DOCUMENT RESUME

ED 280 897 TM 870 230

AUTHOR Chacko, Harsha E.
TITLE An Example of the Use of Canonical Correlation Analysis.
PUB DATE Nov 86
NOTE 11p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Memphis, TN, November 19-21, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; Higher Education; Interest Inventories; *Mathematical Models; *Multivariate Analysis; Predictor Variables; Rating Scales; Statistical Studies; Undergraduate Students; Vocational Interests
IDENTIFIERS *Strong Vocational Interest Inventory

ABSTRACT

Canonical correlation analysis is a multivariate statistical model which facilitates the study of interrelationships among multiple dependent variables and multiple independent variables. It identifies components of one set of variables that are most highly related linearly to the components of the other set of variables. The underlying logic of canonical correlation analysis involves the derivation of a linear conbination of variables from each of the two sets of variables so that correlation between the two sets is maximized. Few research studies that use canonical correlation are reported in the literature because of: (1) prohibitive calculations prior to the use of computers; (2) limited awareness of canonical methods; (3) a multitude of mathematical symbolism used in discussions of the technique in textbooks; and (4) difficulty in interpreting canonical results. Greater use of the technique will be facilitated as computer packages become more readily available and the technique becomes more familiar. An illustration of the technique examines the relationship between the academic comfort and introversion/extraversion scores composite with the composite of the six interest areas of the Strong Vocational Interest Inventory. (LMO)

AN EXAMPLE OF THE USE OF CANONICAL CORRELATION ANALYSIS

Harsha E Chacko

University of New Orleans 70148

Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, November 19, 1986

USES OF CANONICAL CORRELATION ANALYSIS

Presented by HARSHA E. CHACKO

Canonical correlation analysis is the most general case of the general linear model, and all other parametric tests (ie. multiple regression, MANOVA, ANOVA) can be treated as special cases of canonical correlation analysis. Canonical correlation analysis is a multivariate statistical model which facilitates the study of interrelationships among multiple dependent variables and multiple independent variables. It is a useful tool in social science research since the reality we try to explain often consists of many interdependent variables.

Canonical correlation analysis identifies components of one set of variables that are most highly related linearly to the components of the other set of variables. The underlying logic of canonical correlation analysis involves the derivation of a linear combination of variables from each of the two sets of variables so that correlation between the two sets is maximized. The derivation of canonical functions is similar to the procedure used in principal component factor analysis. In factor analysis, the first factor extracted accounts for the maximum amount of variance in the set of variables. The second factor is computed so that it accounts for as much as possible of the variance not accounted for by the first factor, and so forth. Canonical correlation analysis follows a similar procedure, but tries to account for the maximum amount of correlation between the <u>two</u> sets of variables rather than within a single set of variables.

Thus the first canonical function is derived so as to have the highest intercorrelations possible between the two sets of variables. The second canonical function will exhibit the maximum amount of relationship between the two sets, that was not accounted for by the first function. The maximum number of canonical functions is equal to the number of variables in the smaller set--either independent or dependent. In the following example, two canonical functions will be derived since there are two dependent and six independent variables.

In general, there are few research studies reported in the literature that use canonical correlation. Prior to the use of computers, calculations were prohibitive. Other reasons for lack of use are: limited awareness of canonical methods, a multitude of mathematical symbolism used in discussions of the technique in textbooks, and difficulty in interpreting canonical results. However, as computer packages become more readily available, and the technique becomes more familiar to researchers, greater use of the technique will be facilitated.

THE STUDY

Data for this study consisted of scores from the various subscales of the Strong Vocational Interest Inventory (SVII). One hundred and eleven undergraduate students at the University of New Orleans completed the SVII and the variables selected for analysis were:

Dependent Variables

1. Academic Comfort Score

This scale is an indicator of a degree of comfort in an academic environment, a degree of interest in intellectual exercises and a strong orientation toward theoretical or research problems.

2. Introversion/Extroversion Score

This scale differentiates between people who prefer to work alone, to complete projects independently (i.e., "introverts") and people who enjoy working with others, in groups, and like being the center of extraction (i.e., "extroverts").

Independent Variables

The General Occupational Themes

The General Occupational Themes are scales that measure the six vocational types described by John L. Holland in his theory of careers. Holland's theory states the people can be assigned to one of six broad interest areas: Realistic, Investigative, Artistic, Social, Enterprising and Conventional. The independent variables were the scores obtained for each subject in each of these six interest areas.

Purpose of the Study

The purpose of the study is to examine the relationship between the academic comfort and introversion/extroversion scores composite with the composite of the six interest areas. Data were analyzed using the SPSS-X package.

Results and Interpretion

1. Canonical Functions and Significance

Eigenvalues and Canonical Correlations

Root No.	Eigenvalue	Pct.	Cum. Pct.	Canon. Cor.	Squared Cor.
1	3.92951	73.03280	73.03280	.89283	.79714
2	1.45096	26.96720	100.00000	.76941	.59200

Dimension Reduction Analysis

Roots	Wilks Lambda	F	Hypoth. DF	Error DF	Sig. of F
1 TO 2	.08277	42.50334	12.00	206.00	.000
2 TO 2	.40800	30.18002	5.00	104.00	.000

Since there are two dependent variables, two canonical functions were derived (Roots 1 and 2). Both canonical functions are statistically significant (Dimension Reduction Analysis) with the first function having a squared canonical correlation of 0.79 and the second a squared canonical correlation of 0.59. These may interpreted similar to squared multiple correlations obtained in regression analysis. Wilk's Lambda is a statistic for testing the statistical significance canonical correlations and lambda may range from a value of zero to one. The closer lambda is to zero the more likely canonical correlations will be statistically significant.

Structure Correlation Coefficients

Correlations between DEPENDENT and canonical variables

	Function No.	
Variable	1	2
AC	.96256	.27108
IE	−0.62220	.78286

Correlations between COVARIATES and canonical variables

	Can. Var.	
Covariate	1	2
REAL	.39286	−0.08492
INV	.86098	.28193
ART	.72394	−0.20265
SOC	.63063	−0.50693
ENT	.25324	−0.71755
CON	.29143	.08849

Variable	Str. 1	Str. 2	Sq.Str. 1	Sq.Str. 2	Communality
A.C.	.96256	.27108	.9265	.0735	1.0000
IE	−0.62220	.78286	.3871	.6129	1.0000
REAL	.39286	−0.08692	.1543	.0072	0.1615
INV	.86098	.23193	.7413	.0852	0.8265
ART	.72396	−0.20265	.5241	.0411	0.5652
SOC	.63063	−0.50693	.3977	.2570	0.6547
ENT	.25324	−0.71755	.0641	.5149	0.5790
CON	.29143	.08849	.0849	.0078	0.0927

Structure coefficients are analagous to factor structure coefficients in principal components analysis. A structure coefficient is the correlation between the predictor or dependent variable composites and the variables used to create the composites. Structure coefficients are helpful in interpreting canonical results in terms of each variable's contribution to the canonical solution. Each coefficient tells the reader what contribution a single variable makes to the explanatory power of the set of variables, therefore providing independent contributions of the variables to the variance of the composites. It is recommended that these coefficients be utilized for interpretation rather than function coefficients, because a function coefficient may be unstable due to multicollinearity (some of the variance may be explained by other variables and therefore the function coefficient may have artificially distorted values). A squared canonical structure coefficient indicates the proportion of the variance linearly shared by a variable with the variable's canonical composite. In this example, the squared structure correlation between the Academic Comfort Variable and the first canonical function is 0.9265 while that of the Introversion/Extroversion variable is 0.3871. This may be interpreted that the first function has more to do with Academic Comfort rather that with Introversion/Extroversion. On the other hand, the squared structure correlation between Academic Comfort and the second canonical function is 0.0735 while that of Introversion/Extroversion is 0.6129. This means that the second canonical function has more to do with Introversion/Extroversion. The same interpretation can be made on examining the structure correlations of the independent variables. Thus the canonical correlation analysis shows that Academic Comfort is related to the Investigative, Artistic and Social types while extroversion is related to Enterprising and Social types. The communality coefficient is the sum of all of a variable's squared structure coefficients (therefore one must square the structure coefficients for one variable and add up the values). It is an indication of what proportion of each variable's variance is reproducible from the total canonical results. These coefficients indicate how useful each variable is in the analysis. The highest communalities here are for the Investigative, Social, Enterprising and Artistic variables.

Variate adequacy coefficient

The average of all the squared structure coefficients for the variables in one set with respect to one function, is a canonical variate adequacy coefficient. This indicates how adequately a given set of canonical variate scores perform with respect to representing all the variance of the original unweighted variables in the set. In this analysis, they are:

Variance explained by canonical variables of DEPENDENT variables

Can. Var.	Pct Var DEP	Cum Pct DEP	Pct Var COV	Cum Pct COV
1	65.68240	65.68240	52.35807	52.35807
2	34.31760	100.00000	20.31591	72.67399

Variance explained by canonical variables of the COVARIATES

Can. Var.	Pct Var DEP	Cum Pct DEP	Pct Var COV	Cum Pct COV
1	26.12581	26.12581	32.77442	32.77442
2	8.95352	35.07932	15.12426	47.89868

In other words, the first canonical function represents 65.68% of the variance in the dependent variables and 32.77% of the variance in the independent variables. The second function represents 34.32% of the variance in the dependent variable and 15.12% of variance in the independent set.

3. Redundancy coefficient

The redundancy coefficient is an index of the average proportion of variance in the variables in one set that is reproducible from the variables in the other set. It is an evaluation of the adequacy of prediction and not association because it is not fully affected by all the intercorrelations of the variables. The redundancy coefficient is equal to 1.00 only when two variates share exactly 100% of their variance and a variate perfectly represents the original variables in its domain. This almost never expected to be the case and so these coefficients may not usually be very useful.

In this analysis, for the first function, on the average 52.36% of the variance of the dependent variables is reproducible by the independent variables. For the second function, an average of 20.32% of the variance of the dependent variables is reproducible by the independent variables. For the independent variables, only averages of 26.13% and 8.95% of their variance can be reproducible by the first and second functions, respectively. The pooled redundancy coefficients for a given set of variables equals the average multiple correlation for the variables in the set when they are predicted by all the variables in the other set. In this analysis, they are 72.67% and 35.08% for the dependent and independent variables, respectively.

In Multivariate Data Analysis , Hair et al(p.206) write :

"In sum, it seems reasonable to use canonical correlation coefficients to test for the existence of overall relationships between sets of variables, but for a measure of the magnitude of the relationships, redundancy may be more appropriate."

However, Thompson(1984) has disagreed with this position:

> Redundancy analysis makes the most sense when the researcher's primary interest is in deriving functions which "capture" variance in the original, unweighted variables, that is, when the primary concern is function "adequacy".

References

1. Thompson, Bruce.(1984) Canonical Correlation Analysis. Beverly Hills,CA : Sage Publications, Inc.
2. Hair et al.(1979) Multivariate Data Analysis. Tulsa, Okhlahoma Petroleum Publishing Company.

ESSENTIALS IN PRINTOUT FROM SPSSX RUN

Eigenvalues and canonical correlations - look at eigen value, percentage, and canonical correlation value.

Dimension reduction analysis - look at wilks lambda, F, significance of F.

Correlations between dependent and canonical variables (structure coefficients).

Correlations between covariates and canonical variables (structure coefficients).

Variance explained by canonical variables of dependent variables (variate adequacy coefficients, redundancy coefficients and pooled redundancy coefficients of dependent set).

Variance explained by canonical variables of the covariates (variate adequacy coefficients, redundancy coefficients and pooled redundancy coefficients of independent set).