Canonical Correlation And Hotelling's $T^2 Analysis$ On Students' Performance In Science And Non Science Subjects

Mustapha Usman Baba[1] , **Nafisa Muhammad**[1] ,**Ibrahim Isa**[1] , **Rabiu Ado Inusa**[1] and **Usman Hashim Sani**[1]

1. Department of Statistics, School of Technology, Kano State Polytechnic, Kano, Nigeria.

Correspondence e-mail: mustyubaba2@gmail.com

Abstract: This study is aimed at examining the relationship between students' performance in science subjects and non-science subjects. And to also test for the homogeneity of variances of the two sets. The statistical tools used are canonical correlation, Bartlett-Box test and hotelling's T^2. It was observed that the maximum correlation between students' performance in science subjects and their performance in non-science subjects has a value 0.538236. And the Wilk's lambda test confirmed that the correlation is statistically significant. Furthermore, the homogeneity of variance test using Bartlett-Box test show that the variances of the two sets of scores are unequal, so the alternative to T^2 was used to compare the two mean vectors. The comparison show that the mean vectors differ, which means that students' performance in science subjects differs from their performance in non-science subjects.

Keywords: Canonical correlation, Hotelling's T^2 , Students

1.0 Introduction

Canonical correlation analysis deals with the association between composites sets of multiple dependent and independent variables. In doing so, it develops a number of independent canonical functions that maximize the correlation between the linear composites, also known as canonical variates, which are sets of dependents and independents variables.

The Hotelling's T^2 distribution is a multivariate statistical technique used for test of hypothesis concerning mean vectors. It is the multivariate equivalent of the t-test used in univariate test of hypothesis. The study is focused on the 2017 students' performance in NECO, were the results of 100 students From Government College Kano on six selected subjects were considered. Three of these subjects are Physics designated X_1, Chemistry designated X_2 and Biology designated X_3 are called Science subjects. The remaining three variables are Accounting designated Y_1 and Commerce designated Y_2 and Government designated Y_3 are called non-science Subjects. In Nigeria, education remains the largest industry and government continues to ensure that funds, instructional material and teaching personnel are made available for the sector. Government has also continuously encouraged secondary education by adopting the social demand approach towards planning the sector and by subsidizing the senior school certificate examinations (SSCE) fee in the states over a long period of time. Of course despite the efforts being made towards ensuring that citizens have equal educational opportunities as well as making other training facilities readily accessible to the users so as to improve students' academic performance in both internal and external examinations, it has been observed that all is not well with the system as a result of poor performance of the students recorded in the public examinations in the recent years; the persistent poor performance of secondary school students in public examinations such as the senior school certificate examinations (SSCE) in the country. Nigeria in the recent times has made the development of secondary education in the country a difficult task. Parents, guardians and other stakeholders in education industry have variously commented on the performances of secondary school students particularly in basic science subjects. In an attempt to ensure that their children perform better in SSCE and consequently, gain admission to universities of their choice, some parents and guardians have made a particular choice of type of secondary school they want for their children not minding the location and the cost implication of the school chosen.

2.0 Material and methods

The method used in collecting the data is documentary method via secondary source. It was collected from the National Examination Council state office, Kano. The data consist of scores of 100 students for 2017 NECO exams in six subjects which were grouped into Set-Z_1 = {Physics, Chemistry and Biology} and Set Z_2 = {Accounting, Commerce and Government}.

2.1 Statistical tools

The statistical tools used to achieve the aims of this research are canonical correlation analysis and Hotelling's T^2, NCSS is used to make the work easier.

2.1.1 Canonical Correlation Analysis

Canonical correlation analysis is the study of the linear relations between two sets of variables. It is the multivariate extension of correlation analysis.

Suppose you have given a group of students two tests of ten questions each and wish to determine the overall correlation between these two tests. Canonical correlation finds a weighted average of the questions from the first test and correlates this with a weighted average of the questions from the second test. The weights are constructed to maximize the correlation between these two averages. This correlation is called the first canonical correlation coefficient.

You can create another set of weighted averages unrelated to the first and calculate their correlation. This correlation is the second canonical correlation coefficient. This process continues until the number of canonical correlations equals the number of variables in the smallest group.

The methodology of canonical correlation was originally development by Harold Hotelling in 1935. The canonical correlation is the maximum correlation between linear functions of the two vector variables. However, after that pair of linear functions that maximally correlates has been located, there may be an opportunity to locate additional pairs of functions that correlate subject to the restriction that the functions in each new pair may be uncorrelated with all previously located functions in both domains that is Orthogonality. This simply means that for any two vector variables, we can have more than one canonical correlation.

2.1.2 Basic Issues

Some of the issues that must be dealt with during a canonical correlation analysis are:

1. Determining the number of canonical variate pairs to use. The number of pairs possible is equal to the smaller of the number of variables in each set.
2. The canonical variates themselves often need to be interpreted. As in factor analysis, you are dealing with mathematically constructed variates that are usually difficult to interpret. However, in this case, you must relate two constructed variates to each other.
3. The importance of each variate must be evaluated from two points of view. You have to determine the strength of the relationship between the variate and the variables from which it was created. You also need to study the strength of the relationship between the corresponding X and Y variates.
4. Do you have a large enough sample size? In social science work you will often need a minimum often cases per variable. In fields with more reliable data, you can get by with a little less.

2.1.3 Technical Details

As the name suggests, canonical correlation analysis is based on the correlations between two sets of variables which we call Y and X.

The correlation matrix of all the variables is divided into four parts:

1. The correlations among the X variables. R_{xx}
2. The correlations among the Y variables. R_{yy}
3. The correlations between the X and Y variables.R_{xy}
4. The correlations between the Y and X variables.R_{yx}

$$R = \begin{pmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{pmatrix}$$

Canonical correlation analysis may be defined using the singular value decomposition of a matrix C where:

$$C = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy} \qquad 3.1$$

The diagonal matrix of the singular values of C is made up of the eigenvalues of C. The eigenvalue of the matrix C is equal to the square of the canonical correlation which is called r_{ci}^2. Hence, the canonical correlation is the square root of the eigenvalue of C.

An investigator frequently has two vector variables, $Z_1$and Z_2 for a sample subjects, each vector variable representing measurements from a particular domain. It may be that Z_1 would be a set of independent (or predictor) measures and Z_2 would be a set of dependent (or response) measures. Another possibility is that the two domains are conceptually different although they are measured concurrently on the subjects. Thus Z_1 is a set of scores in three science subjects (Physics, Chemistry and Biology) and Z_2 is a set of scores in three non-science subjects (Accounting, Commerce and Government). The research question then would be to display the Inter — relation between the two (2) sets.

Of course, the bivariate correlation between pairs of measures taking one from each domain are of interest, but there may be a great many more of these e.g. if $Z_1$and Z_2 each have ten variables; then there are 100 bivariate correlations between pairs of variables in Z_1 and Z_2. To try to think about all these correlations simultaneously is very difficult if one is trying to generalize about the extent and nature of inter-relationship of the domains.

The actual computation of the canonical correlations involves the solution of a complicated eigen structure problem, which can be expressed in terms of the partitions of the correlation matrix for Z_1 and Z_2 together as:

$$(R_{22}^{-1} R_{21} R_{11}^{-1} R12 - \lambda I\ ) = 0 \qquad 3.2$$

The eigenvalues (characteristic root), λ_j is the square of the canonical correlation coefficient R_{cj}. In assigning Z_1 and Z_2, we assume that $p \leq\ q$. the smallest possible example of canonical correlation is when $p =\ q =\ 2$.

3.0 Result and Discussion

The data below shows the score of students in NECO Exams 2013 in six selected subjects.

X_1 (Physics)
X_2 (Chemistry)
X_3 (Biology)
Y_1 (Accounting)
Y_2 (Commerce)
Y_3 (Government)

Table 3: 1: Scores of Students in NECO Exams

X_1	X_2	X_3	Y_1	Y_2	Y_3
60	65	70	62	50	52
64	59	68	60	49	50
55	54	52	60	62	40
70	60	74	71	74	58
57	47	47	60	65	47
40	45	60	40	50	52
52	42	47	52	52	52
52	42	57	52	52	57
57	47	47	52	52	52
52	52	52	62	57	52
62	47	57	57	57	62
57	42	42	57	52	62
57	47	52	47	57	47
52	50	47	62	57	47
67	52	52	57	52	47
62	65	60	62	57	47
57	47	47	52	57	47
62	47	52	67	57	52
52	47	47	62	52	52
67	57	57	67	52	57
62	42	52	67	52	62
62	62	57	67	62	42
62	52	62	67	52	67
47	52	47	67	57	47
72	67	52	67	68	52
67	52	52	67	57	52
58	42	47	52	47	52
75	65	70	65	70	65
70	68	72	60	70	50
52	47	40	30	52	57

62	70	75	62	70	57
62	70	60	62	62	52
62	47	52	62	52	52
57	52	47	67	52	57
47	47	52	47	35	40
52	54	52	52	42	52
52	47	52	57	62	52
50	47	52	30	40	45
62	47	52	42	52	52
52	52	47	47	52	47
57	52	52	47	52	52
42	47	52	57	57	57
52	42	42	47	57	52
52	47	42	57	52	52
87.5	62	50	47	62	52
52	47	42	47	52	52
62	52	42	60	52	57
70	60	70	47	6	52
57	47	52	52	47	57
52	47	52	65	57	62
52	52	47	52	52	57
65	60	55	52	52	62
47	47	52	47	57	52
57	52	57	52	52	47
67	52	52	57	52	47
67	52	52	57	52	47
67	52	52	52	52	52
57	47	52	52	47	47
67	52	52	62	52	52
57	57	52	52	47	52
57	42	52	62	52	52
52	42	52	52	52	47
52	67	19.5	62	60	52
52	47	52	52	52	47
72	47	70	57	60	52
52	52	42	52	65	52
52	42	42	52	70	47
57	52	47	60	47	52
52	52	42	52	52	56
65	52	64	52	52	47
52	52	42	52	70	56
52	47	52	52	57	47

52	47	52	57	52	52
57	47	52	57	52	47
62	52	52	57	80	52
62	52	52	67	62	52
62	52	52	62	60	52
52	47	47	57	52	47
52	47	47	57	52	47
60	52	52	52	80	67
60	52	52	52	57	67
57	52	52	62	60	67
50	42	52	52	52	62
47	52	47	57	52	57
62	47	47	62	58	62
52	42	42	57	57	67
60	52	67	52	70	67
57	52	47	52	57	67
70	70	73	65	60	67
52	57	52	57	52	62
67	60	52	57	62	67
60	52	65	62	60	67
57	57	52	57	52	67
62	70	52	52	57	67
52	52	47	52	52	52
62	47	57	57	57	67
52	42	57	57	52	62
57	52	52	52	57	67
67	52	62	52	52	62
57	47	47	47	62	52

Source: National Examination Council, Kano Office

3.1 Canonical Correlation Analysis

An initial step in canonical correlation analysis is an inspection of the correlation matrix of the given data.

Let S denote the data such that

$$S = \{setA, setB\}$$

Where:

$$SetA = \{Physics, Chemistry, Biology\}$$
$$SetB = \{Accounting, Commerce, Government\}$$

Table 3.2: Canonical Correlation Section of SetA and SetB

	Y_1	Y_2	Y_3	X_1	X_2	X_3
Y_1	1.00000	0.220845	0.123406	0.425078	0.431572	0.395379
Y_2	0.220845	1.00000	0.219736	0.238211	0.306340	0.119900
Y_3	0.123406	0.219736	1.00000	0.094322	0.231798	0.230272
X_1	0.425078	0. 233211	0.094322	1.00000	0.588491	0.479250
X_2	0.431572	0.306340	0.231798	0.588491	1.00000	0.658924
X_3	0.395379	0.119900	0.230272	0.479250	0.658924	1.000000

From the above table of simple correlations between pairs of variables, Chemistry and Biology turn out to have the highest correlation followed by Physics and Chemistry. While correlation between Physics and Hausa is the weakest. We therefore went further to calculate the correlation between the two groups which is in the next table.

Table 3.3: Canonical Correlation Section

Canonical Function	Canonical Correlation	Eigen Values
1	0.538236	0.289697
2	0.219313	0.048103
3	0.145370	0.021132

The result of the canonical correlation reveals that the first canonical correlation gives the maximum correlation between students' performance in science and non-science subjects, followed by the second then the third; we then have to test further for the significance of each of the canonical correlation.

Table 3.4: Probability values and Wilk's Lamda values

S/N	N	DF	P-value	Wilk's Lambda
1	100	9	0.000001	0.661847
2	100	4	0.149845	0.931781
3	100	1	0.153227	0.978868

The result shows that for the first canonical correlation with p-values less than the alpha $(\alpha = 0.05)$, we reject the null hypothesis and conclude that there is a significant correlation between students' performance in science subjects and non-science subjects.

3.2 Hotelling's T^2

Hotelling's T^2 Test for the difference in students' performance in science subjects and non-science subjects. We first check for the homogeneity of variance between the two sets. The result is shown in the table below.

Homogeneity of Variance Tests

Table 3.5: Bartlett-Box Result

Variables Tested (Box's M Test)	Test Value	DF1	DF2	F-approx.	F-Prob.	Chi-Sq Approx	Chi-Sq. Prob.
ALL	40.928	6	284044	6.709	0.00001	40.256	0.00001

The result using Box' M test reveals that there is unequal variance between the two groups. Therefore we use an alternative to T^2.

Table 3.3: Hotelling's T^2 Test Section

Covariance Assumption	T^2	DF1	DF2	Prob. Level
Equal	27.161	3	198.0	0.00001
Unequal	27.161	3	193.3	0.00001

Since the p-value < 0.05 we reject the null hypothesis and conclude that there is a significant difference in the students' performance in science and non-science subjects.

4.0 Summary/Conclusion and Recommendation

In canonical correlation, a linear combination of the variables in a set defines each dimension measured by the set. Each of these linear combinations is called canonical variate. Canonical correlations are product moment correlations between pairs of canonical varieties, each pair consisting of the canonical variate from each set. Thus, each canonical correlation is measure of the degree of relationship between two dimensions, one measured by each set of variables. The maximum number of variables in the smaller set. In this analysis, we have two set of observations, set of X_s variables and set of Y_s variables each consists of only three variables. That is $p = q = 3$. The NCSS result gives the simple correlations between the variables with chemistry and Biology having the highest correlation followed by Physics and Chemistry, while Physics and Commerce have the weakest relationship. The canonical correlation gives the maximum correlation between the two sets as 0.5382 which shows a fairly strong relationship exists between students' performance in science subjects and non-science subjects. A test for the statistical significance of these canonical correlations carried out using Wilk's λ criterion reveals only the first is statistically significant.

Hotelling's T^2 test is used to compare the two mean vectors from the two sets to ascertain if there is difference in the students' performance in science and non-science subjects.

Test for quality of variance using Box's M test revealed that the variance of the students' scores in the two sets are not equal, therefore an alternative to T^2 is used which shows that there is actually difference in the students' performance in science subjects and non-science subjects.

4.1 Conclusion

In canonical correlation analysis, one (1) value for the canonical correlation obtained is $R_{C1} = 0.538236$. Test of statistical significance of these canonical correlations using Wilk's λ criterion reveals that only R_{c1} is statistically significant. Homogeneity of variance test using Box's M test was carried out which shows that the variances of the two sets are unequal, so an alternative to T^2 is used to compare the mean vectors of the two sets to see whether performance in science subjects differs from performance in non-science subjects.

Thus, from these findings we have an overwhelming evidence to infer that, students that perform better in science subjects also perform better in non-science subjects. But generally, students pay more attention to their science subjects than they do in other subjects.

4.2 Recommendation

Having analyzed the data critically and draw reasonable conclusions, it is important to give some recommendations that can be used in correcting the lapses.

The following recommendations were made:
First, school management should pay equal attention to the provision of teaching and learning materials for non-science subjects as they do in the area of science.

Secondly, school management as well as teachers should encourage students to pay more attention to others subjects such as economics because it will help them to become successful entrepreneurs at this time when white collar jobs are difficult to get.

References

Aliyu, U.A., (2001). Statistical Methods for Biometric and Medical Research. Published by Millenium Printing and Publishing Company Limited Kaduna State Nigeria.

Abdullahi, M.L., (2012). Methodology in Research Report Writing, Kaduna: Yambe Enterprise No. 2 Engineer Road Gonigora Kaduna.

Anderson, T.W. (1971). An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons.

Adepoju, T.L. (2008). Department of Educational Administration and Planning, Obefemi Awolowo University, Ile-Ife, Osun State-Nigeria. A Study of Secondary School Students' Academic Performance at the Senior School Certificate Examinations and Implication

Barkett, M.S (1963). Multivariate Analysis Journal of the Royal Statistics Society Series B.

Bhatia, V.K. (2007). Canonical Correlation Analysis. url iasri.res.in/eBook/EBADAT/4-Applications of Multivairate Techniques/3-canonical correlation.pdf

Cooey, W.W. and P.R. Lohnes (1971). Multivariate Data Analysis. John Wiley and Sons Inc.

Deemer, R. (2004). School as Agent of Education Can Influence Students Vol. 8: Iss1, Article 28.

Dillon, W.R. and Goldstein, W. (1984). Multivariate Analysis Method and Applications. New York: Wiley.

Fisher, R.A. (1936). The Use of Multiple Measurement in Taxonomic Problem Annals of Eugenics.

Johnson, R.A. and Wichern, D.W. (2002). Applied Multivariate Statistical Analysis (5th ed.): New Jersey: Prentice Hall

Murphy, R.J.L. (1981): Symposium: Examination "O" Level Grades and Teachers estimate as predictors of "A" Level results of UUCA applicatints. British Journal of Education Psychology. 51(1):1-9

Nbina, Jacobson Barineka (2000). Analysis of Poor Performance of Senior Secondary Students in Chemistry in Nigeria. Department of Curriculum Studies and Education