MULTIPLE PREDICTORS AND MULTIPLE CRITERIA[1]

W. L. Bashaw
Florida State University

This paper is an introductory discussion of a technique for studying in a single analysis the relationship of several predictors to several criteria. Canonical analysis is a method for the determination of the over-all linear relationship between two sets of variables. No attempt will be made to show the calculations involved in a canonical analysis; however, the rationale and applications of the method will be discussed.

The study of linear relationships between variables is one of the most important concerns in any area of scientific investigation. Several approaches have been discussed in the papers presented earlier. The simplest approach, which in some cases can be the least valuable, is the study of the variables under consideration by pairs. The appropriate measure of linear relationship between any pair of variables is some form of correlation coefficient. A more sophisticated approach is multiple correlation analysis. By the use of the multiple correlation technique, the investigator determines the joint relationship between a single criterion variable and several predictors. This technique involves the determination of the unique linear function of the predictors that correlates most highly with each criterion.

However, if the investigator is concerned with more than one criterion, multiple correlation analysis may not yield the most valuable results. An investigator who employs the multiple correlation technique restricts himself to the separate study of each of the criteria. Canonical analysis is the proper statistical technique for the study of the relationships between two sets of variables. An example of an appropriate application is the investigation of the relationships between the ninth grade test battery and the twelfth grade test battery. Another example is the investigation of the relationships between a set of socio-economic factors and a set of school achievement measures. A third example is the over-all relationship between two test batteries that are constructed to measure achievement of the same content.

The understanding of canonical analysis might best be developed by a comparison of it with regression methods and factor analysis. The models for these various analyses will be illustrated by

---

[1]Paper presented to the Eigth Annual FERA Testing Conference, Tallahassee, Florida, January 25, 1964.

a common example. Suppose that it is desired to study the relationship between two criteria and three predictors. The criteria to be predicted are senior year mathematics grades (Y_1) and achievement on a mathematics test (Y_2). The predictors are ninth grade mathematics grades (X_1), mathematics aptitude scores (X_2), and intelligence test scores (X_3).

The simplest model is that of the standard regression analysis that was presented in a previous paper. The purpose of multiple regression is to obtain the linear function of a set of predictors that correlates more highly with a criterion than does any other linear function. The correlation of the criterion with the linear function of predictors is the multiple correlation coefficient. The multiple regression model is

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \text{error}, \qquad (1)$$

where the X's are the observed predictor values, the Y is the criterion to be predicted, and the b's are the regression weights to be estimated. Separate equations must be determined for each of the two criteria.

Factor analysis methods can also be used in the solution to the example problem. Factor analysis is a technique for the determination of independent linear functions of a set of variables. The technique can be used to reduce the set of predictor variables to a set of new, mutually uncorrelated predictor variables. Usually, the number of independent linear functions of a set of variables is smaller than the number of variables in the set. Because of this, factor analysis can be used to reduce the number of predictor variables. The results of a factor analysis of a set of predictor data can be used to calculate factor scores for each subject. These factor scores have the important property of being uncorrelated with each other. The regression of the factor scores on the criterion can be determined by multiple regression methods. The concept of independent functions of the predictors is a basic concept of canonical analysis as well as factor analysis.

The example can be used to illustrate the role of factor analysis in prediction. Suppose a factor analysis of the predictor data yielded two factors (i.e., two independent linear functions of the predictors). Factor scores corresponding to the two uncorrelated factors can be written as

$$F_1 = a_{01} + a_{11}X_1 + a_{21}X_2 + a_{31}X_3 \text{ and} \qquad (2)$$

$$F_2 = a_{02} + a_{12}X_1 + a_{22}X_2 + a_{32}X_3, \qquad (3)$$

where the X's are observed predictor variables of values and the a's are weights determined by the factor analysis. The multiple regression model for relating the factor scores to the criterion is

$$Y = b_0 + b_1F_1 + b_2F_2 + \text{error}, \qquad (4)$$

where the b's are the regression weights to be estimated, the F's are the observed, uncorrelated factor scores,and Y is the criterion to be predicted. The correlation between Y and the function defined by the right-hand side of equation (4) is the multiple correlation coefficient. Separate equations must be determined for both Y_1 and Y_2.

It will be helpful to compare the model defined by equations (2), (3), and (4) with the standard regression model defined by equation (1). Note the similarity between the formulae for the factor scores and equation (1). If equations (2) and (3) are substituted into equation (4), and the appropriate simplifications are made, then the result would be quite similar to equation (1), the regular regression model.

Canonical analysis differs from the regression models given by equations (1) and (4) primarily because multiple criteria are involved. Similar to the factor score regression model (equation (4)), canonical analysis involves the reduction of the data to new, mutually independent variates. However, independent functions of the criteria, as well as the predictors, are obtained. The technique involves the determination of the unique composite of the predictors and the unique composite of the criteria that correlate maximally with each other. The composites are called "canonical variates" and the correlation between the canonical variates is called the "canonical correlation."

A complete canonical analysis consists of finding as many mutually independent pairs of canonical variates as are possible for the given set of data. The maximum number of canonical variate pairs is equal to either the number of predictors or the number of criteria, whichever is smaller. In general, the number of pairs is smaller than the maximum.

The canonical analysis model can be illustrated by the example used previously. Suppose it is desired to investigate the relationship of the three predictors to the scores on the mathematics test and the senior mathematics grades simultaneously.

The first step of the canonical analysis is to obtain two sets of weights that define the pair of canonical variates that yield

the largest canonical correlation for all possible pairs of variates. The models for the first pair of canonical variates can be written as

$$P_1 = a_{01} + a_{11}X_1 + a_{21}X_2 + a_{31}X_3, \text{ and} \quad (5)$$

$$Q_1 = b_{01} + b_{11}Y_1 + b_{21}Y_2, \quad (6)$$

where the a's and b's are determined so that the correlation between P_1 and Q_1 (the first canonical correlation) is as large as possible.

The second step will complete the analysis since there are only two criteria. A new pair of canonical variates is determined. The models for the new variates are

$$P_2 = a_{02} + a_{12}X_1 + a_{22}X_2 + a_{32}X_3, \text{ and} \quad (7)$$

$$Q_2 = b_{02} + b_{12}Y_1 + b_{22}Y_2. \quad (8)$$

The solution for the a's and b's for this second pair of variates makes the correlation between P_2 and Q_2 (the second canonical correlation) a maximum, subject to the restrictions that P_2 has a correlation of zero with Q_1 and P_1 and Q_2 has a correlation of zero with Q_1 and P_1. These restrictions are the same as the factor analysis restriction in equation (4) that requires F_1 and F_2 to be uncorrelated, except the restriction is extended to apply to the criteria variables as well as the predictors.

It is important to point out that canonical analysis does not lead to statistical systems for predicting each specific criterion. One does obtain predictions of "the most predictable" <u>functions</u> of the set of criteria. The usual application of the analysis is the study of relationships, not the determination of prediction systems.

A brief review of applications of canonical analysis can give some insight into the utility of the technique. Cooley and Lohnes (2) reviewed two studies. The first of these was an investigation of the relationship between early home environment and present orientation to people (6). The authors hypothesized that certain childhood environmental factors (predictors) are correlated with the ways that adults relate to other persons (criteria). Seven predictor measures and eight measures of orientation towards people were used. The canonical analysis resulted in one significant way in which the two sets of variables are related.

The second application discussed by Cooley and Lohnes is a study of the relationship between several group administered aptitude tests and several individually administered aptitude tests. One significant canonical correlation was obtained between the two sets of test scores, $R_c = .62$.

A third application relates intellectual variables to bio-chemical measures (5). This study was an investigation of the hypothesis that certain biological and chemical substances found in the human body are related to performance on cognitive tests. There were five measures of bio-chemical factors and seven measures of intellect. One significant canonical correlation was obtained ($R_c = .60$). This finding indicates a significant relationship between the set of bio-chemical factors and the set of cognitive test scores.

A generalization of canonical analysis has been made and is important to mention (3). The technique that has been discussed involves the relationship between two sets of variables. Horst has generalized the technique to include the single analysis of more than two sets of variables.

Horst, in addition to his generalization of canonical analysis, also discussed applications of the techniques. Horst points out that the techniques can be used to study the relationships between independent groups of persons who have taken the same test battery. This is the same problem as the identification of similarity of factor analyses for independent groups of subjects.

Another broad area of usefulness is the analysis of experiments or investigations involving multiple criteria. Examples of this are the relationship of pre- and post-tests for an experiment, the relationships between results of test batteries administered under different sets of instructions or conditions, and the analysis of Q-methodology data.

A final word about the calculations involved in a canonical analysis is appropriate. The solution to the analysis is not presented since an extensive discussion of matrix algebra is necessary for complete clarity. Although calculations by hand certainly are possible, they are so lengthy and complex that it is doubtful that one should attempt them. A few satisfactory programs for electronic computers are available in many computer program libraries. The programs usually require a fairly large amount of computer storage space; therefore, large computers such as those at the state universities would be necessary.

References

1. Anderson, T. W. An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons, 1958, Chap. 12.

2. Cooley, W. W., and Lohnes, R. R. Multivariate Procedures for the Behavioral Sciences. New York: John Wiley and Sons, 1962, Chap. 3. (Basic reference for calculations and computer program).

3. Horst, Paul "Relations Among m Sets of Measures," Psychometrika, 26 (1961), 129-149.

4. Hotelling, Harold. "Relations Between Two Sets of Variates," Biometrika, 28 (1936), 321-377.

5. King, F. J., Bowman, B. H. and Moreland, H. J. "Some Intellectual Correlates of Biochemical Variability," Behavioral Science, 6 (1961), 297-302.

6. Roe, A., and Siegleman, M. A Study of the Origin of Interests. In press.