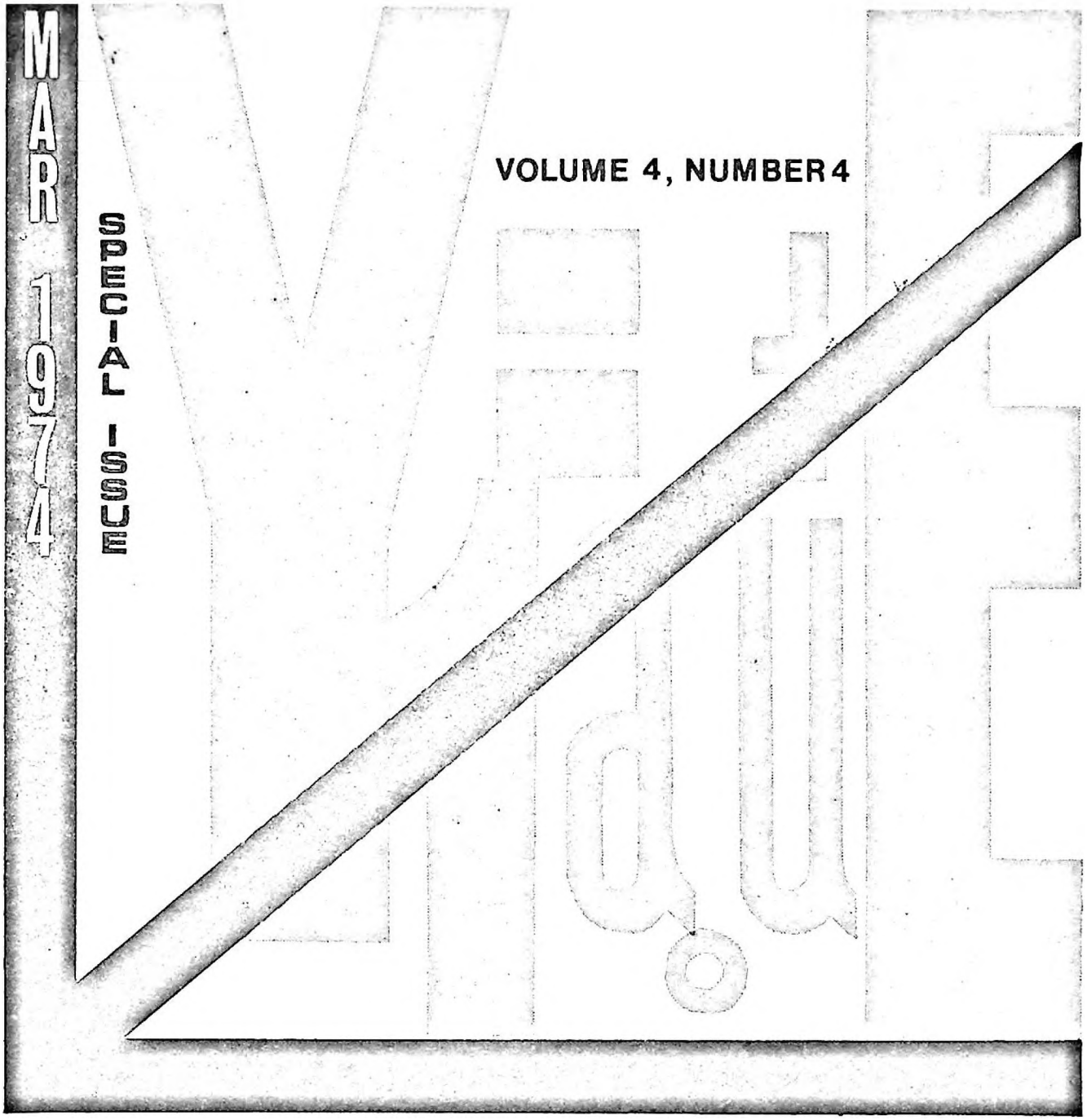
MAR 1974

SPECIAL ISSUE

VOLUME 4, NUMBER 4



MULTIPLE LINEAR REGRESSION VIEWPOINTS

A publication of the Special Interest Group
on Multiple Linear Regression
of The American Educational Research Association.

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

Title
Author and Affiliation
Single-spaced indented abstract
Introduction (purpose—short review of the literature, etc.)
Method
Results
Discussion (conclusion)
References using APA format

All manuscripts should be sent to the editor at the above address.

A C K N O W L E D G E M E N T

I would like to thank the College of Education, University of Akron, for their support in the publications of "Viewpoints" this year. I would also like to especially thank Dr. Caesar Carrino who will be assuming the position of Dean of the Evening College, for his help.

Isadore Newman, Editor
Multiple Linear Regression Viewpoints

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

A NOTE ON CONTRAST CODING VS. DUMMY CODING

John D. Williams
The University of North Dakota

Abstract--A comparison is made between the contrast coding system for solution to the analysis of variance design presented by Lewis and Mouw (1973), and the use of dummy coding for solution to the analysis of variance designs. Some of the limitations and advantages of each approach are given.

Lewis and Mouw (1973), in an earlier issue of Viewpoints, argued for the use of contrast coding for the solution to the analysis of variance (ANOVA) and the analysis of covariance (ANCOVA) designs. They list three major advantages of contrast coding: (1) the number of predictor variables in a model accurately reflects the degrees of freedom for the analysis; (2) the use of contrast coding allows one to ask more specific questions of interest than the overall main effect; and (3) the main effect in a two-way ANCOVA can be tested without pooling the error term. The inference that might be made by reading their article is that the use of "dummy coding" (i.e., 1 if a characteristic is present, 0 if it is not) inherently has these three difficulties.

Actually, dummy coding can be made to accommodate most of the concerns listed by Lewis and Mouw. In their article, Lewis and Mouw discuss four particular statistical tests reformulated in a contrast coding format: the t-test, the one-way ANOVA, the two-way ANOVA and the two-way ANCOVA. Rather than duplicate each of the statistical tests presented by Lewis and Mouw, the focus of the present note is in regard to the one-way ANOVA and two-way ANOVA.

One-Way Analysis of Variance

Lewis and Mouw consider a one-way ANOVA using a contrast coding scheme. A "dummy" coding scheme that is simpler to execute than the use of the orthogonal system is

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + e_1, \qquad (1)$$

where

Y = is the criterion variable,

$b_0 - b_3$ = regression coefficients,

X_1 = 1 if from Treatment 1; 0 otherwise,

X_2 = 1 if from Treatment 2; 0 otherwise,

X_3 = 1 if from Treatment 3; 0 otherwise, and

e_1 = the error in prediction with equation 1.

It should be noticed that Treatment 4 is apparently omitted from this scheme; actually, the coding procedure in equation 1 makes treatment 4 an intrinsic part of the solution. As is shown in Williams (1971), $b_0 = \bar{X}_4$, $b_1 = \bar{X}_1 - \bar{X}_4$, $b_2 = \bar{X}_2 - \bar{X}_4$ and $b_3 = \bar{X}_3 - \bar{X}_4$. It should be noticed that not only does the approach given by equation 1 result in identifying the correct degrees of freedom, the researcher automatically receives Dunnett's (1955) test for comparing one group to several other groups.

If Treatments 1, 2, 3 and 4 are properly viewed as representing equal units of some treatment, then the linear, quadratic and cubic trends can be measured. Three easily generated variables can be formed:

X_5 = 1 if from Treatment 1, 2 if from Treatment 2, 3 if from Treatment 3 and 4 if from Treatment 4,

$X_6 = X_5^2$, and

$X_7 = X_5^3$.

The linear, quadratic and cubic trends can be measured respectively by

$$Y = b_0 + b_5 X_5 + e_2 \quad \text{(for linear)}, \qquad (2)$$

$$Y = b_0 + b_5 X_5 + b_6 X_6 + e_3 \quad \text{(for quadratic), and} \qquad (3)$$

$$Y = b_0 + b_5 X_5 + b_6 X_6 + b_7 X_7 + e_4 \quad \text{(for cubic).} \qquad (4)$$

If the interest is in representing the accounted variance (in terms of R^2), then

R_2^2 (i.e., from equation 2), gives the linear effect,

$R_3^2 - R_2^2$ gives the quadratic effect, and

$R_4^2 - R_3^2$ gives the cubic effect.

While the test of the one-way ANOVA is specifically related to Dunnett's test, it has been shown (Williams, in press) that other multiple comparison procedures (Tukey's test, Dunn's test and Scheffé's test) can be accommodated to coding by the dummy coding approach given in equation 1.

The Two-Way Analysis of Variance

Lewis and Mouw present a contrast coding procedure for analyzing a 2 X 3 factorial design where there are two levels of the A factor and three levels of the B factor; they give the standard orthogonal coefficients for constructing their two-way model.

However, the 2 X 3 design can also be accomplished using the binary (1 or 0) coding format. The full model is given by

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + e_5, \qquad (5)$$

where

Y = is the criterion variable,

X_1 = 1 if from Row 1; 0 otherwise,

X_2 = 1 if from Column 1; 0 otherwise,

X_3 = 1 if from Column 2; 0 otherwise,

$X_4 = X_1 \cdot X_2$,

$X_5 = X_1 \cdot X_3$, and

e_5 = the error in prediction with this model.

It can be noticed that X_1 corresponds to the A effect, X_2 and X_3 correspond to the B effect and X_4 and X_5 correspond to the interaction effect. Three more models can be defined to complete the analysis:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + e_6, \quad (6)$$

$$Y = b_0 + b_1 X_1 + e_7, \text{ and} \quad (7)$$

$$Y = b_0 + b_2 X_2 + b_3 X_3 + e_8. \quad (8)$$

If the interest is in generating a summary table, $SS_A = SS_T(R_6^2 - R_8^2)$, $SS_B = SS_T(R_6^2 - R_7^2)$, $SS_{AB} = SS_T(R_5^2 - R_6^2)$ and $SS_w = (1 - R_5^2)$. It should be emphasized that this procedure will generate the two-way analysis of variance even when disproportionality occurs.

Comparison of Contrast Coding and Dummy Coding

There are logically three situations that the two coding procedures can be compared: (1) an equal number of subjects per cell; (2) a proportionate (but not always equal) number of subjects per cell; and (3) disproportionate cell frequencies.

When there are an equal number of subjects per cell, the contrast coding procedure yields a satisfying and direct solution that can, if the researcher so wishes, pinpoint the source of variation with each degree of freedom. The dummy coding procedure is somewhat more cumbersome, and for designs larger than a 2 X 3 table, no simple solution seems to exist for alloting the variation per degree of freedom for the interaction portion; the main effects can be accounted for through the use of the equations like equations 5, 6 and 7.

If the data is either proportional or disproportional, the contrast coding procedure will generally fail for the main effects portion of the analysis; in both cases, the interaction found by restricting the full model by the non-inclusion of the interaction terms does yield a correct solution. On the other hand, the model given in this paper does yield a useful solution (called the fitting constants solution) and is shown in Williams (1972).

Thus, it can be seen that the contrast coding scheme is particularly useful in the following circumstances: (1) there are an equal number of subjects in each cell; (2) there is interest in the variance due to each degree of freedom; (3) there is no interest in any comparisons among the means that are not orthogonal (such as Tukey's test or Dunnett's test); and (4) the user is sufficiently familiar with the use of orthogonal coding systems such as are described in Hays (1963). That contrast coding can be a useful addition to the repertoire of the applied statisticians is unquestioned; on the other hand, the flexibility of the use of the various dummy coding schemes should be fully exploited.

References

Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 1955, 50, 1096-1121.

Hays, W. L. Statistics for Psychologists. New York: Holt, Rinehart and Winston, 1963.

Lewis, E. L., and Mouw, J. T. The use of contrast coding to simplify ANOVA and ANCOVA procedures in multiple linear regression. Multiple Linear Regression Viewpoints, 1973, 2, 27-44.

Williams, J. D. A multiple regression approach to multiple comparisons for comparing several treatments to a control. Journal of Experimental Education, 1971, Spring, 93-96.

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

Williams, J. D. A simplified regression formulation of Tukey's test. Journal of Experimental Education, (in press).

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

Estimated Parameters of Three Shrinkage Estimate Formuli

Micheal Klein
and
Isadore Newman

Abstract

This paper examines the shrinkage formuli of Wherry, McNemar and Lord in relation to overcorrection. A table is given which shows the number of times that each formula resulted in a negative value of $\bar{R}^2$ and the lowest R^2 which produced a positive R^2 for different numbers of variables and sample sizes.

In an earlier paper, Newman (1973) discussed three shrinkage estimation formuli.

$$\hat{R}^2 = 1 - (1-R^2)\frac{N-1}{N-K} \qquad \text{(Wherry)}$$

$$\hat{R}^2 = 1 - (1-R^2)\frac{N-1}{N-K-1} \qquad \text{(McNemar)}$$

$$\hat{R}^2 = 1 - (1-R^2)\frac{N+K+1}{N-K-1} \qquad \text{(Lord)}.$$

where:

$\hat{R}$ = the corrected estimate of the multiple correlation.

R = the actual calculated multiple correlation.

K = the number of independent variables.

N = the number of independent observations.

A study was run to determine what estimates would be given by the three formuli, when the three were used with 2, 3, 5, or 8 variables and with each set having a N equal to 10, 20, 50, 100 and 200 (see Table 1).

In Case 1 where there were two variables and ten replicates, Wherry's formula produced four negative R^2 estimates, McNemar's formula produced seven negative R^2 estimates and Lord's formula produced 14 negative R^2 estimates. The lowest R^2 that would not produce a negative number when entered into Wherry's shrinkage estimate was .1333, for McNemar's formula it was .2333 and for the Lord's formula it was .4666 (see Table 1). As Case 1 indicates, where you have two variables and ten subjects, Lord's

Table 1

Estimated R^2's from Wherry, McNemar and Lord Shrinkage Formuli

Number of Variables (K)	Number of Subjects (N)				
	10	20	50	100	200
	4, 7, 14	2, 4, 8	1, 2, 4	1, 1, 2	1, 1, 1
	.1333	.0666	.0333	.0333	.0333
	.2333	.1333	.0666	.0333	.0333
	.4666	.2666	.1333	.0666	.0333
	(Case #1)	(Case #2)	(Case #3)	(Case #4)	(Case #5)
	7, 11, 18	4, 5, 11	2, 2, 5	1, 1, 3	1, 1, 2
	.2333	.1333	.0666	.0333	.0333
	.3666	.1667	.0666	.0333	.0333
	.6000	.3667	.1666	.1000	.0666
	(Case #6)	(Case #7)	(Case #8)	(Case #9)	(Case #10)
	14, 17, 28	7, 8, 14	3, 4, 7	2, 2, 4	1, 1, 2
	.4666	.2333	.1000	.0666	.0333
	.5666	.2666	.1333	.0666	.0333
	.7666	.4666	.2333	.1333	.0666
	(Case #11)	(Case #12)	(Case #13)	(Case #14)	(Case #15)
	24, 27, 29	12, 13, 19	5, 5, 10	3, 3, 5	2, 2, 3
	.8000	.4000	.1666	.1000	.0666
	.9000	.4333	.1666	.1000	.0666
	.9667	.6333	.3333	.1666	.1000
	(Case #16)	(Case #17)	(Case #18)	(Case #19)	(Case #20)

Note:

n_1, n_2, n_3
R_1^2
R_2^2
R_3^2

n_1, n_2, n_3 = number of terms <0 for Wherry, McNemar, Lord respectively.

R_1^2, R_2^2, R_3^2 = next R >0 for Wherry, McNemar, and Lord respectively.

The procedure used in this study was simply to generate values of R^2 from 0 to 1 in steps of 0.0333. For each step the three shrinkage formulae were applied and the resulting $\hat{R}^2$ tabulated in Table 1.

formula tends to overshrink twice as much as McNemar's and over three times as much as Wherry's. This is assuming that a negative R^2 is due to an overestimation of the shrinkage.

In Case 20, where there are eight variables and 200 subjects, Wherry's and McNemar's formula both produced two negative R^2 and Lord's formula produced three negative R^2 (see Table 1). The lowest R^2 that would not produce a negative shrinkage estimate for Wherry's and McNemar's formula was .0666 and for Lord's formula it was .1000 (see Table 1).

Based on the results of the study presented in Table 1, it appears that when there are 100 subjects for each variable (Case 5) all three formuli produce the same estimates. When the ratio is less than that, Lord's formula is consistently more conservative, that is, it shrinks more. As the variables increase, there seems to be a tendency for McNemar and Wherry to produce more similar results.

Since it is conceptually meaningless to interpret negative R^2, and since the lowest possible R^2 one can legitimately obtain is 0, it seems that these formuli need a correction factor added so that they are bounded on the low end by $R^2 = 0.0$ and on the high end by $R^2 = 1.0$. It is therefore suggested that if one uses any of these three shrinkage estimates that any negative R^2 be interpreted as if it were $R^2 = 0$.

This study was performed simply to obtain an idea of the range of the parameters as they relate to these three shrinkage estimates. It was not intended to be definitive. We would like to suggest that further research needs to be conducted to obtain the following information:

1. Mathematically determine and empirically test a correction for the limits.

2. Study the relationship between these three estimates and Type I and Type II errors as a function of the number of subjects and number of variables.

3. Further investigate the accuracy of these three formuli in predicting from a sample to a population and from one sample to another as a function of the number of variables and number of subjects.

We believe shrinkage estimates are extremely important to consider when one is dealing with multiple regression since they are more likely to improve our ability to accurately generalize the estimated relationships of the studies being done. However, we can do this more accurately if we know more about the parameters of these formuli.

Reference

Kelley, F. J., Beggs, D., McNeil, A.K., Eichelberger, and Lyon. Research Design in the Behavioral Sciences: Multiple Regression Approach. Carbondale, Ill. Southern Illinois Univ. Press, 1969.

Lord, F. M. "Efficiency of Prediction When A Regression Equation From One Sample Is Used In A New Sample." Research Bulletin, 50, 40. Princeton, N.J: Educational Testing Service, 1950.

McNewmar, Q. Psychological Statistics (3rd. ed.) New York: Wiley & Sons, 1962

Newman, Isadore. "Variations Between Shrinkage Estimation Formula And The Appropriatness of Their Interpretation." Multiple Linear Regression Viewpoints. Vol. 4, 2, 45-48, Aug. 1973.

Nunnally, J. Psychometric Theory. New York: McGraw-Hill Book Co., 1967

Uhl, N. and Eisenberg, T. "Predicting Shrinkage In The Multiple Correlation Coefficient." Educational and Psychological Measurement. 30, 487-489, 1970.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

Complexity in Behavioral Research, as Viewed Within the Multiple Linear Regression Approach

Keith McNeil
Educational Monitoring Systems
3449 Rentz Road
Ann Arbor, Michigan

Michael McShane
Southern Illinois University
Carbondale, Illinois

The present paper will attempt to clarify the notion of complexity in research. Perhaps, in the past, researchers have over simplified the variables and their interrelationships. Thinking of interaction variance as bad variance or as producing an undesired result, or lamenting the complexity of the phenomena under consideration does little or nothing to advance research in the behavioral sciences. The present paper will examine two views of complexity which seem to exist: (1) complexity as indicated by the number of predictor variables needed to account for a criterion behavior, and (2) complexity as indicated by the nature of the predictor variables.

The remaining discussion will attempt to demonstrate that the second view of complexity is not valid, and that the multiple linear regression technique provides an easy way to index the first view.

Complexity of the Variables Themselves

One view of complexity in research seems to be concerned with how complex the variables themselves are. For example, interaction and polynomial variables are veiwed by most researchers as more complex than linear terms. Complex terms such as these are often willfully omitted by researchers. When such terms are found to be significant, researchers often shy away from interpreting those "complex variables". Perhaps these

variables are viewed as more complex because: (1) researchers are not as familiar with them, (2) additional effort must be put forth to obtain these variables, or (3) the variables reflect a different state of affairs than the usually-investigated, not-too-realistic linear relationship.

Unfamiliarity should not be grounds for complexity. Nor should additional effort. Interaction variables and polynomial variables require the multiplication of one variable times another. The important point is that a single number results, and therein lies the basis for the argument that such variables are no more complex than are the originally scored variables. For a given polynomial or interaction variable, only one weighting coefficient needs to be calculated from the data.

That a state of affairs existing in the data is different than the "expected" one should not be grounds for complexity either. Researchers have investigated linear relationships so often and for so long that often they forget that that is the relationship being investigated, or that other relationships could in fact exist in the data. Pearson correlation only investigates the linear relationship. When specific trends are investigated in ANOVA, most textbook authors support the notion that the linear trend should be looked at more so than the other, "more complex trends". It is the position of the present authors that more complex models should be developed to represent those more complex trends.

Complexity of the Predictive System

Each predictor variable in a multiple linear regression analysis has an associated weighting coefficient. Since these weighting coefficients are obtained from the sample data, any test of significance must take into account how many weighting coefficients were determined. In terms of the "Goals of Research" (McNeil, 1970) researchers are striving for models which produce high R^2 values (the Goal of Predictability) while utilizing

score. When this is the case, researchers will come to better "understand" IQ-squared since it maps the construct which is of interest better than does the variable IQ.

We often forget that every variable in the behavioral sciences is arbitrarily scaled. Until we can see inside of our subjects, we only have approximations of the variables. The initial arbitrary way of measuring a variable often takes on a presence of finality of interpretability. This can be unfortunate, as in our hypothetical case with IQ. While using IQ in several situations over a number of years we have developed some expectations about that variable. These expectations should remain flexible, especially when the variable is introduced as a predictor of a new criterion. If IQ-squared is more predictive of a new criterion than is IQ, then it is ridiculous to retain the IQ measure.

Now suppose that we had initially developed the IQ-squared measure. Since we did not realize that it was the square of what other people were calling IQ, we called it QI. In investigating the "new criterion", we found QI to be quite predictive, but when the "original criterion" is investigated we find that the square root of QI is more predictive than is QI. This example should indicate the arbitrariness of the scaling of measures.

Summary

The thesis of this paper is that the time and effort of computing each predictor variable should not determine the complexity of the resulting model. Complexity in the behavioral sciences should be viewed solely as the number of predictor variables needed to satisfactorily account for criterion variance.

a small number of predictor variables (the Goal of Parsimony). Since some models use more predictor information than others, and since some models yield higher R^2 than others, the differing predictability must be evaluated in terms of the number of predictor variables used. Indeed, this is the rationale of the general F test:

$$F(m_1 - m_2, N - m_1) = \frac{(R^2_f - R^2_r)/(m_1 - m_2)}{(1 - R^2_f)/(N - m_1)}$$

Where:

R^2_f = the proportion of variance accounted for in the full model,

R^2_r = the proportion of variance accounted for in the restricted model,

m_1 = the number of linearly independent pieces of information in the full model, and

m_2 = the number of linearly independent pieces of information in the restricted model.

A Discussion of the Complex Variable of IQ

Most researchers would consider an IQ score to be a single, simple variable, although measuring possibly a very complex phenomenon. In reality, an IQ score is defined as an interaction between Mental (MA) and the reciprocal of Chronological Age (CA). Whether the score is represented as IQ or as MA/CA should make no difference as to how complex we think the score is. Furthermore, we usually do not take into consideration the number of items that go together to yield the mental age. Some IQ measures have more items and take more time to administer, but all yield, in the final (data) analysis, one number for each person.

In addition, it may be that in a given situation, a variable of IQ-squared may be more predictive of the criterion than is the original IQ

Modification of Multiple Regression when an Independent Variable is Subtracted from the Dependent Variable

Grace Wyshak, Ph.D.
Yale University, Department of Epidemiology and Public Health

Behavioral scientists are often concerned with regressing some dependent or outcome variable, Y, on a number of independent or explanatory variables, X_i, i = 1,2,...,k. (Model 1). If Y is a final score or measurement and X_1 an initial score, the investigator may be interested in some measure of change, say $Y - X_1$, and its relation to the several explanatory variables including X_1. (Model 2). Analyses would be based on two multiple regression equations, one relating to the regression of Y on $X_1, X_2, \ldots, X_k$; and the other to the regression of $(Y - X_1)$ on the same X's.

In this note we call attention to the fact that one analysis would suffice for the two models because the regression coefficients, the total sums of squares, deviations sums of squares and regression sums of squares are readily obtained for Model 2 once the calculations have been made under Model 1.

The relation between the regression coefficients is as follows:

$$b_1' = b_1 - 1$$

$$b_i' = b_i \qquad i = 2,3,\ldots,k$$

where b_i denotes the coefficients under Model 1 and b_i' under Model 2.

Assume we have n observations of the form $Y, X_1, X_2, \ldots, X_k$.

Let $\sum_{\nu=1}^{n} (X_{i\nu} - \bar{X}_{i.})(X_{j\nu} - \bar{X}_{j.})$ by $\sum x_i x_j$

$\sum_{\nu=1}^{n} (X_{i\nu} - \bar{X}_{i.})(Y_{\nu} - \bar{Y}.)$ by $\sum x_i y$

$\sum_{\nu=1}^{n} (Y_{\nu} - \bar{Y}.)^2$ by $\sum y^2$ where $i,j = 1,2,\ldots,k$.

To find the least squares estimates for $(b_1, b_2, \ldots, b_k)$, minimize

$$S_1 = \sum (y - \beta_1 X_1 - \ldots - \beta_k X_k)^2 \qquad \text{Model 1.}$$

Under Model 2 we have $(Y-X_1)$, $X_1, X_2, \ldots X_k$. To find the least squares estimate for $(b_1', b_2', \ldots, b_k')$, minimize

$$S_2 = \sum (y - x_1 - \beta_1' X_1 - \ldots - \beta_k' X_k)^2$$

$$= \sum (y - (\beta_1' + 1) x_1 - \ldots - \beta_k' X_k)^2$$

The minimum of S_2 is the same as the minimum of S_1 where $b_1 = b_1' + 1$, and $b_i = b_i'$ for all $i \neq 1$.

The relations between the sums of squares can be shown using matrix notation.

Let $X'X = \begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 & \cdots & \sum x_1 x_k \\ \sum x_2 x_1 & \sum x_2^2 & \cdots & \sum x_2 x_k \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sum x_k x_1 & \sum x_k x_2 & & \sum x_k^2 \end{bmatrix}$

and $(X'X)^{-1} = C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ c_{k1} & \cdot & \cdots & c_{kk} \end{bmatrix}$ $G = \begin{bmatrix} \sum x_1 y \\ \sum x_2 y \\ \cdot \\ \cdot \\ \sum x_k y \end{bmatrix}$ $B = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_k \end{bmatrix}$

Then B = C G, and C(X'X) = I, where I is the identity matrix. (Snedecor and Cochran, 1967).

The relations between the regression coefficients, b_i and b_i' can be derived using matrix notation although they were shown more simply above.

Under Model 2, let Z = Y - X_1. Then, $\Sigma x_i z = \Sigma x_i y - \Sigma x_1 x_i$.

Thus,

$$\begin{bmatrix} b_1' \\ b_2' \\ \vdots \\ b_k' \end{bmatrix} = C \cdot \begin{bmatrix} \Sigma x_1 y - \Sigma x_1^2 \\ \Sigma x_2 y - \Sigma x_1 x_2 \\ \vdots \\ \Sigma x_k y - \Sigma x_1 x_k \end{bmatrix} = \begin{bmatrix} b_1 - 1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$$

The last step follows;

$$b_1' = c_{11}\left[\Sigma x_1 y - \Sigma x_1^2\right] + c_{12}\left[\Sigma x_2 y - \Sigma x_1 x_2\right] + \ldots + c_{1k}\left[\Sigma x_k y - \Sigma x_1 x_k\right]$$

$$b_i' = c_{i1}\left[\Sigma x_1 y - \Sigma x_1^2\right] + c_{i2}\left[\Sigma x_2 y - \Sigma x_1 x_2\right] + \ldots + c_{ik}\left[\Sigma x_k y - \Sigma x_1 x_k\right]$$

Rearranging terms, $b_1' = \left[c_{11}\Sigma x_1 y + c_{12}\Sigma x_2 y + \ldots + c_{1k}\Sigma x_k y\right]$
$- \left[c_{11}\Sigma x_1^2 + c_{12}\Sigma x_1 x_2 + \ldots + c_{1k}\Sigma x_1 x_k\right]$

The first term in square brackets is equal to b_1 under Model 1; the second term is equal to 1, being equal to a diagonal term of the identity matrix resulting from the multiplication C(X'X). Thus, $b_1' = b_1 - 1$.

$b_i' = \left[c_{i1}\Sigma x_1 y + c_{i2}\Sigma x_2 y + \ldots + c_{ik}\Sigma x_k y\right]$
$- \left[c_{i1}\Sigma x_1^2 + c_{i2}\Sigma x_1 x_2 + \ldots + c_{ik}\Sigma x_1 x_k\right]$ $\quad i = 2,3,\ldots,k.$

The first term in square brackets is equal to b_i of Model 1; the second is equal to zero, being equal to an off-diagonal term of the identiy matrix resulting from the multiplication C(X'X). Thus, $b_i' = b_i$, $i \neq 1$.

II The relations between the sums of squares follows.

When Y is regressed on $X_1, X_2 \ldots, X_k$, Model 1,

Regression S.S. = $\Sigma b_i \Sigma x_i y$;

Total S.S. $= \sum y^2$;

Deviation SS = Total S.S. - Regression S.S.

$$= \sum y^2 - \sum_{i=1}^{k} b_i \sum x_i y$$

Under Model 2, $Z = Y - X_1$ is regressed on $X_1, X_2, \ldots, X_k$.

Regression S.S. $= (b_1 - 1)\left(\sum x_1 y - \sum x_1^2\right) + \sum_{j=2}^{k} b_j \left(\sum x_j y - \sum x_1 x_j\right)$

$$= \sum_{i=1}^{k} b_i \sum x_i y - \sum_{i=1}^{k} b_i \sum x_1 x_i - \sum x_1 y + \sum x_1^2.$$

Total S.S. $= \sum y^2 + \sum x_1^2 - 2\sum x_1 y.$

Deviation S.S. is obtained by subtraction.

Now, it can be shown that

$$\sum_{i=1}^{k} b_i \sum x_1 x_i = \sum x_1 y$$

Hence, Regression S.S. is

$$\sum_{i=1}^{k} b_i \sum x_i y + \sum x_1^2 - 2\sum x_1 y.$$

Deviation S.S. $= \sum y^2 - \sum_{i=1}^{k} b_i \sum x_i y$, and is unchanged under Model 2.

The difference between the Total S.S. and the Regression S.S. is the same quantity, being $\sum x_1^2 - 2\sum x_1 y$, when the results obtained under Model 1 are subtracted from those obtained under Model 2

We now show that $\sum_{i=1}^{k} b_i \sum x_1 x_i = \sum x_1 y.$

$$\sum_{i=1}^{k} b_i \sum x_1 x_i = b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \ldots + b_k \sum x_1 x_k$$

$$= \left[c_{11} \sum x_1 y + c_{12} \sum x_2 y + \ldots + c_{1k} \sum x_k y\right] \sum x_1^2$$

$$+ \left[c_{21} \sum x_1 y + c_{22} \sum x_2 y + \ldots + c_{2k} \sum x_k y\right] \sum x_1 x_2$$

$$+ \quad . \qquad . \qquad . \qquad .$$

$$+ \quad . \qquad . \qquad . \qquad .$$

$$+ \left[c_{k1} \sum x_1 y + c_{k2} \sum x_2 y + \ldots + c_{kk} \sum x_k y\right] \sum x_1 x_k$$

$$= \left[c_{11} \sum x_1^2 + c_{21} \sum x_1 x_2 + \ldots + c_{k1} \sum x_1 x_k\right] \sum x_1 y$$

$$+ \left[c_{12} \sum x_1^2 + c_{22} \sum x_1 x_2 + \ldots + c_{k2} \sum x_1 x_k\right] \sum x_2 y$$

$$+ \quad . \qquad . \qquad . \qquad .$$

$$+ \quad . \qquad . \qquad . \qquad .$$

$$+ \left[c_{1k} \sum x_1^2 + c_{2k} \sum x_1 x_2 + \ldots + c_{kk} \sum x_1 x_k\right] \sum x_k y$$

On the right hand side of the equation the first term in square brackets is equal to unity; the other terms in square brackets are equal to zero.

Thus,

$$\sum_{i=1}^{k} b_i \sum x_1 x_i = \sum x_1 y.$$

References

Snedecor, G.W. and Cochran, W.G.: Statistical Methods, Sixth Edition, 1967, Ames, Iowa, The Iowa State University Press.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

Robert G. St.Pierre
224 Newtonville Avenue
Newton, Massachusetts 02158

VARGEN: A Multiple Regression Teaching Program

Robert G. St.Pierre
BOSTON COLLEGE

VARGEN (Variable Generator) creates sets of data with known statistical properties by generating user-specified variables which are functions of uniformly distributed random numbers for each of a group of subjects. The user specifies the relative size and location of from one to nine predictor variables and one criterion variable within a ten by ten matrix (hereafter called the "universe") and, therefore, the amount of variance accounted for by each variable. A visual display is produced showing the size and location within the universe of any five variables.

While a multiple correlation and regression routine (Cooley and Lohnes, 1971) has been built into VARGEN to help the student explore the relationships between the generated variables, the program has the option to punch the generated data on cards permitting its input to any data analysis program. Since the user has specified the relationships between variables, many statistical characteristics of the data are known, and certain results can be expected from an analysis enabling the data to be used as part of a learning process.

Effects of adding or deleting variables and varying relationships between variables can be observed. By changing the amount of overlap between variables the user changes the amount of variance that the variables have in common, and thus changes the regression analysis. Effects of varying sample size can also be investigated in terms of stability of the data. Will replications using sample sizes of 100 give the same results - or do we need 1,000 cases to get stable answers? Given the correlations between variables, the variable means and the regression analysis, the student can attempt to construct a picture (perhaps a Venn diagram) showing the relationships between the variables and then compare that structure with the visual display.

VARGEN will create up to ten variables (nine predictors and one criterion) for each of up to 9,999 subjects in the following manner. For each subject, 100 random numbers are drawn from a rectangular distribution which ranges from zero to one. The random numbers are placed in the cells of the universe, with the numbers contained in the cells that are covered by each variable being summed. A variable is thus defined as the sum of equally potent, equally likely, independent elements, which are either present or absent (McNemar, 1969; Garrett, 1946). For example, if a variable is defined as being eight rows by four columns and located at location (2,4) within universe, the value of that variable is computed by:

$$\sum_{i=2}^{9} \sum_{j=4}^{7} a_{ij}$$

where a_{ij} is the random number in the i^{th} row and the j^{th} column of the universe.

Since the expected value of a number drawn from a rectangular distribution with a range of zero to one is .5, and the variable being discussed is defined as the sum of (8)X(4)=32 random numbers, the expected value of the variable is (.5)X(32)=16.

Once each variable has been calculated for the first subject, a new set of random numbers is generated and the value of each variable calculated for the second subject. Similarly, data for up to 9,999 subjects can be generated.

A visual display of any five variables will be produced showing the size and location of each variable within the universe. While it would be convenient to have the capability of displaying ten variables, the resulting visual display is extremely cluttered. It is felt that while a maximum display of five variables may be a hindrance to some, the readability of the display more than makes up for the missing variables. The user can either draw in any remaining variables by hand, or, as might be the case in a learning situation, confine the experimentation to problems involving five or less variables.

The main purpose of VARGEN is as an educational tool to help students deepen their understanding of regression, correlation and the statistical properties of data.

How To Use VARGEN

A. Description of Input

VARGEN requires two input cards. The first card specifies the number of variables, the number of subjects, and which variables will be displayed. Up to five variables may be displayed with the following stipulations:

1) the criterion must be displayed;
2) the criterion must be the last variable displayed.

Card 1

Col. 1-2	Number of variables (maximum of 10)
Col. 3-6	Number of subjects (maximum of 9,999)
Col. 10-11	The number of the first variable to be displayed
Col. 12-13	The number of the second variable to be displayed
Col. 14-15	The number of the third variable to be displayed
Col. 16-17	The number of the fourth variable to be displayed
Col. 18-19	The number of the fifth variable to be displayed
Col. 20	1 to have generated data output to logical unit 7
	0 otherwise
Col. 25-29	Random number generator initialization number (Must be an odd number)
Col. 30	1 to have generated data listed
	0 to suppress listing

The numbers of each variable to be displayed refer to the order in which the variables are defined in Card 2. The user may

display as few as one, or as many as five variables. In any case, please note that the criterion must be displayed and must be the last variable displayed.

The second input card defines the size and location of each variable. The size of a variable is determined by the number of rows and columns of the universe that are covered by the variable. The location is determined by the upper left hand coordinates of the variable within the universe.

Card 2

Col. 1-2 Number of rows in the first variable

Col. 3-4 Number of columns in the first variable

Col. 5-6 Upper left hand row coordinate of the first variable

Col. 7-8 Upper left hand column coordinate of the first variable

Repeat the format of columns 1-8 for each variable to be defined. Again note that the criterion must be the last variable defined. Eight columns are used to define the size and location of each variable, giving the user room for the maximum of ten variables on one card.

As an example of input card setup and use of the visual display suppose we want to create a data set with the following characteristics:

a) five variables (four predictors plus a criterion)

b) one hundred subjects

c) display all variables

d) do not output data to logical unit 7

e) random number generator initialization number = 72347

f) list the generated data

g) variable 1 is a 3X3 predictor located at (3,3)
variable 2 is a 3X3 predictor located at (3,6)
variable 3 is a 3X3 predictor located at (6,3)
variable 4 is a 3X3 predictor located at (6,6)
variable 5 is an 8X4 criterion located at (2,4)

Such a data set will be created by the following input cards.

```
          Column -
          1         10         20         30
          |         |          |          |
Card 1    050100    0102030405      723471

Card 2    03030303030303060303060303030606080402 04
```

Card 1 calls for five variables on 100 subjects and displays all variables. Note that the numbers of the variables to be displayed, 01, ..., 05, refer to the order in which they are defined on Card 2. See Figure 1 for the actual visual display.

Card 2 will define the size and location of each of the five variables called for on Card 1.

VARGEN will create a data set as defined by these two cards, generate a visual display, and perform a regression analysis. More than one problem may be entered by simply adding on additional sets of input cards.

B. Description of Output

1) Number of variables and number of subjects

2) Size and location of each variable, and whether it is a criterion or predictor.

VISUAL DISPLAY OF UNIVERSE

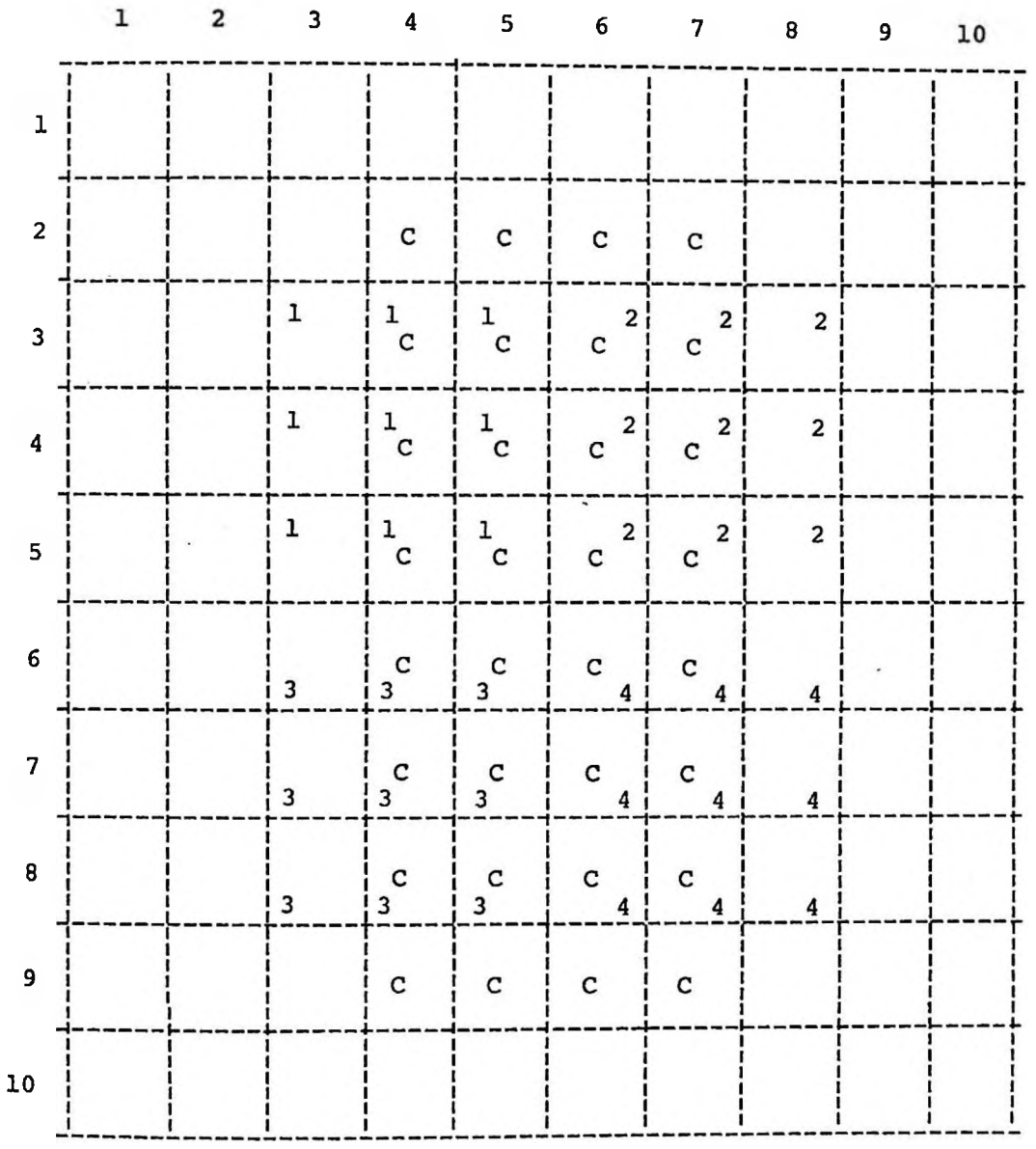


Figure 1

3) Visual display of any five variables. Gives a representation of the universe and shows the area covered by each of the specified variables and the criterion.

4) Listing of the data created (optional)

5) Output of data created to logical unit 7 (optional)

6) The mean and standard deviation of each variable

7) Correlation matrix

8) Predictor Inverse

9) Multiple regression analysis

10) Formulas used in regression analysis

C. Language

VARGEN is written in IBM FORTRAN IV G and requires the subroutine RANDU from the IBM Scientific Subroutine Package. It was developed on an IBM 370 and has been successfully adapted to a Systems Engineering Laboratory 7200. VARGEN can be adapted to other computers which have FORTRAN IV G compilers, but the random number generating subroutine must be specific to the word length of the particular computer.

D. Availability

A users manual, program listing, sample output and source deck may be obtained from Robert G. St.Pierre, 224 Newtonville Avenue, Newton, Massachusetts 02158.

References

Cooley, W. and Lohnes, P. Multivariate Data Analysis. New York: Wiley, 1971.

Garrett, H. Statistics in Psychology and Education. New York: Longmans, Green & Company, 1946.

McNemar, Q. Psychological Statistics. New York: Wiley, 1969.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

REGRESSION COMPUTER PROGRAMS FOR SETWISE REGRESSION AND THREE RELATED ANALYSIS OF VARIANCE TECHNIQUES

John D. Williams and Alfred C. Lindem
The University of North Dakota

A B S T R A C T

Four computer programs using the general purpose multiple linear regression program have been developed. Setwise regression analysis is a stepwise procedure for sets of variables; there will be as many steps as there are sets. COVARMLT allows a solution to the analysis of covariance design with multiple covariates. A third program has three solutions to the two-way disproportionate analysis of variance: (a) the method of fitting constants, (b) the hierarchical model and (c) the unadjusted main effects solution. The fourth program yields three solutions to the two-way analysis of covariance, with or without proportionality, and with multiple covariates. The three solutions are similar to those described for a two-way analysis of variance with disproportionate cell frequencies.

Four different specialized programs have been developed from the utilization of a general purpose multiple linear regression program. The programs that have been developed by these authors are described, together with an indication of the program availability and a description of the statistical technique.

Setwise Regression Analysis

Setwise regression analysis is a technique which was developed (Williams and Lindem, 1971a) to allow a stepwise solution when the interest is in sets of variables rather than in single variables. Thus, the setwise regression procedure bears a strong resemblance to the stepwise regression analysis, and a disadvantage of the stepwise procedure is overcome.

The usual stepwise procedure becomes inappropriate when there are more than two categories being binary coded. A simple example can be made with religious affiliation. Four categories might be used: Catholic, Protestant, Jewish, and other. Three binary predictors can be made with the first three

religious affiliations, and the fourth category can be represented as not having membership in the first three categories. If religious affiliation were used in conjunction with other information, the stepwise procedure would not yield a valid indication of the importance of the set of religious variables. The setwise procedure, on the other hand, would allow a direct approach to such a situation.

The setwise procedure drops one set of variables at a time in a stepwise fashion. There will be as many steps as there are sets. The solution is accomplished by an iterative procedure that allows the R^2 (multiple correlation coefficient squared) term to be maximized at each step in a backward stepwise manner. Once a set is discarded, the set is no longer considered at later steps. One set is discarded at each step, until there is only one set remaining.

As a recent issue of VIEWPOINTS has included a complete solution to a setwise problem (Williams, 1973), an example is omitted here. The documentation for the setwise program is given in Williams and Lindem (1971b).

Analysis of Covariance with Multiple Covariates (COVARMLT)

Analysis of covariance programs are typically available, but many of these programs severely limit the number of covariates, usually to one or two covariates. This limitation is wholly unnecessary. The analysis of covariance can be conceptualized as being completed through the use of two linear models, and a multiple linear regression solution follows in a straight-forward manner.

It is helpful to look at the process of the analysis of covariance as it can be generated through the use of linear models. Before the linear models are developed it is useful to set forth a concrete example. Suppose 15 students are split into three groups of five students each and are assigned to three different methods of learning beginning typewriting. Prior to

beginning the instructional period, the students are given an intelligence test and a test of manual dexterity. After the conclusion of the experiment a timed typing test is given. Table 1 contains the information for this analysis.

TABLE 1

ANALYSIS OF COVARIANCE WITH TWO COVARIATES

Post-Test	Intelligence Score	Manual Dexterity	Group 1 = 1 0 otherwise	Group 2 = 1 0 otherwise
35	120	38	1	0
27	98	28	1	0
32	102	32	1	0
29	106	22	1	0
27	94	30	1	0
38	123	43	0	1
25	96	31	0	1
36	108	46	0	1
35	115	40	0	1
31	128	35	0	1
27	90	27	0	0
35	110	31	0	0
19	94	25	0	0
17	95	24	0	0
32	116	33	0	0

Table 1 is constructed so that it might be easily transferred to IBM cards for a solution through the use of multiple regression. The group identifiers are binary coded and are found in columns 4 and 5. The group 1 identifier is given by a 1 in column 4, and the group identifier for group 2 is given by a 1 in column 5. A member of group 3 can be identified by having a 0 in both columns 4 and 5. (If there are k groups, then there will be k-1 binary predictors for the group identifiers.)

To accomplish an analysis of covariance by regression it is first necessary to construct a full model. A full model is essentially a model that contains all the information relevant to a data analysis. The full model for the present situation is:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + e_1, \qquad (1)$$

where

Y = the post-test score,

X_1 = the intelligence test score,

X_2 = the manual dexterity score,

X_3 = 1 if the score is from a member of group 1, 0 otherwise,

X_4 = 1 if the score is from a member of group 2, 0 otherwise,

b_0 = the Y-intercept,

$b_1 - b_4$ = the regression coefficients for $X_1 - X_4$, and

e_1 = the error in prediction with the full model.

If this model is solved using a multiple linear regression routine, part of the output will include the multiple correlation coefficient (R). For the present usage, since a full model is being used, the R value found from the use of equation 1 can be labeled R_{FM}.

A restricted model can be developed using only the covariates as predictor variables:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e_2, \qquad (2)$$

where

Y = the post-test score,

X_1 = the intelligence test score,

X_2 = the manual dexterity score,

b_0 = the Y-intercept (this b_0 value will, in general, be different than the b_0 value from equation 1),

$b_1 - b_2$ = the regression coefficients for X_1 and X_2 (these regression coefficients will, in general, be different from the b_1 and b_2 values in equation 1), and

e_2 = the error in prediction with the restricted model.

The restricted model also yields an R value, and it can be labeled R_{RM}.

The F test for the analysis of covariance is given by:

$$F = \frac{(R^2_{FM} - R^2_{RM})/(k-1)}{(1 - R^2_{FM})/(N-C-k)}, \qquad (3)$$

where

k is the number of groups,

N is the number of subjects, and

C is the number of covariates.

Using the full model, an R_{FM} value of .88021 is found. Then, $R^2_{FM} = .77478$. For the restricted model, $R_{RM} = .83961$, so that $R^2_{RM} = .70495$.

Using equation 3,

$$F = \frac{(.77478 - .70495)/2}{(1 - .77478)/(15-3-2)} = 1.55.$$

This F value can be interpreted in the usual way with degrees of freedom equal to 2 and 10.

Finding the Adjusted Means

For two covariates the adjusted mean can be found for each group using equation 4:

$$\overline{Y}_k(adj) = \overline{Y}_k - b_1(\overline{X}_{1k} - \overline{X}_{1T}) - b_2(\overline{X}_{2k} - \overline{X}_{2T}), \qquad (4)$$

where

$\overline{Y}_k(adj)$ = the adjusted criterion mean of the k^{th} group,

$\overline{Y}_k$ = the criterion mean of the k^{th} group,

b_1 = the regression coefficient for the first covariate in the full model,

$\overline{X}_{1k}$ = the overall mean on the first covariate,

b_2 = the regression coefficient for the second covariate in the full model,

$\overline{X}_{2k}$ = the mean of the k^{th} group on the second covariate, and

$\overline{X}_{2T}$ = the overall mean of the second covariate.

Additional covariates can be added with no difficulty in an analogous manner. For the present data, $\overline{Y}_1 = 30$, $\overline{Y}_2 = 33$, $\overline{Y}_3 = 26$, $\overline{X}_{11} = 104$, $\overline{X}_{12} = 114$, $\overline{X}_{13} = 101$, $\overline{X}_{1T} = 106.33$, $\overline{X}_{21} = 30$, $\overline{X}_{22} = 39$, $\overline{X}_{23} = 28$, and $\overline{X}_{2T} = 32.33$.

Also, $b_1 = .19514$ and $b_2 = .63027$ (their values are found directly from the printout for the full model).

$$Y_1(adj) = 30 - (.19514)\ [104 - 106.33] - (.63027)\ [30 - 32.33] = 31.92.$$

$$Y_2(adj) = 33 - (.19514)\ [114 - 106.33] - (.63027)\ [39 - 32.33] = 27.30.$$

$$Y_3(adj) = 26 - (.19514)\ [101 - 106.33] - (.63027)\ [28 - 32.33] = 29.77.$$

The process of adjusting the means can be seen as a way to "control" to some extent the difference on the covariates.

Forming a Summary Table

Forming a summary table for the analysis of covariance when using a regression approach is a relatively straight-forward process. The sum of squares within is found directly from the printout from the full model and is 118.32. The adjusted sum of squares total is given by $SS_T(adj) = SS_T(1 - R^2_{RM})$ where R_{RM} is the multiple correlation between Y and the covariates (the restricted model) which, in the present case, is $R_{RM} = .83961$; also $R^2_{RM} = .70495$. With $SS_T = 525.33$, $SS_T(adj) = 525.33\ (1 - .70495) = 155.00$. The adjusted sum of squares among $SS_A(adj)$ can be found as a residual and is 155.00 - 118.32 = 36.68. The summary table is given in Table 2.

TABLE 2

SUMMARY TABLE FOR THE ANALYSIS OF COVARIANCE WITH TWO COVARIATES

Source of Variation	df	SS	MS	F
Among	2	36.68	18.34	1.55
Within	10	118.32	11.83	
Total	12	155.00		

It should be clear from this presentation that any number of covariates could be employed in an analysis of covariance. Potential researchers should be cautioned against using the "slop bucket" approach to using a large number of covariates simply because it is possible. In addition to being non-scientific, the use of each covariate does entail the loss of one degree of freedom in the adjusted sum of squares within term. A person could use 25 covariates with ease; he should be familiar enough with the data to make a reasonable interpretation of that data after the adjustments, however. A program has been prepared (Williams and Lindem, 1974a) to accommodate up to 20 covariates (which can be redimensioned to include more covariates if necessary); the program prints out summary tables for the analysis of variance for the criterion scores and an analysis of covariance with the multiple covariates and the adjusted means.

Two-Way Fixed Effects Analysis of Variance with Disproportionate Cell Frequencies

The solution to the disproportionate case of the two-way fixed effects analysis of variance is complicated by the existence of more than one solution, the different solutions being dependent upon the assumptions of the researcher. The present program (Williams and Lindem, 1972) allows for the selection of any (or all) of the following least squares solutions: (a) the method of fitting constants, a commonly accepted solution, described in Scheffé (1959) and Anderson and Bancroft (1952), a method that adjusts each main effect for the other main effect; (b) the hierarchical model (Cohen, 1968), which allows for one effect to take precedence over the second effect; the first main effect is unadjusted, and the second main effect is adjusted for the first main effect; and (c) the unadjusted main

effects method, in which neither main effect is adjusted for the other main effect. In all three methods, the interaction effect is adjusted for the two main effects. The three least squares methods and the previously mentioned approximate solutions are compared by Williams (1972).

As an example of the solutions to the disproportionate two-way situation, consider the following data in Table 3.

TABLE 3

DATA FOR DISPROPORTIONATE TWO-WAY ANALYSIS OF VARIANCE

Effect	Effect B_1	Effect B_2	Effect B_3
A_1	8 6 4	1 1	6 2
A_2	10	7 5 4 4 3	10 9 7 5 4

To solve for any of the three solutions, four linear models are necessary:

Model I: $Y = b_0 + b_1X_1 + e_1$, (5)

Model II: $Y = b_0 + b_2X_2 + b_3X_3 + e_2$, (6)

Model III: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e_3$, (7)

Model IV: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_4$, (8)

where

Y = the criterion,

X_1 = 1 if the score is from a member of row 1, and 0 otherwise;

X_2 = 1 if the score is from a member of column 1, and 0 otherwise;

X_3 = 1 if the score is from a member of column 2, and 0 otherwise;

$X_4 = X_1 \cdot X_2$;

$X_5 = X_1 \cdot X_3$;

$b_0 - b_5$ = are regression coefficients (The values for b_0, b_1, b_2 and b_3 will, in general, be different for Models I-IV), and

$e_1 - e_4$ = are the errors in prediction with their respective models.

Table 4 contains a formulation for the regression solutions to the two-way fixed effects analysis of variance with disproportionate cell frequencies.

TABLE 4

REGRESSION FORMULATION FOR THE TWO-WAY ANALYSIS OF VARIANCE

Y	X_1	X_2	X_3	X_4	X_5
8	1	1	0	1	0
6	1	1	0	1	0
4	1	1	0	1	0
1	1	0	1	0	1
1	1	0	1	0	1
6	1	0	0	0	0
2	1	0	0	0	0
10	0	1	0	0	0
7	0	0	1	0	0
5	0	0	1	0	0
4	0	0	1	0	0
4	0	0	1	0	0
3	0	0	1	0	0
10	0	0	1	0	0
9	0	0	0	0	0
7	0	0	0	0	0
5	0	0	0	0	0
4	0	0	0	0	0
				0	0

The values from the regression program that are useful for completing the analysis of variance are: the sum of squares attributable to regression for Models I, II, III, and IV, and the sum of squares for deviation from regression for Model IV. The R^2 values are also included in Table 5. The total sum of squares is of course available from all four models.

The A effect for the method of fitting constants is the difference between the sum of squares for attributable to regression for Model III and Model II: $SS_A = 80.25 - 37.43 = 42.82$.

Essentially, this process amounts to finding that part of the A effect that is independent of the B effect.

The B effect for this method is the difference between the sum of squares for attributable to regression for Model III and Model I:

$SS_B = 80.25 - 20.36 = 59.89$.

TABLE 5

VALUES FOUND FROM THE REGRESSION ANALYSIS

	df	SS	R^2
Model I (A effect) Attributable to Regression	1	20.36	.15427
Model II (B effect) Attributable to Regression	2	37.43	.28355
Model III (Combined A & B effects) Attributable to Regression	3	80.25	.60796
Model IV (Full Model) Attributable to Regression	5	80.80	.61212
Deviation from Regression	12	51.20	
Total Sum of Squares	17	132.00	

Similarly, this second calculation yields that part of the B effect that is independent of the A effect.

And finally, the interaction is found as the difference between Model IV and Model III: SS_{AB} = 80.80 - 80.25 = .55. Thus, the effect found in this manner is the AB effect independent of the A and B effects.

The sum of squares for within is equal to the deviation from regression for Model IV. This information for the data in Table 4 can be put into a summary table (Table 6).

TABLE 6

SUMMARY TABLE FOR THE METHOD OF FITTING CONSTANTS

Source of Variation	df	SS	MS	F
A	1	42.82	42.82	10.03**
B	2	59.89	29.95	7.01**
AB	2	.55	.28	.07
Within	12	51.20	4.27	

**$p < .01$

The method of fitting constants is not a partitioning model. That is, if the sum of squares is totaled, it does not equal the total sum of squares of 132.00 (The total is 154.46).

The Hierarchical Model

The hierarchical model (Cohen, 1968) is a method similar to the method of fitting constants. With this approach, a researcher is required to order the variables in relation to their research interest. For example, a

researcher may be most interested in the A, or row effect, less interested in the B, or column effect, and may have little interest in the interaction effect. With this approach, each effect is adjusted only for those effects preceding it in the ordering. Thus, the A effect is found directly, the B effect is adjusted for the A effect, and the AB effect is adjusted for the combined A and B effect. Unlike the previous model, this model is additive in the sense that the sum $SS_A + SS_B + SS_w$ is equal to SS_T. The values for SS_A, SS_B, and SS_w can be found from Table 5: $SS_A = 20.36$, the unadjusted A effect: $SS_B = 80.25 - 20.36 = 59.89$, that part of the B independent of A; $SS_{\overline{AB}} = 80.80 - 80.25 = .55$, as previously, and $SS_w = 51.20$.

These values are placed in a usual summary table (Table 7).

TABLE 7

SUMMARY TABLE FOR THE HIERARCHICAL MODEL

Source of Variation	df	SS	MS	F
A	1	20.36	20.36	4.77*
B	2	59.89	29.95	7.01**
AB	2	.55	.28	.07
Within	12	51.20	4.27	
Total	17	132.00		

*$p < .05$
**$p < .01$

The results from this analysis are identical to the fitting constants method except for the SS_A term. The interpretation would be somewhat different however, because of the decrease in size of the SS_A term. If, on the other hand, the researcher had chosen his order of experimental interest

as B, A, AB, then the F values for the A effect and the AB effect would be unchanged from the fitting constants method, but the B effect would be smaller.

The Unadjusted Main Effects Method

A solution similar to the two previous least squares solutions can be called the unadjusted main effects method. Using this approach, both the A and B effects are found directly, with the interaction found in the same manner as the method of fitting constants and the hierarchical model. The error term (mean square within) is of course the same. The values for SS_A, SS_B, SS_{AB}, and SS_w can be found from Table 5: SS_A = 20.36, the unadjusted A effect; SS_B = 37.43, the unadjusted B effect; SS_{AB} = 80.80 - 80.25 = .55, as previously; and SS_w = 51.20.

Table 8 contains the unadjusted main effects method analysis.

TABLE 8

SUMMARY TABLE FOR THE UNADJUSTED MAIN EFFECTS METHOD

Source of Variation	df	SS	MS	F
A	1	20.36	20.36	4.77*
B	2	37.43	18.72	4.88*
AB	2	.55	.28	.07
Within	12	51.20	4.27	

*$p < .05$

If the sum of squares is totaled for Table 8, the total is less than 132.00 because of the suppressor relationship between A and B (the total for Table 8 is actually 109.54). The unadjusted main effects

method is identical, as a solution, to the one proposed by Jennings (1967). That Jennings' approach and the unadjusted main effects method yield the same results was shown by Halldorson (1969).

Two-Way Analysis of Covariance with Multiple Covariates and Proportionate or Disproportionate Cell Frequencies

The present program (Williams and Lindem, 1974b) is a generalized two-way fixed effects analysis of covariance program that will allow multiple covariates and/or disproportionality of the cell frequencies. Because the program is general, it can be used whether or not there are multiple covariates and whether or not disproportionality of the cell frequencies exists. As was true of the program documented for the two-way fixed effects analysis of variance with disproportionate cell frequencies, three distinct solutions exist for this analysis of covariance situation: (1) the method of fitting constants, a solution that adjusts each main effect for the covariates and the other main effect; (2) the hierarchical model, which allows one main effect to take precedence over the second main effect; the first main effect is adjusted only for the covariates, and the second main effect adjusted for both the first main effect and the covariates, and (3) the unadjusted main effects method, in which the main effects are adjusted only for the covariates. In all three solutions, the interaction effect is adjusted for the covariates and the two main effects. These three solutions are analogous to the previously documented solutions for the fixed effects analysis of variance with disproportionate cell frequencies.

As an illustrative example, suppose the data is cast in a 2 X 3 table with two covariates. Then the following models could be generated:

Model V: $Y = b_0 + b_6 X_6 + b_7 X_7 + e_5$, (9)

Model VI: $Y = b_0 + b_1 X_1 + b_6 X_6 + b_7 X_7 + e_6$, (10)

Model VII: $Y = b_0 + b_2 X_2 + b_3 X_3 + b_6 X_6 + b_7 X_7 + e_7$, (11)

Model VIII: $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_6 X_6 + b_7 X_7 + e_8$, (12)

Model IX: $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + b_6 X_6 + b_7 X_7 + e_9$. (13)

where

Y, X_1, X_2, X_3, X_4, X_5 and b_0 - b_5 are defined as previously given in the solution for disproportionate cell frequencies for a two-way analysis of variance,

X_6 = the score on the first covariate for each subject,

X_7 = the score on the second covariate for each subject,

b_6 - b_7 = are regression coefficients for X_6 and X_7 respectively, (b_0 - b_7 will, in general, be different for Models V-IX), and

e_5 - e_9 = the errors in prediction for Models V-IX.

Then, for the fitting constants solution,

SS_A = the SS for attributable to regression for Model VIII - the SS for attributable to regression for Model VII, (14)

SS_B = the SS for attributable to regression for Model VIII - the SS for attributable to regression for Model VI, (15)

SS_{AB} = the SS for attributable to regression for Model IX - the SS for attributable to regression for Model VIII, (16)

and

SS_w = the SS for deviation from regression for Model IX. (17)

For the hierarchical solution with primary interest in the A effect;

SS_A = the SS for attributable to regression for Model VI - the SS for attributable to regression for Model V, (18)

SS_B = same as equation 15,

SS_{AB} = same as equation 16, and

SS_W = same as equation 17.

For the unadjusted main effects solution:

SS_A = same as equation 18,

SS_B = the SS for attributable to regression for Model VII - the SS for attributable to regression for Model V, (19)

SS_{AB} = same as equation 16, and

SS_W = same as equation 17.

The fitting constants solution for the analysis of covariance can be seen as analogous to the fitting constants solution for the two-way analysis of variance, except that the covariates are also removed as a source of variation; thus, the A effect in the fitting constants solution is that portion independent of both the B effect and the covariates. In the hierarchical solution, the effect of primary research interest is adjusted for the covariates only; in the unadjusted main effects solution, the main effects are adjusted for the covariates only, and not adjusted for the other main effect. The interaction effect and within term are the same for all three solutions.

The solutions for COVARMLT (the analysis of covariance with multiple covariates) and the two-way analysis of covariance described here do not include a test for the homogeneity of the regression on the covariates. Future revisions of these two programs will include options for running these tests if the user so desires.

REFERENCES

Anderson, R. L., and Bancroft, T. A. Statistical theory in research. New York: McGraw-Hill, 1952.

Cohen, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Halldorson, M. H. An iterative multiple regression technique for the factoral analysis of variance with unequal and disproportionate cell frequencies. Unpublished doctoral dissertation, The University of Northern Colorado, 1969.

Jennings, E. Fixed effects analysis of variance by regression analysis. Multivariate Behavioral Research, 1967, 2, 95-108.

Scheffé, H. The analysis of variance. New York: Wiley, 1959.

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

Williams, J. D. Applications of setwise regression analysis. Multiple Linear Regression Viewpoints, 1973, 4, No. 2, 1-7.

Williams, J. D., and Lindem, A. C. Setwise regression analysis--A stepwise procedure for sets of variables. Educational and Psychological Measurement, 1971a, 31, 747-748..

Williams, J. D., and Lindem, A. C. Setwise Linear Regression, UND Computer Center Special Report, No. 30, November, 1971b.

Williams, J. D., and Lindem, A. C. A computer program for two-way fixed effects analysis of variance with disproportionate cell frequencies. UND Computer Center Special Report, No. 35, October, 1972.

Williams, J. D., and Lindem, A. C. A computer program for the analysis of covariance with multiple covariates. UND Computer Center Special Report, March, 1974a.

Williams, J. D., and Lindem, A. C. A computer program for the two-way analysis of covariance with multiple covariates and proportionate or disproportionate cell frequencies. UND Computer Center Special Report, March, 1974b.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

MASHIT - FOR EASE IN REGRESSION

PROGRAM COMMUNICATION

Robert L. Mason

Science Applications, Incorporated
2109 W. Clinton Avenue
Huntsville, Alabama 35805

Keith A. McNeil

Educational Monitoring Systems

Paper Presented to 1974 Annual Meeting Of
American Education Research Association
Chicago, Illinois

ABSTRACT

MASHIT - For Ease In Regression Program Communication

Robert L. Mason
Science Applications, Incorporated
2109 W. Clinton Avenue
Huntsville, Alabama 35805

This regression system is an intermediate result of a project to develop a comprehensive regression computer system as a foundation for a complete statistical man-machine interface. The outstanding features of the system can be condensed into two principle concepts. First, the program dynamically allocates core resulting in no limits on title cards, question cards, etc. Secondly, "English type" user commands are used in a free format mode to save computer instruction time. The resulting system is two phase constructed in such a manner that additional capabilities can be added efficiently.

MASHIT (Mason's Automatic Statistical Hypothesis Interpreter and Tester) is a computer program written in high level programming languages to facilitate the interaction between the computer and the researcher. The primary unique aspect of MASHIT is that the program performs regression analysis based on conversation language research hypotheses. Ultimately, other statistical techniques will be incorporated into the system; however, the current version handles that wide range of research hypotheses that can be tested using MLR. Indeed, any least squares hypotheses concerned with a single criterion can be tested with MASHIT. The program readily handles "analysis of covariance" questions.

After studying several available regression programs, a composite of shortcomings was compiled. There was no one system that offered all the features that the researchers and students desired to test their hypotheses. In addition, the only systems that allowed free form input were the interactive terminal programs, whereas the majority of researchers must cope with "batch mode" computer systems. MASHIT is a system directed toward researchers desiring ease and flexibility in accessing a "batch mode" computer system. The following is a list of the outstanding program features:

(a) Analysis of natural language regression questions.
(b) Free format (no column restrictions with the one exception of any optional FORTRAN transformation statements desired).
(c) Virtually unlimited number of variables.
(d) Virtually unlimited number of models.
(e) Virtually unlimited number of research questions.
(f) Virtually unlimited number of title cards.
(h) Any size variable labels.
(i) The program will dichotomize all "A" or "I" field (discrete) variables. The user does not have to keep track of newly created variables.
(j) Any FORTRAN transformation statements allowed.
(k) No parameter card necessary.
(l) All double precision calculations.
(m) Multiple returns from transformation subroutine.

MASHIT was written for the IBM 360/370 series computers. Constructed in two parts, the program first reads and analyzes the researchers instructions, and passes this information to the second stage which performs the regression and test calculations.

The first stage actually creates the second stage program resulting in a "tailored" regression program for each individual user. Storage sizes can be expanded or contracted to fit the researchers program requirement and the desired machine region (core) request. This flexibility of the program does not affect the number of variables which can be processed, however, the machine CPU time is decreased as the core storage size is increased.

In order for a computer program to interpret natural language, established criteria must be met. They are (1) variables must be prelabeled if referenced by labels in a hypotheses, (2) only specific phrases from a list of keyword can be used and (3) certain syntactical rules must be followed. These rules are discussed in a later section.

The flexibility of the program allows the user to input his "control deck" as though it were written in a manner similar to a paragraph. There are no column restrictions with the one exception of FORTRAN trnasformations. MASHIT searches for keywords and labels; therefore, blanks are placed between coded words.

The program reads the entire "control deck" as if it were one long card. Therefore, coding can skip from card to card, even with the option of inserting blank cards in the control deck. Slashes, periods, or question marks are delimiters indicating the end of one type of control information within the control deck. There are presently eight types of control cards as follows:

(a) Title Card(s)
(b) Label Card(s)
(c) Transformation Card(s)
(d) Special Command Card(s)
(e) Question Card(s)
(f) Model Card(s)
(g) Test Card(s)
(h) Format Card(s)

All the control cards are optional with the exception of a format. If the format is the only card included, MASHIT prints the means, standard deviations and the correlations. A summary of the rules for each type control card is included as a "mini reference guide" at the end of this paper. To facilitate the remainder of the discussion, example deck setups are shown to illustrate the features of MASHIT.

Example 1

This is a rather limited application of MASHIT. One question is asked without the use of labels. The program recognizes the word "PREDICT" and uses the variable that follows as the criterion. The first variable and the unit vector are employed as predictor variables. Since there are no covariates, the restricted model is inferred to have an R^2 value of zero. The number of subject and numbers of linearly independent vectors are calculated and the F test is evaluated. The next page contains the printout as generated by the program.

```
// (Job Card)
// EXEC  MASHIT
Does X1 PREDICT VAR 2?
(2F3.0)
          Data Cards
/*
```

NUMBER OF OBSERVATIONS---------- 12

NUMBER OF VARIABLES READ---------- 2

NUMBER OF VARIABLES AFTER TRANSFORMATION-- 2

NUMBER OF VARIABLES - CONTINUOUS---------- 2

NUMBER OF VARIABLES - DISCRETE---------- 0

INPUT UNIT NUMBER---------- 5

FORMAT (2F3.0)

VARIABLE NUMBER	TYPE OF VARIABLE	NUMBER OF DIFFERENT VALUES	MEAN ****	STANDARD DEVIATION	VARIABLE NAME
1	CONTINUOUS		31.50000	19.80951	
2	CONTINUOUS		56.63333	26.73273	

CORRELATION MATRIX

		1	2
VARIABLE	1	1.00000	
VARIABLE	2	-0.48373	1.00000

****

DOES X1 PREDICT VAR 2 ?

****

CRITERION NUMBER = 2 INDEPENDENT VECTORS = 2

MODEL R-SQUARE...= 0.2339974 ********** NUMBER OF ITERATIONS = 1

VARIABLE NUMBER	RAW SCORE WEIGHTS	VARIABLE NAME
1	-0.65279252	
REGRESSION CONSTANT =	72.39629787	

****

FULL MODEL.....= MODEL FROM ABOVE

RESTRICTED MODEL...= ZERO RSQ MODEL

$$F = \frac{(RSQ\ F - RSQ\ R\ ) / DF1}{(1.0 - RSQ\ F\ ) / DF2} = \frac{(0.23400 - 0.0\ ) / 1}{(1.0 - 0.23400\ ) / 10} = 3.054787$$ ********

MONODIRECTIONAL PROBABILITY = 0.1084337

DIRECTIONAL PROBABILITY = 0.0542168
(IN HYPOTHESIZED DIRECTION)

****

Example 2

This example illustrates the use of the "title", "model", "label", and "test" cards. Note that the title is two cards in length and each variable label is placed on a seperate card. This is not necessary as a title can be any number of cards in length and labels only have to be seperated by a delimiter.

```
// (Job Card
// EXEC   MASHIT
THIS STUDY USES LABEL CARDS, TWO MODEL CARDS, AND
A TEST CARD /
LABELS:
  RUNNING SPEED FOR 100 YARD DASH:
  AGE:
  MOTIVATION /
MODEL A: AGE AND MOTIVATION PREDICTING SPEED FOR
         100 YARD DASH /
MODEL B: AGE PREDICTING RUNNING SPEED FOR 100 YARD
         DASH /
TEST MODEL A AGAINST MODEL B /
(3F3.0)
        Data Cards
/ *
```

The structure in this example is similar to that used in other hypothesis testing regression programs with the exception that the variables have been labeled and referenced in natural language in the delineation of the models. This technique is used when testing many restricted models against the same full model. A simpler structure is available especially if only one F test were being computed. By use of the "question" card as in example 1 with the attachment of a covariate phase, one question can replace two models and a test as the following:

```
DOES MOTIVATION PREDICT RUNNING
SPEED FOR 100 YARD DASH OVER AND ABOVE AGE?
```

Since all least squares hypothesis can be phrased in "covariance" terminology as above, the "question" card has much potential for researchers. The printed output is shown on the next two pages.

THIS STUDY USES LABEL CARDS, TWO MODEL CARDS, AND
A TEST CARD /

NUMBER OF OBSERVATIONS—————————— 12
NUMBER OF VARIABLES READ———————————— 3
NUMBER OF VARIABLES AFTER TRANSFORMATION— 3
NUMBER OF VARIABLES - CONTINUOUS———————— 3
NUMBER OF VARIABLES - DISCRETE———————— 0
INPUT UNIT NUMBER———————————————— 5

FORMAT (3F3.0)

VARIABLE NUMBER	TYPE OF VARIABLE	NUMBER OF DIFFERENT VALUES	MEAN	STANDARD DEVIATION	VARIABLE NAME
1	CONTINUOUS		490.16667	2d5.29570	RUNNING SPEED FOR 100 YARD DASH
2	CONTINUOUS		484.41667	265.58597	AGE
3	CONTINUOUS		642.08333	274.91831	MOTIVATION

CORRELATION MATRIX

	1	2	3
VARIABLE 1	1.00000		
VARIABLE 2	0.26136	1.00000	
VARIABLE 3	0.07457	0.21081	1.00000

MODEL A : AGE AND MOTIVATION PREDICTING RUNNING SPEED FOR 100 YARD DASH /

CRITERION VARIABLE NAME = ... RUNNING SPEED FOR 100 YARD DASH

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 3
MODEL R-SQUARE...= 0.06855198 NUMBER OF ITERATIONS = 2

VARIABLE NUMBER	RAW SCORE WEIGHTS	VARIABLE NAME
2	0.27560396	AGE
3	0.02121645	MOTIVATION
REGRESSION CONSTANT =	342.93861958	

MODEL B : AGE PREDICTING RUNNING SPEED FOR 100 YARD DASH /

CRITERION VARIABLE NAME = ... RUNNING SPEED FOR 100 YARD DASH

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 2
MODEL R-SQUARE...= 0.06815249 NUMBER OF ITERATIONS = 1

VARIABLE NUMBER	RAW SCORE WEIGHTS	VARIABLE NAME
2	0.28043419	AGE
REGRESSION CONSTANT =	354.31967218	

TEST MODEL A AGAINST MODEL B /

FULL MODEL.....= MODEL A

RESTRICTED MODEL...= MODEL B

$$F = \frac{(RSQ\ F - RSQ\ R\ ) / DF1}{(1.0 - RSQ\ F\ ) / DF2} = \frac{(0.06855 - 0.06815\ ) / 1}{(1.0 - 0.06855\ ) / 9} = 0.003860$$

NONDIRECTIONAL PROBABILITY = 0.9505624

DIRECTIONAL PROBABILITY = 0.4752812
(IN HYPOTHESIZED DIRECTION)

Example 3

Here is featured the input of nominal data, FORTRAN transformations and a "covariate" question. Three variables are read and a fourth is created as the square of the third variable.

```
// (Job Card)
// EXEC   MASHIT
       X (4) = X (3) ** 2
IS VAR 1 PREDICTED BY X 2
GIVEN KNOWLEDGE OF X (3), X 4 ?
(F5.0, A1, 4X, F5.0)
          Data Cards
/ *
```

Also, the A-format for variable two implies that it is discrete. MASHIT then automatically constructs and maintains the mutually exclusive group membership vectors. When variable two is referenced in the research question, the group membership vectors are substituted.

Notice in the printout (next pages) that the program reports the number of observations, how many variables were read, how many were created by transformations and the number of mutually exclusive vectors that resulted.

Further Documentation

MASHIT was developed by Robert L. Mason as part of a doctoral dissertation under the direction of Dr. Keith McNeil. The dissertation (Mason, 1973) has complete documentation. Also, a 65 page "MASHIT) user's guide is available. The following is a summary of the users manual.

NUMBER OF OBSERVATIONS---------- 15

NUMBER OF VARIABLES READ---------- 3

NUMBER OF VARIABLES AFTER TRANSFORMATION-- 4

NUMBER OF VARIABLES - CONTINUOUS---------- 3

NUMBER OF VARIABLES - DISCRETE---------- 1

INPUT UNIT NUMBER---------- 5

NUMBER OF DICHOTOMIZED VARIABLES---------- 2

FORMAT (F5.0,A1,4X,F5.0)

VARIABLE NUMBER	TYPE OF VARIABLE	NUMBER OF DIFFERENT VALUES	MEAN ****	STANDARD DEVIATION	VARIABLE NAME
1	CONTINUOUS		41.66667	29.12883	
2	DISCRETE	2			
3	CONTINUOUS		53.20000	21.22326	
4	CONTINUOUS		3280.66667	2415.94944	

VARIABLE NUMBER	TYPE OF VARIABLE	CREATED BY VARIABLE NUMBER	MEAN ****	STANDARD DEVIATION	VARIABLE VALUE
5	DICHOTOMOUS	2	0.66667	0.47140	A
6	DICHOTOMOUS	2	0.33333	0.47140	B

CORRELATION MATRIX

		1	2	3	4	5	6
VARIABLE	1	1.00000					
VARIABLE	2	********	********				
VARIABLE	3	-0.38035	********	1.00000			
VARIABLE	4	-0.32130	********	0.96756	1.00000		
VARIABLE	5	0.03560	********	-0.10662	-0.17225	1.00000	
VARIABLE	6	-0.03560	********	0.10662	0.17225	-1.00000	1.00000

****

IS VAR 1 PREDICTED BY X 2 GIVEN KNOWLEDGE OF X (3) , X4 ?

****

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 4
MODEL R-SQUARE...= 0.18111067 NUMBER OF ITERATIONS = 5
****

VARIABLE NUMBER	RAW SCORE WEIGHTS	VARIABLE NAME
2 - 5	3.08248870	
2 - 6	0.0	VALUE = A
		VALUE = B
3	-1.55778456	
4	0.00946970	
REGRESSION CONSTANT =	91.41887092	

****

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 3
MODEL R-SQUARE...= 0.17083439 NUMBER OF ITERATIONS = 2
****

VARIABLE NUMBER	RAW SCORE WEIGHTS	VARIABLE NAME
3	-1.49366987	
4	0.03882175	
REGRESSION CONSTANT =	92.18867231	

****

FULL MODEL.....= MODEL FROM ABOVE

RESTRICTED MODEL...= MODEL FROM ABOVE

$$F = \frac{(RSQ\ F - RSQ\ R) / DF1}{(1.0 - RSQ\ F) / DF2} = \frac{(0.18111 - 0.17883) / 1}{(1.0 - 0.18111) / 11} = 0.030577$$
****

NONDIRECTIONAL PROBABILITY = 0.8583112

DIRECTIONAL PROBABILITY = 0.4291556
(IN HYPOTHESIZED DIRECTION)

****

"MASHIT" MINI GUIDE

Program MASHIT (Mason's Automatic Statistical Hypothesis Interpreter and Tester) was developed to aid the researcher in his ever losing battle with the computer. The program is written in SNOBOL and FORTRAN IV to run on the IBM 360-370 series computers. Presently only regression type questions can be interpreted.

This Mini reference guide is intended to be a summary for the computer user. If questions arise, the user should refer to the "MASHIT" users manual, which is complete with examples (Mason, 1973).

CONTROL CARD RULES

The program looks for keywords in the statement made to the computer. Care with the spelling and spacing of input words is necessary. Space must be maintained between words and data input can be continued from card to card freely as the computer thinks the deck is one long card. The following are general types of cards with the limited rules necessary. An important point to note is that the only card absolutely necessary besides the data is a FORMAT card. Just placing the FORMAT card before the data results in the printing of means, standard deviations, and correlations.

A. TITLE CARD(S) - Optional
   1. Any number of cards
   2. Must use keyword "TITLE" or "PROJECT" or "STUDY".
   3. Must end with slash or period.

B. TRANSFORMATIONS - Optional
   1. The rules of FORTRAN apply.
   2. Refer to variables in the Array X.

      EXAMPLE: X(22) = X(1) *X(3)

C. LABEL CARD(S) - Optional
   1. Must start with keyword "LABEL" or "LABELS" followed by a colon or semicolon.
   2. Separate names by colons or semicolons.
   3. Can indicate or change variable number. Otherwise they are thought to be sequential. This example labels variables 1, 2, 10 and 11.

      LABEL: SEX RACE:
      VAR 10: GRADE POINT: EDUCATION /

   4. Must end entire variable labeling series with a slash or period.

D. SPECIAL INSTRUCTIONS - Optional
   1. The following are instructions that the program understands:
      a. READ IN 20 VARIABLES
      b. READ DATA FROM UNIT 4
      c. READ IN 120 OBSERVATIONS
      d. REGION SIZE = 132K
      e. THERE ARE 514 VARIABLES AFTER TRANSFORMATIONS.
   2. Must end with slash or period.

E. MODEL(S) - Optional
   1. Must have a model name that includes keyword "MODEL".
   2. Model name must be followed by colon or semicolon.
   3. Model structure follows the colon delimiter. Model structure is discussed later.
   4. Model structure must end with a period or slash.

F. TEST CARD(S) - Optional
   1. Must have keywords "TEST" and "AGAINST" or "WITH".
   2. Must have at least two model names that were previously defined.
   3. Must end with period or slash.

G. QUESTION(S) - Optional
   1. Must use the model structure that is defined elsewhere.
   2. Must end with period or slash.

H. FORMAT CARD(S) - Necessary
   1. Can start the card with "(" or "FORMAT (".
   2. Must have balanced parentheses.
   3. Program will count the number of variables by the format.
   4. F type variables considered continuous variables.
   5. A type or I type variables are considered discrete and are dichotomized for the researcher.

MODEL STRUCTURE

All questions and models must be formed in regression terms. That is, all predictor (independent) variables and the criterion (dependent) variables must be separated by a key word or phrase. Examples of this are:

A. DOES SEX, EDUCATION LEVEL, AND IQ PREDICT GRADE POINT AVERAGE?

B. VAR 5, VAR 7 AND VAR 9 PREDICTING X1.

C. IS THE VARIANCE IN X1 ACCOUNTED FOR BY SEX?

The keywords are underlined in the examples. The present list of keyphrases consists of:

PREDICT	EXPLAIN
PREDICTS	EXPLAINS
PREDICTING	EXPLAINING
PREDICTIVE OF	EXPLAINED BY
PREDICTION OF	ACCOUNT FOR
PREDICTED BY	ACCOUNTED FOR BY
INFLUENCED BY	

OTHER CONCEPTS

Two other concepts regarding the structure must be mentioned. The first is that of covariates, and the second is naming and grouping variables. In asking a question, input can include covariate variable(s) in the question by the use of one of the following keyphrases:

OVER AND ABOVE
WITH KNOWLEDGE OF
GIVEN KNOWLEDGE OF
IN CONJUNCTION WITH
IN THE PRESENCE OF
WITH - list - AS COVARIATES

An example is:

DOES GROUP MEMBERSHIP PREDICT VAR 1 OVER AND ABOVE SEX?

The other concept is the naming and grouping of variables. Variables can be denoted by an assigned name, or using "X" or "VAR" notations. A comma or the word "AND" between two variables names means to use only those two variables. A dash (-) or one of the keywords (TO, THRU, THROUGH) indicates the use of those variables and all variables between. The example:

VAR 7, VAR 9 - 11

refers to the four variables 7, 9, 10, 11. For further information and examples, the user is referred to "MASHIT" users manual (Mason, 1973) describing the program.

DECK SETUP

The following card order is used in making a run on program MASHIT.

```
// (Job Card)
// EXEC   MASHIT
                    Consists of the following types of cards
                    in any order:
                    A.   TITLE
                    B.   TRANSFORMATIONS
CONTROL DECK        C.   LABELS
(All Optional)      D.   INSTRUCTIONS
                    E.   MODELS
                    F.   TESTS
                    G.   QUESTIONS
FORMAT CARD(S)
        Data Cards
/ *
```

Example

```
// (JOB CARD)
// EXEC   MASHIT
THIS STUDY MEASURES THE CURVILINEAR EFFECT OF AGE
ON RUNNING SPEED  /
LABELS :
  RUNNING SPEED FOR 100 YARD DASH :
  AGE :
  AGE SQUARED, USED FOR CURVILINEAR TEST /
      X(3) - X(2) ** 2
MODEL A : VAR 2, VAR 3 PREDICTING VAR 1 /
MODEL B : VAR 2        PREDICTING VAR 1 /
TEST MODEL A AGAINST MODEL B /
DOES X3 PREDICT VAR 3 OVER AND ABOVE AGE?
(2F3.0)
        Data Cards
/ *
```

NOTE: The one question does the same thing as the two models and test combined.

REFERENCES

Mason, R. L. "The Development of an Automated Statistical Hypothesis Interpretor and Tester". Unpublished doctor's dissertation, Southern Illinois University Carbondale, Illinois, 1973.

Mason, R. L. "MASHIT" User's Manual, unpublished manuscript, 1973.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

THE DEVELOPMENT AND DEMONSTRATION OF MULTIPLE REGRESSION MODELS FOR OPERANT CONDITIONING QUESTIONS

Fred Fanning
Isadore Newman
University of Akron

Paper Presented to 1974 Annual Meeting Of
American Education Research Association
Chicago, Illinois

The Development and Demonstration of Multiple Regression Models for Operant Conditioning Questions

Fred Fanning
Isadore Newman
The University of Akron

Abstract...based on the assumption that inferential statistics can make the operant conditioner more sensitive to possible significant relationships, regression models were developed to test the statistical significance between slopes and Y intercepts of the experimental and control group subjects. These results were then compared to the traditional operant conditioning eyeball technique analysis.

Fred Fanning
The University of Akron
College of Education
Akron, Ohio 44325

The Development and Demonstration of Multiple Regression Models for Operant Conditioning Questions

Fred Fanning
Isadore Newman
The University of Akron

Summarization of research in operant psychology has relied predominately upon descriptive statistics. Probably the main reason inferential statistics has been given little attention is that early operant research yielded such clear-cut distinctions that it was not necessary to resort to tests of statistical significance. A second reason may be the lack of advice from statisticians regarding limitations of single subject data.

Presently, much research in operant psychology is being done in the natural environment outside the laboratory, as applied behavior modification. In these settings, the control of extraneous variables is more difficult to achieve. As a result, data may fail to exhibit the clear magnitude of effects observed in data from a laboratory manipulation. When this occurs, significant results may not be immediately obvious even though the expected trend seems to be present. When some doubt exists concerning the outcome of an experimental manipulation using behavior modification procedures, consideration should be given to the use of inferential statistics. A number of inferential statistical models are currently available that may assist the operant researcher in analyzing his data. These models are essentially specific applications of the generalized analysis of variance using multiple regression procedures to partial variance.

The purpose of this paper is to develop and demonstrate regression models that may be useful to operant conditioners for statistically analyzing their data. A comparison will be presented between a regression approach to answering operant conditioning questions and traditional operant analysis and interpretations of the same data.

The research questions dealt with here are only examples of the many possible kinds of questions which can be dealt with effectively using multiple regression procedures. Models will be developed to test the following questions:

1. Is there a significant mean difference between Control Group 1 and Control Group 2?

2. Is there a significant difference between the slope of Control Group 1 and Control Group 2 above and beyond individual differences?

3. Is there a significant mean difference between Control Group 1 and Control Group 3?

4. Is there a significant difference between the slope of Control Group 1 and Control Group 3 above and beyond individual differences?

5. Is there a significant mean difference between Control Group 2 and Control Group 3?

6. Is there a significant difference between the slope of Control Group 2 and Control Group 3 above and beyond individual differences?

7. Is there a significant mean difference between Control Group 1 and Experimental Group 1 above and beyond individual differences?

8. Is there a significant second degree curvilinear relationship for Control Group 1 and Experimental Group 1 above and beyond a linear relationship and any individual differences?

9. Is there a significant mean difference between Control Group 2 and Experimental Group 2 above and beyond individual differences?

10. Is there a significant second degree functional relationship for Control Group 2 and Experimental Group 2 above and beyond a linear relationship and any individual differences?

11. Is the mean of Control Group 3 significantly different from the mean of Experimental Group 3 above and beyond any individual differences?

12. Is there a significant second degree functional relationship for Control Group 3 and Experimental Group 3 above and beyond a linear relationship and any individual differences?

13. Is there a significant difference between the slope of Control Group 1 and Experimental Group 1 above and beyond any individual differences?

14. Is there a significant difference between the slope of Control Group 2 and Experimental Group 2 above and beyond any individual differences?

15. Is there a significant difference between the slope of Control Group 3 and Experimental Group 3 above and beyond any individual differences?

METHOD

Subjects. The total subject group consisted of twelve male and female students selected from a pool of names referred for chronic tardiness behavior by the school psychologist, teachers, and counselors at Westland High School, population 1,700, near Columbus, Ohio. Selection was made on the basis of the highest reported frequency of tardiness behavior.

The sample included one male freshman, four male and one female sophomores, two male and one female juniors, and three male seniors. All subjects were white, from approximately middle class socioeconomic background.

Material. The behavioral instruction program used in this design was a modification of Hall's book (1971, Pt. II) describing the basic principles of behavior modification.

The content of the control group instruction for both the teacher's daily lesson plans and the course outline, was taken from the general psychology text (Engle and Snellgrove, 1969), which students were given to use during this instruction.

PROCEDURE

During the initial phase of this design, the control period, the twelve subjects were assigned to three groups, four subjects to each group. Groups 1, 2, and 3 received a control treatment consisting of classroom instruction in general psychology. Immediately following the control period, the four students in Group 1 began receiving behavioral instruction treatment, consisting of classroom instruction in behavioral principles and their application. Group 2 continued receiving classroom instruction in general psychology, and Group 3 received general psychology instruction. When a decelerating trend in Group 1's tardiness behavior was noted, following instruction in behavioral principles, then Group 2 began receiving instruction in behavioral principles, and no longer received instruction in general psychology. When a decelerating trend in Group 2's tardiness behavior was noted, general psychology instruction was terminated with Group 3, and they began receiving instruction in behavioral principles. Group 1 and 2 continued receiving behavioral instruction throughout the remainder of the four week class.

ANALYSIS

The data was analyzed using two techniques:

(1) A multiple baseline design was used to demonstrate the effectiveness of the group instruction in behavioral principles (independent variable) on decelerating tardiness behavior (dependent variable). The multiple baseline design used for analysis of data is illustrated in Figure 1. Further information concerning the use of this type of design can be obtained by reference to Baer, et. al., (1968); Hall, et. al., (1970); and Hall, (1971, Pt. I). Additional data representing the total frequency of tardiness for the 3 groups is illustrated in Table 1.

THE EFFECTS OF BEHAVIORAL INSTRUCTION ON THE MEAN FREQUENCY OF TARDINESS BEHAVIOR OF THREE GROUPS

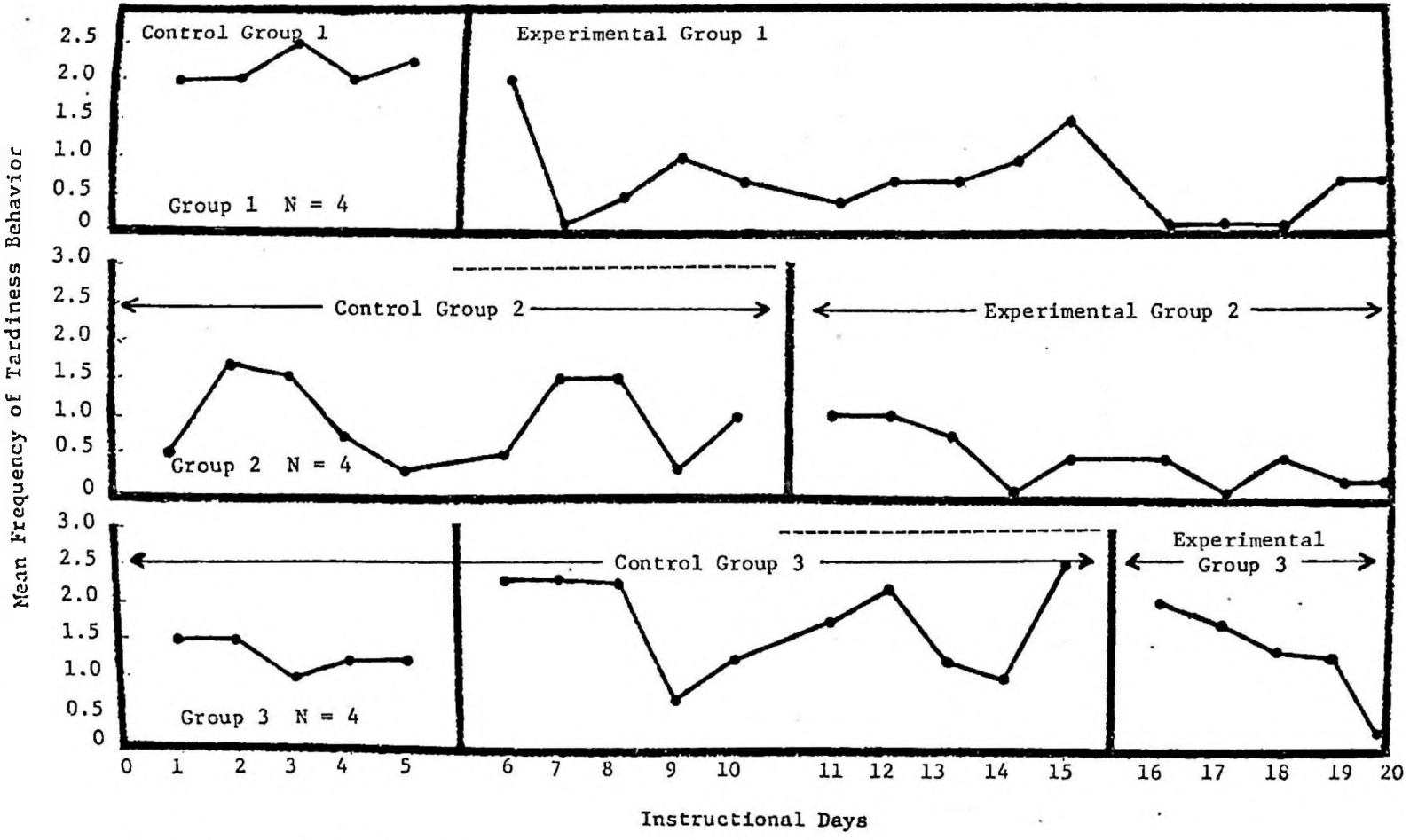


Figure 1 A Multiple Baseline Design indicating the mean frequency of tardiness behavior to school and class for three groups. N = 4 in all groups.

TABLE 1

TOTAL FREQUENCY OF TARDINESS BEHAVIOR TO SCHOOL AND CLASS FOR THREE GROUPS

	Week 1	Week 2	Week 3	Week 4
Group 1	Control 41	Beh. Inst. 17	Beh. Inst. 18	Beh. Inst. 6
Group 2	Control 17	Control 18	Beh. Inst. 13	Beh. Inst. 6
Group 3	$Baseline_1$ 26	Control 31	Control 35	Beh. Inst. 23

(2) Multiple regression was used to test the same hypothesis as the above traditional method for analyzing the data (see 1 above). For an example of how the data is set up, Figure 2 presents the hypothesis and models used to test them.

EXAMPLE MODELS

Research Hypothesis 1: The control group mean ($\bar{C}_1$) is significantly higher than the experimental group mean ($\bar{E}_1$) above and beyond person differences (P) + E.

Model 1: $Y_1 = a_0U + a_1(C_1) + a_2(E_1) + a_3(P_1) + a_4(P_2) + a_5(P_3) + a_6(P_4) + E$

$$a_1 = a_2$$

Model 2: $Y_1 = a_0U + a_3(P_1) + a_4(P_2) + a_5(P_3) + a_6(P_4) + E$

Research Hypothesis 2: The slopes of the experimental group (D_{e_1}) is significantly different than the slope of the control (D_{c_1}) group above and beyond person differences (P).

Model 3: $Y_1 = a_0U + a_1(C_1) + a_2(E_1) + a_3(D_{c_1}) + a_4(D_{e_1}) + a_5(P_1) + \ldots + a_6(P_4) + \text{Error}$

$$a_3 = a_4$$

Model 4: $Y_1 = a_0U + a_1(C_1) + a_2(E_1) + a_7(\text{Day}) + a_5(P_1) + \ldots + a_6(P_4) + \text{Error}$

In this example there were four persons (P_1, P_2, P_3, and P_4). During the control condition (C_1) each was measured on three consecutive days (D). The same four persons were again measured on three consecutive days during the experimental condition (E_1).

Figure 2

$$Y_1 = a_0U + a_1(C_1) + a_2(E_1) + a_3(D_{c_1}) + a_4(D_{e_1}) + a_5(P_1) + a_6(P_2) + a_7(P_3) + a_8(P_4) + \text{Error}$$

CONTROL GROUP											
Person 1	9	1	1	0	$1_{(1)}$	0	1	0	0	0	?
2	6	1	1	0	$1_{(1)}$	0	0	1	0	0	.
3	13	1	1	0	$1_{(1)}$	0	0	0	1	0	.
4	12	1	1	0	$1_{(1)}$	0	0	0	0	1	.
Person 1	12	1	1	0	$2_{(2)}$	0	1	0	0	0	.
2	10	1	1	0	$2_{(2)}$	0	0	1	0	0	?
3	9	1	1	0	$2_{(2)}$	0	0	0	1	0	.
4	7	1	1	0	$2_{(2)}$	0	0	0	0	1	.
Person 1	6	1	1	0	$3_{(3)}$	0	1	0	0	0	.
2	7	1	1	0	$3_{(3)}$	0	0	1	0	0	.
3	8	1	1	0	$3_{(3)}$	0	0	0	1	0	?
4	9	1	1	0	$3_{(3)}$	0	0	0	0	1	.
EXPERIMENTAL GROUP											
Person 1	7	1	0	1	0	$1_{(4)}$	1	0	0	0	.
2	9	1	0	1	0	$1_{(4)}$	0	1	0	0	
3	8	1	0	1	0	$1_{(4)}$	0	0	1	0	?
4	3	1	0	1	0	$1_{(4)}$	0	0	0	1	.
Person 1	5	1	0	1	0	$2_{(5)}$	1	0	0	0	.
2	7	1	0	1	0	$2_{(5)}$	0	1	0	0	.
3	2	1	0	1	0	$2_{(5)}$	0	0	1	0	
4	1	1	0	1	0	$2_{(5)}$	0	0	0	1	.
Person 1	5	1	0	1	0	$3_{(6)}$	1	0	0	0	?
2	1	1	0	1	0	$3_{(6)}$	0	1	0	0	.
3	2	1	0	1	0	$3_{(6)}$	0	0	1	0	.
4	3	1	0	1	0	$3_{(6)}$	0	0	0	1	.

An example of data set up. (N=24) There are four persons. Each is measured 6 times, 3 times for the control condition and 3 times for the experimental condition.

Figure 2 Continued

Where: C_1 = 1 if S_s was in control condition when measured zero otherwise

E_1 = 1 if S_s was in experimental condition when measured zero otherwise

D_{c_1} = day when measured for control group zero otherwise

D_{e_1} = day when measured for experimental group zero otherwise

Error = $Y - \hat{Y}$

U = 1 for each replicate in the study

$a_0 \ldots a_8$ = partial regression weights

RESULTS AND DISCUSSION

Table 2 presents the results of the regression analysis testing each of the fifteen questions. The operant analysis of these questions is presented in Table 3. In comparing these tables one should note that there is only disagreement on question five.

One major advantage of using the regression procedure, rather than the traditional eyeball technique is that probability estimates can be attributed to the accuracy of the statements.

Another advantage of the regressions procedure used is ability to test the curvilinear relationships above and beyond linear ones, which is not feasible with the eyeball technique on multiple baseline analysis. Similarily, one cannot test to see if the slopes of the control group are significantly different statistically.

In addition, as demonstrated in this paper we can also test to see if the functional relationship of one treatment is significantly different from the functional relationship of some other treatment (across some area of interest).

These advantages represent only some of the additional information which can be obtained through statistical analysis of operant data.

Table 2

RESULTS OF REGRESSION ANALYSIS

RESEARCH QUESTION	MODEL	R^2	df_n	df_d	α	F	P
Is there a significant mean difference between Control Group 1 and Control Group 2?							
$Y_1 = a_0U + a_1(C_1) + a_2(C_2) + E$	Full	.66	1	58	.05	94.28	.00001
$Y_1 = a_0U + E$	Restricted	0					
Is there a significant difference between the slope of Control Group 1 and Control Group 2 above and beyond individual differences?							
$Y_1 = a_0U + a_1(D_{c_1}) + a_2(D_{c_2}) + a_3(P_1) + a_4(P_2) + a_5(P_3) + \ldots + a_{10}(P_8) + E$	Full	.66	1	50	.05	.03	.84
$Y_1 = a_0U + a_{12}(D_{c_{1+2}}) + a_{13}(P_1) + a_{14}(P_2) + \ldots + a_{20}(P_8) + E$	Restricted	.66					

Table 2 Continued

RESEARCH QUESTION	MODEL	R^2	df n	df d	α	F	P
Is there a significant mean difference between Control Group 1 and Control Group 3?							
$Y_1 = a_0U + a_1(C_1) + a_2(C_3) + E$	Full	.46	1	78	.05	65.71	.00001
$Y_1 = a_0U + E$	Restricted	0					
Is there a significant difference between the slope of Control Group 1 and Control Group 3 above and beyond individual differences?							
$Y_1 = a_0U + a_1(D_{c_1}) + a_2(D_{c_3}) + a_3(P_1) + \ldots + a_6(P_4) + \ldots + a_{11}(P_9) + \ldots + a_{14}(P_{12}) + E$	Full	.46	1	70	.05	.0001	1.0000
$Y_1 = a_0U + a_1(D_{c_{1+c_3}}) + a_2(P_1) + \ldots + a_5(P_4) + \ldots + a_{10}(P_9) + \ldots + a_{13}(P_{12}) + E$	Restricted	.46					

Table 2 Continued

RESEARCH QUESTION	MODEL	R^2	df n	df d	α	F	P
Is there a significant mean difference between Control Group 2 and Control Group 3?							
$Y_1 = a_0U + a_1(C_2) + a_2(C_3) + E$	Full	.29	1	98	.05	41.42	.00001
$Y_1 = a_0U + E$	Restricted	0					
Is there a significant difference between the slope of Control Group 2 and Control Group 3 above and beyond individual differences?							
$Y_1 = a_0U + a_1(D_{c_2}) + a_2(D_{c_3}) + a_3(P_5) + \ldots + a_{10}(P_{12}) + E$	Full	.49	1	90	.05	.31	.57
$Y_1 = a_0U + a_{11}(D_{c_{2+3}}) + a_{12}(P_{24}) + \ldots + a_{20}(P_{31}) + E$	Restricted	.49					

Table 2 Continued

RESEARCH QUESTION	MODEL	R^2	df n	df d	α	F	P
Is there a significant mean difference between Control Group 1 and Experimental Group 1 above and beyond individual differences?							
$Y_1 = a_0U + a_1(C_1) + a_2(E_1) + a_3(P_1) + \ldots + a_6(P_4) + E$	Full	.56	1	71	.05	46.31	.00001
$Y_1 = a_0U + a_1(P_1) + \ldots + a_4(P_4) + E$	Restricted	.27					
Is there a significant second degree curvilinear relationship for Control Group 1 and Experimental Group 1 above and beyond a linear relationship and any individual differences?							
$Y_1 = a_0U + a_1(D_{c_1}) + a_2(D_c)^2 + a_3(D_{e_1}) + a_4(D_{T_1})^2 + a_5(P_1) + \ldots + a_8(P_4) + E$	Full	.58	2	68	.05		.258
$Y_1 = a_0U + a_1(D_{c_1}) + a_2(D_{T_1}) + a_3(P_1) + \ldots + a_6(P_4) + E$	Restricted	.57					

Table 2 Continued

RESEARCH QUESTION	MODEL	R^2	df n	df d	α	F	P
Is there a significant mean difference between Control Group 2 and Experimental Group 2 above and beyond individual differences?							
$Y_1 = a_0U + a_1(C_2) + a_2(E_2) + a_3(P_5) + \ldots + a_6(P_8) + E$	Full	.59	1	71	.05	17.17	.0001
$Y_1 = a_0U + a_1(P_5) + \ldots + a_3(P_8) + E$	Restricted	.50					
Is there a significant second degree functional relationship for Control Group 2 and Experimental Group 2 above and beyond a linear relationship and any individual differences?							
$Y_1 = a_0U + a_1(D_{c_2}) + a_2(D_{c_2})^2 + a_3(D_{T_2}) + a_4(D_{T_2})^2 + a_5(P_5) + \ldots + a_8(P_8) + E$	Full	.63	2	68	.05	.82	.44
$Y_1 = a_0U + a_1(D_{c_2}) + a_2(D_{T_2}) + a_3(P_5) + \ldots + a_6(P_8) + E$	Restricted	.62					

Table 2 Continued

RESEARCH QUESTION	MODEL	R^2	df n	df d	α	F	P
Is the mean of Control Group 3 significantly different from the mean of Experimental Group 3 above and beyond any individual differences?							
$Y_1 = a_0U + a_1(C_3) + a_2(E_3) + a_3(P_9) + \ldots + a_7(P_{13}) + E$	Full	.42	1	71	.05	4.67	.03
$Y_1 = a_0U + a_1(P_9) + \ldots + a_4(P_{12}) + E$	Restricted	.39					
Is there a significant second degree functional relationship for Control Group 3 and Experimental Group 3 above and beyond a linear relationship and any individual differences?							
$Y_1 = a_0U + a_1(D_{c_3}) + a_2(D_{c_3})^2 + a_3(D_{T_3}) + a_4(D_{T_3})^2 + a_5(P_9) + \ldots + a_9(P_{13}) + E$	Full	.45	2	68	.05	1.18	.31
$Y_1 = a_0U + a_1(D_{c_3}) + a_2(D_{T_3}) + a_3(P_9) + a_7(P_{12}) + E$	Restricted	.44					

Table 2 Continued

RESEARCH QUESTION	MODEL	R^2	df n	df d	α	F	P
Is there a significant difference between the slope of Control Group 1 and Experimental Group 1 above and beyond any individual differences?							
$Y_1 = a_0U + a_1(D_{c_1}) + a_2(D_{T_1}) + a_3(P_1) + \ldots + a_7(P_5) + E$	Full	.57	1	70	.05	8.19	.005
$Y_1 = a_0U + a_1(D_{c_{1+T_1}}) + a_2(P_1) + \ldots + a_6(P_5) + E$	Restricted	.52					
Is there a significant difference between the slope of control Group 2 and Experimental Group 2 above and beyond any individual differences?							
$Y_1 = a_0U + a_1(D_{c_2}) + a_2(D_{T_2}) + a_3(P_5) + \ldots + a_7(P_9) + E$	Full	.63	1	70	.05	4	.04
$Y_1 = a_0U + a_1(D_{c_{2+T_2}}) + a_2(P_5) + \ldots + a_5(P_8) + E$	Restricted	.61					

Table 2 Continued

RESEARCH QUESTION	MODEL	R^2	df n	df d	α	F	P
Is there a significant difference between the slope of Control Group 3 and Experimental Group 3 above and beyond any individual differences?							
$Y_1 = a_0U + a_1(D_{c_3}) + a_2(D_{T_3}) + a_3(P_1)$ $+a_7(P_{12}) + E$	Full	.48	1	70	.05	7.14	.01
$Y_1 = a_0U + a_1(D_{c_{3+T_3}}) + a_2(P_9) + a_6(P_{12}) + E$	Restricted	.43					

Table 3

THE FIFTEEN TESTED QUESTIONS

Hypothesis Number	
1.	There appears to be a significant mean difference between Control Group 1 and Control Group 2.
2.	There is no apparent slope difference between Control Group 1 and Control Group 2 above and beyond individual differences.
3.	There appears to be a significant mean difference between Control Group 1 and Control Group 3.
4.	There is no apparent slope difference between Control Group 1 and Control Group 3 above and beyond individual differences.
5.	There is no apparent mean difference between Control Group 2 and Control Group 3.
6.	There is no apparent slope difference between Control Group 2 and Control Group 3 above and beyond individual differences.
7.	There appears to be a significant mean difference between Control Group 1 and Experimental Group 1 above and beyond individual differences.
8.	Not applicable.
9.	There appears to be a significant mean difference between Control Group 2 and Experimental Group 2 above and beyond individual differences.
10.	Not applicable.
11.	There appears to be a significant mean difference between Control Group 3 and Experimental Group 3 above and beyond individual differences.
12.	Not applicable.
13.	There appears to be a significant difference between the slope of Control Group 1 and Experimental Group 1 above and beyond individual differences.

Table 3 Continued

Hypothesis Number	
14.	There appears to be a significant difference between the slope of Control Group 2 and Experimental Group 2 above and beyond individual difference.
15.	There appears to be a significant difference between the slope of Control Group 3 and Experimental Group 3 above and beyond individual differences.

REFERENCES

Baer, D. D., Wolf, M. M. and Risley, T. R. "Some Current Dimensions of Applied Behavior Analysis." Journal of Applied Behavior Analysis, 1968, 1, 91-97.

Engle, T. L. and Snellgrove, L. Psychology: Its Principles and Applications. Harcourt, Brace, and World, Inc., 1969.

Hall, R. V. Managing Behavior, Behavior Modification: The Measurement of Behavior. Part I, pp. 24. H. & H. Enterprises, Inc., P. O. Box 3342, Lawrence, Kansas 66044.

Hall, R. V. Managing Behavior, Behavior Modification: Basic Principles. Part II, H. & H. Enterprises, Inc., P. O. Box 3342, Lawrence, Kansas 66044.

Hall, R. V., Cristler, C., Cranston, S. S. and Tucker, B. "Teachers and Parents as Researchers Using Baseline Designs." Journal of Applied Behavior Analysis, 1970, 3, 247-55.

Kelly, F. J., Beggs, D. L., McNeil, K. A., Eichelberger, T. and Lyon, J. Research Design in the Behavioral Sciences: Multiple Regression Approach. Carbondale, Illinois: Southern Illinois University Press, 1969.

Kelly, F. J., Newman, I. and McNeil, K. A. Suggested inferential statistical models for research in behavioral modification. The Journal of Experimental Education, (in press).

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Vol. 4, No. 4, 1974

APPLYING THE GENERAL LINEAR MODEL TO REPEATED MEASURES PROBLEMS

John T. Pohlman

and

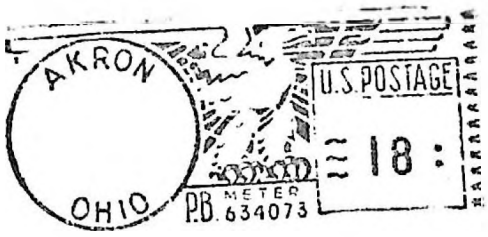
Michael G. McShane

*NOTE: this article was not submitted on time for publication.

It is the policy of the SIG=Multiple Linear Regression and of Viewpoints to consider for publication articles dealing with the theory and the application of Multiple Linear Regression. Manuscripts should be submitted to the Editor as an original, double-spaced typed copy. 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.

Judy T. McNeil, Chairman
SIG=Multiple Linear Regression
Isadore Newman, Editor
Multiple Linear Regression Viewpoints

The University of Akron
Akron, Ohio 44325



BOOKS - SPECIAL 4TH CLASS RATE

TABLE OF CONTENTS