.605\times1.098 - 3.243\times0 + 4.217\times0 = -2.823$$

Using the group centroids table 10 means the student would be in group 3.

And the corresponding group centroids are as shown in the table 10 in the appendix. It must be noted that the cases are assigned to groups based on their discriminant scores and an appropriate decision rule based on the group centroids in appendix table 10. The interpretation of the discriminant weights or coefficients is similar to that in multiple regression analysis. The value of the coefficient for a particular predictor depends on the other predictors included in the discriminant function. The signs of the coefficients are arbitrary, but they indicate which variable values result in large and small function values and associate them with particular groups.

5. Conclusion

It has become imperative to depend on more scientific and data based models to predict and assess students' performances at the technical and vocational based courses in the polytechnics rather than the current systems that are modeled along the lines of the prevailing conditions at the universities. The predictive models as described above could be used to predict the performance of prospective applicants into the courses. The fact that the elective subjects have more discriminating power is also revealing. The current admission numbers could be improved if this method is considered in place of the present policy.

Appendix

Table 1: Transformation table

Old Grade	New Grade	Weight	Ln
A	A1	6	1.7918
B	B2	5	1.6094
C	B3	4	1.3863
D	C4,C5,C6	3	1.0986
E	D7, D8	2	0.6931
F	F9	1	0.0000

Table 2 Wilks lambda for Stats

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1 through 2	.167	101.092	20	.000
2	.638	25.378	9	.003

Table 3 Wilks's lambda for Sec

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1 through 2	.068	248.616	20	.000
2	.838	16.313	9	.061

Table 4 Eigen Value for Stats

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2.819a	83.3	83.3	.859
2	.567a	16.7	100.0	.602

Table 5 Eigen Value for Sec

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	11.322a	98.3	98.3	.959
2	.193a	1.7	100.0	.402

Table 6 Standardized Canonical Coefficients for Stats

	Function	
	1	2
SS	.616	.379
CM	.123	.173
EG	.324	-.980
IS	-.953	.113
EM	1.166	-.272
PH	.294	-.145
CH	.331	.684
BI	-.424	.029
EC	.181	-.115
GO	.777	.626

Table 7 Standardized Canonical Coefficients for Sec

	Function	
	1	2
SS	.081	.260
CM	.192	-.338
EG	.159	.003
IS	.250	.149
TY	.464	.409
EC	.427	-.536
BM	.242	.439
HI	.336	-.282
AC	.345	-.009
CR	.359	-.015

Table 8: Classification table for Stats

				Predicted Group Membership			
			CLS	1	2	3	Total
Cases Selected	Original	Count	1	21	1	1	23
			2	0	22	0	22
			3	0	3	16	19
		%	1	91.3	4.3	4.3	100.0
			2	.0	100.0	.0	100.0
			3	.0	15.8	84.2	100.0
Cases Not Selected	Original	Count	1	6	1	0	7
			2	0	7	0	7
			3	0	2	6	8
		%	1	85.7	14.3	.0	100.0
			2	.0	100.0	.0	100.0
			3	.0	25.0	75.0	100.0

			CLS	Predicted Group Membership 1	2	3	Total
Cases Selected	Original	Count	1	21	1	1	23
			2	0	22	0	22
			3	0	3	16	19
		%	1	91.3	4.3	4.3	100.0
			2	.0	100.0	.0	100.0
			3	.0	15.8	84.2	100.0
Cases Not Selected	Original	Count	1	6	1	0	7
			2	0	7	0	7
			3	0	2	6	8
		%	1	85.7	14.3	.0	100.0
			2	.0	100.0	.0	100.0
			3	.0	25.0	75.0	100.0

Table 9 : Classification table for Sec

			CLS	Predicted Group Membership 1	2	3	Total
Cases Selected	Original	Count	1	38	1	0	39
			2	2	35	0	37
			3	0	0	24	24
		%	1	97.4	2.6	.0	100.0
			2	5.4	94.6	.0	100.0
			3	.0	.0	100.0	100.0
Cases Not Selected	Original	Count	1	9	1	0	10
			2	2	7	0	9
			3	0	3	8	11
		%	1	90.0	10.0	.0	100.0
			2	22.2	77.8	.0	100.0
			3	.0	27.3	72.7	100.0

Table 10 Group Centroid

Statistics		Sec	
Group	Centroid	Group	Centroid
1	3.050	1	2.059
2	0.364	2	-1.800
3	-5.517	3	-0.409

REFERENCES

Albaum, G., R. and D. Hawkins (1975), “Applying Discriminant Analysis to Unipolar Semantic Scaling Data” American Institute of Decision Sciences Western Meetings.

Evans, F. B. (1959), “Psychological and Objective Factors in the Prediction of Brand Choice: Ford vs. Chevrolet,” Journal of Business 32 (October), pp.340-369

Friendly, M. (1991). Statistical Graphics for Multivariate Data. Paper presented at the SAS User's Group International Conference. New Orleans.
Gazette Notification. (1993). Polytechnic Law, P.N.D.C, L 321 (Publication No. A66/1,500/1/93). Accra: Ghana. Ghana Publishing Corporation.
Gazette Notification. (2007). Polytechnic Act, Act 745 (Publication No. GPC/A156/350/4/2007). Accra: Ghana. Ghana Publishing Corporation.
Guney, Y. (2009). Exogenous and Endogenous Factors Influencing Students' Performance in Undergraduate Accounting Modules. Accounting Education, 18 (1), 51-73.
Kendall, M. G. & Stuart, A. (1966). The Advanced Theory of Statistics In Three volumes. (Volumes 3). Design and Analysis, And Time-Series. New York: Hafner Publishing Company.
Klecka, W. R. (1980) Discriminant Analysis. New York: Sage Publication Inc.
Malhotra, N. K. (2004). Applied Orientation In Marketing Research. (4 ed.) New York: Pearson Prentice-Hall.
Manly, B. F. J. (1994). Multivariate Statististical Methods. A Primer (2nd , ed.) New York: Champman & Hall.
Moroney, M. J. (1968). Facts from Figures. Middlesex: Penguin Books Ltd.
Seber, G. A.F. (1984). Multivariate Observations. New York: Wiley.
Sharma, S. (1996). Applied Multivariate Techniques. Ney York: John Wiley & Sons Publications.
Skinner, C. E. (1945) Educational Technology, (Revised Edition). New York: Prentice-Hall Inc.
Stubbs, B. F. (1998) Students Performance Analysis. (MPhil dissertion, Bath University, 1998) Retrieved March 21, 2009, from http://www.bstubbs.co.uk//mphil
Wilks, D. S. (2006) Statistical Methods in The Atmospheric Science (2nd ed.) New York: Academic Press ELSEVIER.