IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 12, Issue 6 Ver. VI (Nov. - Dec.2016), PP 13-17
www.iosrjournals.org

Canonical Correlation A Veritable Tool for Decision Making

Okeke, Evelyn Nkiruka and Okeke, Joseph Uchenna
Department of Mathematics and Statistics, Federal University Wukari, Nigeria

I. Introduction

Canonical correlation is a multivariate tool that measures the strength of association between two sets of variables. It extends bivariate correlation, allowing more than two independents variables with more than two dependent variables to be analyzed together at once. This correlation tells us how the best linear combinations of predictor variables related to the best linear combinations of the dependent variable. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets.

Suppose you are given a group of old women suffering from three different ailments and are placed under four different types of drugs and you wish to determine the overall correlation between these two set of factors. On one hand you have different ailments like malaria parasite, high blood pressure, and stomach ulcer. But you also might have chloroquin, Macalum, Anderna and Waipa ACT on the other hand. Canonical correlation finds a weighted linear function between the ailments and correlates this with a weighted linear function of the different drugs. The weights are constructed to maximize the correlation between these two functions. An important property of canonical correlation is that they are invariant with respect to affine transformation of the variables Borga (2001). This is the most important difference between canonical correlation and ordinary correlation analysis which highly depend on the basis in which the variables are described.

In this article we are going to study the relationship between Unified Tertiary Matriculation Examination (UMTE) and post- Unified Tertiary Matriculation Examination (post-UMTE) as a way of finding whether the decision of the Federal Government of Nigeria on scrapping of post- UMTE has substantial numerical backing.

1.1 Joint Admission and Matriculation Board in Nigeria

University education deal with course instructions in numerous disciplines, some target at specific professional courses like Medicine, Law, Engineering, Accounting etc, while some people enter diverse occupational courses such as business and administration courses. In Nigeria students gain admission into these disciplines if they finish writing Unified Tertiary Matriculation Examinations UMTE and Post- UMTE and pass both examinations.

The establishment of Joint Admission and Matriculation Board (board in charge of conducting UMTE) was done in stages. In 1974, there were several universities in the country; every one of them was conducting its own entry examination. In February 1978 the legal instruction establishing the board was promulgated by the act No 2 of 1978 of the Federal Military Government. In August 1998, the Federal Executive council amendments were codified into degree No 33 Section 2 1989 empowering the Joint Admission and Matriculation Board (JAMB) to:

- Conduct matriculation examination for entry into all universities, polytechnic and colleges of education in Nigeria.
- Place suitable and qualified candidates into the tertiary institutions after having taken into account the vacancies available in each tertiary institution, which the registrar and chief Executive of the board approve.

The JAMB examination was made compulsory for only the student who finished Ordinary level (or O level) and wanted to enter higher institution of learning. Universities use the JAMB score gotten by students to place them to their different specialization or course of study. In 2005, the Ministry of Education introduced an examination called Post-JAMB examination, which was being organized by the universities themselves. The Ministry added that a student who took JAMB examination and meet the cutoff will take Post JAMB to enable him enter University using the assessment from the university too.

In June this year 2016, the Federal Government scrapped Post-Unified Tertiary Matriculation Examination, UTME (or Post-JAMB), as a pre-condition to gaining admission into universities in the country. According to the Ministry of Education the federal government and stakeholders have confidence in the examinations conducted by the joint Admission and Matriculation Board, JAMB, and so there was no need for other examinations to be conducted by the universities after JAMB exams (Akinboade-Orire 2016).

2.2 **Canonical correlation**

In canonical correlation analysis (CCA), the correlation is between the linear combinations created for two sets of variables. By creating single variable that represents the Xs and another single variable that represents the Ys variables, CCA tries to see that the correlations between the projected single variables are mutually maximized, that is, if

$$F = f_1x_1 + f_2x_2 + \cdots + f_px_p = x'f \qquad 1$$

and

$$G = g_1y_1 + g_2y_2 + \ ... + g_py_p = y'g$$

the function to be maximized is

$$\rho = \frac{E(xy)}{\sqrt{E(x^2)E(y^2)}} = \frac{E(x'f)'(y'g)}{\sqrt{E(x'f)^2E(y'g)^2}}$$
$$= \frac{E(f'xy'g)}{\sqrt{E(f'xx'f)E(g'yy'g)}}$$
$$= \frac{f'C_{xy}g}{\sqrt{f'C_{xx}fg'C_{yy}g}} \qquad 2$$

The maximum of ρ with respect to f and g is the maximum canonical correlation. The subsequent canonical correlations are uncorrelated for different solutions, that is,

$$E(x_ix_j) = E(f_i'xx'f_i) = f_i'C_{xx}g_j = 0$$
$$E(y_iy_j) = E(g_j'yy'g_j) = g_j'C_{yy}g_j = 0 \qquad for\ i = j \qquad 3$$
$$E(x_iy_j) = E(f_i'xyg_j) = f_i'C_{xy}g_j = 0$$

where f_i and g_j are the x_i and y_i canonical loadings. The projections onto f and g are called canonical variates.

The sample canonical correlation coefficient r_c is

$$r_c = \frac{f'S_{xy}g}{\sqrt{(f'S_{xx}f)(g'S_{yy}g)}} \qquad 4$$

subject to $\|f\| = \|g\| = 1$; where $f$and g are two canonical vectors that maximizes r_c.
The first canonical correlation coefficient is the maximum correlation between F and G, for all F and G, that is,

$$r_{c1} = argmax\left(\frac{f'S_{xy}g}{\sqrt{(f'S_{xx}f)(g'S_{yy}g)}}\right) \qquad 5$$

The single variables, like F and G, which are linear combination of the original variables which is the same as the projections onto f and g is referred to as canonical variates.
The sample canonical correlations between x and y can as well be found by solving the eigenvalue equations

$$S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}\hat{f} = \lambda^2\hat{f}$$
$$S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}\hat{g} = \lambda^2\hat{g} \qquad 6$$

The eigenvalues λ^2 are the squared canonical correlation and the eigenvectors$\hat{f}\ and\ \hat{g}$ are the normalized canonical variates for x and y variables respectively. The number of non-zero solution to these equations are limited to the smallest dimensionality of x and y. Example if the dimensionality of x and y are 4 and 3 as in our own case, the maximum number of canonical correlations will be 3. Only one of the eigenvalue equations needs to be solved to get the eigenvalues since the two equations are related by

$$S_{xy}\hat{g} = \lambda\vartheta_xS_{xx}\hat{f}$$
$$S_{yx} = \lambda\vartheta_yS_{yy}\hat{g}$$

where

$$\vartheta_x = \vartheta_y{}^{-1} = \sqrt{\frac{\hat{g}'S_{yy}\hat{g}}{\hat{f}'S_{xx}\hat{f}}} \qquad 7$$

Canonical correlation can as well be calculated from the correlation matrices of the two sets of variables X and Y. Canonical correlation calculates the correlation from the raw data and uses this as an input data to calculating canonical correlation. It is the product of four correlation matrices, between dependent variables (inverse of it), independent variables (inverse of it) and between dependent variables and independent variables, that is,

$$R = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy} \qquad 8$$

where $R_{xx} = \begin{bmatrix} r_{x_1x_1} & \cdots & r_{x_1x_p} \\ \vdots & & \vdots \\ r_{x_px_1} & \cdots & r_{x_px_p} \end{bmatrix}$; $R_{yy} = \begin{bmatrix} r_{y_1y_1} & \cdots & r_{y_1y_p} \\ \vdots & & \vdots \\ r_{y_py_1} & \cdots & r_{y_py_p} \end{bmatrix}$; and $R_{yx} = \begin{bmatrix} r_{y_1x_1} & \cdots & r_{y_1x_p} \\ \vdots & & \vdots \\ r_{y_px_1} & \cdots & r_{y_px_p} \end{bmatrix}$

To get R_{xy}, wherever you see y in R_{yx} change to x and vice versa.
It also can be thought of as a product of regression coefficients for predicting Xs from Ys, and Ys from Xs. The eigenvalues of the R matrix

$$R = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{bmatrix} \qquad 9$$

represent the percentage of overlapping variance between the canonical variate pairs. To get the canonical correlation, you get the eigenvalues of R and take the square root, that is, the canonical correlation

$$r_{ci} = \sqrt{\lambda_i} \qquad 10$$

There will be as many canonical correlations as there are variables in the smaller set. Not all will be statistically significant or meaningful even if statistically significant. The eigenvector corresponding to each eigenvalue is transformed into the coefficients that specify the linear combination that will make up a canonical variate. The coefficients will be of two sets, a set will combine the independent variables while the other one will combine the dependent variable. Those coefficients have the same interpretation as regression coefficient. Canonical coefficients are multiplied by the standardized scores of the cases and summed to yield the canonical score for each case

$$\boldsymbol{X} = Z_x \beta_x \qquad 11$$
$$\boldsymbol{Y} = Z_y \beta_y \qquad 12$$

Where $\beta_y = R_{yy}^{-1/2} \hat{\beta}$ and $\beta_x = \Lambda R_{xx}^{-1} R_{xy} \beta_y$ are the canonical coefficients for X and Y respectively; $\hat{\beta}$ and Λ are from single value decomposition of $R = U' \Lambda \hat{\beta}$ and $Z_x$and Z_y are the standardized score for X and Y respectively.

1.21 Canonical variate

Each canonical variate is interpreted with canonical loadings, the correlation of the individual variable and their respective variates. Canonical loadings are similar to the factor loadings of each variable as were described in factor analysis. A unique characteristic of canonical correlation is that it develops multiple canonical functions. Each function is independent (orthogonal) from the other canonical function so that they represent different relationship found among the sets of dependent and independent variables. The canonical loadings of the individual variables differ in each canonical function and represent the variable's contribution to the specific relationship being depicted. Each canonical function consists of a different pair of canonical variates (one for the independent variables and the other for the dependent variables), each function representing a different relationship between the sets of variables. The researcher retains and interprets only the statistically significant canonical functions. Canonical functions are somewhat analogous to the discriminant functions in discriminant analysis in which each represents a different dimension of disimination in the dependent variable.

1.22 Canonical loadings

A measure of how well the variate(s) on either side relate to their own set of measured variables is provided by the canonical loadings (or structure coefficients). Canonical loadings represent the correlation between each variable and its own canonical variate. This equals the canonical coefficient if all the variables were uncorrelated with one another. By squaring each of the canonical loadings, one obtains a measure of the amount of variation in each of the variables explained by the canonical variate. This is obtained by multiplying the correlation matrix of the variable by the matrix of the canonical coefficient.

The canonical loading for each case

$$L_x = R_{xx} \beta_x \qquad 13$$
$$L_y = R_{yy} \beta_y \qquad 14$$

1.23 Canonical communality coefficient which is the sum of the squared loadings across all variates for a given variable measures how much of a given original variable's variance is reproducible from the canonical variates.

1.24 Adequacy coefficient

The amount of shared variance explained by a set of variables by a canonical variate of the set called the canonical variate adequacy coefficient (AC) is obtained by the sum of the squared loadings divided by the number of the variables in the set. This is a measure of how well a given canonical variable represents the variance in the set of original variables.

$$AC_{xk} = \frac{L^2_{xk1} + \cdots + L^2_{xki}}{i} \qquad 15$$

$$AC_{yk} = \frac{L^2_{yk1} + \cdots + L^2_{yki}}{j} \qquad 16$$

where:

AC_{xt} is the shared variance of the k[th] variate of independent (X) variables

AC_{yk} is the shared variance of the k[th] variate of dependent (Y) variables

L^2_{xki} is the squared loading of the i[th] independent variable in the k[th] variate of X variables

L^2_{ykj} is the squared loading of the j[th] dependent variable in the k[th] variate of Y variables

1.25 Canonical root

An estimate of the account of shared variance between the dependent and independent variates is called canonical root. This is the squared correlation between the independent canonical variate and the dependent canonical variate. The canonical root **r** is calculated by

$$r_i = 100(r^2_{ci}) \qquad 17$$

1.26 Redundancy index

Redundancy has to do with assessing the effectiveness of the canonical analysis in capturing the variance in the original variable. It is used when exploring relationship between the independent and the dependent variables. Redundancy measures how much of the average proportion of the variance of the original variables of one set may be predicted from the variables in the other set. High redundancy suggests, perhaps, high ability to predict. To have a high redundancy index, one must have a high canonical correlation and a high degree of shared variance by its own variate. A high canonical correlation alone does not ensure a valuable canonical function. Redundancy indices are calculated for both the dependent and the independent variate, although in most instances the researcher is concerned only with the variance extracted from the dependent variable set, which provides a much more realistic measure of the predictive ability of canonical relationship. The redundancy index of a variate is obtained by multiplying adequacy coefficient (shared variance of the variate) and squared canonical correlation, that is,

$$RI = AC_{xk} \times {r_c}^2 \text{ for the predictor variable} \qquad 18$$

and

$$RI = AC_{yk} \times {r_c}^2 \text{ for the dependent variable} \qquad 19$$

Redundancy analysis from some packages (like SPSS and others) gives a total of four measures

- The percentage of variance in the set of original individual dependent variables explained by the independent canonical variate
- A measure of how well the independent canonical variate predicts the values of the original dependent variables (AC_{xk})
- A measure of how well the dependent canonical variate predicts the values of the original dependent variables (AC_{yk})
- A measure of whether the dependent canonical variate predicts the values of the original independent variables. (Hair et. al. 1998)

II. Numerical Application

2.1 Source of Data

The data we used were obtained from unpublished degree project submitted at the Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria by Onyenanu (2007). The data is from the files of post JAMB and JAMB results of students of 2006/2007 academic session. Out of the total number of student that took post JAMB that session 50 students were randomly selected for the study. The results selected are in two sets, one a set of four subjects taken by students from JAMB and another set of three subjects from post JAMB.

The subjects taken by the students from JAMB were English (X_1), Mathematics (X_2), Economics (X_3), and Agric (X_4) while the ones taken from post JAMB were English (Y_1), Mathematics (Y_2), and Economics (Y_3).

2.2 Calculation of Canonical correlation.

Since our data do not have different scales of measurement the covariance matrices of the two sets of data S_{xx}, S_{yy} and S_{xy} were computed and used as an input to calculating canonical correlation. The canonical correlations of our data were computed using eigenvalue equations (16). The eigenvalues were first computed and from there we got the canonical correlation using the relationship stated at equation (10).

III. Result and discussion

The results of the analysis give three canonical correlations $r_{c1} = 0.4589$, $r_{c2} = 0.2152$, and $r_{c3} = 0.1229$. $r_{c1} = 0.4589$ is the best overall measures of association while $r_{c2} = 0.2152$, and $r_{c1} = 0.4589$ provide measures of supplementary dimension of linear relationship between x and y. The eigenvalues which is the squared canonical correlations are 0.2106, 0.0463, and 0.0151.

From the first canonical correlation we observed a weak association between the JAMB and post JAMB result. The canonical variates F_i for the JAMB result are

$$F_1 = 0.7025x_1 + 0.0011x_2 - 0.6651x_3 + 0.2533x_4$$
$$F_2 = -0.3673 - 0.2863x_2 - 0.7242x_3 - 0.5079x_4$$
$$F_3 = 0.6169x_1 + 0.5339 + 0.2463x_3 + 0.5232x_4$$
$$F_4 = 0.4830x_1 - 0.8154x_2 - 0.2816 + 0.1502x_4$$

The canonical variates G_i for the post JAMB result are

$$G_1 = -0.8618y_1 + 0.5028y_2 + 0.0675y_3$$
$$G_2 = 0.4104y_1 + 0.5349y_2 - 0.7386y_3$$
$$G_3 = 0.0106y_1 + 0.6597y_2 + 0.7514y_3$$

The results also showed that 21.1% of the total variation of F_1 is accounted for by its relationship with G_1.Hence $r_{c1} = 0.4589$ gives the strongest relationship between the JAMB and post JAMB examination. From the canonical variates we observed that English language and Economics are subjects that strongly determine students' performance in JAMB that session. The canonical variate of post JAMB revealed that English and mathematics determined the performances of student in that session. Since the aim of this research has been achieved through these few analyses we wish to end our result discussion here, other information can be obtained using the equations in this article.

IV. Conclusion

The maximum value of the canonical correlations $r_{c1} = 0.4589$ and the maximum percentage of 21.1% of total variation in post JAMB examination that was accounted for by JAMB examination revealed that there is weak association between the JAMB and post JAMB examination. This is to say that student's performance in post JAMB may not be attributed by his performance in JAMB. This goes a long way to say that a student who performs better in JAMB may not perform well in post JAMB. This could be that the two examination bodies may be using completely different standard.

References

[1]. Onyenanu O.A. (2007). Canonical correlation analysis on the comparison of Jamb results and post jamb result 2006-2007 session, Unpublished B.Sc research project, Nnamdi Azikiwe University, Awka, Nigeria.

[2]. Akinboade- Orire L. (2016).Why FG of Nigeria Scrapped post UMTE tests, Vanguard News June 3, Follow@vanguardngrnews, http://www.vanguardngr.com/2016/06/fg-scrapped-post-utme-tests/

[3]. Borga M. (2001). Canonical correlation a Tutorial, www.imt.liu.se/ magnus/ cca / tutorial/ tutorial.pdf

[4]. Hair J. F, Anderson Jr. R.E, Tatham R. L, and Black W. C (1995). Canonical correlation a supplement to multivariate data analysis, Multivariate Data Analysis, Pearson Prentice Hall Publishing, 1-44.