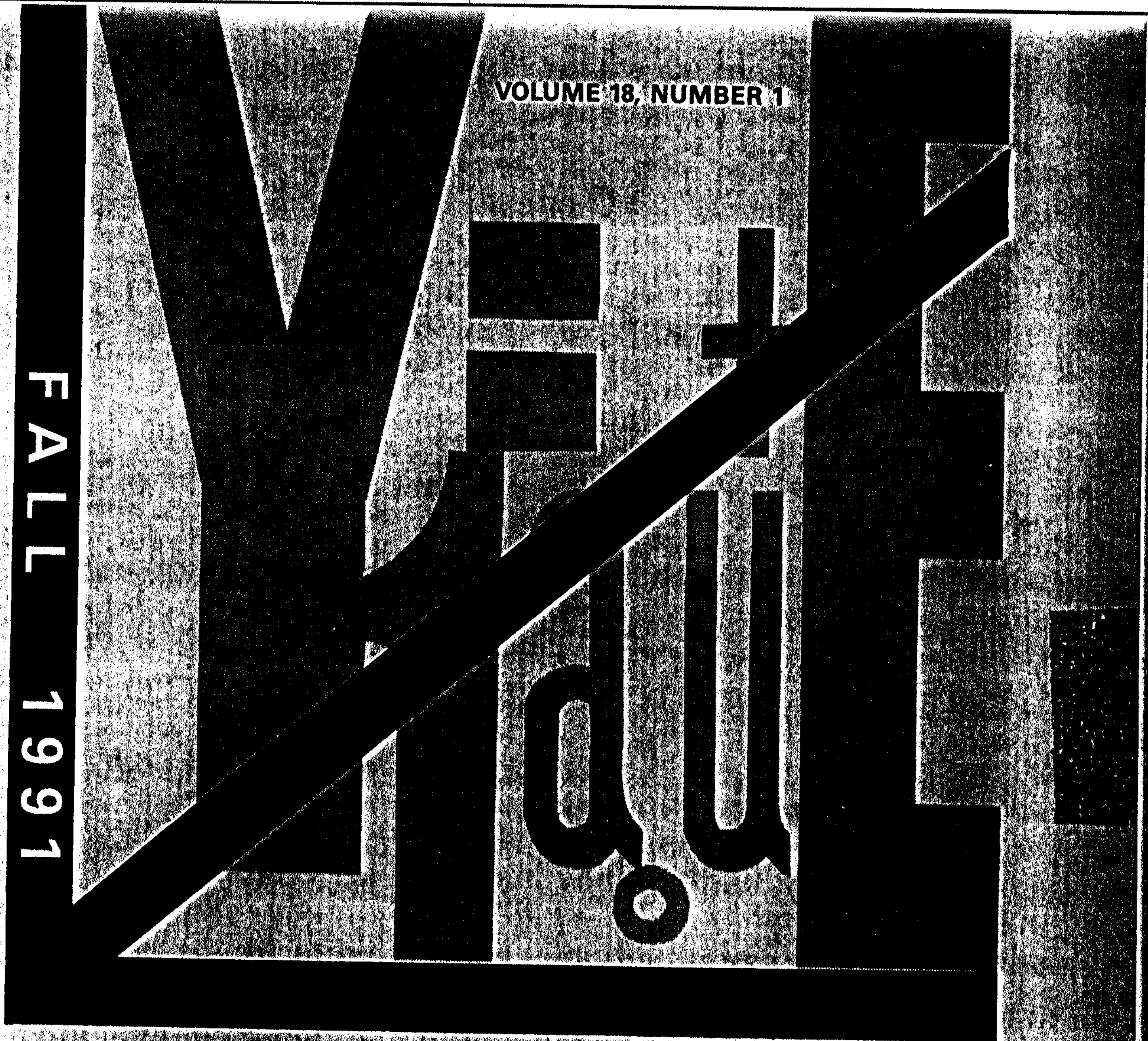


MULTIPLE LINEAR REGRESSION VIEWPOINTS

A publication of the Special Interest Group
on Multiple Linear Regression

MLRV Abstracts appear in CIJE, the ERIC System,
and microform copies are available from
University Microfilms International
MLRV is listed in EBSCO Librarians Handbook.

ISSN 0195-7171

MULTIPLE LINEAR REGRESSION VIEWPOINTS

Chairperson

Editor .... Isadore Newman
The University of Akron
Akron, OH 44325-4205

Assistant Editor .... Keith McNeil
New Mexico State University
Las Cruces, NM 88001

Cynthia Boone-Hawkins
The University of Akron
Akron, OH 44325-4205

Executive Secretary

Cover Artist .... David G. Barr

EDITORIAL BOARD

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

Comparison of Conjoint Analysis, Multiple Regression Models with Person Vectors and Profile Analysis to Assess Important Factors Used to Select Colleges

Isadore Newman
The University of Akron

John Frass
Ashland University

The purpose of this study was to investigate the relative effectiveness of the traditional conjoint analysis approach to the multiple regression approach that includes person vectors profiles analysis. It was expected that the more sophisticated models would increase the effectiveness in terms of its shrinkage estimates and the accuracy of its predictability of two holdout groups. The data source consisted of a sample of 100 students who rated eight colleges on five attributes--quality of education, financial aid, quality of dorm life, student/faculty relations, and social aid.

PRESENTED AT THE AMERICAN EDUCATIONAL RESEARCH ASSOCIAITON
1990 ANNUAL MEETING, BOSTON, MASSACHUSETTS.

Comparison of Conjoint Analysis, Multiple Regression Models with Person Vectors and Profile Analysis to Assess Important Factors Used to Select Colleges

Introduction

In recent years, many colleges and universities have faced increased competition for students. Thus, it has been increasingly important for an institution of higher education to be able to identify what factors are important to the students who chose to enroll in the institution.

Marketing research (Cattin & Wittink, 1982) has identified conjoint analysis as a very useful statistical technique in which one is interested in having the clients, students, or consumers prioritize a variety of items. Two other approaches also seem to be appropriate to use when attempting to assess the selection process of college-bound students: (1) multiple regression models with person vectors (Frass & Newman, 1989); and (2) profile analysis.

Objectives

This paper attempted to compare the ability of conjoint analysis, multiple regression models with person vectors, and profile analysis to produce information that could be used by college and university personnel to determine which factors were important to students when selecting a university or what type of students selected a given type of university.

Data Collection

The research instrument used to collect the data analyzed in this study focused on five institutional attributes reported to be of significance to students who martriculated to Ashland University. This list of attributes was developed through literature reviews (Tiernry 1980; Traynor, 1981; Kuh, Coomers, & Lindquist, 1984; Conant, Brow, & Mokwa, 1985), discussion with program advisors and students, and from the past experiences of admissions recruiters.

The five attributes included in this study were financial aid, social life, quality of dorm life, student-faculty relationships, and quality of education. Each of the five attributes had two levels. The two levels that were formed for each attribute were assigned a value of 0 or 1 in order to allow the researchers to quantitatively form hypothetical universities with various combinations of attribute levels. The attributes, levels, and values assigned to each level were as follows:

1. Quality of education
   a) reputation is not well known = 0
   b) reputation is well known = 1
2. Student/Faculty relationships
   a) faculty are accessible if sought = 0
   b) faculty are extremely accessible = 1
3. Quality of dorm life
   a) below my expectations = 0
   b) above my expectatons = 1

4. Financial aid

a) little financial need is met = 0
b) most financial need is met = 1

5. Social Life

a) few social activities are available = 0
b) many social activities are available = 1

Five attributes with two levels each would allow 32 different university profiles to be formed. With the assumption that interaction effects are negligible, the main effects could be estimated with only eight orthogonal arrays. The eight orthogonal arrays used in this study which were formed with the aid of the computer software entitled Conjoint Designer (Bretton-Clark, 1987), were listed in Table 1.

In addition to the eight orthogonal arrays, two arrays were designed to provide a means of assessing the degree of predictive validity. (See Table 1.) These two arrays were referred to as the "holdout universities" because they were not included in the estimation procedures.

The questionnaire was administered during the second week of the fall term of 1987 to freshman students enrolled in a freshman seminar course. The responses of 100 of the students were used in this study. See Fraas and Paugh (1989) for

Conjoint Analysis

The analysis conducted by the use of a software package (Bretton-Clark, 1987) produces a set of five regression coefficients plus a constant term for each student. That is, a separate regression analysis was performed on the data of each of the 100 students.

Each of the regression coefficients generated by the conjoint analysis for a given student indicated what would happen to the respondent's ratings of the universities when the attribute changed from the "zero" level to the "one" level. To illustrate the point, consider the regression coefficient value of 2.0 recorded for the financial attribute for respondent 1. If financial aid was to increase from the "little need being met" category to the "most need being met" category, the respondent's ratings of the universities would increase by 2.0 points on the 1 to 10 scale used on the questionnaire.

A relative importance figure was calculated for each attribute by dividing the sum of the five average regression coefficients into each of the average regression values. The five relative importance figures generated by this procedure were expressed as percentages.

Table 1

Orthogonal Arrays Used for Conjoint Analysis and Multiple Linear Regression Models

Universities	Quality of Education	Student/ Faculty Relation-ships	Quality of Dorm Life	Financial Aid	Social Life
A	0	0	0	0	0
B	1	0	0	1	1
C	1	1	1	1	0
D	0	1	1	0	1
E	0	1	0	1	0
F	1	1	0	0	1
G	1	0	1	0	0
H	0	0	1	1	1
Holdout Universities					
I	1	1	1	0	1
J	1	1	0	1	0

<u>Note</u>. Each characteristic is composed of two levels. The zero value indicates the presence of the lower of the two levels.

Results of the Conjoint Analysis

The relative importance figures indicated that financial aid was the most important attribute with a value of 26.24%. Financial aid was followed in importance by the quality of dorm life (21.29%), the quality of education (20.84%), the student/faculty relationships (16.63%), and the social life (15%). (See Table 2.)

Predictive Validity

The observed and predicted ratings for the holdout universities were used to provide two estimates of the ability of the results of the conjoint analysis to predict student ratings. The first estimate was a correlation coefficient for the predicted and observed ratings. The second estimate was an average absolute difference value for the difference between the predicted and observed ratings. The correlation coefficient value and the average absolute difference for the observed and predicted ratings were .37 and 1.87, respectively.

Multiple Linear Regression Model With a Surrogate Person Variable

Model Structure

The second approach used to analyze the survey information required the construction of a multiple linear regression model that included a surrogate person variable. Before such a model is presented, however, a discussion of a model that includes the actual person variables may prove helpful. The variables included in the model that used person variables (Model 1) were as follows:

Y = ratings of the eight hypothetical universities (values ranged from 1 to 10)

X1 = quality of education
0 = "low" level; 1 = "high" level)

X2 = student/faculty relationship
(0 = "low" level; 1 = "high" level)

X3 = quality of dorm life
(0 = "low" level; 1 = "high" level)

X4 = financial aid
(0 = "low" level; 1 = "high" level)

X5 = social life
(0 = "low" level; 1 = "high" level)

P1 = respondent 1

(1 if from respondent 1; 0 otherwise)

P2 = respondent 2

(1 if from respondent 2; 0 otherwise)

P99= respondent 99

(1 if from respondent 99; 0 otherwise)

The structure of the regression model with person variables was:

$$Y = aU + b1X1 = b2X2 = b3X3 = bb4X4 = b5X5 = b6P1 = b7P2 = \ldots b104P99 = e \quad \text{(model 1)}$$

The use of the person variable required by Model 1 is not practical due to their large number. Thus a multiple linear regression model designed to include a surrogate person variable was used. This surrogate person variable measured the impact of the 99 person variables required by Model 1.[1]

Table 2

Conjoint Analysis Results

Characteristic	Average Regression Coefficient	% of Relative Importance
Financial Aid	1.775	26.24
Quality of Dorm Life	1.440	21.29
Quality of Education	1.410	20.84
Student/Faculty Relationships	1.125	16.63
Social Life	1.015	15.00

Correlation coefficient between the predicted and observed ratings of the holdout universities = .37

Average absolute difference between the predicted and observed ratings of the holdout universities = 1.87

The value of the surrogate person variables was composed of an average rating for each person. The surrogate variables was represented in Model 2 by "X6." The values for this variable ranged from 2.625 to 8.5 for the 100 students.

The multiple regression model with the surrogate person variable (Model 2) used to analyze the survey information was as follows:

$$Y = aU = b1X2 = b2X2 = b3X3 = b4X4 = b5X5 = b6X6 = e \quad \text{(Model 2)}$$

The regression coefficients for the university attributes that were generated by Model 2 were equal to the average regression coefficients for the conjoint analysis (See Table 3).

Before the regression coefficients could be statistically tested, the standard errors had to be corrected for the appropriate degrees of freedom. The number of degrees of freedom was 695, which was equal to the sample size of 100 (number of students) multiplied by 8 (number of colleges) minus 6 (number of attributes plus one). Each of the regression coefficients for the university attributes was statistically significant at the .01 level. The multiple correlation coefficient was .764; and the R^2 value was .58.

__Predictive Validity__

The regression coefficients generated by Model 2 were used to predict the ratings of the holdout universities. The correlation coefficient for the predicted and observed ratings was .76. The average absolute difference between the predicted The same procedure applied to the second half of the data set resulted in a correlation coefficient value of .74 between the observed and predicted ratings. Again, this value shows little shrinkage (1.7%) from the multiple correlation coefficient of 753 for Model 2.

[1] Refer to Pedhazur (1977), Williams (1977; 1980), Fraas McDougall (1983), and Williams and Williams (1985a; 1985b) for discussions of a surrogate variable used to measure the amount of variation in the dependent variable associated with a set of person variables.

Comparison of the Results

The estimated impact of the university attributes on the student ratings by the conjoint analysis, and the multiple linear regression model with a surrogate person variable were identical. For both procedures, the order of importance was as follows: (1) financial aid, (2) quality of dorm life, (3) quality of education, (4) student/faculty relationships, and (5) quality of social life.

The multiple linear regression model with the surrogate person variable, however, produced a correlation coefficient value of .76 for the predicted and observed ratings of the holdout universities, as compared to the value of only .37 for the conjoint analysis.

The multiple linear regression model with the surrogate person variable also produced a lower average absolute difference between the predicted and observed ratings for the holdout universities than did the conjoint analysis. The average absolute difference values were 1.50 and 1.87.

The low R^2 values of the regression models that used the clusters as the independent variables indicated that the clusters were unable to explain the variation in the university ratings to any high degree. For this data set, the cluster information was of little assistance in identifying the importance of university characteristics as viewed by various groups of students.

Quannal Analysis

The following description of quannal analysis is heavily based on Vantubergen (1966) and Newman and Carolyn Benz (1988). The third data analysis procedure applied to the data set was quannal analysis. The purpose of using this procedure was to determine whether certain types of people could be identified that favored different types of schools.

The factor analysis computer program used in this study was QUANNAL Vantubergen, 1966). This program places squared multiple correlation values in the principle diagonal as commonality estimates and conducts a Q-analysis. This approach is appropriate for the purpose of differentiating between people int erms of the shape of their profiles.

Five steps are used in a Q factor analysis.

Step 1 - An intercorrelation matrix is formed by correlating every person's ratings of the items with every other person's rating of items.

Thus, the eight ratings for respondent 1 were correlated with the ratings of the other 99 respondents. The same procedure was followed for each respondent.

Step 2 - The matrix of intercorrelations if submitted to factor analysis so that "persons" are variables and items are observations. A principal axis solution is obtained. This result is submitted to a varimax rotation which produces orthogonal factors. On this basis, a factor represents a grouping of persons around a common pattern of sorting the items. Hence, a factor represents a type of "person" (Vantubergen, 1966).

Sub. No.	Two Factor Solution I	II	h	Sub. No.	Three Factor Solution I	II	III	h^2
1.	.22	.83	.75	1.	.30	.87	.03	.85
2.	.92	.17	.88	2	.87	.16	.39	.93
3.	.98	-.13	.97	3.	.84	-.16	.50	.98
4.	.75	.49	.81	4.	.33	.37	.86	.99
5.	.82	.19	.71	5.	.37	.05	.90	.95
6.	-.06	.90	.82	6.	-.04	.91	.03	.83
7.	.86	.09	.76	7.	.97	.14	.14	.99
8.	.17	.92	.88	8.	-.02	.87	.39	.91
% Total Var.	48	34	82	% Total Var.	34	32	27	93

The factor analytic model constructs hypothetical types of "persons" based on the way the actual people interviewed rated the items. One can group people by assigning them to the type that they are most like, i.e., the factor on which they have the highest loading.

Step 3 - Each pattern of items associated with each by weighting each item response of each item response of each of the persons most highly associated with a given factor by the degree to which they are loaded on that factor, the greater is the weight. These weighted responses are summed across each item separately. This procedure produces an item array of weighted responses for each factor in the rotated factor analysis solution selected. The arrays of weighted repsonses are then converted to z-scores (Vantubergen, 1966).

Hypothetical types constructed by the factor analytic model is based on a weighted pattern of the items (hypothetical types). The more a person's rating is like the hypothetical type, the more weight it receives in the average. The specific weight given is calculated as follows:

$$\text{weight} = \frac{r}{1 - r_2} \qquad \text{where: } r = \text{loading}$$

The weighted average is called an item factor array.

The persons used to estimate an array are highly associated with that type, but they are not associated to a high degree with any of the other types. For the persons selected, the square of the loading on that factor should approach the communality h_2. The arrays of weighted item ratings are converted to z scores. The array of z scores for each type is called the factor array.

Step 4 - The arrays of item z - scores for each factor (factor arrays) are ordered from most rejected for each factor. This provides a hierarchy of item acceptance for each factor or type of "persons" (Vantubergen, 1966).

The following are examples of hypothetical types of "persons" that the factor analytic model would construct:

Items	Types I	II	III
University 1	1.02	- .24	.72
University 2	1.53	1.03	1.54
University 3	.42	.31	-1.03
University 4	- .06	.32	- .51
University 5	-1.08	-1.35	-1.54
University 6	.80	1.20	.5
University 7	-1.20	.02	- .6
University 8	.70	1.50	2.0

When ordered in terms of the z-scores, the factor array becomes a hierarchy of items that are rated for each of the factors or types. The following is an example of the first typology (Type I):

Z-Score	Item
1.53	University 2
1.02	University 1
.80	University 6
.70	University 8
.42	University 3
- .06	University 4
-1.08	University 5
-1.20	University 7

Similar results were obtained for each type.

Step 5 - The arrays of item z-scores (factor arrays) for each type are compared by subtraction for each pair of factors. This produces arrays of difference scores for each pair of factors. This provides the basis for differentiating one factor or type of person from another Vantugergen, 1966).

This is accomplished by comparing the types dealing with the following questions:

1. What items differentiate one type from another type?
2. What items differentiate one type from all other types?
3. What items or areas of agreement seem to cut across all of the types?

Question 1 is dealt with by comparing the array for all types taken two at a time. The Z-scores for each pair of universities are subtracted and ranked according to absolute differences. To illustrate, consider the following:

	Type II	Type I-Type II	
1.02	- .24	1.26	University 1
-1.20	.02	1.22	University 7
.70	1.80	.80	University 8
1.53	1.03	.50	University 2
.80	1.20	.40	University 6
-1.08	-1.35	.27	University 5
- .06	.32	.38	University 4
.43	.31	.12	University 3

Similar analyses are conducted for all other comparisons.

Question 2. Question 2 was addressed by examining those items that are higher (or lower) in the array for one type than they are in the arrays for all other types. This process is similar to the process followed in Question 1. That is, the Z scores of Type I are compared to the average Z scores for Types II and III.

Question 3. To the extent that the z-scores for all types are nearly equal, one assumes agreement. A consensus item would be one in which the difference between the largest z-score given that item by one of the types and the smallest z score is less than 1.00. In our example, the consensus items would be the following:

Rating of Universities	Maximum Difference	Average Z-Scores Across Types
University 5	.46	1.32
University 2	.50	1.37
University 6	.70	.83
University 4	.83	.08

The average Z-scores of the consensus items and the Z-scores of the differentiation items, which resulted from addressing Questions 1, 2, 3, are used to describe the types. That is, the universities corresponding to the aforementioned Z-scores are used to identify types.

Results of Quannal Analysis

Three Q-factor analyses were computed. One analysis was based upon the ratings of the eight universities, the second on demographic variables, and the third on the university and demographic variables together. On all three of the Q-factor analyses, only one typology emerged.

In the first analysis, all of the 100 subjects were identified in Type I. In the second analysis, 99 of the 100 were identified in Type I. In the analysis combining the universities and demographic variables, 98 of the subjects were identified in Type I. As one can see form these results, only one type consistently emerged; therefore, we were unable to use differences in types as predictor variables. A multiple regression analysis by Fraas on the impact of the demographic variable of the data further validates the homogeneity of this sample.

Since we were in a desperate search for more than one type, it was suggested that we try a cluster approach, which tends to produce more than one type. Ward's (1963) clustering program takes a set of N objects, which are measured on a number of different variables, and attempts to optimally group them from N to N-1, etc. The groupings are based upon maximizing the average intergroup distance, while minimizing the average intragroup distance.

The approach begins by defining each object as a group. These N groups are then reduced by one, until all persons have been classified into one of two groups. More detail of this approach can be found in SAS, as well as Veldman (1967).

Using the clustering program, three cluster analyses were completed. When using a cluster analysis, one has to decide on the number clusters one wants in the solution. The decision used for this study was that no cluster would contain less than five people.

The first cluster analysis, using the universities' ratings and the three demographics, produced four clusters with 27 people in cluster one, 56 in cluster two, 11 in cluster three, and 6 in cluster four. These four clusters accounted for 61% of the variance for all groupings. The second cluster analysis, based upon universities' ratings, produced three clusters with an R^2 equal to .55, with 58 individuals in cluster one, 36 in cluster two, and 7 in cluster three. The third cluster analysis, based upon demographics alone, produced only two clusters with almost everyone loading on cluster one. Therefore, it was not considered.

The four clusters produced by the first cluster analysis were used as predictor variables to predict the ratings of each of the eight universities, the eight regression equations

Table 3

Multiple linear Regression Results for Model 2

Variable	Regression Coefficients	T Value
X_1	1.410	12.21*
X_2	1.125	9.74*
X_3	1.440	12.47*
X_4	1.775	15.37*
X_5	1.015	8.79*
X_6	1.000	
Constant	-3.38	

$n = 800$
$R^2 = .58$
$df_d = 695$

* Statistically significant at the .01 level.

produced the following values: .12, .27, .17, .18, .18, .26, .34, and .28. When the clusters from the second cluster analysis containing three clusters, were used as predictor variables, they yielded the following R^2 values: .03, .18, .14, .15, .16, .20, .30, and .18. Since the use of cross-validation procedures would produce even lower values, those procedures were not implemented.

Discussion

The conjoint analysis and the multiple regression model with a surrogate person vector produced identical estimates for the five university attributes. The multiple regression procedure that incorporated a surrogate person vector was better able to predict the holdout universities. Thus, these results seem to imply that if a university administration wants to obtain information on which university attributes are most important to their students, either conjoint analysis or a multiple regression model with a surrogate variable is an appropriate procedure.

With this data set the Q-factor analysis failed to provide useful information. The classifying of student by type did not allow for a high degree of explanation of the ratings of the various hypothetical universities. The use of Q-factor analysis, however, may provide insight into the university selection process by students if various groups are identifiable.

Three points should be noted with regard to future research. First, a multiple linear regression model with a surrogate person vector is a valuable procedure to use to determine which university attributes are important to students when selecting a university. The inclusion of the surrogate person variable did improve the researchers' ability to predict the ratings of the holdout universities. Further studies in this area with more detailed attributes would be informative.

Second, unless various groups of students rate the universities differently, Q-factor analysis obviously will not provide useful informaiton. If such groups exist, however, the information may provide university administrators with some insight into what type of students prefer their particular university.

Third, the conjoint and regression analyses are really asking different questions that the Q-factor analysis. The conjoint and regression analyses are attempting to determine which of the university characteristics are most important. The Q-factor analysis attempts to determine if there are various typologies based on the students' university ratings. This third point leads to an often discussed conclusion. Determining the preferable research method is dependent upon the question of interest. In other words, the research question has to dictate the methodology.

References

Bretton-Clark. (1987). Conjoint analyzer. New York.

Conant, J., Brown, J., & Mokwa, M. (1985), Summer). Students are important consumers: Assessing satisfaction in a higher educaton context. Journal of Marketing Education, 13-20.

Fraas, J. W. & Paugh, R. (1989). Student perceptions of the relative importance of selected attributes of an institution of higher education: a conjoint approach. Ashland, Ohio: Ashland University, 1989. (ERIC Document Reproduction Service No. ED 312 960)

Fraas, J. W., & McDougall, W. R. (1983). The use of one full full MLR movel to conduct multiple comparisons in a repeated measures design: An industrial application. Multiple Linear Regression Viewpoints, 12(1), 42-55.

Fraas, J. W. & Newman, I. Conjoint Analysis: A study of the effects of using person variables. Ashlan, Ohio: Ashland University, 1989. (ERIC Document Reproduction Service No. ED 312 961)

Green, P., Goldbert, S., & Montemajor, M. (1981). A hybrid utility estimation model for conjoint analysis. estimation model for conjoint analysis. Journal of Marketing, 45, 33-41.

Green, P., Carroll, J. D., & Goldberg, M. (1981). A general approach to product design optimization via conjoint analysis. Journal of Marketing, 45, 17-37.

Green, P., & Strinivasan, V. (1978). Conjoinjt analysis in consumer research: Issues and outlook. Journal of Consumer Research, 5, 103-124.

Kuh, G., Comes, M., & Lundquist, I. (1984, Winter). What prospective students really need to know about institutional quality. College and University, 167-175

Leigh, T., Mackay, D., & Summer, J. (1981). Realiability and validity of conjoint analysis and self-explicated weights: A comparison. Journal of Marketing Research, 21, 456-462.

McNeil, K., Kelly, F., & McNeil, J. (1975). Testing research hypotheses using multiple linear regression. Carbondale, IL: Southern Illinois University Press.

Newman, I. & Benz, C. R. Multivariate evaluation design: A suggested technique for mental health systems. April 25, 1987. Ohio Academy of Science, Canton, Ohio.

Pedhazur, E. J. (1977). Coding subjects in repeated measures designs. Psychological Bulletin, 84, 298-305.

Tierney, M. L. (1984, November). Ten questions to ask when choosing a college. Money, 133-134.

Vantubergen, G. N. (1966). Quannal: A computer program for Q analysis. Mass Communication Research Bureau, School of Journalism, University of Iowa. Mimeo.

Veldman, D. (1967). Fortran Programming for the behavioral sciences. New York: Holt, Rinehart, & Winston.

Ward, J. d. (1963). Hierarchical grouping to potimize an objective function. American Statistical Association Journal, 58, 236-244.

Williams, J. D. (1989). Multiple comparisons in higher dimension designs. Monograph Series #5. Multiple Linear Regression Viewpoints, 10(3).

Williams, J. D., & Williams, J. A. (1985a). Testing hypotheses in a repeated measures design on employee attitudes with large samples. Multiple Linear Regression Viewpoints, 13(2), 1-20.

Wliams, J. D., & Williams, J. A. (1985b). Testing hypotheses in a repeated measures design: An example. Multiple Regressions Viewpoints, 13(2), 35-46.

rev.042291

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

Relationship Between Multiple Regression, Path, Factor, and LISREL Analyses

Randall E. Schumacker
University of North Texas

Abstract

A basic knowledge of multiple regression concepts permits further understanding of path, factor, and lisrel analyses. Specifically, standardized partial regression coefficients (beta weights) as applied in path, factor, and lisrel analyses are presented. The multivariable methods have in common the general linear model and are the same in several respects. First, they identify, partition, and control variance. Second, they are based upon a linear combination of variables. And third, the linear weights can be computed based on standardized partial regression coefficients.

Multiple regression or the general linear model approach to the analysis of experimental data in educational research has become increasingly popular since 1967 (Bashaw and Findley, 1968). In fact today, it has become recognized as an approach that bridges the gap between correlational and analysis of variance thought in answering research hypotheses (McNeil, Kelly, & McNeil, 1975). Statistical textbooks in psychology and education often present the relationship between data analysis with multiple regression and analysis of variance (Draper & Smith, 1966; Williams, 1974a; Roscoe, 1975; Edwards, 1979). Graduate students taking an advanced statistics course are therefore provided with the multiple linear regression framework for data analysis. Given their knowledge of multiple linear regression techniques applied to univariate analysis (one dependent variable), their understanding can be extended to the relationship of multiple linear regression to various multivariate statistical techniques (Kelly, Beggs, McNeil, with Eichelberger & Lyon, 1969, pps 228-248; Newman, 1988). The article therefore expands upon this understanding and indicates the importance of the standardized partial regression coefficient (beta weight) in multiple linear regression as it is applied in path, factor, and lisrel analyses.

MULTIPLE REGRESSION

Multiple regression techniques require a basic understanding of sample statistics (n, mean, and variance), standardized variables, correlation (Pedhazur, 1982, pp 53-57), and partial correlation (Cohen & Cohen, 1975; Houston & Bolding, 1974). In standard score form the multiple regression equation is:

$$\hat{z}_y = b z_x$$

The relationship between the correlation coefficient, the unstandardized regression coefficient and the standardized regression coefficient is:

$$b = \frac{S z_x z_y}{S z_x^2} = b \frac{s_x}{s_y} = r_{xy}$$

For two independent variables, the regression equation with standard scores is:

$$\hat{z}_y = b_1 z_1 + b_2 z_2$$

And the standardized partial regression coefficients are computed by:

$$b_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2} \qquad b_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}$$

The correlation between the original and predicted scores is given the special name Multiple Correlation Coefficient. It is indicated as:

$$R_{y\hat{y}} = R_{y.12}$$

And the Squared Multiple Correlation Coefficient is related as follows:

$$R_{y\hat{y}}^2 = R_{y.12}^2 = b_1 r_{y1} + b_2 r_{y2}$$

MULTIPLE REGRESSION EXAMPLE

A multiple linear regression example using a correlation matrix as input (SPSSX User's Guide, 3rd Edition, 1988, Chapter 13) is in the appendix. The results are:

$$R^2_{y.123} = b_1 r_{y1} + b_2 r_{y2} + b_3 r_{y3}$$

$$= (.423)\ .507 + (.363)\ .481 + (.040)\ .276$$

$$R^2_{y.123} = .40$$

A systematic determination of the most important set of variables can be accomplished by setting the partial regression weight of each variable to zero. This approach and other alternative methods are presented by Kelly, Beggs, & McNeil et al (1969) and Darlington (1968).

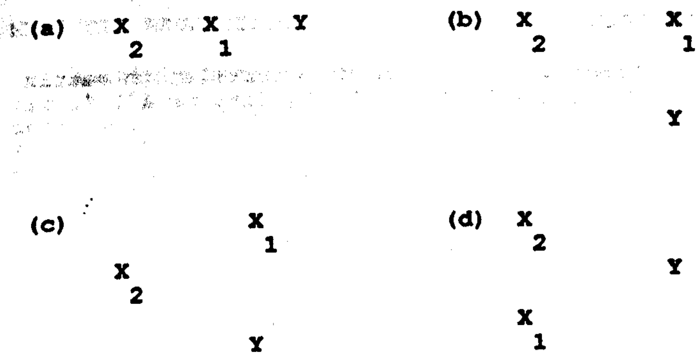
In summary, regression techniques have been shown to be robust (Bohrnstedt & Carter, 1971); applicable to contrast coding (Lewis & Mouw, 1978); dichotomous coding (McNeil, Kelly, & McNeil, 1975); and ordinal coding (Lyons, 1971) research situations. Multiple regression can also be viewed as a special case of path analysis.

PATH ANALYSIS

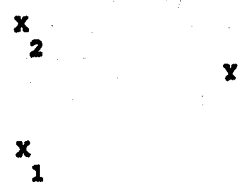
Sewall Wright is credited with the development of path analysis as a method for studying the direct and indirect effects of variables (Wright, 1921, 1934, 1960). Path analysis is not a method for discovering causes, rather it tests theoretical relationships called "causal modeling". The specified model establishes causal relationships among the variables when:

a. temporal ordering exists
b. covariation (correlation) is present
c. controlled for other causes

Model specification is necessary in examining multiple variable relationships. In the absence of a model, many different relationships among variables can be postulated with many different path coefficients being selected. For example, in a three variable model the following four relationships could be postulated:



The four different models have been considered without reversing the order of the variables. How can one decide which model is correct? Path analysis doesn't provide a way to specify the model, but rather estimates the effects once the model has been specified "a priori". Path coefficients in path analysis take on the values of a product-moment correlation and/or standardized regression coefficients in a model (Wolfle, 1977). For example given model (d):



THEN:

$b_1 = p_{y1}$ $\quad$ $b_2 = p_{y2}$ $\quad$ $r_{12} = p_{12}$

A path model is specified by the researcher based on theory or prior research. Variable relationships once specified, in standard score form, become standardised regression coefficients. In multiple regression, a dependent variable is regressed in a single analysis on all the independent variables. In path analysis one or more multiple regression analyses are performed. Path coefficients are computed based upon only the particular set of independent variables that lead to the dependent variable under consideration. As in regression analysis, path analysis can use dichotomous and ordinal data in the causal model (Boyle, 1970; Lyons, 1971).

MODEL SPECIFICATION

Path models permit diagramming how a particular set of independent variables lead to a dependent variable under consideration. How the paths are drawn determine whether the independent variables are correlated causes (unanalyzed), mediated causes (indirect), or independent causes (direct). The model can be tested for the significance of path coefficients (Pedhazur, 1982, pp 58-62) and a goodness-of-fit criteria (Marascuilo & Levin, 1983, pp 169-172; Tatsuoka & Lohnes, 1988, pp 98-100) which reflects the significance between the original and reproduced correlation matrix. This process is commonly called decomposing the correlation matrix (Asher, 1976, pp 32-34) according to certain rules (Wright, 1934).

PATH ANALYSIS EXAMPLE

A four variable path analysis program is in the appendix. In order to calculate the path coefficients for the model, two regression analyses were performed. The model with the path coefficients is:

X_1 $\qquad$ $p_{Y1} = .423$

$p_{31} = -.071$

$p_{21} = .224$ $\qquad$ X_3 $\qquad$ $p_{Y3} = .040$ $\qquad$ Y

$p_{32} = .593$

$p_{Y2} = .363$

X_2

The original and reproduced correlations are presented in matrix form. The upper half represents original correlations and the lower half the reproduced correlations which include the regression of paths linking independent variables to the dependent variable.

VARIABLE	Y	X1	X2	X3
Y	1.000	.507	.481	.276
X1	.423	1.000	.224	.062
X2	.362	.224	1.000	.577
X3	.040	-.070	.593	1.000

Original Correlations

Reproduced Correlations

The original correlations can be completely "reproduced" if all effects: direct (DE), indirect (IE), spurious (S) and correlated (C) are included. For example:

$$r_{12} = \underset{C}{p_{12}} = .224$$

$$r_{13} = \underset{DE}{p_{31}} + \underset{IE}{p_{32}p_{21}} = .062$$

$$r_{23} = \underset{DE}{p_{32}} + \underset{S}{p_{31}p_{21}} = .577$$

$$r_{1Y} = \underset{DE}{p_{Y1}} + \underset{IE}{p_{Y2}p_{21}} + \underset{IE}{p_{Y3}p_{31}} + \underset{IE}{p_{Y3}p_{32}p_{21}} = .507$$

$$r_{2Y} = \underset{DE}{p_{Y2}} + \underset{IE}{p_{Y3}p_{32}} + \underset{S}{p_{Y1}p_{21}} + \underset{S}{p_{Y3}p_{31}p_{21}} = .481$$

$$r_{3Y} = \underset{DE}{p_{Y3}} + \underset{S}{p_{Y1}p_{31}} + \underset{S}{p_{Y2}p_{32}} + \underset{S}{p_{Y1}p_{21}p_{32}} + \underset{S}{p_{Y2}p_{21}p_{31}} = .276$$

In summary, path analysis can be carried out within the context of ordinary regression analysis and does not require the learning of any new analysis techniques (Asher, 1976, p32; Williams, 1974b). The advantage of path analysis is that it enables one to specify direct and indirect effects among independent variables. In addition, path analysis enables us to decompose the correlation between any two variables into simple and complex paths of which some are meaningful. Path coefficients and the relationship between the original and reproduced correlation matrix can also be tested for significance.

FACTOR ANALYSIS

Path models and the associated test of significance between original and reproduced correlations are used in confirmatory factor analysis. Factor analysis assumes that the observed (measured) variables are linear combinations of some underlying source variable (factor). In practice, one estimates population parameters of the measured variables from a sample (with the uncertainties of model specification and measurement error). A linear combination of weighted variables relates to multiple regression in a single factor model and to a linear causal system (path analysis - "multiple" multiple regressions) in multiple factor models. Path diagrams therefore permit representation of the causal relationships among factors and observed (measured) variables in factor analysis.

In general, the first step in factor analysis involves the study of interrelationships among variables in the correlation matrix. Factor analysis will address the question of whether these subsets can be identified by one or more factors (hypothetical constructs). Confirmatory factor analysis is used to test specific hypotheses regarding which variables correlate with which constructs (Long, 1983).

FACTOR MODELS

Factor analysis assumes that some factors, which are smaller in number than the number of observed variables, are responsible for the covariation among the observed variables. For example, given a unidimensional trait in a single factor model with four variables the diagram would be (Kim & Mueller, 1978a, p 35):

	$b_Y = .677$	Y	$d_Y = .735$	U_Y
	$b_1 = .402$	X_1	$d_1 = .917$	U_1
F	$b_2 = .800$	X_2	$d_2 = .600$	U_2
	$b_3 = .535$	X_3	$d_3 = .843$	U_3

WHERE:

b_i = standardized regression coefficient

The variance of each observed variable is therefore comprised of the proportion of variance determined by the common factor and the proportion determined by the unique factor, which together equal the total variance of each observed variable. Therefore:

$$S_i^2 = b_i^2 + d_i^2 = 1$$

The correlation between a common factor and a variable is:

$$r_{F,X_i} = b_i$$

The correlation between a unique factor and a variable is:

$$r_{U,X_i} = d_i$$

The correlation between observed (measured) variables sharing a common factor is:

$$r_{X_i,X_j} = b_i b_j$$

And finally, the variance attributed to the factor as a result of the linear combination of variables is:

$$h^2 = \frac{S b_i^2}{M} = R^2_{F.1234}$$

Where: M = number of variables

b_i^2 = squared factor loadings

Note: $S b_i^2$ = eigenvalue

b_i^2 = communality

FACTOR ANALYSIS EXAMPLE

A single factor analysis program with four variables in a correlation matrix format is in the appendix. The path diagram is the same as above (Kim & Mueller, 1978a, p 35) with the weights as follows:

$$b_Y = .677 \quad b_1 = .402 \quad b_2 = .800 \quad b_3 = .535$$

And, factor scores computed as:

$$F = b_Y Y + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Multiplying the coefficients between pairs of variables gives the following correlation matrix:

VARIABLE	Y	X1	X2	X3
Y	b_1^2	.27	.54	.36
X1	.27	b_2^2	.32	.22
X2	.54	.32	b_3^2	.43
X3	.36	.22	.43	b_4^2

The common factor variance is:

$$R^2_{F.1234} = \frac{8\, b_i^2}{M} = \frac{.46 + .16 + .64 + .29}{4} = .39$$

The unique factor variance is:

$$1 - R^2_{F.1234} = \frac{8\,(1 - b_i^2)}{M} = \frac{.54 + .84 + .36 + .71}{4} = .61$$

In summary, factor loadings (variable weights) are standardized regression coefficients. As such, linear weighted combinations of variables loading on a factor are used to compute factor scores (Kim & Mueller, 1978b p 60). The weights are also the correlation between the observed (measured) variables and the factor (hypothetical construct). If the variable correlations (weights) are squared and summed, they describe the proportion of variance determined by that factor. This is traditionally known as an eigenvalue, but termed communality in factor analysis. When all variables are standardized, then the linear weights are called standardized regression coefficients (regression analysis), path coefficients (path analysis), or factor loadings (factor analysis). The factor analysis approach is distinquished from regression or path analysis in that observed variable correlation is explained by a common factor (hypothetical construct). In factor analysis therefore the correlation between observed variables is the result of sharing a common factor rather than a variable being the direct cause (path analysis) or predictor of another (regression analysis).

LISREL

Linear structural relationships (lisrel) are often diagrammed by using multiple factor path models where the factors (hypothetical contructs) are viewed as latent traits (Joreskog & Sorbom, 1986, pp I.5-I.7). The lisrel model consists of two parts: the measurement model and the structural equation model. The measurement model specifies how the latent variables or hypothetical constructs are measured in terms of the observed (measured) variables and describes their measurement properties (reliability and validity). The structural equation model specifies the causal relationship among the latent variables and is used to describe the causal effects and the amount of unexplained variance. The lisrel model includes or encompasses a wide range of models, for example: univariate or multivariate regression models, confirmatory factor analysis, and path analysis models (Joreskog & Sorbom, 1986, pp I.3, I.9-I.12). Cuttance (1983) presents an overview of several lisrel submodels with diagrams and explanations. Wolfle (1982) presents an indepth presentation of a single model to introduce and clarify lisrel analysis. The lisrel program therefore permits regression, path, and factor analysis whereby model specification and measurement error can be assessed.

MEASUREMENT ERROR

Fuller (1987) extensively covers lisrel and factor analysis models and especially extends regression analysis to the case where the variables are measured with error. Wolfe (1979, pp 48-51) presents the relationship between lisrel, regression and path analysis especially in regards to how measurement error effects the regression coefficient (path coefficient). Errors of measurement in statistics have been studied extensively (Wolfe, 1979). Cochran (1968) studied it from four different aspects: (1) types of mathematical models, (2) standard techniques of analysis which take into account measurement error, (3) effect of errors of measurement in producing bias and reduced precision and what remedial procedures are available, and (4) techniques for studying error of measurement. Cochran (1970) also studied the effects of error of measurement on the squared multiple correlation coefficient.

LISREL-FACTOR ANALYSIS EXAMPLE

A LISREL factor analysis program with a correlation matrix as input is in the appendix. The factor analytic model in matrix notation is:

$$X = L_x x + q_d$$

Where:
- X = observed variables
- L = structural weights (factor loadings)
- x = latent trait (factor)
- q_d = error variance (unique variance)

The LISREL results are:

a. L = LAMBDA X (structural weights-factor loadings)

$Y = .677$ $X_1 = .402$ $X_2 = .800$ $X_3 = .535$

b. q_d = THETA DELTA (unique factor variance)

$Y = .54$ $X_1 = .84$ $X_2 = .36$ $X_3 = .71$

c. b^2 = LAMBDA X^2 (common factor variance)

$Y = .46$ $X_1 = .16$ $X_2 = .64$ $X_3 = .29$

The concept of model specification and goodness of fit pertains to the original correlation matrix and the estimated correlation matrix. The estimated correlation matrix is:

$$\sigma = \begin{matrix} .272 & & \\ .542 & .321 & \\ .362 & .215 & .427 \end{matrix}$$

The original correlation matrix is:

$$S = \begin{matrix} .507 & & \\ .481 & .224 & \\ .276 & .062 & .577 \end{matrix}$$

The goodness of fit index (GFI) using the unweighted least squares approach (ULS) is then computed as:

$$GFI = 1 - 1/2 \text{ trace } (S - \sigma)^2$$

$$GFI = 1 - 1/2 \, (1.308 - 1.02)^2$$

$$GFI = 1 - .041$$

$$GFI = .959$$

LISREL-REGRESSION ANALYSIS EXAMPLE

A LISREL regression program with a correlation matrix as input is in the appendix. The regression model in matrix notation is:

$$Y = G\,X + z$$

Where: Y = dependent variable
G = gamma matrix (beta weights)
X = independent variables
z = errors of prediction (error variance)

The LISREL results are the same as in the previous regression program:

$$R^2_{y.123} = G_1 r_{y1} + G_2 r_{y2} + G_3 r_{y3}$$

$$R^2_{y.123} = (.423)\ .507 + (.363)\ .481 + (.040)\ .276$$

$$R^2_{y.123} = .40$$

CONCLUSION

The appropriate statistical method to use is often an issue of debate. It sometimes requires more than one approach to analyzing data. The rationale for choosing between the alternative methods of analysis is usually guided by research hypotheses or questions.

The multivariable methods discussed have in common the general linear model and are the same in several respects. First, they identify, partition, and control variance. Second, they are based on linear combinations of variables. And third, the linear weights can be computed based on standardized partial regression coefficients.

The multivariable methods however have different applications. Multiple regression seeks to identify and estimate the amount of variance in the dependent variable attributed to one or more independent variables (prediction). Path analysis seeks to identify relationships among a set of variables (explanation). Factor analysis seeks to identify subsets of variables from a much larger set (common/shared variance). Lisrel determines the degree of model specification and measurement error. The different methods were derived because of the need for prediction, explanation, common variance, model and measurement error assessment type applications.

Multiple regression techniques are robust except for model specification and measurement errors (Borhnstedt & Carter, 1971). Multiple regression techniques are also useful in understanding path, factor, and LISREL applications. LISREL permits regression, path, and factor analyses whereby model specification and measurement error can be assessed. Lisrel also permits univariate or multivariate least squares analysis in either single sample or multiple sample (across populations) research settings. An understanding of multiple regression and general linear model techniques can therefore greatly facilitate ones understanding of the testing of research questions in multivariable situations.

APPENDIX

MULTIPLE REGRESSION PROGRAM

```
TITLE REGRESSION WITH CORRELATION MATRIX INPUT
COMMENT VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507   1.000
 .481    .224   1.000
 .276    .062    .577  1.000
END DATA
REGRESSION MATRIX=IN(*)/
           MISSING=LISTWISE/
           VARIABLES=Y X1 X2 X3/
           DEPENDENT=Y/
           ENTER X1 X2 X3/
FINISH
```

PATH ANALYSIS PROGRAM ONE

A. VARIABLE 3 REGRESSED ON VARIABLES 1 AND 2

```
TITLE PATH ANALYSIS EXAMPLE WITH CORRELATION MATRIX INPUT
COMMENT  VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507  1.000
 .481   .224  1.000
 .276   .062   .577  1.000
END DATA
REGRESSION MATRIX=IN(*)/
           MISSING=LISTWISE/
           VARIABLES=Y X1 X2 X3/
           DEPENDENT=X3/
           ENTER X1 X2/
FINISH
```

PATH ANALYSIS PROGRAM TWO

B. VARIABLE Y REGRESSED ON VARIABLES 1, 2, AND 3

```
TITLE PATH ANALYSIS EXAMPLE WITH CORRELATION MATRIX INPUT
COMMENT VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507  1.000
 .481   .224  1.000
 .276   .062   .577  1.000
END DATA
REGRESSION MATRIX=IN(*)/
           MISSING=LISTWISE/
           VARIABLES=Y X1 X2 X3/
           DEPENDENT=Y/
           ENTER X1 X2 X3/
FINISH
```

FACTOR ANALYSIS PROGRAM

```
TITLE FACTOR ANALYSIS EXAMPLE WITH CORRELATION MATRIX INPUT
COMMENT VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507  1.000
 .481   .224  1.000
 .276   .062   .577  1.000
END DATA
FACTOR VARIABLES=Y X1 X2 X3/
     MATRIX=IN(COR=*)/
     CRITERIA=FACTORS(1)/
     EXTRACTION=ULS/
     ROTATION=NOROTATE/
     PRINT CORRELATION DET INITIAL EXTRACTION ROTATION/
     FORMAT SORT/
     PLOT EIGEN/
     FINISH
```

LISREL FACTOR ANALYSIS PROGRAM

```
TITLE 'LISREL FACTOR ANALYSIS WITH CORRELATION MATRIX INPUT'
INPUT PROGRAM
NUMERIC DUMMY
END FILE
END INPUT PROGRAM
USERPROC NAME=LISREL
DATA FOR GROUP ONE
DA NG=1 NI=4 NO=100
LA
'Y' 'X1' 'X2' 'X3'
KM SY
1.000
 .507 1.000
 .481  .224 1.000
 .276  .062  .577 1.000
MO NX=4 NK=1 TD=DI,FR PH=ST
LK
'FACTOR'
PA LX
4 * 1
OU ULS SE TV PC RS VA FS SS MI
END USER
```

LISREL REGRESSION ANALYSIS PROGRAM

```
TITLE 'LISREL REGRESSION ANALYSIS WITH CORRELATION MATIRX'
INPUT PROGRAM
NUMERIC DUMMY
END FILE
END INPUT PROGRAM
USERPROC NAME=LISREL
DATA FOR GROUP ONE
DA NG=1 NI=4 NO=100
LA
'Y' 'X1' 'X2' 'X3'
KM SY
1.000
 .507 1.000
 .481  .224 1.000
 .276  .062  .577 1.000
MO NY=1 NX=3 PS=DI
OU ULS SE TV PC RS VA SS MI TO
END USER
```

REFERENCES

Asher, H.B. (1976). Causal modeling. Sage Publications: Beverly Hills: CA.

Bashaw, W.L. & W.G. Findley (1968). Symposium on general linear model approach to the analysis of experimental data in educational research. Project No. 7-8096. U.S. Department of Health, Education, and Welfare, Washington, D.C.

Bohrnstedt, G.W. & T.M. Carter (1971). Robustness in regression analysis. In H.L. Costner (Ed), Sociological Methodology, Jossey-Bass, pp 118-146.

Boyle, R.P. (1970). Path analysis and ordinal data. American Journal of Sociology, vol. 75(4), 461-480.

Cohen, J. & P. Cohen (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erlbaum: NJ.

Cochran, W.G. (1968). Errors of measurement in statistics. Technometrics, 10(4), 637-666.

Cochran, W.G. (1970). Some effects of errors of measurement on multiple correlation. Journal of the American Statistical Association, 65(329), 22-34.

Cuttance, P.F. (1983). Covariance structure and structural equation modelling in research: a conceptual overview of LISREL modelling. Multiple Linear Regression Viewpoints, 12(2), 1-63.

Darlington, R.B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69, 161-182.

Draper, N.R. & H. Smith (1966). Applied regression analysis. John Wiley & Sons: New York, NY.

Edwards, A.L. (1979). Multiple regression and the analysis of variance and covariance. W.H. Freeman: San Francisco, CA.

Fuller, W.A. (1987) Measurement error models. John Wiley & Sons: New York, NY.

Houston, S.R. & J.T. Bolding, Jr. (1974). Part, partial, and multiple correlation in commonality analysis of multiple regression models. Multiple Linear Regression Viewpoints, 5(3), 36-40.

Joreskog, K.G. & D. Sorbom (1986). LISREL VI USER'S GUIDE: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods. Scientific Software: Mooresville, Indiana.

Kelly, F.J, D.L. Beggs, K.A. McNeil, T. Eichelberger, & L. Lyon (1969). Multiple regression approach. Southern Illinois University Press: Carbondale, IL.

Kim, J. & C. Mueller (1978a). Introduction to Factor Analysis. Sage Publications, Beverly Hills: CA.

Kim, J. & C. Mueller (1978b). Factor Analysis. Sage Publications, Beverly Hills: CA.

Lewis, E.L. & J.T. Mouw (1978). The use of contrast coefficients. Southern Illinois University Press, Carbondale, IL.

Long, J.S. (1983). Confirmatory Factor Analysis. Sage Publications, Beverly Hills: CA.

Lyons, M. (1971). Techniques for using ordinal measures in regression and path analysis. In H.L. Costner (Ed), Sociological Methodology, Jossey-Bass, pp 147-171.

Marascuilo, L.A. & J.R. Levin (1983). Multivariate statistics in the social sciences: a researcher's guide. Brooks/Cole Publishing: Belmont: CA.

McNeil, K.A., F.J. Kelly, J.T. McNeil (1975). Testing research hypotheses using multiple linear regression. Southern Illinois University Press: Carbondale, IL.

Newman, I. (1988). There is no such thing as multivariate analysis: all analyses are univariate. Presidents Address at Mid-Western Educational Research Association, October 15, 1988, Chicago, IL.

Pedhazur, E.J. (1982). Multiple regression in behavioral research: explanation and prediction (2nd ed.). Holt, Rinehart, & Winston: New York, NY.

Roscoe, J.T. (1975). Fundamental research statistics for the behavioral sciences (2nd ed.). Holt, Rinehart, & Winston: New York, NY.

SPSSX Users' Guide, 3rd ed. (1988). McGraw-Hill: New York, NY.

Tatsuoka, M.M & P.R. Lohnes (1988). Multivariate analysis: techniques for educational and psychological research (2nd ed.). Macmillan Publishing Company: New York, NY.

Williams, J.D. (1974a). Regression analysis in educational research. MSS Information Corporation: New York, NY.

Williams, J.D. (1974b). Path analysis and causal models as regression techniques. Multiple Linear Regression Viewpoints, 5(3), 1-20.

Wolfle, L.M. (1977). An introduction to path analysis. Multiple Linear Regression Viewpoints, 8(1), 36-61.

Wolfle, L.M. (1979). Unmeasured variables in path analysis. Multiple Linear Regression Viewpoints, 9(5), 20-56.

Wolfle, L.M. (1982). Causal models with unmeasured variables: an introduction to LISREL. Multiple Linear Regression Viewpoints, 11(2), 9-54.

Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, 20, 557-585.

Wright, S. (1934). The method of path coefficients. Annals of Mathematical Statistics, 5, 161-215.

Wright, S. (1960). Path coefficients and path regression: alternative or complementary concepts? Biometrics, 16, 189-202.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

The Case for Non-Zero Restrictions in Statistical Analysis

Keith McNeil
New Mexico State University

One of the many advantages of MLR is its versatility and its ability to answer a vast array of questions. Unfortunately, most researchers fall into the habit of asking a small subset of very similar questions. The question being tested should be stated first, but can be identified from the full model and the restriction(s) placed on that full model. While the restrictions can take on any numerical value, almost all applications use the "default " value of zero:

1. $a_1 = a_2$ or $(a_1 - a_2 = 0)$ (t-test)
2. $a_1 = 0$ (Correlation)
3. $a_1 = a_2 = a_3 = \ldots a_1$ or $(a_1 - a_2 = a_2 - a_3 = a_3 - a_2 = \ldots 0)$ (F-test)
4. $(a_1 - a_2) = (a_3 - a_4)$ or $((a_1 - a_2) - (a_3 - a_4) = 0)$ (interaction)

The focus of this paper will be on the utility of making a non-zero restriction. Why the zero restriction occurs so frequently will be questioned and hopefully researchers and statisticians will see how the zero restriction limits the conclusions of the research. The argument will be made for making non-zero restrictions, resulting of course, from "non-zero" research hypotheses. The argument will be made for each of these statistical procedures: two group t test, Pearson correlation, single population mean, one-way analysis of variance, and interaction.

Two Group t Test

Perhaps the most widely used design compares the performance of two groups. The research hypothesis takes the following form: Research Hypothesis 1: For a given population, the New treatment is better than the Traditional treatment on Y. (See Note 1 for discussion of directional hypothesis testing.)
Full Model: $Y = a_1N + a_2T + E_1$
Where Y = criterion of interest,
N = 1 if subject in New treatment; 0 otherwise, and
T = 1 if subject in Traditional treatment; 0 otherwise.
The research hypothesis implies that the sample mean for N should be greater than the sample mean for T, or $a_1 > a_2$, or $a_1 - a_2 > 0$.
Restriction: $a_1 = a_2$, or $(a_1 - a_2 = 0)$
Forcing the restriction into the full model results in:
Restricted Model: $Y = a_1N + a_1T + E_2$
But since the two vectors (N and T) are multiplied by the same weights, the vectors can be added first. But N + T equals the Unit vector (or everyone). Therefore:
Restricted Model: $Y = a_1U + E_2$

There are two linearly independent pieces of information in the full model. Forcing the one restriction on the full model results in one linearly independent piece of information in the restricted model. (See Note 2 for test of significance.)

A significant drop in the R^2 from the full model to the restricted model results in a significant F. If the sample means are in accord with the anticipated result, then Research Hypothesis 1 can be held as tenable and the conclusion would be: For the given population, the New treatment is better than the Traditional treatment on Y. But all that has been said is that the New treatment is better than the Traditional treatment. We do not know how much better; all we know is that the two treatments are not equally effective.

But what if the cost of the two treatments is not the same? The Traditional treatment has surely been somewhat effective in the past. The New treatment will surely require some additional cost in the form of special inservice, purchase of new materials, acceptance by teachers, students, and community, etc. Before the Traditional treatment is replaced by the New treatment, perhaps the researcher should demonstrate that there is, say, more than a five-point superiority of the New treatment over the Traditional treatment.

When a non-zero research hypothesis is proposed, other researchers and statisticians often ask for the justification for the actual non-zero value chosen, as they should. But why should more justification be required for a non-zero value than for a zero value? Or looking at the issue form the other side, why are researchers allowed to test a zero value with little or no justification. When one realizes that zero is only one of an infinite number of values, then one realizes that the same amount of justification should be required of a zero value as of a non-zero value. Furthermore, when one attempts to justify the zero value restriction, one may realize that zero is not the value of interest. Those researchers who have been defaulting with zero should know how to choose a value, but may not. It is not the intent of this paper to illustrate how one determines the magnitude of the value tested in the research hypothesis, although a few suggestions will be provided.

In the case where there was an expectation of a five-point superiority, the research hypothesis would be:

Research Hypothesis 2: For a given population, the New treatment is more than five points better than the Traditional treatment on Y.

Full Model: $Y = a_1N + a_2T + E_3$

The research hypothesis implies that the sample mean for the New treatment is more than five units greater than the sample mean for the treatment or, a_1 greater than $(a_2 + 5)$ or $(a_1 - a_2 > 5)$

Restriction: $a_1 = a_2 + 5$, or $(a_1 - a_2 = 5)$ or $(a_2 = a_1 - 5)$

Restricted Model:
$$Y = a_1N + (a_1 - 5)T + E_4$$
$$Y = a_1N + a_1T - 5T + E_4$$
$$(Y + 5T) = a_1(N + T) + E_4$$
$$(Y + 5T) = a_1U + E_4$$

There are two linearly independent pieces of information in the full model. Forcing the one restriction on the full model results in one linearly independent piece of information in the restricted model. (See Note 1 for test of significance.)

Notice that the full model in Research Hypothesis 1 is exactly the same as the full model in Research Hypothesis 2. The number of restrictions is also the same, resulting in the same number of degrees of freedom. What is different, though, is the nature of the restriction and hence the restricted models are different. The two research hypotheses are both "correct" and equally "valid" - they just test two different hypotheses. Research Hypothesis 2 provides a more definitive conclusion.

The actual "cost" of any treatment may be difficult to determine. But one must remember that Research Hypothesis 1 reduces to the default assumption that the "costs" are equal. The choice of a research hypothesis leading to a restriction of ($a_1 - a_2 = 0$) should be defended as much as a research hypothesis leading to a restriction of ($a_1 - a_2$ = some non-zero value). The restriction ($a_1 - a_2 = 0$) has become a widely used default value, but we must realize that it is only one of an infinite number of values.

Pearson Correlation

The usual application of the Pearson correlation hypothesis is:

Research Hypothesis 3: For a given population, the linear correlation between X and Y is greater than zero.

Full Model: $Y = a_0U + a_1X + E_5$

The research hypothesis implies that the slope of the line of best fit in the sample is positive, or $a_1 > 0$.

Restriction: $a_1 = 0$

Restricted Model: $Y = a_0U + 0X + E_6$

$Y = a_0U + E_6$

There are two linearly independent pieces of information in the full model. Forcing the one restriction on the full model results in one linearly independent piece of information in the restricted model.

If the F test is significant, then one concludes that the research hypothesis is tenable, that the linear correlation between X and Y is greater than 0, or that the change in Y per unit change in X is greater than 0; but we do not know how much greater than 0. There may be reasons for wanting to know if the correlation is greater than a particular value. For instance, if the correlation under consideration is either a validity coefficient or a reliability coefficient, then we would definitely want a correlation coefficient above some specified value, such as:

Research Hypothesis 4: For a given population, the linear correlation between Y and the Retest of Y is greater than .80.

Full Model: $Y = a_0U + a_1R + E_7$

Restriction: Restricted Model R2 = .64

The research hypothesis implies that the restricted model R^2 will be $(.80)^2$ or .64. The formula in Note 2 can be used when testing this hypothesis for significance.

Consider Research Hypothesis 5: For a given population, there is more than a .6 unit change in Y for every unit change in X. In this case the models would be:

Full Model: $Y = a_0U + a_1X + E_9$

Restriction: $a_1 = .6$

Restricted Model: $Y = a_0U + .6X + E_9$

$(Y - .6X) = a_0U + E_9$

Notice that the full model in Research Hypotheses 3 and 4 is exactly the same as Research Hypothesis 5. The number of restrictions is also the same; resulting in the same number of degrees of freedom. What is different, though, is the nature of the restriction and hence the restricted models are different. The three research hypotheses are all "correct" and equally "valid" - they just test three different hypotheses. Research Hypotheses 4 and 5 provide more definitive conclusions.

The desired correlation (reliability, validity, etc.) may be difficult to determine, but should be no more difficult to justify than justifying the default value of 0. Just because $a_1 = 0$ has been used in the past does not justify its use, particularly with hypotheses about reliability and validity.

<u>Single Population Mean</u>

The usual application of the single population mean hypothesis is:

Research Hypothesis 6: For a given population, the population mean is greater than a particular value, S.

Here S is some meaningful value, depending on the given circumstances. Maybe the researcher wants to establish that the population mean height is greater than 72 inches. Or possibly the researcher is concerned that a four-choice, 100 item multiple choice test score is greater than a chance score of 25. Note that in these two examples (and in most hypotheses regarding a single population mean), the value of zero makes no sense. Suppose that a researcher wanted to establish that the population of freshman at a particular University had a mean College Board Score above the national average of 450:

Research Hypothesis 7: The population of freshmen at University X has a mean College Board Score greater than the national mean of 450.

Full Model: College Board Scores = $a_0U + E_{10}$

The research hypothesis implies that the sample mean is greater than 450, or $a_0 > 450$

Restriction: $a_0 = 450$

Restricted Model: (College Board Scores) = $450\ U + E_{11}$, or

(College Board Scores - 450) = E_{11}

(See bottom of Note 2 for test of significance and McNeil, 1973 and McNeil, et al., 1975, p 315 for further details.)

The desired mean may be difficult to determine (i.e., it may require some thought or knowledge of the phenomenon under consideration), but no more difficult than justifying the default mean of 0. Indeed, using a mean of 0 in this example makes absolutely no sense at all, and that is why it doesn't appear in the literature.

<u>One-way Analysis of Variance</u>

The usual application of the multiple group F test (one-way ANOVA) is:

Research Hypothesis 8: There is at least one difference in the means on Y between the i populations.

Full Model: $Y = a_1G_1 + a_2G_2 + \ldots a_iG_i + E_{12}$

The research hypothesis implies that not all the sample means are equal, or that a_1 not equal a_2 not equal $\ldots a_i$ for at least one pair of means, or ($a_1 - a_i$ not equal 0; $a_2 - a_i$ not equal 0; $\ldots a_{i-1} - a_i$ not equal 0 for at least one pair of means.)

Restriction: $a_1 = a_2 = \ldots a_i$; or

$a_1 - a_i = 0;\ a_2 - a_i = 0;\ \ldots\ a_{i-1} - a_i = 0$

By replacing all the coefficients with a common coefficient, a_0, we arrive at the following restricted model:

Restricted Model: $Y = a_0G_1 + a_0G_2 + \ldots a_0G_i + E_{13}$

Restricted Model: $Y = a_{0*}(G_1 + G_2 + \ldots G_i) + E_{13}$

Restricted Model: $Y = a_0U + E_{13}$

When the F test is significant then the restriction is rejected and the research hypothesis is accepted as tenable. But the research hypothesis just indicates that the i means are not all equal. Since most researchers are not satisfied with that information (confirming that the research hypothesis wasn't very interesting in the first place), most researchers turn to post-hoc comparisons to find out where the differences lie. These post-hoc comparisons are basically t-test comparisons and are thus like Research Hypothesis 1. (See Williams 1974). The suggestion here is to avoid asking a research hypothesis that you aren't interested in, and to go directly to non-zero research hypotheses that will yield satisfying information.

<u>Interaction</u>

Interaction is usually viewed only as a potentially contaminating factor when trying to explain main effects. That is, most researchers hope that there is no interaction so that they can proceed with interpreting main effects. But the interaction research hypothesis may be important in and of itself. Indeed, whenever an F has been computed for the interaction, the interaction research hypothesis has been tested. The usual interaction research hypothesis in a 2x2 design is as follows:

Research Hypothesis 9: For a given population, the difference on Y between Treatment 1 and Treatment 2 is not the same on Level 1 as on Level 2.

Full Model: $Y = a_1(T_1*L_1) + a_2(T_1*L_2) + a_3(T_2*L_1) + a_4(T_2*L_2) + E_{14}$

$m_{2)} = (R^2_F - R^2_R) / (1i_F - 1i_R)$

Where $T_1 = 1$ if in Treatment 1; 0 otherwise,

$T_2 = 1$ if in Treatment 2; 0 otherwise,

$L_1 = 1$ if in Level 1; 0 otherwise,

$L_2 = 1$ if in Level 2; 0 otherwise,

$(T_1 * L_1) = 1$ if in Treatment 1 <u>and</u> Level 1, etc.
The research hypothesis implies that the two differences are not the same, and that in the sample $(a_1 - a_3)$ not equal $(a_2 - a_4)$, or $[(a_1 - a_3) - (a_2 - a_4)$ not equal 0].
Restriction: $(a_1 - a_3) = (a_2 - a_4)$, or $[(a_1 - a_3) - (a_2 - a_4) = 0]$.
By placing the one restriction on the full model, one arrives at the following restricted model (See Note 3 and McNeil, et al., 1975):
Restricted Model: $Y = b_1T_1 + b_2T_2 + b_3L_1 + b_4L_2 + E_{15}$

Acceptance of the non-directional research hypothesis leads to a non-directional statement. All that can be concluded is that the differences are not the same. Hence we don't even know if the differences are greater at Level 1 or Level 2, let alone the magnitude of the difference of the differences. We have just conducted a non-directional test of interaction; a directional test of interaction is reflected in the following:
Research Hypothesis 10: For a given population, the difference on Y between Treatment 1 and Treatment 2 is greater at Level 1 than at Level 2.
Full Model: $Y = a_1(T_1*L_1) + a_2(T_1*L_2) + a_3(T_2*L_1) + a_4(T_2*L_2) + E_{16}$
The research hypothesis implies that the difference between T_1 and T_2 is greater at Level 1 than at Level 2, or in the sample $(a_1 - a_3)$ higher than $a_2 - a_4)$ or $[(a_1 - a_3) - (a_2 - a_4) > 0]$.
Restriction: $(a_1 - a_3) = (a_2 - a_4)$ or $[(a_1 - a_3) - (a_2 - a_4) = 0]$
Restricted Model: $Y = b_1T_1 + b_2T_2 + b_3L_1 + b_4L_2 + E_{17}$
A significant F for Research Hypothesis 10 provides more insight than would one for Research Hypothesis 9. We know that the differences are greater at Level 1, but again we do not know how much greater. If cost or theory dictate, say, a difference greater than six before a decision is made, the following Research Hypothesis would be appropriate:
Research Hypothesis 11: For a given population, the difference on Y between Treatment 1 and Treatment 2 is more than 6 units at Level 1 than at Level 2.
Full Model: $Y = a_1(T_1*L_1) + a_2(T_1*L_2) + a_3(T_2*L_1) + a_4(T_2*L_2) + E_{18}$
The research hypothesis implies that the difference between T_1 and T_2 is greater at Level 1 than at Level 2 by more than 6 units, or in the sample $(a_1 - a_3)$ higher than $(a_2 - a_4 + 6)$ or $[(a_1 - a_3) - (a_2 - a_4) > 6]$.
Restriction: $(a_1 - a_3) = (a_2 - a_4) + 6$; or
$(a_1 - a_3) - (a_2 - a_4) > 6$
Restricted Model: $(Y - 6) = b_1T_1 + b_2T_2 + b_3L_1 + b_4L_2 + E_{19}$

Research Hypotheses 9, 10, and 11 all test an interaction question, but in slightly different ways. In all three hypotheses, there are four linearly independent pieces of information in the full model. Forcing the one restriction on the full model results in three linearly independent pieces of information in the restricted model. Notice that the full models in Research Hypotheses 9, 10, and 11 are exactly the same. The number of restrictions is also the same; resulting in the same number of degrees of freedom. What is different, though, is the nature of the restriction and hence the restricted models are different. The three research hypotheses are all "correct" and equally "valid" - they just test three different hypotheses. Research Hypothesis 11, though, provides a more definitive conclusion, because as in the previous examples, a non-zero restriction was made.

Note 1. All the Research Hypotheses in this paper (except the one-way ANOVA) are directional Research Hypotheses. This follows the author's contention that a directional Research Hypothesis provides conclusive information whereas a non-directional Research Hypothesis provides no conclusive information. The Full and Restricted models are the same for the directional and non-directional hypotheses. The non-directional Research Hypothesis allows the researcher to conclude that a_1 does not equal 0, while the directional Research Hypothesis allows the researcher to conclude that $a_1 > 0$ (McNeil & Beggs, 1971). With reference to the non-zero restriction, of, say 6, the non-directional Research Hypothesis allows the conclusion that a_1 not equal to 6, while the directional Research Hypothesis allows the conclusion that $a_1 > 6$. The directional Research Hypothesis allows a more definitive conclusion using the same data and the same degrees of freedom.

Note 2. The general F test for testing two regression models is

$$F(m_1, m_2) = \frac{(R^2_F - R^2_R) / (li_F - li_R)}{(1 - R^2_F) / (N - li_F)}$$

Where: $R^2_F = R^2$ of the full model,
$R^2_R = R^2$ of the restricted model,
li_F = pieces of linearly independent information in the full model,
li_R = pieces of linearly independent information in the restricted model,
$m_1 = (li_F - li_R)$, and
$m_2 = (N - li_F)$.

This test cannot be used when either the restricted model has no predictors, when the criterion variable is different in the two models, or when the Unit vector is not in the restricted Model. In these cases, the F test must rely upon the sum of the squared scores in the error vector, E in both the full model (ESS_F) and the restricted model (ESS_R):

$$F = \frac{(ESS_R - ESS_F) / (li_F - li_R)}{(ESS_F) / (N - li_R)}$$

Note 3. The interaction examples all assumed equal N. The concepts still apply to the unequal N situation, although the restricted models will be different. (See Williams, 1972.)

SUMMARY

SIGNIFICANCE TEST	USUAL RESTRICTION	SUGGESTION
Pearson Correlation	zero	non-zero based on theory or cost
difference between two means	zero	non-zero based on theory or cost
difference between means (one-way f)	only zero	ignore omnibus F, go with planned comparisons based on theory or cost
interaction	almost always zero	non-zero based on theory or cost
single population mean	always non-zero	use more often

Bibliography

McNeil, K. A. Testing an hypothesis about a single population mean with multiple linear regression. Multiple Linear Regression Viewpoints, 1973, 4(1), 7-14.

McNeil, K. A., & Beggs, D. L. Directional hypotheses with the multiple linear regression approach. Paper presented at the meeting of the American Educational Research Association, New York, February 1971.

McNeil, K. A., Kelly, F. J., & McNeil, J. T. Testing Research Hypotheses Using Multiple Linear Regression. Cardondale: Southern Illinois University Press, 1975.

Williams, J. D. Two way fixed effects analysis of variance with disproportionate cell frequencies. Multivariate Behavioral Research, 1972, 7, 67-83.

Williams, J. D. Regression Analysis In Educational Research. New York: MSS Information Corporation, 1974.

C:CASENON

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

Considerations, Issues and Comparisons in Variable Selection and Interpretation in Multiple Regression

Susan Tracz, Ric Brown, and Rebecca Kopriva
California State University, Fresno

The selection of independent variables when utilizing multiple linear regression in a study is an involved and complex process. The availability of a variety of computer programs usually referred to as "stepwise" procedures affords users numerous options about which they often have little understanding. The purpose of this paper, then, is twofold: first, to present the major uses of regression analyses, the advantages and disadvantages of selection procedures and some caveats for researchers and those who teach statistics, and secondly, to present, compare and contrast several variable selection techniques using two data set.

Huberty (1989) suggests that the concept of variable selection may have some worth in terms of parsimony, explaining relationships, lowering the cost of data collection, and, sometimes, parameter estimation. Variable selection procedures called stepwise procedures are available on all the major statistical computing packages including SAS, SPSS, and BMDP. Even novice researchers can easily run numerous stepwise procedures. Huberty (1989), however, continues by saying that stepwise analyses have been basically used for three purposes: 1) selection and deletion of variables, 2) assessing relative variable importance, and 3) a combination of selection and variable ordering.

Given this information, it is not surprising to find numerous articles in the literature and theses and dissertations in university libraries that have used and misused stepwise procedures despite the many published caveats concerning its appropriateness. Perhaps one reason for the frequent misuse of stepwise procedures is the mistaken perception that the results of a stepwise procedure will yield the "best" equation. According to Hocking (1983), "there is not likely to be a best equation in multiple regression" (p. 226). This is because the use of differing criteria may result in the selection of different sets of variables (Draper & Smith, 1981). Pedhazur (1982) more specifically stated that such methods as all possible regressions, forward selection, backward elimination, stepwise selection and blockwise selection can be utilized with differing criteria which will result in differing solutions depending on the criteria. Morris (1989) sums up these ideas by saying that "there is little theoretical justification for expecting any stepwise procedure to be best" (p. 2).

The goal of stepwise regression is to choose a subset of variables from a larger set for the purpose of parsimony, prediction, explanation, and/or theory-building. However, since the criteria used in selecting variables are statistical, measurement error or randomness may lead to the selection of one variable instead of an equally viable alternative variable. Cohen and Cohen (1975) expounded on this issue saying that "problems include capitalization on chance because of simultaneous tests, sample specificity and trivial differences in partial relationships leading to choosing one variable over another" (p. 103).

When predictor variables are intercorrelated, "there is no satisfactory way to determine relative contributions of the variables on R squared" (Edwards, 1984, p. 107) and "the idea of independent contribution to variance has no meaning" (Darlington, 1968, p. 169). Huberty (1989) reiterates these points by noting that various subsets of a given size can yield nearly the same R^2 value. Pedhazur (1982) states that the R^2 in variance partitioning is sample specific and that nearly identical regression equations can have radically different R^2 values. Furthermore, an incremental R^2 may be statistically significant but substantially meaningless. Pedhazur (1982) argues that the incremental partitioning of variance may be used to control one variable while studying another variable only in causal modeling, and even then the results are of limited value in determining policy.

Another problem to be dealt with is the interpretation of the regression coefficients. Huberty (1989) cautions that the order in which a variable is entered into a model should not be used to assess its relative importance. "The interpretation of regression coefficients as indices of effects of independent variables on the dependent variable appeals to researchers because it holds the promise for unraveling complex phenomena. Examination, however, is important because the apparent simplicity is deceptive" (Pedhazur, 1982, p. 221). Pedhazur (1982) warns that the absence of a theoretical model makes the meaningful interpretation of the estimated regression coefficients impossible. The types of specification errors that can occur are numerous including omission of relevant variables, inclusion of irrelevant variables, interactions among variables, and the hierarchy of polynomial terms (Cohen & Cohen, 1975; Pedhazur, 1982; Peixoto, 1990).

When so many caveats against it have been published, the continued wide usage of stepwise procedures is difficult to understand. Variable selection techniques in regression analysis can be discussed in terms of parsimony, prediction, explanation and theory-building, and selection techniques are problematic in all of these areas.

Parsimony involves finding "a smaller set of predictor variables that do an accurate job of predicting, nearly as well as the total set of variables" (Morris, 1984, p. 1). Obviously, parsimony is helpful to researchers who reap benefits in terms of economy of data collection costs and time. However, the criteria for the selection of the best variables must be weighed on a continuum between internal (parsimony) and external (cross validation) accuracy (Morris, 1984). A prior decision made in the name of parsimony can have a tremendous impact on the results of regression analyses used for prediction, explanation and theory-building.

Pedhazur (1982) states that "for prediction, the goal of regression is to optimize prediction of criteria" (p. 136). The selection of variables for this purpose should account for as much of the variance as possible while balancing practical considerations such as cost and ease of administration. While Morris (1989) finds "particularly 'pernicious' ... a situation with a naive researcher ascribing the best prediction equation from the results of a stepwise program" (p. 1), Pedhazur (1982) argues that "prediction may be accomplished in the absence of theory, but explanation is inconceivable without theory" (p. 174).

The goals of many researchers in terms of explanation have been to identify major variables and determine their relative importance (Pedhazur, 1982). This suggests that stepwise techniques may be plausible initially. The stepwise programs basically perform a hypothesis formulation function (McNeil, Kelly, & McNeil, 1975). However, "problems arise with the stepwise approach, since a great many hypotheses are being tested the resulting best model will most likely be drastically overfit with replication relatively unlikely" (p. 364).

Cohen and Cohen (1975) state that "a research strategy of treating all independent variables simultaneously is most appropriate when no logical or theoretical basis for considering any variable to be prior to any other either causal or relevant in terms of research goals" (pp. 97-98). However, despite this seeming endorsement, they continue by saying "a dim view is taken of stepwise in exploratory research because orderly advance is more likely in the social sciences when researchers use theory to provide hierarchical ordering formed by causal hypotheses rather than computers ordering independent variables" (p. 103).

Given all the problems of sample specificity, interpretation of regression weights, and varying R values, the question arises when is it actually appropriate to use stepwise procedures. Huberty (1989) says that in cases where a large ratio of sample size to variables exists, generalizability of stepwise regression is enhanced, but an external analysis or a cross validation should also be conducted. Thorndike (1978) agrees arguing that "when a fairly large number of predictor variables are available it is advisable to use a stepwise approach, but cross validate" (p. 167). Finally, Cohen and Cohen (1975) state that the distrust of stepwise procedures decreases if: "1) the research goal is predictive not explanatory; 2) N is very large for a given number of independent variables (40 to 1); and, 3) cross validate" (p. 104). Perhaps, Huberty (1989) offers the best advise when he says that "thorough study and sound judgement are suggested for choosing variables at the outset" (p. 62), and that "the data analyst should allow the findings at each stage to influence the direction through subsequent stages" (Allen & Cody, cited in Huberty, 1989, p. 65).

The numerous stepwise procedures available in the major statistical computing packages are so easy to execute, however, that users quickly learn to rely on them, and there is a great temptation for researchers, especially novice researchers, to assume that a stepwise procedure will yield the best model which will stand up to the test of cross validation. Again, this is simply not true. Stepwise procedures actually yield many best models depending on the procedure used and the criteria employed, and it is up to the researcher to decide which one to use and why. In short, stepwise procedures are no substitution for thinking and theorizing. This paper, will now present, compare and contrast several variable selection techniques using two data sets. In the first example, the results of various stepwise techniques from the SAS package will be compared. In the second example, the results of several stepwise regressions used to answer various research questions will be compared.

The first example consists of a dummy data set of 30 subjects used for classroom teaching purposes. The dependent variable is graduate grade point average [GPA], and the four independent variables are the Graduate Record Exam Quantitative subscale [GREQ], the Graduate Record Exam Verbal subscale [GREV], the Miller's Analogy Test [MAT], and a faculty rating of graduate student performance [RAT]. (This data set is available from the authors upon request).

The intercorrelations among these variables and the associated probabilities are presented in Table 1.

Table 1 Correlations and probabilities (N = 30)

Variables		GREQ	GREV	MAT	RAT
GPA	(r)	.61	.58	.60	.62
	(p)	.0003	.0008	.0004	.0003
GREQ	(r)		.47	.27	.51
	(p)		.009	.15	.004
GREV	(r)			.43	.41
	(p)			.02	.03
MAT	(r)				.52
	(p)				.003

As can be seen the dependent variable GPA is highly correlated with all of the independent variables. All the independent variables are also highly correlated with each other except for the combination of GREQ and MAT (r = .27) and possibly GREV and RAT (r = .41). Therefore, pairs of unique information have been set up between GREQ and MAT and between GREV and RAT.

Five different analyses were run using this data set. The first was a full model with all four dependent variables using the forced solution, PROG REG. This model was significant ($F_{4, 25}$ = 11.13, p <.0001, R^2 = .64, adjusted R^2 = .58). The parameter estimates, t values and probabilities appear in Table 2. In this model the t values for GREQ and MAT are significant, while those for GREV and RAT are not.

Table 2 Results of full model using the forced solution in PROC REG to predict GPA from all independent variables

Variable	Parameter Estimate	t	p
Intercept	-1.738	-1.83	.08
GREQ	.004	2.18	.04
GREV	.002	1.45	.16
MAT	.021	2.19	.04
RAT	.144	1.28	.21

The next analysis which was performed was a forward selection. This program identifies a subset of variables which will be as efficient as the entire set of variables for predicting GPA. In this case, the significance level for entering a variable into the model has been set on the lenient side to .15. The variables were entered into the model in the following order: RAT, GREV, MAT, and GREQ. The R^2 values for each new model and the change in R^2 are presented in Table 3. The R^2 for the full stepwise model is .64, as in the full model, since all the variables were entered into the model.

Table 3 Resulting R^2s and changes in R^2s from the forward selection method to predict GPA from allindependent variables

Variable Entered into the Model	R^2	Change in R^2
RAT	.39	-
GREV	.52	.13
MAT	.57	.05
GREQ	.64	.07

The third analysis was a backward elimination. The procedure starts with all the variables entered into the model and then eliminates variables. The significance level for retaining a variable in the model has been set to .05. Again the full model had an R^2 of .64. The variable, RAT, was removed first (R^2 = .62) and then GREV (R^2 = .58), so the best model with GREQ and MAT only included has an R^2 of .58. The results appear in Table 4.

Table 4 Resulting R^2s and changes in R^2s from the backward elimination method to predict GPA from all independent variables

Variables Included	Variables Removed	R^2	Change in R^2
GREQ, GREV, MAT, RAT	-	.64	-
GREQ, GREV, MAT	RAT	.62	.02
GREQ, MAT	RAT, GREV	.58	.04

The fourth analysis used the stepwise method. This rocedure differs from the forward selection method in that ariables entered on earlier steps do not necessarily remain in he model on subsequent steps. After a variable is added, other ariables in the model are inspected to determine if they still roduce a significant F statistic. If the F is not significant, he variable is deleted from the model on that step. For this ase, the significant level for entry into the model was set to 15, and the significance level for remaining in the model was et to .05. The results for this analysis appear in Table 5. he variable, RAT, was entered into the model first (R^2 = .39), hen GREV (R^2 = .52) and then MAT (R^2 = .57). Finally, MAT (R^2 = 52) was removed from the model because the F value for that ariable was not significant, so the resulting best model ncluded RAT and GREV (R^2 = .52).

Table 5 Resulting R^2s and changes in R^2s from the stepwise procedure to predict GPA from all independent variables

Step	Variable Entered	Variable Removed	R^2	Change in R^2
1	RAT	-	.39	-
2	GREV	-	.52	.13
3	MAT	-	.57	.05
4	-	MAT	.52	.05

Finally, the last stepwise procedure used was the maximum R^2 thod. This procedure adds variables that maximizes R^2. The sults of this procedure are presented in Table 6. This ocedure went through five steps and arrived at a model which cluded all four independent variables (R^2 = .64). However, it uld be argued that the best model is determined on the basis of e C(P) statistic. The optimal model is the one for which the P) statistic approaches the number of predictors. In this se, the researcher should stop at step 4 since the C(P) atistic is then equal to 4.63 which is closest to the number of edictor variables or four.

Table 6 Resulting R^2 and C(P) from the maximum R^2 method to predict GPA from all independent variables

Step	Variables in the model	R^2	C(P)
1	RAT	.39	16.74
2	GREV, RAT	.52	9.69
3	GREV, Mat, Rat	.57	7.77
4	GREQ, GREV, MAT	.62	4.63
5	GREQ, GREV, MAT, RAT	.64	5.00

Table 7 presents a summary of the results of all the procedures. The full model, forward selection, and maximum R^2 method all include all four predictor variables and give an R^2 of .64. What is curious is that for the procedures which select only two variables the solutions are quite different. The stepwise procedure ends up with RAT and GREV (R^2 = .52), while the backward elimination ends up with GREQ and MAT (R^2 = .58). The forward, stepwise and maximum R^2 methods all enter RAT into the model first because this variable has the highest correlation with GPA (4 = .62). The next variable entered is GREV. The correlation between RAT and GREV is .41. In the other "best" two variable solutions the correlation between the two predictors, GREQ and MAT is .27. It is important to note that these are the lowest two correlations among all the predictor variables. When variables are highly intercorrelated and one variable is entered into a model first, the next variable entered will add the most unique information, i.e., has the lowest correlation with the first variable. In other words, variables are really entered as pairs (GREQ & MAT, R^2 = .58; GREV & RAT, R^2 = .52). Also, in some situations the procedures, namely forward selection, stepwise, maximum R^2, did not produce the maximum R^2 for the two variable models even though most users think they do. This is because the algorithms in these procedures don't really check <u>all</u> the possibilities.

Table 7 Comparison among the best models of the full model and stepwise results

Procedure	Variables in the model	R^2
Full model	GREQ, GREV, MAT, RAT	.64
Forward selection	RAT, GREV, MAT, GREQ	.64
Backward elimination	GREQ, MAT	.58
Stepwise procedure	RAT, GREV	.52
Maximum R^2	GREQ, GREV, MAT, RAT	.64

In light of this information, what advise can be given to researchers using stepwise procedures? First of all, users of computer packages should know the limitations of the procedures they use. Secondly, researchers should always study the correlation matrix before looking at other results. A thorough knowledge of the intercorrelations may lead researchers to force certain variables into their models first.

In the next example, the results of stepwise regressions are used to answer different research questions. In this example, data from 65 first time, post-myocardial infarction and first time, post-coronary bypass patients were used to study attributions, self-efficacy, and outcome expectations as predictors of depression. The dependent variable was a 20 item scale called the Center for Epidemiological Studies - Depression [CES-D]. Attribution was measured by two instruments: a 9 item behavioral attribution scale [BEHATT] measuring the causes of heart disease that an individual can change, such as smoking, drinking, etc., and an 8 item nonbehavioral attribution scale [NONBATT] measuring the causes of heart disease that are less controllable, such as heredity, luck, etc. The self-efficacy scale [SELFEFF] has 19 items and measures behaviors that individuals have some degree of confidence that they can change. Outcome expectancy 1 [OUTEXP1] was a 19 item scale rating how important patients believe changing particular behaviors are in preventing future heart attacks. Outcome expectancy 2 [OUTEXP2] was a 19 item scale rating the extent of a patient's belief that if behaviors are changed future heart disease will be prevented. A series of four research questions was asked by individual members of a group of researchers and medical practitioners who each advocated a different modelling approach. The data was then analyzed using combinations of forced and stepwise procedures.

In the first analysis, the question was asked whether the set of attribution or the set of self-efficacy and outcome expectation yielded the largest R^2. The results of this analysis consisting of two regression models which entered all variables simultaneously appears in Table 8. These two regression models produce very similar R^2 values (.28 for the attribution variables and .32 for the self-efficacy and outcome expectation variables), and the weights for four of the five variables were significant. In general, it was found that individuals were less depressed about their heart condition if they believed they had some control in the matter.

Table 8 Analysis 1 - A comparison of outcome expectancy/self-efficacy and attribution regression analyses to predict depression

Variable	Beta	F
Regression Model 1		
OUTEXP2	-.48	15.6*
SELFEFF	-.26	4.6*
OUTEXP1	-.12	1.0
R^2 = .32		
Regression Model 2		
BEHATT	-.37	9.6*
NONBATT	.31	6.5*
R^2 = .28		

*$p < .05$

In the second analysis, the question was asked which set of variables explains the most variance after one set was already forced into the model. When the self-efficacy and outcome expectation variables were entered into the model first, the R^2 was .32. After the attribution measures were added the R^2 increased by .08 to .40. When the attribution measures were forced into the model first, the R^2 was .28. After the self-efficacy and outcome expectation variables were added, the R^2 increased by .12 to .40. The results of both analyses were fairly similar.

The third analysis was a forward stepwise regression using all five independent variables. These results appear in Table 9. In this case, the two behavioral attributions added significantly to outcome expectancy 2 in predicting depression.

Table 9 Analysis 3 - Resulting R^2s and changes in R^2s using stepwise regression to predict depression with all independent variables

Variables	R^2	Change in R^2
OUTEXP2	.19	.19*
BEHATT	.31	.12*
NONBATT	.37	.06*
SELFEFF	.40	.03
OUTEXP1	.40	.00

* $p< .05$

The fourth analysis took a more theoretical approach. Some theory suggests that attributions precede behaviors. Following this reasoning two analyses were performed. For the first model, the behavioral attribution variable was forced into the model followed by the stepwise addition of the self-efficacy and outcome expectation variables. For the second model, the nonbehavioral attribution scale was forced into the model followed by the stepwise addition of the self-efficacy and outcome expectation variables. The results appear in Table 10. Only the significant additions of the stepwise procedures are reported. In both cases, outcome expectancy 2 was the only significant contribution to the attribution variable in predicting depression. Again the resulting R^2 values (.31 and .27) from these two models are quite similar.

Table 10 Analysis 4 Two combinations of forced attribution and stepwise outcome expectancy/self-efficacy regression analyses to predict depression

Variables	R^2	Change in R^2
Regression Model 1		
BEHATT	.18	.18
OUTEXP2	.31	.13
Regression Model 2		
NONBATT	.14	.14
OUTEXP2	.27	.13

In summary, although one could argue in favor of each of hese four analyses, the last analysis seems most reasonable ince it was based on theory. This example does show, once gain, that the research question must dictate the research ethodology.

It is hoped that researchers will realize that although ultiple linear regression is a powerful and flexible statistical echnique and although stepwise computer procedures are otentially useful and facilitative, using these techniques and rocedures to meaningfully explain data is a complex process.

For non-experimental research, it is difficult if not impossible o untangle the effects of various variables. Sound thinking, heoretical framework and understanding of the analytical methods re necessary to avoid illogical or unwarranted conclusions" Pedhazur, 1982, p. 175). "Any meaningful analysis applied to omplex problems is never routine. The clarifying of ontroversies in social science research will not be enhanced by pplying all sorts of techniques" (Pedhazur, 1982, p. 171).

References

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, New Jersey: John Wiley & Sons.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69(3), 161-182.

Draper, N., & Smith, H. (1981). Applied regression analysis (2nd Ed.). New York: Wiley.

Edwards, A. L. (1984). An introduction to linear regression and correlation (2nd Ed.). New York: W. H. Freeman and Company.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics, 32, 1-49.

Huberty, C. J. (1989). Problems with stepwise methods -- better alternatives. Advances in Social Science Methodology, 1, 43-70.

McNeil, K. A., Kelly, F.J., & McNeil, J.T. (1975). Testing research hypotheses using multiple linear regression. Carbondale, IL: Southern Illinois University Press.

Morris, J. D. (1989). Alternative variable selection strategies in classification problems. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Nie, N. H., Huss, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). Statistical package for social sciences (2nd Ed.). New York: McGraw Hill.

Pedhazur, E. J. (1982). Multiple regression in behavioral research. (2nd Ed.) New York: Holt, Rinehart and Winston.

Peixoto, J. (1990). A property of well-formulated polynomial regression models. American Statistician, 44(1), 26-30.

Thorndike, R. M. (1978). Correlational procedures for research. New York: Gardner Press, Inc.

Case Influence Statistics Available in SAS Version 5

John T. Pohlman
Southern Illinois University, Carbondale

Abstract

Case influence statistics are a useful diagnostic tool for identifying high leverage cases in a sample. A case's influence on a solved regression model depends on that case's residual and its location in the distribution of the predictor variables. Cases with large residuals and located in extreme ranges of the predictor variables' distributions will be most influential. Case influence is illustated with an SAS analysis of a simple data set.

The REG program in version 5 of the Statistical Analysis System (SAS) provides a collection of case influence statistics described by Belsley, Kuh and Welsch (1980), and Freund and Littell (1986). Influence statistics are designed to aid in the detection of cases which are highly influential in the estimation of the regression coefficients. A case's influence on the regression solution is estimated by deleting that case from the sample and recomputing the coefficients. If the coefficients change considerably upon deleting a case, that case is deemed influential. Generally, cases which have large residuals and are in extreme ranges of the predictor variables' distributions will be most influential.

Figure 1 presents a scatter diagram which illustrates case influence for a simple linear regression model in which a dependent variable (Y) is regressed on one predictor (X). The ten data points denoted with the symbol (•) yield the regression equation

$$Y' = 1 + 1X.$$

The ten data points denoted with the letters A to J are then used, one at a time, to augment the original sample of ten observations. Ten augmented samples of size 11 are thus created. The first augmented sample is composed of the 10 original data points plus point A. The second augmented sample consists of the 10 original observation plus point B, and so on to the tenth augmented sample using case J along with the original observations. The influence of the ten lettered data points is determined by comparing the regression coefficients obtained when the lettered data point is included in the analysis with the coefficients obtained after deleting that data point. Table 1 shows the results of this analysis.

Insert Figure 1 About Here

The second and third columns in Table 1 contain the regression coefficients obtained when cases A to J augment the original sample of 10 cases. The last two columns of the table show the change in the regression coefficients due to the presence of each lettered case. Note that the largest change in the slope coefficient occurs for cases F and J. Cases F and J have the largest deleted residuals and are the most disparate cases in the distribution of X. Cases F and J are the most influential cases. Case J has a strong positive influence on the slope coefficient, since case J's presence in the sample causes the slope coefficient to be .231 units higher than it would be if case J were not in the sample. Case F, to the contrary, has an identically strong negative influence on the slope coefficient.

Insert Table 1 About Here

INFLUENCE STATISTICS AVAILABLE IN PROC REG

The influence statistics described here are available in the SAS REG procedure as options. SAS provides the statistics HAT DIAG H, DFBETA and DFFITS. For this illustration assume that the general linear model is fit to a data set, namely

$$Y = XB + E$$

where Y is a vector of values on the response variable, X is an $n \times (p+1)$

matrix of values on the independent variables with a leading unit vector, B is the vector of regression coefficients and E is a residual vector. Letting X^T denote the transpose of X, the ordinary least squares regression coefficients are given by

$$B = (X^TX)^{-1}X^TY,$$

and the predicted values of Y are produced by

$$\begin{aligned} Y' &= XB \\ &= X(X^TX)^{-1}X^TY \end{aligned}$$

Letting $H = X(X^TX)^{-1}X^T$, then

$$Y' = HY.$$

The matrix H is the projection matrix for the predictor space in that it operates on Y to yield Y', and is termed the hat matrix. H is of order n×n and of the same rank as X. The main diagonal values of H, h_{ii}, are measures of the dispersion of case i from the centroid of the predictor variable space. Two cases with the same value of h_{ii} are on the same probability contour of the multivariate distribution of the predictor variables. In fact, h_{ii} is a linear transformation of the Mahalanobis distance of case i from the centroid of X (Weisberg, 1980, p. 105). The h_{ii} values are labeled HAT DIAG H by the REG program. The h_{ii} values measure the potential for a case to be influential. The actual influence exerted by a case will also depend on that case's residual.

The DFBETA statistics are measures of the influence each case has on each of the regression coefficients. For each case the will be a separate DFBETA value for each regression coefficient in the model, including the intercept. The DFBETA for case i on coefficient j is

$$DFBETA_{j(i)} = \frac{b_j - b_{j(i)}}{\left[S^2_{(i)} (X^TX)^{jj}\right]^{1/2}}$$

where b_j is the regression coefficient for predictor j estimated from the total sample, $b_{j(i)}$ is the regression coefficient for variable j estimated in the sample with case i deleted, $S^2_{(i)}$ is the error variance estimate from the sample with case i deleted and $(X^TX)^{jj}$ is the j-th diagonal element of $(X^TX)^{-1}$.

The DFFITS statistic is a scaled measure of the influence of case i on the predicted value of Y. Since all of the regression coefficients are used to produce a predicted Y value, DFFITS becomes an aggregate measure of the influence of case i on the entire regression equation. The DFFITS statistic for case i is given by

$$DFFITS_{(i)} = \frac{Y'_i - Y'_{i(i)}}{[S^2_{(i)} h_{ii}]^{1/2}}$$

where Y'_i is the predicted Y for case i based on the total sample, $Y'_{i(i)}$ is the predicted Y based on the regression equation estimated without case i in the sample, and h_{ii} is the i-th diagonal value of H. The DFFITS statistic is very similar to Cook's D (Cook, 1979), another measure of influence available in the REG program and also in the SPSSx regression program. Cases with DFFITS values greater than $2[(p+1)/n]^{1/2}$ are considered to be high leverage cases (Belsley et al., 1980, p. 28).

ILLUSTRATION WITH A DATA SET

Appendix A provides a SASLOG and LISTING for a sample regression model based on 24 cases. Page 1 in Appendix A contains the model statement (SASLOG line 30) which requests the regression of attitudes toward school (ATTSCH) on INCOME and IQ. The INFLUENCE option is requested for the model.

Page 2 in the Appendix contains the parameter estimates for the model, followed by the influence statistics. The studentized residuals (RSTUDENT) and the HAT DIAG H present the two important sources of case influence. Case 6 has the largest studentized residual (2.9823) and case 14 also has a large studentized residual (-1.5497). The DFFITS value for case 14 is (-1.5747), and this is the largest value, in absolute terms, in the sample. The negative value of DFFITS for case 14 means that the predicted Y for case 14 is increased when case 14 is deleted from the sample. Conversely, the presence of case 14 in the sample causes that case's predicted value to be reduced.

The DFBETA statistics are then presented for each regression coefficient, for each case. Case 14 is also the most influential case for estimating each of the regression parameters individually: INTERCEP DFBETA = -.5455, INCOME DFBETA = -1.4997 and IQ DFBETA = .9250. As with the DFFITS statistic, the sign of the DFBETAs indicate the direction of influence on the regression coefficients for case 14. Case 14's presence in the sample causes the y-intercept to decrease, the regression coefficient for INCOME to decrease and the coefficient for IQ to increase. On page 5 of the Appendix the regression equation is estimated with case 14 deleted from the sample, and indeed the changes in the coefficients are as suggested by the DFBETA diagnostics for case 14.

HANDLING INFLUENTIAL CASES

Once the influential cases have been identified the analyst must decide what to do with them. The first step should be to determine if the influential cases are correctly coded. Typographical errors made while entering the data can produce highly influential cases. If data errors are detected, clearly the proper course of action is to correct the data values. If the correct data values are not available then deletion of such cases is reasonable.

However, if the analyst determines that a case is correctly coded and still highly influential, three alternatives are available: 1. delete the case from the sample, 2. retain the case in the sample but note that the case is influential, or 3. revise the model to accommodate the influential case.

It is a questionable practice to delete cases from a sample simply because they are unusual. In fact, unusual cases often point to weaknesses in our models and may suggest improvements in our theories. For example, if a researcher fit a linear model to a nonlinear relationship many of the data points would be found to have large residuals and therefore might be highly influential. Deletion of unusual cases in this example would lead to the interpretation of an incorrect model. When a case is deleted from a sample it is presumed that the model is correct and the offending case is invalid. Our models should be burdened to fit our data; our data should not be obliged to fit our models. Data should not be deleted to better fit our models unless we have compelling evidence that the data is wrong.

The least squares criterion can itself be the cause of an influence problem. A case's influence is proportional to the square of its residual when OLS estimation is used. A researcher might try fitting a model using a criterion other than OLS. The SAS version 5 package has a procedure that fits models using the least absolute value error (PROC LAV). Unfortunately, this procedure is not available in version 6 of SAS. This program minimizes the sum of the absolute deviations from the model, thereby tempering the influence of high residual cases. If the coefficients estimated with OLS and LAV criteria are comparable, the model may be considered sufficiently robust for interpretation. Page 4 in the Appendix shows the LAV solution for the same model estimated earlier using OLS. The only coefficient that is changed markedly is the y-intercept. The coefficients for INCOME and IQ are approximately the same as their OLS counterparts. One might, therefore, conclude that the OLS estimates are fairly robust in this sample.

REFERENCES

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). Regression diagnostics. New York: John Wiley and Sons, Inc.

Cook, R. D. (1979). Influential observations in linear regression. Journal of the American Statistical Association, 74, 169-174.

Freund, R. J. and Littell, R. C. (1986). SAS system for regression 1986 edition. Cary, N. C.: SAS Institute Inc.

Weisberg, S. (1980). Applied linear regression. New York: John Wiley and Sons, Inc.

Figure 1. Scatter Diagram Illustating Influence

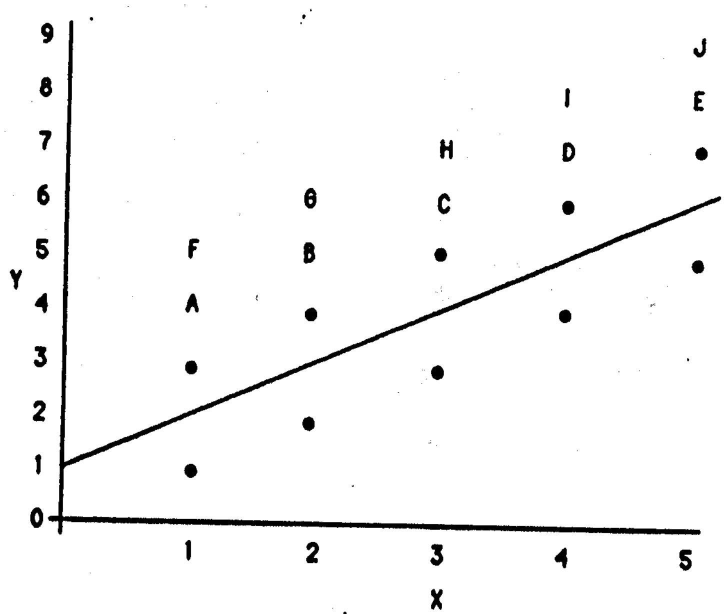


Table 1. Influence of Cases A-J on Model Coefficients

Case	Regression Coeficients		Influence of Case on	
	Intercept	Slope	Intercept	Slope
A	1.625	.846	.625	-.154
B	1.435	.913	.435	-.087
C	1.182	1.000	.182	.000
D	.913	1.087	-.087	.087
E	.692	1.154	-.308	.154
F	1.920	.769	.920	-.231
G	1.652	.870	.652	-.130
H	1.273	1.000	.273	.000
I	.870	1.130	-.130	.130
J	.538	1.231	-.462	.231

Note: The regression equation for the original 10 cases is Y' = 1 + 1X.

Appendix Page 1

SASLOG FOR THE INFLUENCE ILLUSTRATION

```
 1 DATA ONE;
 2 OPTIONS LS = 70 NUMBER;
 3 INPUT SUBID GENDER IQ HEALTH GRADE INCOME ATTSCH;
 4 CARDS;

NOTE: DATA SET WORK.ONE HAS 24 OBSERVATIONS AND 7 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.07 SECONDS AND 84K.

 29 PROC REG;
 30 MODEL ATTSCH = INCOME IQ /INFLUENCE;
NOTE: THE PROCEDURE REG USED 0.15 SECONDS AND 416K
      AND PRINTED PAGES 1 TO 2.

 31 PROC LAV;
 32 MODEL ATTSCH = INCOME IQ;

NOTE: LAV IS NOT SUPPORTED BY THE AUTHOR OR BY SAS INSTITUTE INC.
NOTE: THE PROCEDURE LAV USED 0.16 SECONDS AND 3020K
      AND PRINTED PAGE 3.

 33 DATA TWO;
 34 SET ONE;
 35 IF SUBID NE 14;

NOTE: DATA SET WORK.TWO HAS 23 OBSERVATIONS AND 7 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.04 SECONDS AND 424K.

 36 PROC REG;
 37 MODEL ATTSCH = INCOME IQ;
NOTE: THE PROCEDURE REG USED 0.10 SECONDS AND 440K
      AND PRINTED PAGE 4.
```

DEP VARIABLE: ATTSCH
ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2	4964.02906	2482.01453	42.469	0.0001
ERROR	21	1227.30427	58.44306054		
C TOTAL	23	6191.33333			

ROOT MSE	7.644806	R-SQUARE	0.8018
DEP MEAN	34.66667	ADJ R-SQ	0.7829
C.V.	22.05238		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	-0.31262577	8.23978063	-0.038	0.9701
INCOME	1	1.13695467	0.16067102	7.080	0.0001
IQ	1	0.11410777	0.08624277	1.323	0.2000

OBS	RESIDUAL	RSTUDENT	HAT DIAG H	COV RATIO	DFFITS	INTERCEP DFBETAS
1	9.5254	1.3285	0.0864	0.9832	0.4086	-0.1528
2	-2.4677	-0.3309	0.0738	1.2295	-0.0934	-0.0266
3	-10.2098	-1.4342	0.0892	0.9476	-0.4490	-0.1037
4	14.4111	2.0795	0.0462	0.6760	0.4680	-0.0780
5	-1.3467	-0.1798	0.0811	1.2597	-0.0534	-0.0137
6	18.9446	2.9823	0.0500	0.4041	0.6841	0.0167
7	-6.4506	-0.8926	0.1151	1.1695	-0.3219	-0.0891
8	5.0227	0.6935	0.1246	1.2314	0.2616	0.2274
9	-4.8311	-0.6718	0.1303	1.2564	-0.2601	0.1845
10	2.6169	0.3459	0.0620	1.2123	0.0890	0.0613
11	-2.6141	-0.3441	0.0541	1.2023	-0.0823	-0.0180
12	2.8344	0.3845	0.1080	1.2694	0.1338	0.1125
13	-12.9953	-1.8851	0.0880	0.7771	-0.5856	0.3392
14	-8.0459	-1.5497	0.5080	1.6743	-1.5747	-0.5455
15	-3.0505	-0.4636	0.2842	1.5861	-0.2921	0.1620
16	-1.8238	-0.2494	0.1266	1.3131	-0.0950	-0.0821
17	1.4782	0.1933	0.0494	1.2110	0.0441	0.0134
18	-3.0891	-0.4588	0.2536	1.5029	-0.2674	0.1857
19	-1.9172	-0.2548	0.0745	1.2387	-0.0723	-0.0343
20	3.3009	0.4419	0.0818	1.2246	0.1319	0.1080

OBS	RESIDUAL	RSTUDENT	HAT DIAG H	COV RATIO	DFFITS	INTERCEP DFBETAS
21	7.3604	1.1207	0.2530	1.2909	0.6522	-0.3998
22	-1.5951	-0.2199	0.1402	1.3367	-0.0888	-0.0779
23	-3.8772	-0.5110	0.0495	1.1714	-0.1166	-0.0628
24	-1.1594	-0.1528	0.0605	1.2279	-0.0388	0.0140

OBS	INCOME DFBETAS	IQ DFBETAS
1	-0.2443	0.2673
2	0.0560	-0.0075
3	0.3081	-0.0658
4	-0.0875	0.1724
5	0.0345	-0.0061
6	0.2325	0.0099
7	0.2376	-0.0408
8	-0.0245	-0.1678
9	-0.0158	-0.1829
10	-0.0059	-0.0401
11	0.0383	-0.0121
12	0.0554	-0.1049
13	0.0461	-0.3850
14	-1.4997	0.9250
15	-0.1526	-0.1098
16	-0.0410	0.0778
17	-0.0150	0.0001
18	0.1655	-0.2399
19	0.0353	0.0092
20	0.0088	-0.0830
21	0.2792	0.3027
22	-0.0399	0.0745
23	-0.0249	0.0463
24	0.0111	-0.0216

Appendix Page 4

LAV REGRESSION PROCEDURE FOR DEPENDENT VARIABLE ATTSCH

VARIABLE	LAV COEFFICIENT
INTER	-0.23255814
INCOME	1.13953488
IQ	0.09302326

(NOTE: THE COEFFICIENT ESTIMATES ARE UNIQUE.)

RESIDUAL SUM OF ABSOLUTE VALUES = 120.74418605
ADJUSTED TOTAL SUM OF ABSOLUTE VALUES = 272.00000000
NUMBER OF OBSERVATIONS IN DATA SET = 24

DEP VARIABLE: ATTSCH
ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2	4425.92197	2212.96099	40.392	0.0001
ERROR	20	1095.73020	54.78650996		
C TOTAL	22	5521.65217			

ROOT MSE	7.401791	R-SQUARE	0.8016
DEP MEAN	33.56522	ADJ R-SQ	0.7817
C.V.	22.05197		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	4.03922071	8.45765999	0.478	0.6381
INCOME	1	1.37254205	0.21674866	6.332	0.0001
IQ	1	0.03686642	0.09724565	0.379	0.7086

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

Some Parallels Between Predictive Discriminant Analysis and Multiple Regression

John D. Morris
Florida Atlantic University

Carl J. Huberty
University of Georgia

Some Parallels Between Predictive Discriminant Analysis and Multiple Regression

The purpose of this paper is to outline some important similarities in, and differences between, predictive discriminant analysis (DA) and multiple regression (MR). The areas covered, chosen for their importance and need for clarification, are estimates of model accuracy, hypothesis testing, and non-least squares models. Some of the parallels are well known, some are less well known, and some appear to have not yet been considered at all.

It is well known that when 1) only two groups are involved, 2) the two population predictor covariance matrices are assumed equal, and 3) the two prior probabilities of group membership are taken to be equal, the popular "minimum chi-square rule" (Tatsuoka, 1971, p. 218) associated with discriminant analysis (DA) is equivalent to predicting a dichotomous criterion variable via multiple regression (MR) methods and classifying a subject into the group for which the predicted criterion is nearer the actual. An especially enlightening examination of this and some other multivariate techniques from the general perspective of MR is provided by Flury and Riedwyl (1985).

However, a precaution about the equivalence of two-group classification and multiple regression with a dichotomous criterion is appropriate. In a two-group situation, there is one linear discriminant function (LDF) and there are two linear classification functions (LCFs); an LDF and an LCF are simply linear composites of the predictors. It is true in a two-group context that the regression weights are proportional to the single set of LDF weights. When a linear regression function (LRF) or an LDF is used for classification purposes a cut-off criterion needs to be determined — with an LRF it is midway between the two values by which the dichotomous criterion is coded, with an LDF it is midway between the LDF means for the two groups. With the use of LCFs, there is no cut-off per se; rather a unit is classified into the group with which is associated the larger LCF score. It turns out that the respective LCF weight differences are proportional to the corresponding LDF and (therefore) the LRF weights.

Input scores for an LRF, an LDF, and an LCF are typically predictor variable measures. [As stated above, any of the three linear composite types may be used for a two-group classification problem.] It turns out that another, still equivalent, approach to two-group classification may be employed. Here, one uses LDF scores for each unit as input for an LCF; we thus have, in essence, a single predictor score for each unit.

When generalizing from a two-group problem to a k-group problem, it is advisable to forget the LRF and LDF approaches and focus on the LCF approach, with predictor measures as input scores.

Estimates of Model Accuracy

Estimation of the cross-validated accuracy of a prediction model offers similarities and differences between MR and DA methods. In both DA and MR the researcher must decide what type of cross-validated accuracy is of concern. For instance, is interest in simply estimating an accuracy index parameter from the associated statistic, that is, estimating the index of accuracy (R^2 or percent of "hits", respectively) that would obtain in the population from that same index in the sample, or is interest in the accuracy that would obtain on application of sample optimized weights to alternate samples from the same population? The concern in this paper will be with the latter type of accuracy.

As in an estimate of cross-validated R^2 in MR, a judgment of DA "hit-rate" based on the calibration sample is optimistically biased in reference to application to alternate samples. To estimate a cross-validated result in MR, another decision that must be made is whether interest is in relative accuracy, as manifested in the correlation of Y and $\hat{Y}$, or in absolute accuracy, as manifested in the MSE. In either case, several formula estimates are available (see Huberty & Mourad, 1980; Rozeboom, 1978). It is probable that in most predictive uses of MR in the behavioral sciences, such as in personnel selection, concern is with relative accuracy.

Unlike in MR, the concern in predictive DA is in classification accuracy; this is implicitly a concern of absolute accuracy. A formula estimate for cross-validated hit-rate in the general *k*-group case has largely eluded methodologists. However, a useful, although complicated, formula estimate for cross-validated hit-rate in the two-group case was derived by McLachlan (1975). According to that estimator, the hit rate, P_g for group g, where g = 1 or 2 is:

$$\hat{P}_g = 1 - F(-D/2) - f(-D/2)\{(p-1)/(Dn_g) + D[4(4p-1) - D^2]/(32m) + (p-1)(p-2)/(4Dn_g^2) + (p-1)[-D^3 + 8D(2p+1) + 16/D]/(64mn_g) + D[3D^6 - 4D^4(24p+7) + 16D^2(48p^2 - 48p - 53) + 192(-8p+15)]/(12288m^2)\},$$

where F is the standard normal distribution function, i.e., F(-D/2) is the area to the "left" of -D/2, f is the standard normal density function, D is the Mahalanobis distance, p is the number of predictor variables, n_g is the number of subjects in group g, and $m = n_1+n_2-2$. While the formula looks formidable, with patience, it is calculable with a hand-held calculator. Moreover, as the last term in the multiplier for f(-D/2) is usually very small, one may choose to ignore it, making the formula even more tractable. If the researcher with an orientation toward MR notes that $D^2 = R^2N(N-2)/(1-R)^2n_1n_2$, then the McLachlan estimator of cross-validated hit-rate can be obtained from the R^2 resulting from regressing the dichotomous criterion on the predictors.

One slightly "unnerving" aspect of the McLachlan estimator is that it can yield estimated hit-rates that are <u>larger</u> than those that are estimated from the known positively biased process of reclassifying the calibration sample (Morris & Huberty, 1986; 1987). This is unlike the case in MR where the "shrunken" multiple correlation is necessarily less than the value of the multiple correlation derived from the calibration sample. The explanation for this apparent paradox between methods is that estimators of the cross-validated multiple correlation are functions of the corresponding calibration sample multiple correlation, and are therefore guaranteed to yield smaller values than the sample value. In this sense, the McLachlan hit-rate estimator is not parallel to the MR formula estimators. While it is an estimator of cross-validation hit-rate, it is not a function of the calibration sample generated hit-rate. Rather, it is a function of the Mahalanobis distance between groups, as well as other variables. That is, it does not simply estimate a parameter from a function of the corresponding statistic as do MR formula estimators.

An alternate nonparametric approach to estimating cross-validated hit-rate, which has a wide following in the DA literature, is the "leave-one-out" procedure (Huberty, 1984; Huberty & Mourad, 1980; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968). In this method, a subject is classified by applying the rule derived from all Ss except the one being classified. This process is repeated "round-robin" for each subject with a count of the overall classification accuracy used to estimate the cross-validated accuracy.

Clearly the same "round-robin" procedure can be used to estimate either relative or absolute accuracy in the use of MR, and has appeared in that context, with perhaps the earliest reference due to Gollob (1967). In a system intended to select optimal MR predictor variable subsets, Allen (1971) coined the procedure "PRESS," and he appears to be the source most often cited in the MR literature.

The apparent computational difficulties due to the inversion of N matrices can be avoided in both MR and DA by using a matrix identity due to Bartlett (1951). This identity is cited and used explicitly in introducing the technique in the DA context by Lachenbruch and Mickey (1968), but was not mentioned by Allen in the first introduction of PRESS (1971) nor in its presentation in a later text (Allen & Cady, 1982, p. 254), although the same identity was implicitly used. Moreover, Allen doesn't cite the DA literature and the parallel application of the PRESS procedure. It appears that this resampling process was "invented" independently in the MR and DA literatures.

Full vs Restricted Model Hypothesis Testing

A technique that is well known and widely used by MR researchers is that of hypothesis testing through contrasting full and restricted prediction models. The power of this method, its generality, and its applicability to a very wide arena of theoretical questions in science is no doubt part of the reason for the establishment of the MLRSIG within AERA.

The same types of model contrast "explanatory increment" questions can be asked and seem to be at least as much potential interest when the criterion is classification accuracy. However, we know of no examples of this technique being used in the literature. There seems to be no reason not to test the difference in proportion of correct classifications (hit-rate) between full and restricted models to examine meaningful hypotheses, just as is done using the R^2 in MR. The appropriate test statistic is McNemar's (1947) contrast between correlated proportions. Moreover, as the index, "I", of increase in classification accuracy over chance (see Huberty, 1984, p. 168) is distributed similarly, it becomes apparent that such a test would also be applicable to that statistic.

An example of such a test from a study in which the subsequent high school dropout of a sample of 76 children was predicted from data available in fifth grade will now be presented. The six predictor variables were gender, race (two levels), number of elementary schools in which the child had been a student, the number of grades the child had repeated, the family structure (living with at least one natural parent and no other adult, or not), and the child's total number of fifth grade absences. As we have evidence of the relationship between both gender and race and the criterion of high school dropout, the hypothesis to be tested concerned the significance of the increment to classification accuracy afforded by adding the four "non-organismic" variables (number of elementary schools, number of grades repeated, family structure, and the total number of fifth grade absences) to the prediction model containing only gender and race.

Classifying the calibration sample, the proportion of correct classifications for the total model was 75% and for the model including only gender and race it was 63%. A 2x2 table illustrating the number of hits and misses for both models is:

All Predictors

		MISS	HIT
Gender and Race	HIT	9	39
	MISS	10	18

The test statistic, z = 1.73, would typically be considered non-significant (P = .08) and therefore offers no evidence that the additional variables add to the classi [illegible] accuracies for these two three predictor variable models (number of elementary schools, number of repeats, and family structure, 79%; number of elementary schools, number of repeats, and number of absences, 79%) were each greater than for the total six variable predictor model. Thus, unlike the multiple correlation coefficient in MR, even with non-cross-validated "internal" estimates of classification hit-rate, accuracy does not necessarily monotonically increase as one adds predictor variables. A different perspective concerning contrasting reduced and full model predictor variable subsets may therefore be necessary for DA applications.

One may argue, however, that the cross-validated estimate of accuracy should be used in any case. An illustration of the impact that using a cross-validated estimator might have is that the leave-one-out estimator for the hit rates involved in the hypothesis tested above were 64% for the full six-variable model, and 49% for the three variable model, with a resulting test statistic of $z = 2.45$, which is of course significant at the .02 level. Therefore, the researcher would most likely come to a different conclusion concerning the significance of the increment due to these additional variables using cross-validated estimates.

Nonleast Squares Models

Non least-squares prediction strategies, particularly ridge regression, have received a great deal of attention in the MR literature (e.g., Darlington, 1978; Morris, 1982, 1986; Pagel & Lunneborg, 1985; Rozeboom, 1979), and some attention in DA (Campbell, 1980; DiPillo, 1976, 1977, 1979; Morris & Huberty, 1987). As the benefit to predictive accuracy of such methods is a function of whether the concern is relative or absolute accuracy, the results for DA tend to be a subset of those for MR. They appear to be largely parallel to the case of absolute accuracy in the MR case (Morris & Huberty, 1987); enhanced predictive accuracy is available under certain limited circumstances, however, decrements in accuracy are just as likely to occur without an informed decision about when to use the technique. Ridge methods are far from the panacea that they have been purported to be for either the MR or DA case. A suggested method for choosing between alternate predictor weighting algorithms, including ridge and least squares, has been advanced for the DA case by Morris and Huberty (1987), and for the MR case by Morris (1986). Computer programs for both analysis types are available.

References

Allen, D. A. (1971). The prediction sum of squares as a criterion for selecting predictor variables (Tech. Rep. No. 23). University of Kentucky, Department of Statistics.

Allen, D. A., & Cady, F. B. (1982). Analyzing experimental data by regression. Belmont, CA: Wadsworth.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. Annals of Mathematical Statistics, 22, 107.

Campbell, N. A. (1980). Shrunken estimates in discriminant and canonical variate analysis. Journal of the Royal Statistical Society, 29, 5-14.

Darlington, R. B. (1978). Reduced variance regression. Psychological Bulletin, 85, 1238-1255.

DiPillo, P. J. (1976). The application of bias to discriminant analysis. Communications in Statistics, A5, 843-854.

DiPillo, P. J. (1977). Further applications of bias to discriminant analysis. Communications in Statistics, A6, 933-943.

DiPillo, P. J. (1979). Biased discriminant analysis: Evaluation of the optimum probability of misclassification. Communications in Statistics, A8, 1447-1457.

Flury, B., & Riedwyl H. (1985). T^2 tests, the linear two-group discriminant function and their computation by linear regression. The American Statistician, 39, 20-25.

Gollob, H. F. (1967, September). Cross-validation using samples of size one. Paper presented at the meeting of the American Psychological Association, Washington D.C.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. Psychological Bulletin, 95, 156-171.

Huberty, C. J, & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 40, 101-112.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10, 1-11.

McLachlan, G. J. (1975). Confidence intervals for the conditional probabilities of misallocation in discriminant analysis. Biometrics, 31, 161-167.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 12, 153-157.

Morris, J. D. (1982). Ridge regression and some alternate weighting techniques: A comment on Darlington. Psychological Bulletin, 91, 203-210.

Morris, J. D. (1983). Stepwise ridge regression: A computational clarification. Psychological Bulletin, 91, 363-366.

Morris, J. D. (1986). Microcomputer selection of a predictor weighting algorithm. Multiple Linear Regression Viewpoints, 15, 53-68.

Morris, J. D., & Huberty, C. J. (1986, April). A comparison of three methods of classification hit-rate estimation. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Morris, J. D., & Huberty, C. J. (1987). Selecting a two-group classification weighting algorithm. Multivariate Behavioral Research, In press.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey, & E. Aronson (Eds.), Handbook of social psychology: Vol 2. Reading Mass.: Addison-Wesley.

Pagel, M. D., & Lunneborg, C. E. (1985). Empirical evaluation of ridge regression. Psychological Bulletin, 97, 342-355.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. Psychological Bulletin, 85, 1348-1351.

Rozeboom, W. W. (1979). Ridge regression: Bonanza or beguilement? Psychological Bulletin, 86, 242-249.

Tatsuoka, M. M. (1971). Multivariate analysis. New York: John Wiley.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

The Use of Regression Diagnostics to Improve Model Fit: A Case of Role Strain and Job Stress

Susan Tull Beyerlein and Michael Martin Beyerlein
University of North Texas

Abstract

This paper illustrates the importance of using regression diagnostics to improve model fit when using standard multiple regression statistical packages such as SASPC. This study examined the relationship between employee perceptions of their work environments and perceived job stress. The analysis was theory driven rather than exploratory in nature, and was performed using SASPC multiple regression procedures. Variables were coded to reduce possible collinearity. Various regression diagnostics were examined to detect the presence of outliers, influential observations, residual correlation, and collinearity (e.g., VIFs, DFFITS, the C_p criterion, HAT (leverage) values, and the Durbin-Watson test). These values, coupled with the various regression procedures yielded a final, best nine-variable model of $R^2 = .48$, significantly larger than the initial value of $R^2 = .27$. Future research in this area could be strengthened through 1) an examination of the path analytic and LISREL models in the literature that attempt to model indirect effects, 2) possible incorporation of select, higher-order terms from these studies, and 3) utilization of the regression diagnostic procedures outlined in this paper.

Introduction

Role conflict and role ambiguity are two stressors that have been linked to various health and physical outcomes. Role conflict involves conflicting task assignments initiated by superiors of equal rank and authority. Role ambiguity concerns the lack of clarity regarding job assignments, work objectives, and others' expectations. Kahn and others (1964) found that men who experience role conflict and role ambiguity on the job exhibit more tension and less job satisfaction than men whose roles are congruent or unambiguous. Research shows that role conflict correlates with a number of other outcomes including poor performance (Liddel & Slocum, 1976), poor peer relationships (French & Caplan, 1972), and turnover (Brief & Aldag, 1976; Hamner & Tosi, 1974). Role ambiguity has been linked to ineffective coping, as well as turnover.

Underutilization and job future ambiguity are two additional job stressors that have been shown to impact perceived job stress (Caplan, Cobb, French, Harrison, & Pinneau, 1980). Underutilization of abilities involves the lack of opportunity on the job to use skills and knowledge acquired in school or from previous experience and training. Job future ambiguity concerns levels of certainty regarding future career plans, opportunities for promotion, future value of current job skills, and future job responsibilities. These four variables, that is, role conflict, role ambiguity, underutilization, and job future ambiguity, plus years on the job and gender were chosen from a larger set of variables because of strong theoretical connections to stress, and after correlation analysis suggested they were the best set for predicting perceived job stress.

Method

The present study involved a survey of staff members at a large, southwestern university. Respondents were white collar workers in various clerical, secretarial, and administrative positions. A total of 660, 14 page surveys were sent through the campus mail system, and 134 were returned for a response rate of 20.3 percent. Twenty-three cases were omitted because of missing data. The initial predictor variables used in the study were as follows: gender (D1), years on job (X1), role conflict (X2), role ambiguity (X3), underutilization (X4), and job future ambiguity (X5). The criterion variable was perceived job stress (Y). Gender (D1) was represented by dummy coding (i.e., 0 males, 1 females). Selected interaction terms were then created based on developed theory in the literature, that is, years on job times underutilization (X6), role conflict times role ambiguity (X7), and years on job times job future ambiguity (X8). Due to the fact that stress has often shown nonlinear relationships to other variables, several squared, higher order terms were included in the analysis, that is, role conflict (X2X2), role ambiguity (X3X3), underutilization (X4X4), and job future ambiguity (X5X5). Finally, all the predictor variables with the exception of gender (D1), were coded in order to reduce the likelihood of rounding errors in regression coefficients leading to collinearity (Mendenhall & Sincich, 1989, p. 343). Thus, to denote coded variables, "U" replaces "X" for all variables except D1 and the criterion variable Y.

Results

The analysis was performed using SASPC and involved a number of procedures. First, the initial set of predictors was included in the general regression procedure, PROC REG (i.e., D1, X1-X5). This analysis yielded an R^2 = .27. Next, this procedure was repeated with these variables and the additional interaction and higher order terms (i.e., D1, X1-X8, X2X2, X3X3, X4X4, X5X5). This yielded an R^2 = .34. The correlation procedure, PROC CORR, was also run at this point in order to obtain means and standard deviations for the predictor variables.

The twelve variables (excluding D1 and Y) were then coded and analyzed using the general regression procedure, PROC REG with the INFLUENCE option, (i.e., U1-U8, U2U2, U3U3, U4U4, U5U5). This analysis yielded an R^2 = .31. The subsequent inclusion of D1 (gender) raised the R^2 value to .34. The DFFITS values were then examined in order to identify possible influential observations. The SAS User's Guide: Statistics (1985) describes the DFFITS statistic as "a scaled measure of the change in the predicted value of the ith observation (which is) calculated by deleting the ith observation" (p. 677). The difference, $y_i - y_{(i)}$, has been divided by its standard error so that the differences can be more easily compared. The investigator is interested in values that are considerably larger relative to the other differences in predicted values. For most purposes, a value of 1.0 is considered to be sufficiently large to warrant attention. In the present study, influence diagnostics revealed five DFFITS values greater than 1.0. These were subsequently deleted from the analysis leaving a remaining sample of n=106.

The regression procedure, PROC STEPWISE, was then utilized, specifically, the FORWARD, BACKWARD, and MAXR options. The PROC STEPWISE procedure is a good choice when there are a number of independent variables to consider. The various options do not always isolate the model with the highest R^2 but rather seek the best one-variable model, two-variable model, and so forth (SAS User's Guide: Statistics, 1985). The FORWARD option requests the forward selection technique, BACKWARD requests the backward elimination technique, and MAXR requests the maximum R^2 improvement technique. MAXR looks at all possible regression equations, however, as with the other options it outputs only the best models, for example, the best ten-variable, nine-variable, eight-variable models, and so forth.

After examining the output from the PROC STEPWISE analyses it was decided that the following was the best model: $R^2 = .38$, D1, U1, U2, U3, U4, U8, U3U3, U4U4, $C_p = 7.87$, with all variables significant at the 0.10 level. The C_p criterion is gleaned from the FORWARD and BACKWARD procedures (rather than MAXR) and is used to select the best subset model with a small total mean square error (C_p), and a value of C_p near $p + 1$, which indicates that slight or no bias exists $[E(C_p) \approx p + 1]$. In this case, the C_p value was slightly less than the number of parameters in the model (i.e., eight).

This model was then analyzed using the general regression procedure, PROC REG, with the VIF, P, R, DW, and INFLUENCE options. VIF prints variance inflation factors with the parameter estimates; variance inflation is the reciprocal of tolerance; P calculates predicted values from the estimated model and input data; R analyzes the residual and includes the Cook's D statistic which is an overall measure of influence for each observation, the standard errors of the predicted and residual values, and the studentized residual; DW calculates the Durbin-Watson statistic; INFLUENCE prints the following diagnostics used in the present study for each observation: the residual, studentized residual, HAT or leverage value (h_i), and the DFFITS statistic.

Examination of the plot RESID*PRED (residuals times predicted scores) revealed a value greater than +2 standard deviations, that is, a possible outlier. This value was subsequently deleted leaving a sample of n=105. A rerun of the general regression procedure, PROC REG, using the above best model yielded an R^2 = .41, and a Durbin-Watson, D = 2.21, suggesting the residuals were slightly negatively correlated (Mendenhall & Sincich, 1989, p. 307). However, calculation of the average leverage value, $\bar{h} = (k + 1)/n = .17$, and examination of the HAT values revealed four values greater than twice the average value, suggesting that these values were influential observations and should be eliminated from the data set. They were subsequently deleted leaving a final sample of n=101.

Examination of PROC STEPWISE options, that is, FORWARD, BACKWARD, and MAXR revealed significant gains in R^2 values. At this point, a nine-variable model was chosen as the best model for several reasons: 1) the C_p value was only slightly less than the number of predictors (Younger, 1985), whereas it was significantly larger for other models with similar R^2 magnitude; 2) all variables were significant at the 0.10 level; 3) there was a significant drop in R^2 using the BACKWARD procedure as one dropped to the eight variable models; and 4) the Durbin-Watson statistic was close to a value of two, suggesting minor residual correlation. Therefore, the best model chosen was as follows: R^2 = .48, D1, U1, U2, U3, U4, U7, U8, U3U3, U4U4, C_p = 8.85, DW = 2.21. Thus, the final model included the following variables: gender (D1), years on job (U1), role conflict (U2), role ambiguity (U3), underutilization (U4), role conflict times role ambiguity (U7), years on job times job future ambiguity (U8), and the squared, higher-order terms utilizing role ambiguity (U3U3) and underutilization (U4U4). The only conflicting evidence was the value of the variance inflation factors (VIFs) for U3, U4, U3U3, and U4U4. These values were greater than 10, whereas the VIFs for all other variables in the model were approximately 10 or less. VIFs greater than 10 indicate the presence of collinearity where, $(VIF) = 1/(1-R^2_i)$, $i = 1, 2, \ldots, k$ (Mendenhall & Sincich, 1989, p. 237). Values greater than 10 occurred only in those variables used both singularly and squared in the higher-order terms, making them obvious candidates for collinearity. In addition, Mendenhall and Sincich (1989) discuss the need to code the dependent, as well as the independent variables, in order to properly calculate VIFs (p. 236). The criterion variable, perceived job stress (Y), was not coded in this study. Finally, the $R^2 = .48$ representing the best model did not appear to be sufficiently large to indicate the presence of collinearity. Consistent with this finding, the standard errors of the individual beta parameters were not inflated, and the t-tests on the individual beta parameters were significant suggesting lack of evidence for collinearity (Mendenhall & Sincich, 1989, p. 236).

Discussion and Recommendations

Of several hundred studies of stress examined by the authors, it appears that none have used the regression diagnostics discussed in this paper, suggesting that the results of models presented in the literature may be weaker than necessary. Results of the present study illustrate that the use of the various regression diagnostics can improve best model fit considerably. In addition, it should be obvious that investigators cannot depend solely on regression selection options such as MAXR, FORWARD and BACKWARD when searching for the best subset regression model. Options such as MAXR will provide R^2 values for all generated models, however, the final decision as to which is the best model cannot be made without the C_p statistic and other values, for example, regression diagnostics such as HAT values, Cook's D, results of jackknifing procedures such as deleted residuals, DFFITS, DFBETAS, and the Durbin-Watson statistic which are available under the FORWARD and BACKWARD options of the PROC STEPWISE procedure.

The FORWARD and BACKWARD options offer different best models. That is, they each output best models based on the particular programmed criteria embedded in their respective routines, with the R^2 as the salient criterion. However, a strong R^2 value is not inequivocably the last word on model fit. For example, if two models with similar R^2 values are examined, it may be that the model with the slightly lower R^2 will better satisfy the other criteria discussed above and will thus be the better choice overall. Therefore, the investigator needs to utilize the power of these routines coupled with intelligent decision making regarding the various procedures. Coding variables reduces the likelihood of collinearity, and outputting regression diagnostics enables the investigator to experiment with dropping outliers and influential observations to see how their absence affects the variance accounted for by the overall model. In summary, there is nothing automatic about the process. SASPC and other packages will provide the mathematics, but it remains the responsibility of the investigator to examine the output carefully to arrive at truly the best model.

The analysis of the present study was theory-driven rather than exploratory in nature. In other words, because of the authors' preference for confirmatory modelling techniques, a limited number of interaction and higher-order terms were chosen based on the literature. However, the literature is replete with more complex models, that is, path analyses and LISREL models that attempt to model indirect effects. Thus, future analyses could be improved by studying the literature in more depth to arrive at other plausible variables and higher-order terms. Possible variables to be included in additional studies involve two general categories, that is, 1) job design facets such as autonomy, responsibility, feedback, task significance, task wholeness, leadership style; and 2) moderator variables consisting of personality characteristics and other demographics, such as Type-A, locus of control, and growth need strength. In addition, existing studies could be strengthened through replication and utilization of the regression diagnostics detailed in the present study.

References

Brief, A. P., & Aldag, R. J. (1976). Correlates of role indices. Journal of Applied Psychology, 61, 469-472.

Caplan, R. D., Cobb, S., French, J. R. P., Harrison, R. V., & Pinneau, S. R., Jr. (1980). Job demands and worker health. Ann Arbor, MI: Institute for Social Research.

French, J. R. P., & Caplan, R. D. (1972). Organizational stress and individual stress. In A. J. Marrow (Ed.), Failure of success. New York: AMACOM.

Hamner, W. C., & Tosi, H. L. (1974). Relationship of role conflict and role ambiguity to job involvement measures. Journal of Applied Psychology, 39, 497-499.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. (1964). Occupational stress: Studies in role conflict and ambiguity. New York: Wiley.

Liddell, W. W., & Slocum, J. W. (1976). The effects of individual-role compatibility upon group performance: An extension of Schutz's FIRO theory. Academy of Management Journal, 19, 413-426.

Matteson, M. T., & Ivancevich, J. M. (1987). Controlling work stress: Effective human resource and management strategies. San Francisco: Jossey-Bass.

Mendenhall, W., & Sincich, T. (1989). A second course in business statistics: Regression analysis (3rd ed.). San Francisco: Dellen Publishing Company.

SAS Institute, Inc. (1985). SAS User's Guide: Statistics (5th ed.). Cary, NC: Author.

Younger, M. S. (1985). A first course in linear regression (2nd ed.). Boston: Duxbury Press.

If you are submitting a research article other than notes or comments, I would like to suggest that you use the following format if possible:

- Title
- Author and affiliation
- Indented abstract (entire manuscript should be single spaced)
- Introduction (purpose—short review of literature, etc.)
- Method
- Results
- Discussion (conclusion)
- References

All manuscripts should be sent to the editor at the above address. (All manuscripts should be camera-ready.)

It is the policy of the M.L.R. SIG-multiple linear regression and of *Viewpoints* to consider articles for publication which deal with the theory and the application of multiple linear regression. Manuscripts should be submitted to the editor as original, double-spaced, *camera-ready copy*. Citations, tables, figures and references should conform to the guidelines published in the most recent edition of the *APA Publication Manual* with the exception that figures and tables should be put into the body of the paper. A cost of $1 per page should be sent with the submitted paper. Reprints are available to the authors from the editor. Reprints should be ordered at the time the paper is submitted, and 20 reprints will cost $.50 per page of manuscript. Prices may be adjusted as necessary in the future.

A publication of the Multiple Linear Regression Special Interest Group of the American Educational Research Association, *Viewpoints* is published primarily to facilitate communication, authorship, creativity and exchange of ideas among the members of the group and others in the field. As such, it is not sponsored by the American Educational Research Association nor necessarily bound by the association's regulations.

"Membership in the Multiple Linear Regression Special Interest Group is renewed yearly at the time of the American Educational Research Association convention. Membership dues pay for a subscription to the *Viewpoints* and are either individual at a rate of $5, or institutional (libraries and other agencies) at a rate of $18. Membership dues and subscription requests should be sent to the executive secretary of the M.L.R. SIG."

THE UNIVERSITY OF AKRON
AKRON, OH 44325-4205

TABLE OF CONTENTS

ISSN 0195-7171