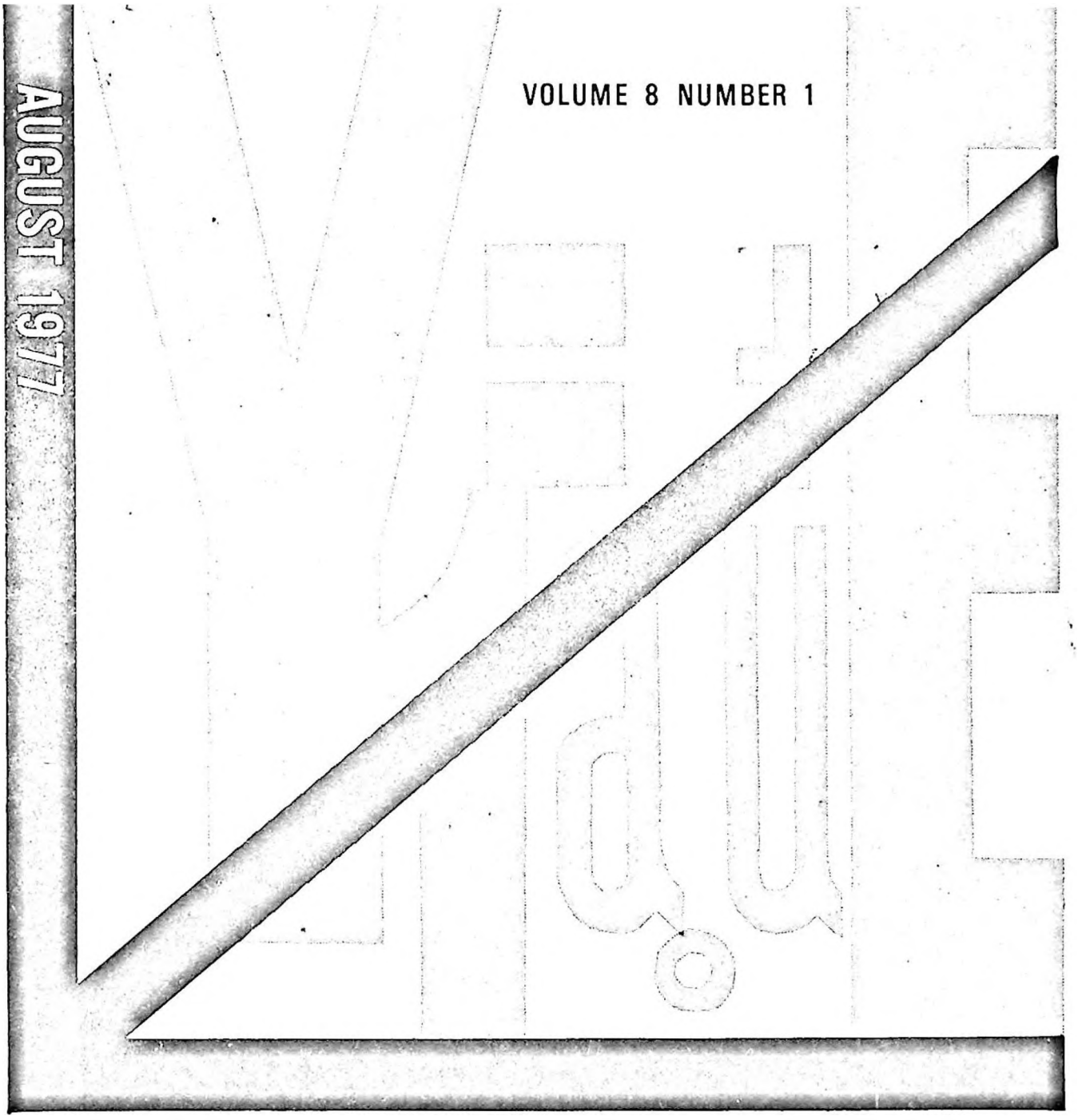
VOLUME 8 NUMBER 1

AUGUST 1977



# MULTIPLE LINEAR REGRESSION VIEWPOINTS

A publication of the Special Interest Group
on Multiple Linear Regression

# TABLE OF CONTENTS

Multiple Linear Regression Viewpoints
Volume 8 Number 1

# FULL RANK AND NON-FULL RANK MODELS WITH CONTRAST AND BJ BINARY CODING SYSTEMS FOR TWO-WAY DISPROPORTIONATE CELL FREQUENCY ANALYSES

JOHN D. WILLIAMS
The University of North Dakota

The two-way non-orthogonal design has been a source of considerable controversy. Several recent publications have emphasized the full rank model solution and discouraged the use of the fitting constants solution, the hierarchical model and the unadjusted main effects solution. By using a cell coding system instead of an effects coding system, a full rank model different from that of Timm and Carlson (1975) is found: this model was first suggested by Jennings (1967). The second full rank solution can be found to be computationally identical to the unadjusted main effects solution.

Speed and Hocking (1976) made a considerable contribution to the two-way disproportionate cell literature; in it, they describe both non-full rank and full rank solutions for the two-way disproportionate cell frequencies situation. They make five points worthy of reiteration:

1) Using the full rank model, the main effect solutions are not always unique (nor always defined);

2) The R's obtained by the two types of models often yield different solutions:

3) The hypotheses being tested are unclear;

4) It is possible to misinterpret what is being tested; and

5) Several types of contrasts are not possible using this approach.

Finally, Speed and Hocking point out that the description of Option 9 of the SPSS program (Nie et al., 1975) for the analysis of variance with disproportionate cells does not correspond to the actual solution executed by the program. As important, and useful, as the Speed and Hocking article is, it fails to clarify another problem with the full rank model; when binary coding is used, the results vary from the contrast coding scheme; also, the main effect $R^2$ (and sums of squares) values vary depending on how the binary coding was accomplished; this difficulty has also been pointed out by Klimko (1976). To demonstrate the various concerns, the data

in Table 1 are utilized; these data were taken from Williams (1972) in describing three non-full rank solutions and were re-considered by Timm and Carlson (1975) in a full rank solution.

TABLE 1

DATA FOR TWO-WAY DISPROPORTIONATE CELL FREQUENCIES

| | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $A_1$ | 8 | 1 | 6 |
| | 6 | 1 | 2 |
| | 4 | | |
| | 10 | 7 | 10 |
| $A_2$ | | 5 | 9 |
| | | 4 | 7 |
| | | 4 | 5 |
| | | 3 | 4 |

First, several vectors are defined that are subsequently used in the various analyses:

$Y$ = the criterion;

BINARY CODING

$X_1$ = 1 if a member of $A_1$, 0 otherwise;

$X_2$ = 1 is a member of $A_2$, 0 otherwise;

$X_3$ = 1 if a member of $B_1$, 0 otherwise;

$X_4$ = 1 if a member of $B_2$, 0 otherwise;

$X_5$ = 1 if a member of $B_3$, 0 otherwise;

$X_6 = X_1 \cdot X_3$;

$X_7 = X_1 \cdot X_4$;

$X_8 = X_1 \cdot X_5$;

$X_9 = X_2 \cdot X_3$;

$X_{10} = X_2 \cdot X_4$; and

$X_{11} = X_2 \cdot X_5$.

CONTRAST CODING

$X_{12}$ = 1 if a member of $A_1$, -1 if a member of $A_2$;

$X_{13}$ = 1 if a member of $B_1$, 0 if a member of $B_2$, -1 if a member of $B_3$;

$X_{14}$ = 0 if a member of $B_1$, 1 if a member of $B_2$, -1 if a member of $B_3$;

$X_{15} = X_{12} \cdot X_{13}$; and

$X_{16} = X_{12} \cdot X_{14}$.

Any of the following models yield identical $R^2$ values and sums of squares (SS):

EQUATIONS FOR FULL MODELS:

$$Y = b_0 + b_1X_1 + b_3X_3 + b_4X_4 + b_6X_6 + b_7X_7 + e_1; \quad (1a)$$

$$Y = b_0 + b_1X_1 + b_3X_3 + b_5X_5 + b_6X_6 + b_8X_8 + e_1; \quad (1b)$$

$$Y = b_0 + b_1X_1 + b_4X_4 + b_5X_5 + b_7X_7 + b_8X_8 + e_1; \quad (1c)$$

$$Y = b_0 + b_2X_2 + b_3X_3 + b_4X_4 + b_9X_9 + b_{10}X_{10} + e_1; \quad (1d)$$

$$Y = b_0 + b_2X_2 + b_3X_3 + b_5X_5 + b_9X_9 + b_{11}X_{11} + e_1; \quad (1e)$$

$$Y = b_0 + b_2X_2 + b_4X_4 + b_5X_5 + b_{10}X_{10} + b_{11}X_{11} + e_1; \quad (1f) \text{ and}$$

$$Y = b_0 + b_{12}X_{12} + b_{13}X_{13} + b_{14}X_{14} + b_{15}X_{15} + b_{16}X_{16} + e_1. \quad (1g)$$

For equations 1a - 1g, $R_1^2 = .61212$, $SS_1 = 80.80$.

EQUATIONS FOR A EFFECT:

$$Y = b_0 + b_1X_1 + e_2; \quad (2a)$$

$$Y = b_0 + b_2X_2 + e_2; \quad (2b) \text{ and}$$

$$Y = b_0 + b_{12}X_{12} + e_2. \quad (2c)$$

For equations 2a - 2c, $R_2^2 = .15427$, $SS_2 = 20.36$.

EQUATIONS FOR B EFFECT:

$Y = b_0 + b_3X_3 + b_4X_4 + e_3$ ; (3a)

$Y = b_0 + b_3X_3 + b_5X_5 + e_3$; (3b)

$Y = b_0 + b_4X_4 + b_5X_5 + e_3$; (3c) and

$Y = b_0 + b_{13}X_{13} + b_{14}X_{14} + e_3$. (3d)

For equations 3a - 3d, $R_3^2 = .28355$, $SS_3 = 37.43$.

EQUATIONS FOR COMBINED A AND B EFFECTS:

$Y = b_0 + b_1X_1 + b_3X_3 + b_4X_4 + e_4$; (4a)

$Y = b_0 + b_1X_1 + b_3X_3 + b_5X_5 + e_4$; (4b)

$Y = b_0 + b_1X_1 + b_4X_4 + b_5X_5 + e_4$; (4c)

$Y = b_0 + b_2X_2 + b_3X_3 + b_4X_4 + e_4$; (4d)

$Y = b_0 + b_2X_2 + b_3X_3 + b_5X_5 + e_4$; (4e)

$Y = b_0 + b_2X_2 + b_4X_4 + b_5X_5 + e_4$; (4f) and

$Y = b_0 + b_{12}X_{12} + b_{13}X_{13} + b_{14}X_{14} + e_4$. (4g)

For equations 4a - 4g, $R_4^2 = .60796$, $SS_4 = 80.25$.

Also, for equations 1a-4g, the values of the $b_I$'s will in general be different from equation to equation. For each equation, $SS_T = 132.00$. Equations like those here are often used to generate the fitting constants solution (see Anderson and Bancroft, 1952 and Overall and Spiegel, 1969), the hierarchical model (see Cohen, 1968) and the unadjusted main effects solution (see Jennings, 1967 and Williams, 1972). All three solutions are integrated into Table 2.

TABLE 2

TWO-WAY SOLUTION FOR DISPROPORTIONATE

CELL FREQUENCY DATA (NON-FULL RANK)

| Source of Variation | df | SS | MS | F | $R^2$ |
|---|---|---|---|---|---|
| A | 1 | 20.36 | 20.36 | 4.77 | .15427 |
| A (independent of B) | 1 | 42.82 | 42.82 | 10.03 | .32441 |
| B | 2 | 37.43 | 18.71 | 4.41 | .28355 |
| B (independent of A) | 2 | 59.89 | 29.95 | 7.01 | .45369 |
| AB | 2 | .55 | .28 | .07 | .00416 |
| Within | 12 | 51.20 | 4.27 | | |
| Total | 17 | 132.00 | | | |

The A (independent of B) effect is found as $SS_4 - SS_3 = 80.25 - 37.43$, or alternatively, as $SS_T (R^2_{A,B} - R^2_B)$.

The B (independent of A) effect is found as $SS_4 - SS_2 = 80.25 - 20.36$, or alternatively as $SS_T (R^2_{A,B} - R^2_A)$. The fitting constants solution uses the following sources of variation: $SS_A$ (independent of B), $SS_B$ (independent of A), $SS_{AB}$ AND $SS_{within}$. The hierarchical model uses either $SS_A$, $SS_B$ (independent of A), $SS_{AB}$, $SS_{within}$ and $SS_T$ or $SS_B$, $SS_A$ (independent of B), $SS_{AB}$, $SS_{within}$ and $SS_T$; the solution used depends upon which effect (A or B) is to be found first. The unadjusted main effects solution uses $SS_A$, $SS_B$, $SS_{AB}$ and $SS_{within}$.

Some authors prefer to include both types of sources of variation [for example, both $SS_A$ and $SS_A$ (independent of B)] in a decision base. Applebaum and Cramer (1974) have given a decision tree for such a situation. Perhaps the most widely known multiple decision base is that of Searle (1971). Searle considers 16 different possible outcomes and gives a table for interpreting those outcomes (p.278).

## THE USE OF FULL RANK MODELS

The publication of Timm and Carlson (1975) gave a major impetus to the usage of full rank models. Overall and Spiegel (1969) originally rejected the use of full rank models because such models failed to yield results that agreed with conventional analyses of proportional but non-equal cell frequency designs. However, Overall, Spiegel and Cohen (1975) reversed their position by recommending the full rank model solution for non-orthogonal designs for two reasons. First, they present data wherein they artificially create non-orthogonality by duplicating all entries in some cells but not others and show that only the full rank models yield the same parameter estimates in both the orthogonal and non-othogonal cases. Also, they point out that non-full rank models implicitly assume the non-existence of interaction effects, whereas the full rank models neccessarily take interaction into account when estimating main effects.

One might hazard a guess that statistical practice, which formerly emphasized the use of the fitting constants solution (see also Rao, 1965, and Winer, 1971), has seen the pendulum begin the swing to the use of full rank models. While one could dismiss the disagreement among statisticians as simply an intra-fraternity squabble, the more important issue is the effect this "squabble" has on non-members of the statistical fraternity who may be more inclined to look for conclusions rather than the reasoning process that lead to a particular conclusion. While it is easy to say, "Let the researcher be aware of the drawbacks of a particular methodology, and choose his/her strategy on the basis of this knowledge," all too often, the important issue regards the effect upon decision-makers in research, i.e., journal editors and referees. Other things being equal, it would appear that an article has a higher probability

of being accepted by a journal in a substantive area if it uses methodologies that agree with "standard statistical practice." While the present writer is in full agreement with the precept that researchers should choose their analytic techniques on the basis of their fully understanding the strengths and weaknesses of the analytic techniques, unfortunately, decision-makers in research (journal editors and referees), under the guise of "maintaining the standards of quality" sometimes deny the researcher this option.

## FULL RANK MODELS FOR A EFFECT

The following models would appear to generate a full rank solution for the row (A) main effect as a restriction of the corresponding full model (1a - 1g):

$$Y = b_0 + b_3X_3 + b_4X_4 + b_6X_6 + b_7X_7 + e_5; \quad (5a)$$

$$Y = b_0 + b_3X_3 + b_5X_5 + b_6X_6 + b_8X_8 + e_6; \quad (5b)$$

$$Y = b_0 + b_4X_4 + b_5X_5 + b_7X_7 + b_8X_8 + e_7; \quad (5c)$$

$$Y = b_0 + b_3X_3 + b_4X_4 + b_9X_9 + b_{10}X_{10} + e_8; \quad (5d)$$

$$Y = b_0 + b_3X_3 + b_5X_5 + b_9X_9 + b_{11}X_{11} + e_9; \quad (5e)$$

$$Y = b_0 + b_4X_4 + b_5X_5 + b_{10}X_{10} + b_{11}X_{11} + e_{10}; \quad (5f) \text{ and}$$

$$Y = b_0 + b_{13}X_{13} + b_{14}X_{14} + b_{15}X_{15} + b_{16}X_{16} + e_{11}. \quad (5g)$$

The following $R^2$s and sums of squares result and are shown in Table 3.

TABLE 3

RESULTS FROM USING DIFFERENT FULL RANK MODELS

FOR FINDING A EFFECT

| Equation | $R^2$ | SS | $R_1^2 - R_5^2$ | $SS_1 - SS_5$ | F |
|---|---|---|---|---|---|
| 5a | .51472 | 67.94 | .09740 | 12.86 | 3.01 |
| 5b | .47186 | 62.29 | .14026 | 18.51 | 4.33 |
| 5c | .52121 | 68.80 | .09091 | 12.00 | 2.81 |
| 5d | .51472 | 67.94 | .09740 | 12.86 | 3.01 |
| 5e | .47186 | 62.29 | .14026 | 18.51 | 4.33 |
| 5f | .52121 | 68.80 | .09091 | 12.00 | 2.81 |
| 5g | .30070 | 39.69 | .31142 | 41.11 | 9.63 |

Clearly, the results from the use of models 5a - 5g are different from one another; only those for 5g (using contrast coding) agree with those of Timm and Carlson (1975); this state of affairs is made more understandable by reference to Speed and Hocking's (1976) paper.

If we concentrate on the regression coefficients for (say) 1a, we can gather an idea regarding the restriction made in 5a.

Equation 1a could be rewritten as

$$Y = (\bar{Y}_{13} - \bar{Y}_{23})X_1 + (\bar{Y}_{21} - \bar{Y}_{23})X_3 + (\bar{Y}_{22} - \bar{Y}_{23})X_4 + (\bar{Y}_{11} - \bar{Y}_{13} - \bar{Y}_{21} + \bar{Y}_{23})X_6 + (\bar{Y}_{12} - \bar{Y}_{13} - \bar{Y}_{22} + \bar{Y}_{23})X_7 + e_1. \quad (6a)$$

Finding this solution for the regression coefficients follows from finding the expected values for each cell and then solving the simultaneous equations:

$E(Y_{11}) = b_0 + b_1 + b_3 + b_6 = \bar{Y}_{11}$;

$E(Y_{12}) = b_0 + b_1 + b_4 + b_7 = \bar{Y}_{12}$;

$E(Y_{13}) = b_0 + b_1 = \bar{Y}_{13}$;

$E(Y_{21}) = b_0 + b_3 = \bar{Y}_{21}$;

$E(Y_{22}) = b_0 + b_4 = \bar{Y}_{22}$;

$E(Y_{23}) = b_0 = \bar{Y}_{23}$.

Thus, $b_0 = \bar{Y}_{23}$, $b_1 = \bar{Y}_{13} - \bar{Y}_{23}$, $b_3 = \bar{Y}_{21} - \bar{Y}_{23}$, $b_4 = \bar{Y}_{22} - \bar{Y}_{23}$, $b_6 = \bar{Y}_{11} - b_0 - b_1 - b_3 = \bar{Y}_{11} - \bar{Y}_{13} - \bar{Y}_{21} + \bar{Y}_{23}$ and $b_7 = \bar{Y}_{12} - b_0 - b_1 - b_4 = \bar{Y}_{12} - \bar{Y}_{13} - \bar{Y}_{22} + \bar{Y}_{23}$.

Using equation 5a as the restricted model is testing the hypothesis $\bar{Y}_{13} - \bar{Y}_{23} = 0$; clearly, this is not the same as testing the A effect. It can be shown that the restrictions for equations 5a - 5g on equations 1a - 1g test the following hypotheses (given in Table 4):

TABLE 4

HYPOTHESES TESTED BY RESTRICTIONS IN 5a - 5g

| Equation | Associated Full Model | Hypothesis Being Tested |
|---|---|---|
| 5a | 1a | $\bar{Y}_{13} - \bar{Y}_{23} = 0$ |
| 5b | 1b | $\bar{Y}_{12} - \bar{Y}_{22} = 0$ |
| 5c | 1c | $\bar{Y}_{11} - \bar{Y}_{21} = 0$ |
| 5d | 1d | $\bar{Y}_{23} - \bar{Y}_{13} = 0$ |
| 5e | 1e | $\bar{Y}_{22} - \bar{Y}_{12} = 0$ |
| 5f | 1f | $\bar{Y}_{21} - \bar{Y}_{11} = 0$ |
| 5g | 1g | $\frac{\bar{Y}_{11} + \bar{Y}_{12} + \bar{Y}_{13}}{3} - \bar{Y}_c = 0$ |

Note: $\bar{Y}_c$ is the unweighted mean of the cell means.

Only equation 5g tests the hypothesis for the row main effect. Equations 5a - 5f test individual multiple comparisons that are related to, but not synonymous with the row main effect. For example, the F value for hypothesis 5a is 3.01; $\sqrt{3.01} = 1.735$ which is both the test of significance for $b_1$, in equation 1a and the test of significance for either Scheffé's test or Dunn's test, following the methodology in Williams (1976). It could be concluded that only equation 5g properly could be considered to be a full rank model; the models given by equations 5a - 5f could be considered to be pseudo-full rank models.

## FULL RANK MODELS FOR COLUMN EFFECT

Models could be written corresponding to equations 1a - 1g as the column (B effect) restrictions. Intuition would lead one to suspect that the restrictions on equations 1a - 1f would lead to pseudo-full rank model restrictions, in a manner similiar to the row effects; in this case, the intuitive judgment would appear to be justified. The column restrictions could be given as:

$Y = b_0 + b_1X_1 + b_6X_6 + b_7X_7 + e_{12}$; (6a)

$Y = b_0 + b_1X_1 + b_6X_6 + b_8X_8 + e_{13}$; (6b)

$Y = b_0 + b_1X_1 + b_7X_7 + b_8X_8 + e_{14}$; (6c)

$Y = b_0 + b_2X_2 + b_9X_9 + b_{10}X_{10} + e_{15}$; (6d)

$Y = b_0 + b_2X_2 + b_9X_9 + b_{11}X_{11} + e_{16}$; (6e)

$Y = b_0 + b_2X_2 + b_{10}X_{10} + b_{11}X_{11} + e_{17}$; (6f) and

$Y = b_0 + b_{12}X_{12} + b_{15}X_{15} + b_{16}X_{16} + e_{18}$. (6g)

Table 5 gives the hypotheses being tested, the $R^2$ values, the sum of squares, the difference in $R^2$s between the full and restricted models and the corresponding sum of squares.

TABLE 5

HYPOTHESES BEING TESTED, $R^2$ VALUES, SUMS OF SQUARES, DIFFERENCES IN $R^2$ VALUES IN FULL AND RESTRICTED MODELS AND CORRESPONDING SUMS OF SQUARES

| EQUATION | HYPOTHESES | $R^2$ | SS | $R^2_1 - R^2_6$ | $SS_1 - SS_6$ |
|---|---|---|---|---|---|
| 6a | $\bar{Y}_{21} - \bar{Y}_{23} = 0$; $\bar{Y}_{22} - \bar{Y}_{23} = 0$ | .38154 | 50.36 | .23058 | 30.44 |
| 6b | $\bar{Y}_{21} - \bar{Y}_{22} = 0$; $\bar{Y}_{23} - \bar{Y}_{22} = 0$ | .38154 | 50.36 | .23058 | 30.44 |
| 6c | $\bar{Y}_{22} - \bar{Y}_{21} = 0$; $\bar{Y}_{23} - \bar{Y}_{21} = 0$ | .38154 | 50.36 | .23058 | 30.44 |
| 6d | $\bar{Y}_{11} - \bar{Y}_{13} = 0$; $\bar{Y}_{12} - \bar{Y}_{13} = 0$ | .38485 | 50.80 | .22727 | 30.00 |
| 6e | $\bar{Y}_{11} - \bar{Y}_{12} = 0$; $\bar{Y}_{13} - \bar{Y}_{12} = 0$ | .38485 | 50.80 | .22727 | 30.00 |
| 6f | $\bar{Y}_{12} - \bar{Y}_{11} = 0$; $\bar{Y}_{13} - \bar{Y}_{11} = 0$ | .38485 | 50.80 | .22727 | 30.00 |
| 6g | $\frac{\bar{Y}_{11} + \bar{Y}_{21}}{2} - \bar{Y}_c = 0$; $\frac{\bar{Y}_{12} + \bar{Y}_{22}}{2} - \bar{Y}_c = 0$ | .18750 | 24.75 | .42462 | 56.05 |

NOTE: The associated full models are respectively 1a - 1g.

Again, only the equations in 6g test the column main effect hypotheses. One conclusion that could be drawn from these findings is that if main effects are of interest in the presence of interaction, then the contrast coding scheme is to be used rather than a simple binary system. On the other hand, it is of at least of passing interest to note that many experimenters prefer not to interpret main effects whenever significant interactions exist.

## RELATIONSHIP TO THE SPSS OPTION 9

The difficulty with Option 9 of the SPSS ANOVA program was alluded to earlier; actually, Option 9 will execute one of the sets of solutions for full rank models. Six different sets of solutions will be possible depending upon the way in which the data is coded. One such solution

would have the row effect given by 5a ($SS_A$ = 12.86) and the column effect given by 6a ($SS_B$ = 30.44). The other five solutions would correspond to 5b, 6b - 5f, 6f. Consequently, the hypotheses tested would correspond to those given here. Unfortunately, these differences in the solutions are not described in any available SPSS publication, nor is there any indication which of the solutions is to be given. While the solutions given may be of some usefulness, the lack of specific information renders them to be of doubtful usefulness to the typical SPSS user.

## ANOTHER LOOK AT THE UNADJUSTED MAIN EFFECT SOLUTION

Most recent authors have shown a preference for either the full rank model solution or the fitting constants solution. Either they fail to consider the unadjusted main effects solution or dismiss it as unworthy of extensive investigation. However, consider the following:

$$Y = b_1X_{11} + b_2X_{12} + b_3X_{13} + b_4X_{21} + b_5X_{22} + b_6X_{23} + e_1 \quad (7)$$

where

$X_{11}$ = 1 if a member of the row 1 column 1 cell, 0 otherwise;

$X_{12}$ = 1 if a member of the row 1 column 2 cell, 0 otherwise;

$X_{13}$ = 1 if a member of the row 1 column 3 cell, 0 otherwise;

$X_{21}$ = 1 if a member of the row 2 column 1 cell, 0 otherwise;

$X_{22}$ = 1 if a member of the row 2 column 2 cell, 0 otherwise;

$X_{23}$ = 1 if a member of the row 2 column 3 cell, 0 otherwise;

$b_1$ - $b_6$ are regression coeficients.

A hypothesis likely to be of interest for testing the row main effects is the following (if the $n_{ij}$ are equal for all cells): $b_1 + b_2 + b_3 = b_4 + b_5 + b_6$. That is, the average effect for the first row equals the average effect for the second row. For unequal $n_{ij}$'s, the restriction can be written

$$\frac{n_{11}b_1 + n_{12}b_2 + n_{13}b_3}{n_{11} + n_{12} + n_{13}} = \frac{n_{21}b_4 + n_{22}b_5 + n_{23}b_6}{n_{21} + n_{22} + n_{23}} \quad (8)$$

If equation 8 is solved in terms of (say) $b_1$, the result is

$$b_1 = \frac{(n_{11} + n_{12} + n_{13})(n_{21}b_4 + n_{22}b_5 + n_{23}b_6)}{n_{11}(n_{21} + n_{22} + n_{23})} - \frac{n_{12}b_2}{n_{11}} - \frac{n_{13}b_3}{n_{11}} \quad (9)$$

If the substitution of the right hand side of equation 9 is made for $b_1$ into equation 8, the result is

$$Y = \left[\frac{(n_{11} + n_{12} + n_{13})(n_{21}b_4 + n_{22}b_5 + n_{23}b_6)}{n_{11}(n_{21} + n_{22} + n_{23})} - \frac{n_{12}b_2}{n_{11}} - \frac{n_{13}b_3}{n_{11}}\right] X_{11}$$

$$+ b_2X_{12} + b_3X_{13} + b_4X_{21} + b_5X_{22} + b_6X_{23} + e_{19}.$$

After rearranging terms,

$$Y = b_2\left(X_{12} - \frac{n_{12}X_{11}}{n_{11}}\right) + b_3\left(X_{13} - \frac{n_{13}X_{11}}{n_{11}}\right)$$

$$+ b_4\left(X_{21} + \frac{n_{21}(n_{11} + n_{12} + n_{13})}{n_{11}(n_{21} + n_{22} + n_{23})} X_{11}\right) + b_5\left(X_{22} + \frac{n_{22}(n_{11} + n_{12} + n_{13})}{n_{11}(n_{21} + n_{22} + n_{23})} X_{11}\right)$$

$$+ b_6\left(X_{23} + \frac{n_{23}(n_{11} + n_{12} + n_{13})}{n_{11}(n_{21} + n_{22} + n_{23})} X_{11}\right) + e_{19}. \quad (10)$$

Except for the change in the notational scheme, equation 10 is identical to Jennings' (1967) Model VII.

Equations 7 and 10 can be implemented directly in programs such as Ward and Jennings' (1973) MODEL or McNeil, Kelly and McNeil's (1975) LINEAR. However, the usual general purpose multiple regression program will fail to invert the matrix due to a dependency relationship caused by the inclusion of the unit vector. A slight modification will allow a solution to occur; if any <u>one</u> of the predictors in both equations 7 and 10 are deleted (i.e., set $b_6 = 0$ in both equations) then a useful result is found. The results are $R^2_7 = .61212$, $SS_7 = 80.80$; $R^2_{10} = .45785$,

$SS_{10} = 60.44$. Thus, $R^2_7 - R^2_{10} = .15427$, $SS_7 - SS_{10} = 20.36$, results identical to those given earlier for the unadjusted main effects solution for rows. It can then be seen that the use of the computational procedure for the unadjusted main effects solution yields $R^2_s$ and sums of squares that are identical to a hypothesis (restriction) that is likely to be of interest.

A similiar finding will also occur from the likely column main effect hypotheses:

$$\frac{n_{11}b_1 + n_{21}b_4}{n_{11} + n_{21}} = \frac{n_{12}b_2 + n_{22}b_5}{n_{12} + n_{22}} = \frac{n_{13}b_3X_{13} + n_{23}b_6}{n_{13} + n_{23}} . \quad (11)$$

Using an algebraic logic as was shown for the row effects, equation 11 results in two restrictions on equation 7 that yield $R^2_{11} = .32857$, $SS_{11} = 43.37$, so that $R^2_7 - R^2_{11} = .28355$, $SS_7 - SS_{11} = 37.43$ (after deleting any one of the remaining predictors), results identical for the unadjusted main effects solution for columns.

## ADVANTAGES AND DISADVANTAGES OF THE DIFFERENT CODING SYSTEMS

To explicate the entire set of ramifications of different coding systems is beyond the scope of the present paper. However, if the researcher is interested only in the usual analysis of variance components and is not interested in the full rank model solution, the coding system is not too important an issue. If, however, a full rank solution is of interest, one fairly easy way of obtaining this solution is through contrast coding. This is not to say the full rank solution is beyond the capabilities of the binary coding system; if the hypotheses in 5g and 6g are incorporated into the model as restrictions on equation 7, the same results occur. The present writer has a preference for binary

coding both for heuristic (that is classroom teaching) and hypothesis testing ease. However, the contrast coding scheme does have a direct interpretation in traditional analysis of variance hypotheses testing and can prove attractive for those situations.

Finally, another note should be made here regarding terminology. The term contrast coding has been used throughout; others, such as Cohen and Cohen (1975) refer to this process as effect coding.

## ADVANTAGES AND DISADVANTAGES OF THE DIFFERENT SOLUTIONS

The full rank model appears to be getting greater acceptance within the research community. Its greatest advantage is that it does not require the assumption of a non-existent interaction. Each main effect is measured in the presence of the interaction effect. Among the disadvantages of the full rank model (as proposed by Timm and Carlson, 1975) are the lack of an additive solution except in the case of equal cell frequencies. This method appears to be in a position to begin dominating the other methods for disproportionate cell data; indeed, Dalton (1977) used the full rank model solution as the criterion to judge the adequacy of other solutions in an article in VIEWPOINTS.

The fitting constants solution has enjoyed the distinction of being considered to be "the" solution to the disproportionate cell frequency case and is widely described in standard advanced statistical methodology texts. Each main effect is found independent of the other main effect. When the cell frequencies are disproportional, the fitting constants solution is non-additive. The hierarchical model (Cohen, 1968) can be useful if the researcher clearly has a preferred order of the main effects. It should be noted that some researchers (e.g. Cohen) recommend

that the most important effect be the first fitted effect, and other researchers (e.g. Applebaum and Cramer) recommend the most important effect be the second fitted effect. Because of this controversy, the hierarchical model is not used to the same extent as the fitting constants solution. The hierarchical model is an additive model. The unadjusted main effects solution appears to suffer from the criticism that the main effects are allowed to be found wherein a dependency relationship may be occurring between them.

If we allow researchers to take a pragmatic view that the unadjusted main effects solution is being used only as a computational convenience, and the actual hypotheses being tested are those given in equations 8 and 11, then a quite different point of view can be taken. If the researcher states the hypotheses as those given by equations 8 and 11, and begins with equation 7, then the researcher can quite properly say that a full rank solution (but different from Timm and Carlson's) has been found. It would seem reasonable to allow a researcher who is fully cognizant of the hypotheses being tested to use a computational shortcut if that shortcut yields a solution equivalent to the hypotheses to be tested. Thus, the unadjusted main effect solution can be seen to be useful as a computational aid in a meaningful test made on a full rank model.

## A FINAL NOTE ON OPTION 9

It appears that a recent (April, 1977) updated version of Option 9 has deleted the previous version and replaced it with a method that gives results that agree with Timm and Carlson's full rank model.

REFERENCES


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

Appelbaum, M.I. and Cramer, E.M. Some problems in the non-orthogonal analysis of variance. Psychological Bulletin, 1974, 81, 335-343.

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

Cohen, J. and Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. New York: Lawrence Erlbaum Associates, 1975.

Dalton, S. Regression approaches and approximate solutions to analysis of variance with disproportionality varied. Multiple Linear Regression Viewpoints, 1977, 7, No. 2, 16-32.

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

Klimko, L. Unbalanced data. The American Statistician, 1976, 30, 205-206.

McNeil, K.A., Kelly, F.J. and McNeil, J.T. Testing research hypotheses using multiple linear regression. Carbondale, Ill.: Southern Illinois Press, 1975.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K.S., and Bent, D.H. Statistical package for the social sciences, 2nd ed. New York: McGraw-Hill, 1975.

Overall, J.E. and Spiegel, D.H. Concerning least squares analysis of experimental data. Psychological Bulletin, 1969, 72, 311-322.

Overall, J.E. Spiegel, D.H. and Cohen, J. Equivalence of orthogonal and nonorthogonal analysis of variance. Psychological Bulletin, 1975, 82, 182-186.

Rao, C.R. Linear Statistical inference and its applications. New York: Wiley, 1965.

Searle, S.R. Linear models. New York: Wiley, 1971.

Speed, F.A, and Hocking, R.R. The use of R ()- Notation with unbalanced data. The American Statistician, 1976, 30, 30-33.

Timm, N.H. and Carlson, J.E. Analysis of variance through full rank models. Multivariate Behavioral Research: Monograph, 1975, 75-1.

Ward, J.H. and Jennings, E. Introduction to linear models. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

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

Williams, J.D. Multiple comparisons by multiple linear regression. Multiple Linear Regression Viewpoints Monograph Series, 2, 1976, 7, No. 1.

Winer, B.J. Statistical prinicples in experimental design. 2nd ed. New York: McGraw-Hill, 1971.

Multiple Linear Regression Viewpoints
Volume 8 Number 1

# INTERVAL ESTIMATION OF THE* POPULATION SQUARED MULTIPLE CORRELATION

JOHN T. POHLMANN & JAMES F. MOORE
Southern Illinois University, Carbondale

Key words: Squared multiple correlation, Regression analysis, Interval estimation

Abstract

A technique is presented which applies the Neyman theory of confidence intervals to interval estimation of the squared multiple correlation. The technique makes use of the equivalence between $R^2$ and F, $\rho^2$ and $\lambda$, the non-centrality parameter of the non-central F distribution. A computer program is also presented which can be used to apply the technique.

The squared multiple correlation ($R^2$) is one of the most frequently used statistics in social science and educational research, yet it can also be a very biased estimator of the population squared multiple correlation ($\rho^2$). The bias in the point estimation of $\rho^2$ is a function of the variables/observations ratio. Formulas have been developed to provide unbiased point estimates of $\rho^2$ given a sample $R^2$ derived from a multiple regression equation with p linearly independent predictors, exclusive of the unit vector, on n observations. Olkin and Pratt (1958) developed an exact formula for the unbiased point estimation of $\rho^2$, and the commonly used shrinkage formula (Cohen & Cohen, 1975, p. 106) is a very good approximation of the Olkin and Pratt formula.

---

*One of the reviewers stated that Formula (6) is the formular for large degrees of freedom for both $V_1$ and $V_2$. We could not find any other formula. If anyone in MLRV audience knows of such an apporopriate formula, we would appreciate receiving their information.

A procedure for interval estimation of $\rho^2$ has not been satisfactorily developed for the non-statistician researcher. Tables and charts for interval estimation of $\rho$, the multiple correlation coefficient, are available in the literature (Ezekiel & Fox, 1959; Kramer, 1963) for selected values of p and n, but the values of p and n in these tables are very limited.

Hypothesis testing with $R^2$s is easily done for hypotheses of the form $\rho^2 = 0$, with the central F distribution, but in many cases this is a meaningless hypothesis to reject. The approach presented here permits the testing of hypotheses of the form $\rho^2 = C$, where $0 \leq C < 1$, since hypothesis testing can be considered as a special application of interval estimation procedures. Finally, interval estimation of $\rho^2$ better allows for the comparison of findings from different studies in which regression analysis was used. If the interval for $\rho^2$ from one study overlaps the confidence interval for $\rho^2$ from another study, one might conclude the studies replicate each other with respect to the parameter $\rho^2$.

## Mathematical Development

The distribution function of $(R^2:\rho^2)$ was formulated by Fisher (1928) as follows:

Let $a = \frac{1}{2}p$, $b = \frac{1}{2}(n-p-1)$ and F denote the hypergeometric function.

$$(1)\quad H(R^2) = (1-\rho^2)^{a+b}(R^2)^a \sum_{p=0}^{b-1} \frac{(a+p+1)!(1-R^2)^p F(-p,-b,a,R^2\rho^2)}{(a-1)!p!(1-R^2\rho^2)^{a+b+p}}$$

Unfortunately (1) is only applicable when (n-p-1) is an even number, and (1) was the formula used by Ezekiel and Fox (1959) and Kramer (1963) to develop their tables. They solved (1) for .95 and reported the resulting R values, i.e. the 95th percentile R value given $\rho^2$. A user who then observes an R value at the 95th percentile on $H(R^2:\rho_1)$, concludes with a confidence level of 95% that $\rho \geq \rho_1$.

If (1) is used to form confidence intervals for $\rho^2$, one would solve for two values of $\rho^2$, $\rho_1^2$ and $\rho_2^2$, for fixed $R^2$, n and p such that formula (1) would equal $\alpha/2$, and $1-\alpha/2$ respectively. According to Neyman (Cramér, 1946) these two values for $\rho^2$ would define a $1-\alpha$ confidence region for $\rho^2$.

Formula (1) could be used to form confidence intervals for $\rho^2$, but it is limited to cases when n-p-1 is an even number, and the hypergeometric function and the factorial terms make it a difficult formula to solve with large n. For these reasons another approach was attempted which capitalized upon the relationship between the F distribution and the distribution of $R^2$. For example, F and $R^2$ can be expressed as direct functions of one another.

$$(2) \quad F_{(v_1,v_2)} = \frac{R^2/v_1}{(1-R^2)/v_2}$$

$$(3) \quad R^2 = \frac{v_1(F_{(v_1,v_2)})}{v_2+v_1F_{(v_1,v_2)}}$$

where $v_1 = p$ and $v_2 = n-p-1$

The relationship between $\rho^2$ and the non-central F distribution can be seen through the following equalities:

$$(4) \quad \lambda = \frac{n\rho^2}{1-\rho^2}$$

$$(5) \quad \rho^2 = \frac{\lambda}{\lambda+n}$$

where $\lambda$ is the non-centrality parameter of the non-central F distribution (Kendall & Stewart, 1973).

Finally, a function which contained these values that was amenable to easy solution was found in a normal approximation to the non-central F distribution (Zelen and Severo, 1960):

$$(6) \quad z = \frac{\left(\frac{v_1 F}{v_1+\lambda}\right)^{1/3}\left(1-\frac{2}{9v_2}\right) - \left[1 - \frac{2(v_1+2\lambda)}{9(v_1+\lambda)^2}\right]}{\left[\frac{2(v_1+2\lambda)}{9(v_1+\lambda)^2} + \frac{2}{9v_2}\left(\frac{v_1 F}{v_1+\lambda}\right)^{2/3}\right]^{1/2}}$$

The use of (6) to form confidence intervals on $\rho^2$ is accomplished as follows:

1. Given: $v_1 = p$

   $v_2 = n-p-1$

   $F_{(v_1,v_2)} = \frac{R^2/v_1}{(1-R^2)/v_2}$

   $1-\alpha$ = the confidence coefficient

2. Solve (6) iterating on $\lambda$ until $z = z_{\alpha/2}$ and again until $z = z_{1-\alpha/2}$. The two values of $\rho^2 = \lambda/(\lambda+n)$ thus obtained define the $1-\alpha$ confidence limits for $\rho^2$.

The accuracy of this approach is primarily dependent on the convergence criterion established for z. If four significant digits on z are used, the reported upper and lower limits on $\rho^2$ will have two significant digits. This was the convergence criterion used in the program presented with this paper. The results using this approach have been checked against the results reported by Ezekiel and Fox (1959) and Kramer (1963) and have been found to be accurate.

Table 1 shows the results of this algorithm applied to selected values of $R^2$, n (the sample size), and p (the number of linearly independent predictors). Inspection of Table 1 reveals the degree of bias involved in estimating $\rho^2$ with $R^2$, especially when p is large relative to n. Consider the case when p = 20, and n = 50. The upper limit of each of the 90% confidence intervals is less than the observed $R^2$ value. This degree of bias makes the use of $R^2$ as an estimator of $\rho^2$ questionable to say the least.

---

insert Table 1 about here

---

## SUMMARY

The approach presented here provides a mechanism for the formal statistical comparison of $R^2$s from similar regression studies. Also, the technique allows for a more refined inferential analysis of the squared multiple correlation. With its use, a researcher can make a more reasonable estimate of $\rho^2$, rather than having to rely on the shrinkage formula to make

only a point estimate of $\rho^2$. The limits on $\rho^2$ provided by this method can also be used to test hypotheses about $\rho^2$ other than the common hypothesis $\rho^2 = 0$. Consider the hypothesis $\rho^2 = C$ $(0 \leq C < 1)$. If the confidence interval reported here does not contain C, then the hypothesis can be rejected.

This method can also be used to estimate confidence intervals for increments in $\rho^2$ when comparing full and restricted regression models, by rewriting (2) and (4) as follows:

$$(7) \quad F = ((R_f^2 - R_r^2)/v_1)/(1 - R_f^2)/v_2$$

$$(8) \quad \lambda = n(\rho_f^2 - \rho_r^2)/(1 - \rho_f^2)$$

where $R_f^2$ = an $R^2$ for a full model

$R_r^2$ = an $R^2$ for a model containing a linear restriction on the model producing $R_f^2$

$\rho_f^2$ and $\rho_r^2$ = the parameter counterparts of $R_f^2$ and $R_r^2$.

Since the fixed effects analysis of variance (ANOVA) is a special case of least squares general linear model analysis, certain statistics formed in the context of ANOVA can also be analyzed with this technique. The $\eta^2$ statistic $(SS_A/SS_T)$ and its unbiased estimator $\omega^2$ (Hays, 1973) are comparable to $R^2$ and the shrunken $R^2$. In fact, $\eta^2$ and $R^2$ possess the same sampling distribution (Kendall & Stewart, 1973). Hence, $\eta^2$ may be substituted for $R^2$ in the preceding development and confidence intervals may be formed for its parameter.

# REFERENCES

Byars, J. A. and Roscoe, J. T. Rational approximations of the inverse gaussian function. Paper presented at the annual meeting of the American Educational Research Assn., Chicago, 1972.

Cohen, J. and Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. New York: John Wiley & Sons, 1975.

Cramér, H. Mathematical methods of statistics. Princeton: Princeton Univ. Press, 1946.

Ezekiel, M. J. B. and Fox, K. A. Methods of correlation and regression analysis: linear and curvilinear. New York: John Wiley & Sons, 1959.

Fisher, R. A. The general sampling distribution of the multiple correlation coefficient. Proceedings of the Royal Society, 1928, 121, 654-673.

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart & Winston. 1973.

Kendall, M. G. and Stuart, A. The advanced theory of statistics, London: Griffin & Co., 1973.

Kramer, K. H. Tables for constructing confidence limits on the multiple correlation coefficient. Journal of the American Statistical Assn., 1963, 58, 1082-1087.

Olkin, I. and Pratt, J. W. Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics. 1958, 29, 201-210.

Zelen, M. and Severo, N. C. Probability functions. In Handbook of Mathematical Functions (National Bureau of Standards Applied Mathematics Series No. 55). Washington: U.S. Government Printing Office, 1964, 932-151

Table 1

LIMITS FOR THE 90% CONFIDENCE INTERVAL
FOR SELECTED VALUES OF $R^2$, N, AND P.

| $R^2$= | .1 | | .3 | | .5 | | .7 | | .9 | |
|---|---|---|---|---|---|---|---|---|---|---|
| p/ N | $\rho_L^2$ | $\rho_U^2$ | $\rho_L^2$ | $\rho_U^2$ | $\rho_L^2$ | $\rho_U^2$ | $\rho_L^2$ | $\rho_U^2$ | $\rho_L^2$ | $\rho_U^2$ |
| 2/ 10 | .00 | .28 | .00 | .50 | .00 | .66 | .14 | .79 | .60 | .93 |
| 2/ 30 | .00 | .24 | .05 | .45 | .23 | .62 | .49 | .78 | .82 | .93 |
| 2/ 50 | .00 | .21 | .11 | .43 | .31 | .60 | .56 | .77 | .85 | .92 |
| 2/100 | .02 | .18 | .17 | .40 | .37 | .58 | .61 | .75 | .87 | .92 |
| 2/200 | .04 | .16 | .21 | .37 | .42 | .56 | .64 | .74 | .88 | .91 |
| 4/ 10 | .00 | .00 | .00 | .35 | .00 | .56 | .00 | .74 | .33 | .91 |
| 4/ 30 | .00 | .17 | .00 | .42 | .17 | .60 | .44 | .76 | .80 | .92 |
| 4/ 50 | .00 | .18 | .08 | .41 | .28 | .59 | .54 | .76 | .84 | .92 |
| 4/100 | .00 | .17 | .15 | .39 | .36 | .57 | .60 | .75 | .86 | .92 |
| 4/200 | .03 | .15 | .20 | .37 | .41 | .56 | .64 | .74 | .88 | .91 |
| 6/ 10 | .00 | .00 | .00 | .01 | .00 | .38 | .00 | .64 | .00 | .88 |
| 6/ 30 | .00 | .09 | .00 | .37 | .10 | .57 | .39 | .75 | .78 | .92 |
| 6/ 50 | .00 | .13 | .04 | .38 | .24 | .57 | .51 | .75 | .83 | .92 |
| 6/100 | .00 | .15 | .14 | .37 | .35 | .56 | .59 | .74 | .86 | .91 |
| 6/200 | .02 | .14 | .19 | .36 | .40 | .55 | .63 | .73 | .88 | .91 |
| 8/ 30 | .00 | .00 | .00 | .32 | .02 | .54 | .32 | .73 | .75 | .91 |
| 8/ 50 | .00 | .09 | .00 | .35 | .20 | .55 | .48 | .74 | .82 | .91 |
| 8/100 | .00 | .13 | .12 | .36 | .33 | .56 | .58 | .74 | .86 | .91 |
| 8/200 | .01 | .13 | .19 | .35 | .40 | .55 | .63 | .73 | .88 | .91 |
| 8/300 | .03 | .13 | .21 | .35 | .42 | .54 | .65 | .73 | .88 | .91 |
| 10/ 30 | .00 | .00 | .00 | .25 | .00 | .49 | .24 | .70 | .71 | .90 |
| 10/ 50 | .00 | .03 | .00 | .32 | .16 | .53 | .45 | .73 | .81 | .91 |
| 10/100 | .00 | .10 | .10 | .35 | .32 | .55 | .57 | .73 | .85 | .91 |
| 10/200 | .01 | .12 | .18 | .35 | .39 | .54 | .63 | .73 | .87 | .91 |
| 10/300 | .03 | .13 | .21 | .34 | .42 | .54 | .64 | .73 | .88 | .91 |
| 15/ 30 | .00 | .00 | .00 | .00 | .00 | .35 | .00 | .62 | .57 | .87 |
| 15/ 50 | .00 | .00 | .00 | .22 | .04 | .47 | .36 | .69 | .77 | .90 |
| 15/100 | .00 | .04 | .05 | .31 | .27 | .52 | .55 | .72 | .84 | .91 |
| 15/200 | .00 | .10 | .16 | .33 | .37 | .53 | .62 | .72 | .87 | .91 |
| 15/300 | .01 | .11 | .19 | .33 | .41 | .53 | .64 | .72 | .88 | .91 |
| 20/ 30 | .00 | .00 | .00 | .00 | .00 | .05 | .00 | .47 | .20 | .82 |
| 20/ 50 | .00 | .00 | .00 | .09 | .00 | .40 | .24 | .65 | .72 | .88 |
| 20/100 | .00 | .00 | .00 | .27 | .23 | .49 | .51 | .70 | .83 | .90 |
| 20/200 | .00 | .07 | .13 | .31 | .36 | .52 | .60 | .72 | .87 | .91 |
| 20/300 | .00 | .09 | .18 | .32 | .40 | .52 | .63 | .72 | .88 | .91 |

PROGRAM COFIN



```
C  ***************************************************************
C
C       SUBROUTINE COFIN
C
C          ALGORITHM DEVELOPED BY JOHN POHLMANN
C           SUPPLEMENTARY CODE BY JAMES MOORE
C            SOUTHERN ILLINOIS UNIVERSITY AT CARBONDALE
C
C
C   CALL COFIN (OBS,PRCT,,RSQF,PRDF,RSQR,PRDR,SRSQ)
C
C       ALL PARAMETERS ARE FLOATING POINT MODE
C
C       1NPUT PARAMETERS ARE
C
C             OBS=NUMBER OF OBSERVATIONS
C             PRCT=THE PERCENT OF CONFIDENCE FOR THE INTERVAL
C             RSQF=R**2 OR THE R**2 ASSOCIATED WITH THE FULL MODEL
C             PRDF=NUMBER OF PREDICTORS OR THE NUMBER OF PREDICTORS
C                    ASSOCIATED WITH THE FULL MODEL
C             RSQR=R**2 FOR THE RESTRICTED MODEL OR 0
C             PRDR=NUMBER OF PREDICTORS IN THE RESTRICTED
C                    MODEL OR 0
C             SRSQ=1 IF SINGLE R**2; OR 2 IF R**2FULL-R**2 RESTRICTED
C
C     KEY VARIABLE NAMES
C
C         REXP- SHRUNKEN R**2
C         V1- NUMERATOR DEGREES OF FREEDOM
C         V2- DENOMINATOR DEGREES OF FREEDOM
C         RLOW- LOWER BOUND OF THE CONFIDENCE INTERVAL
C         RHIGH- UPPER BOUND OF THE CONFIDENCE INTERVAL
C
C
C  ***************************************************************
      SUBROUTINE COFIN(OBS,PRCT,RSQF,PRDF,RSQR,PRDR,SRSQ)
C
C   CHECK R**2 TO BE SURE THEY ARE LESS THAN OR EQUAL TO 1
C
      IF(RSQF.LE.1.AND.RSQR.LE.1)GOTO1
      WRITE(6,25)
25    FORMAT(' R-SQUARE MUST BE LESS THAN OR EQUAL TO 1.0')
      RETURN
C
C   CALCULATE DEGREES OF FREEDOM FOR F-TEST
C
1     V1=PRDF-PRDR
      V2=OBS-PRDF-1.
```

```
      F=((RSQF-RSQR)/V1)/((1.0-RSQF)/V2)
      REXP=1.0-(1.0-(RSQF-RSQR))*((OBS-1.0)/V2)
      P1=(1.0-(PRCT/100.))/2.
      P2=1.-P1
C
C     BR IS BYARS AND ROSCOE'S INVERSE GAUSSIAN FUNCTION
C
      T1-BR(P1)
      T2=BR(P2)
C
C     RSQF ESTABLISHES THE BOUNDS OF THE CONFIDENCE INTERVAL
C
      CALL RSQP(V1,V2,F,T1,RHIGH)
      CALL RSQP(V1,V2,F,T2,RLOW)
      IF(SRSQ-2.)3,4,2
2     RETURN
3     WRITE(6,30)F,RSQF,RSQR,V1,RSQF,V2,REXP,PRCT,RLOW,RHIGH
30    FORMAT(//' F=',F8.4,'=((',F4.3,'-',F4.3,')/',F7.4,')/(1-',
     *F4.3,'/',F7.1,')',/' SHRUNKEN R-SQUARE=',F7.5/
     *' FOR THE ',F4.1,'% CONFIDENCE INTERVAL R-SQUARE LOW=',F6.2,' R-SQ
     *UARE HIGH=',F6.2)
      RLOW=SQRT(RLOW)
      RHIGH=SQRT(RHIGH)
      WRITE(6,35)RLOW,RHIGH
35    FORMAT('                                        MULT-R LOW=',F6.2,'   M
     *ULT-R HIGH=',F6.2)
      RETURN
4     WRITE(6,50)F,RSQF,RSQR,V1,RSQF,V2,REXP,PRCT,RLOW,RHIGH
50    FORMAT(//' F=',F8.4,'=((',F4.3,'-',F4.3,')/',F7.4,')/(1-',
     *F4.3,'/',F7.1,')'/' SHRUNKEN INCREMENT IN R-SQUARE (FULL-RESTRICTE
     *D(=',F7.5/' FOR THE ',F4.1,'% CONFIDENCE INTERVAL FOR THE PARTIAL
     *MULTIPLE R-SQUARE, LOW=',F6.2,/70X,'HIGH=',F6.2)
9     RETURN
      END
      SUBROUTINE RSQP(V1,V2,F,T,RP)
      REAL L
      XN2=V1+V2+1.0
      RP=.5
      D2=.25
      K=0
10    L=XN2*RP/(1.0-RP)
      K=K+1
      A=((V1*F)/(V1+L))**.333333
      B=(1.0-(2.0/(9.0*V2)))
      C=1.0-(2.0*(V1+2.0*L))/(9.0*((V1+L)**2.0))
      D=(2.0*(V1+2.0*L))/(9.0*((V1+L)**2.0))
      E=(2.0/(9.0*V2))*(((V1*F)/(V1+L))**.6666667)
      D=SQRT(D+E)
```

```
      Z=(A*B-C)/D
      C=(ABS(Z-T))
      IF(C.LT..0001)RETURN
      IF(K.EQ.30)RETURN
      IF(Z.GT.T)RP=RP+D2
      IF(Z.LT.T)RP=RP=D2
      D2=D2/2.0
      IF (RP.LT.0.0)RP=0.0
      IF(RP.GE.1.0)RP=.99999
      GOTO10
      END
      FUNCTION BR(P)
      R=P-.5000000
      Q=R*R
      BR=((2.505922+(-15.73223+23.54337*Q)*Q)*R)/(1.0+(-7.337743+
     *(14.97266-6.016088*Q)*Q)*Q)
      RETURN
      END
```

```
C  ***********************************************************************
C
C       SUBROUTINE COFIN
C
C          ALGORITHM DEVELOPED BY JOHN POHLMANN
C           SUPPLEMENTARY CODE BY JAMES MOORE
C            SOUTHERN ILLINOIS UNIVERSITY AT CARBONDALE
C
C
C    CALL COFIN (OBS,PRCT,,RSQF,PRDF,RSQR,PRDR,SRSQ)
C
C       ALL PARAMETERS ARE FLOATING POINT MODE
C
C       INPUT PARAMETERS ARE
C
C             OBS=NUMBER OF OBSERVATIONS
C             PRCT=THE PERCENT OF CONFIDENCE FOR THE INTERVAL
C             RSQF=R**2 OR THE R**2 ASSOCIATED WITH THE FULL MODEL
C             PRDF=NUMBER OF PREDICTORS OR THE NUMBER OF PREDICTORS
C                     ASSOCIATED WITH THE FULL MODEL
C             RSQR=R**2 FOR THE RESTRICTED MODEL OR 0
C             PRDR=NUMBER OF PREDICTORS IN THE RESTRICTED
C                     MODEL OR 0
C             SRSQ=1 IF SINGLE R**2; OR 2 IF R**2FULL-R**2 RESTRICTED
C
C      KEY VARIABLE NAMES
C
C          REXP- SHRUNKEN R**2
C          V1- NUMERATOR DEGREES OF FREEDOM
C          V2- DENOMINATOR DEGREES OF FREEDOM
C          RLOW- LOWER BOUND OF THE CONFIDENCE INTERVAL
C          RHIGH- UPPER BOUND OF THE CONFIDENCE INTERVAL
C
C
C  ***********************************************************************
      SUBROUTINE COFIN(OBS,PRCT,RSQF,PRDF,RSQR,PRDR,SRSQ)
C
C    CHECK R**2 TO BE SURE THEY ARE LESS THAN OR EQUAL TO 1
C
      IF(RSQF.LE.1.AND.RSQR.LE.1)GOTO1
      WRITE(6,25)
25    FORMAT(' R-SQUARE MUST BE LESS THAN OR EQUAL TO 1.0')
      RETURN
C
C    CALCULATE DEGREES OF FREEDOM FOR F-TEST
C
1     V1=PRDF-PRDR
      V2=OBS-PRDF-1.
```

```
      F=((RSQF-RSQR)/V1)/((1.0-RSQF)/V2)
      REXP=1.0-(1.0-(RSQF-RSQR))*((OBS-1.0)/V2)
      P1=(1.0-(PRCT/100.))/2.
      P2=1.-P1
C
C     BR IS BYARS AND ROSCOE'S INVERSE GAUSSIAN FUNCTION
C
      T1=BR(P1)
      T2=BR(P2)
C
C     RSQP ESTABLISHES THE BOUNDS OF THE CONFIDENCE INTERVAL
C
      CALL RSQP(V1,V2,F,T1,RHIGH)
      CALL RSQP(V1,V2,F,T2,RLOW)
      IF(SRSQ-2.)3,4,2
2     RETURN
3     WRITE(6,30)F,RSQF,RSQR,V1,RSQF,V2,REXP,PRCT,RLOW,RHIGH
30    FORMAT(//' F=',F8.4,'=((',F4.3,'-',F4.3,')/',F7.4,')/(1-',
     *F4.3,'/',F7.1,')',/' SHRUNKEN R-SQUARE=',F7.5/
     *' FOR THE ',F4.1,'% CONFIDENCE INTERVAL R-SQUARE LOW=',F6.2,' R-SQ
     *UARE HIGH=',F6.2)
      RLOW=SQRT(RLOW)
      RHIGH=SQRT(RHIGH)
      WRITE(6,35)RLOW,RHIGH
35    FORMAT('                                        MULT-R LOW=',F6.2,'   M
     *ULT-R HIGH=',F6.2)
      RETURN
4     WRITE(6,50)F,RSQF,RSQR,V1,RSQF,V2,REXP,PRCT,RLOW,RHIGH
50    FORMAT(//' F=',F8.4,'=((',F4.3,'-',F4.3,')/',F7.4,')/(1-',
     *F4.3,'/',F7.1,')'/' SHRUNKEN INCREMENT IN R-SQUARE (FULL-RESTRICTE
     *D)=',F7.5/' FOR THE ',F4.1,'% CONFIDENCE INTERVAL FOR THE PARTIAL
     *MULTIPLE R-SQUARE, LOW=',F6.2,/70X,'HIGH=',F6.2)
9     RETURN
      END
      SUBROUTINE RSQP(V1,V2,F,T,RP)
      REAL L
      XN2=V1+V2+1.0
      RP=.5
      D2=.25
      K=0
10    L=XN2*RP/(1.0-RP)
      K=K+1
      A=((V1*F)/(V1+L))**.333333
      B=(1.0-(2.0/(9.0*V2)))
      C=1.0-(2.0*(V1+2.0*L))/(9.0*((V1+L)**2.0))
      D=(2.0*(V1+2.0*L))/(9.0*((V1+L)**2.0))
      E=(2.0/(9.0*V2))*(((V1*F)/(V1+L))**.6666667)
      D=SQRT(D+E)
```

```
      Z=(A*B-C)/D
      C=(ABS(Z-T))
      IF(C.LT..0001)RETURN
      IF(K.EQ.30)RETURN
      IF(Z.GT.T)RP=RP+D2
      IF(Z.LT.T)RP=RP-D2
      D2=D2/2.0
      IF (RP.LT.0.0)RP=0.0
      IF(RP.GE.1.0)RP=.99999
      GOTO10
      END
      FUNCTION BR(P)
      R=P-.5000000
      Q=R*R
      BR=((2.505922+(-15.73223+23.54337*Q)*Q)*R)/(1.0+(-7.337743+
     *(14.97266-6.016088*Q)*Q)*Q)
      RETURN
      END
```

Multiple Linear Regression Viewpoints
Volume 8 Number 1

# A NOTE ON CODING THE SUBJECTS EFFECT IN TREATMENTS X SUBJECTS DESIGNS

JOHN D. WILLIAMS
The University of North Dakota

Abstract - Using a recent innovation described by Pedhazur (1977), a simpler regression solution to the repeated measure design is shown. Instead of coding N-1 vectors to represent the subject effect, the sum of each subject's criterion scores are entered as a vector. This single vector yields the same $R^2$ value as does the N-1 binary coded subject vectors.

Treatments X subjects designs, when executed in a regression framework, have typically had an associated cumbersome coding precess. A design with N subjects and k treatments has, in most formulations, N-1 vectors to represent the subjects effect. When the N becomes at all large, (say N > 50), the process easily gets out of hand.

Fortunately, Pedhazur (1977) has recently shown an alternative procedure that necessitates only one vector to represent the subjects effect. Basically, a vector is formed for each subject such that the entries for that subject are the sum of the subject's Y score values. This sum is entered separately for each Y score. To illustrate this coding process, consider the data in Table 1, taken from Williams (1974, p. 56), which has ten subjects and three treatments.

TABLE 1

Illustration of Design Matrix for Treatments X Subjects Designs

| Y | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 18 | 60 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | 60 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 60 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 55 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | 55 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 55 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 39 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 39 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 39 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 19 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8 | 19 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 6 | 19 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 11 | 35 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 14 | 35 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 10 | 35 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 9 | 29 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 12 | 29 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 8 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 14 | 45 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 16 | 45 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 15 | 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 12 | 38 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 17 | 38 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 9 | 38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 22 | 59 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 21 | 59 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | 59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 10 | 43 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 43 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

The values in Table 1 are defined as follows:

Y = the criterion score;

$X_1$ = the sum of the criterion scores for each subject separately;

$X_2$ = 1 if the score corresponds to Treatment 1, 0 otherwise;

$X_3$ = 1 if the score corresponds to Treatment 2, 0 otherwise;

$X_4$ = 1 if the score is obtained from Subject 1, 0 otherwise;

$X_5$ = 1 if the score is obtained from Subject 2, 0 otherwise;

$X_6$ = 1 if the score is obtained from Subject 3, 0 otherwise;

$X_7$ = 1 if the score is obtained from Subject 4, 0 otherwise;

$X_8$ = 1 if the score is obtained from Subject 5, 0 otherwise;

$X_9$ = 1 if the score is obtained from Subject 6, 0 otherwise;

$X_{10}$ = 1 if the score is obtained from Subject 7, 0 otherwise;

$X_{11}$ = 1 if the score is obtained from Subject 8, 0 otherwise;

$X_{12}$ = 1 if the score is obtained from Subject 9, 0 otherwise;

Note that treatment 3 and subject 10 do not have separate vectors, as they are linearly dependent respectively on $X_2$, $X_3$ and $X_4 - X_{12}$. The analysis in Williams (1974) proceeds as follows: three linear models are defined, one for the treatments effect, one for the subjects effect and one for the combined treatments and subjects effects. These models are given as

$$Y = b_0 + b_2 X_2 + b_3 X_3 + e_1, \quad (1)$$

$$Y = b_0 + b_4 X_4 + b_5 X_5 + b_6 X_6 + b_7 X_7 + b_8 X_8 + b_9 X_9 + b_{10} X_{10} + b_{11} X_{11} + b_{12} X_{12} + e_2, \quad (2)$$

and

$$Y = b_0 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + b_6 X_6 + b_7 X_7 + b_8 X_8 + b_9 X_9 + b_{10} X_{10} + b_{11} X_{11} + b_{12} X_{12} + e_3. \quad (3)$$

The associated $R^2$ values and sums of squares (SS) for equations 1-3 are $R_1^2 = .1784$; $SS_1 = 136.27$; $R_2^2 = .6823$; $SS_2 = 521.20$; $R_3^2 = .8607$; $SS_3 = 657.47$; $SS_T = 763.86$. A complete summary table is shown in Table 2.

Table 2

Summary Table for the Treatments X Subjects Design

| Source of Variation | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | 2 | 136.27 | 68.13 | 11.52 |
| Subjects | 9 | 521.20 | | |
| Error | 18 | 106.39 | 5.91 | |
| Total | 29 | 763.86 | | |

An alternative analysis, using $X_1$, the sum of criterion scores for each subject, would use the following equations:

$$Y = b_0 + b_2X_2 + b_3X_3 + e_1, \quad (1)$$

$$Y = b_0 + b_1X_1 + e_2, \quad (4) \text{ and}$$

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e_3. \quad (5)$$

The results, in terms of sums of squares and $R^2$ values, is identical to that already given, with $R^2_4 = R^2_2$, $SS_4 = SS_2$, $R^2_5 = R^2_3$ and $SS_5 = SS_3$. However, care must be taken with the degrees of freedom. Equation 4 uses only one predictor; thus, the "apparent" degrees of freedom is one. It must be remembered that the actual df = N - 1. Remembering the degrees of freedom is a small price to pay for a much more parsimonious solution to the repeated measures designs.

While only a simple treatments x subjects design has been shown here, the process works as well for higher dimensional completely crossed designs as well as for "mixed" designs.

## REFERENCES

Pedhazur, E.J. Coding subjects in repeated measure designs. Psychological Bulletin, 1977, 84, No. 2, 298-305.

Williams, J.D. Regression analysis in educational research. New York: MSS Publishing Co., 1974.

Multiple Linear Regression Viewpoints
Volume 8 Number 1

# AN INTRODUCTION TO PATH ANALYSIS

LEE M. WOLFLE
Virginia Polytechnic Institute and State University

Many causal analyses in the social sciences are now being conducted within the framework of path analysis, or more appropriately, the analysis of structural equation models. These methods are not new, and are commonly attributed in their development to some early writings of Sewell Wright (1921, 1925, 1934). The method was introduced to the social sciences within the past two decades by Blalock (1962, 1964, 1967, 1971), Boudon (1965, 1968), but most importantly by Duncan (1966, 1968, 1969a, 1969b, 1970, 1972, 1975). Since Duncan's 1966 article, the literature on path analysis has increased nearly exponentially, and owes much to contributions by Blau and Duncan (1967), Duncan, Featherman and Duncan (1972), Finney (1972), Goldberger (1972), Goldberger and Duncan (1973), Hauser and Goldberger (1971), and Heise (1969, 1972, 1975). Land (1969) has written an adequate introduction to the subject; Duncan (1975) and Heise (1975) have now produced good texts.

Yet the literature on path analysis in educational journals remains slim despite its heuristic advantages (but see Anderson and Evans, 1974; Williams and Klimpel, 1975; Wolfle, 1977). My purpose here is to provide an introductory discussion of path analysis, demonstrating how the numeric coefficients may be cal-

culated. The manner in which causal effects may be decomposed will be addressed, followed by a brief discussion of strategies of analysis in structural equation models. Finally, I will discuss some recent applications of path analysis to educational topics. In their several parts, little of what will follow is new--it has all been covered before--but the synthesis is new, and should provide an introduction to the method of path analysis for those who are interested in pursuing the matter further in some of the literature cited herein.

## CONSTRUCTION OF THE MODELS

Path analysis employs diagrams such as the following:

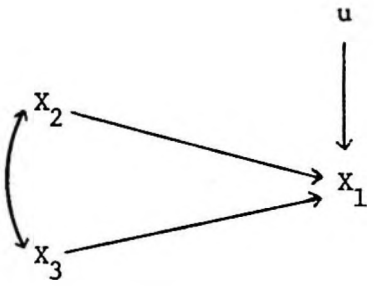



Figure 1.

A straight arrow represents the researcher's hypothesis of a causal effect; the arrowhead points toward the influenced variable. The arrow from $X_2$ to $X_1$, for example, represents the verbal statement "$X_2$ is a cause of $X_1$," or "a change in $X_2$ produces a change in $X_1$." The double-headed curved arrow represents a correlation, in this case between the exogenous variables $X_2$ and $X_3$, to which

we attach no causal interpretation. That is, we allow $X_2$ and $X_3$ to be associated for unknown reasons. The three variables, $X_1$, $X_2$, and $X_3$, are explicit variables which have names--income, occupational prestige, and educational attainment, for example. A fourth variable, u, in the model does not have an explicit name; it is called the "disturbance," or the "residual," or "error." It represents all other sources of variation in $X_1$ not jointly explained by $X_2$ and $X_3$. Such sources may include explicit variables not included in the model, deviations from linearity, random errors, and the like. The expected value of the residual is zero, and the expected covariations of the residual with the independent (or, in this case, exogenous) variables, $X_2$ and $X_3$, are zero. Notice that we have by definition constructed a model which accounts for all of the variation in $X_1$, some by explicitly named and measured variables, some by an unnamed disturbance term.

The diagram above actually represents only one of several models we could have constructed with these variables. We could have constructed a so-called chain model:

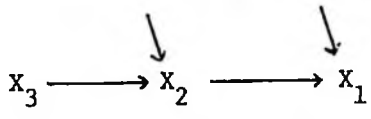



Figure 2.

The diagram represents a hypothesis that $X_3$ causes $X_2$ which causes $X_1$, but that there is no direct effect of $X_3$ on $X_1$. Or, we could allow for such a direct effect by the model:

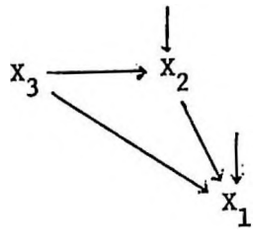



Figure 3.

Yet another alternative would suggest that $X_3$ causes $X_2$ and $X_1$, but that $X_1$ and $X_2$ have no causal effect on each other. Thus,

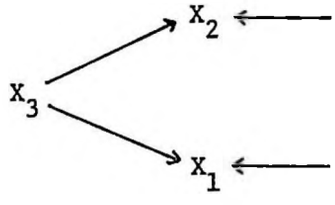



Figure 4.

Notice that we have considered four different models without reversing the inherent order of the three variables. We could also have considered several other models which allowed different inherent orders (see Duncan, 1975: 19). How may we decide which model is the correct one? The answer may come only from *a priori* considerations. Without some basis--temporal order, previous research, or theoretical conjecture, for example--we would become hopelessly enmeshed in a number of models, all of which the data might support as plausible representations of the variables. You should conclude that neither path analysis, nor any other method, provides a way of inferring causality from nonexperimental data. Path analysis does provide a method for attaching quantitative estimates to causal effects though to exist on *a priori* grounds. From

such quantitative estimates, we might conclude that the data are inconsistent with a certain model, and reject the model as implausible. But the data will not suggest an alternative.

## COMPUTING PATH COEFFICIENTS

How do we obtain the quantitative estimates, the path coefficients? Let's denote these coefficients by the symbol, $p_{ij}$, where $i$ represents the variable thought to be caused by $j$; that is, $i$ is the dependent variable. Returning to the model pictured in Figure 1, we know the sample correlation coefficients, $r_{12}$, $r_{13}$, and $r_{23}$; we assume $r_{2u} = r_{3u} = 0$, and want to obtain estimates of $p_{12}$ and $p_{13}$.

Translating the picture in Figure 1 into its corresponding equation, we have

$$X_1 = p'_{12}X_2 + p'_{13}X_3 + p'_{1u}u + a \qquad (1).$$

We may transform these variables into standard form by taking deviations from their respective means and dividing by the appropriate standard deviations. This transformation makes the following presentation more convenient, but is not necessary. Indeed, analysis of variables in their original metric is preferred (Blalock, 1972: 383-385; Duncan, 1975: chap. 4; Kim and Mueller, 1976). Let's indicate the standardized variables by using lowercase symbols. Thus

$$x_1 = p_{12}x_2 + p_{13}x_3 + p_{1u}u \qquad (2).$$

We may multiply this equation by $x_2$, which yields

$$x_1x_2 = p_{12}x_2x_2 + p_{13}x_2x_3 + p_{1u}x_2u \qquad (3).$$

These variable names represent observations on individuals; we may sum the observations, and divide by N. Doing so results in

$$\frac{\Sigma x_1 x_2}{N} = p_{12}\frac{\Sigma x_2 x_2}{N} + p_{13}\frac{\Sigma x_2 x_3}{N} + p_{1u}\frac{\Sigma x_2 u}{N} \qquad (4).$$

Because these are standardized variables, we have

$$r_{12} = p_{12} + p_{13}r_{23} \qquad (5),$$

since $r_{22} = 1$, and $r_{2u} = 0$ by assumption.

Similarly, we could have multipled equation (2) by $x_3$, which would yield

$$r_{13} = p_{12}r_{23} + p_{13} \qquad (6).$$

We now have two equations with two unknowns, and simple algebraic manipulation will show that

$$p_{12} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \qquad (7),$$

$$p_{13} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \qquad (8).$$

Returning to equation (2), we multiply through by u, intermediately yielding

$$r_{1u} = p_{12}r_{2u} + p_{13}r_{3u} + p_{1u}r_{uu} \qquad (9).$$

Because $r_{uu} = 1$, and $r_{2u} = r_{3u} = 0$ by assumption, we have

$$r_{1u} = p_{1u} \qquad (10),$$

which is generalizable to any number of independent variables. Finally, multiplying equation (2) by $x_1$ results in

$$r_{11} = 1 = p_{12}r_{12} + p_{13}r_{13} + p_{1u}r_{1u} \qquad (11),$$

and from equation (10) we find

$$p_{1u} = (1 - [p_{12}r_{12} + p_{13}r_{13}])^{1/2} \qquad (12).$$

The expression inside the brackets is the coefficient of determination, $R^2$, and we have

$$p_{1u} = \sqrt{1 - R^2_{1.23}} \qquad (13),$$

which is also generalizable to any number of independent variables.

Let's look again at the so-called normal equations, (5) and (6). These may conveniently be depicted in matrix notation:

$$\begin{bmatrix} r_{12} \\ r_{13} \end{bmatrix} = \begin{bmatrix} 1 & r_{23} \\ r_{23} & 1 \end{bmatrix} \begin{bmatrix} p_{12} \\ p_{13} \end{bmatrix} \qquad (14),$$

from which we have

$$\begin{bmatrix} p_{12} \\ p_{13} \end{bmatrix} = \begin{bmatrix} 1 & r_{23} \\ r_{23} & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{12} \\ r_{13} \end{bmatrix} \qquad (15)$$

This matrix equation (15) is easily generalized to $k$ independent variables. Thus, any computer program which can invert and multiply matrices can be used to obtain path coefficients. These coefficients are alternatively labeled standardized regression coefficients, or beta weights (BETA in the popular SPSS [Nie, *et al.*, 1975] regression package), and can be obtained from most standard regression programs.

Path analysis (in its recursive form), therefore, is nothing

more than regression analysis with standardized variables. Indeed, as mentioned above, the metric coefficients are now preferred in most cases to the standardized, which means path analysis *is* regression analysis. The advantage of path analysis may not, therefore, be found in the computation of coefficients, but arises from its requiring one to think about cause, particularly systems of intercausal connections. It is this ability to provide an explicit link between theoretical notions of what causes what, and estimates of causal impact, which is the greatest advantage of path analysis.

## DECOMPOSITON OF EFFECTS

Consider the hypothetical model in Figure 5.

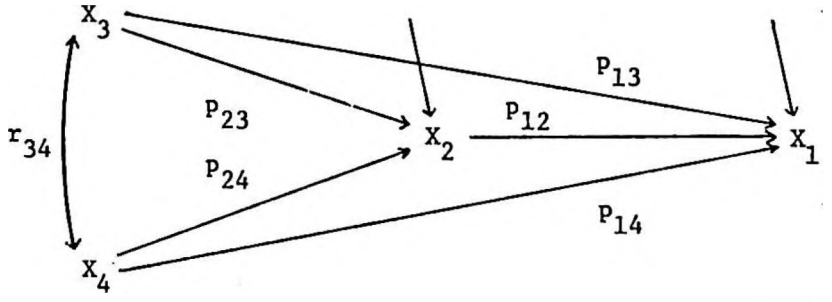



Figure 5.

The model depicts an *a priori* notion that $X_1$ depends on the explicit variables $X_2$, $X_3$, and $X_4$; that $X_2$ depends on $X_3$ and $X_4$; that $X_3$ and $X_4$ are correlated for unknown reasons, and whatever their causes may be, they come from outside the model (hence the phrase, "exogenous").

Path diagrams may be read as follows:

> Read *back* from variable $i$, *then forward to* variable $j$, forming the product of all paths along the traverse; then sum these products for all possible traverses. The same variable cannot be intersected more than once in a single traverse. In no case can one trace back having once started forward. The bidirectional correlation is used in tracing either forward or back, but if more than one bidirectional correlation appears in the diagram, only one can be used in a single traverse. The resulting expression . . . may consist of a single direct path plus the sum of several compound paths representing all the indirect connections allowed by the diagram (Duncan, 1966: 6).

When the variables in the model have been standardized, the sums of such products yield the correlation coefficient, $r_{ij}$. (When the variables are measured in their original metric, the coefficients are partial regression coefficients, and the sum of their products along the appropriate traverses equals the zero-order slope, $b_{ij}$. Of course, one may no longer use the correlation coefficient between exogenous variables, but must use the regression slopes, $b_{34}$ or $b_{43}$, as appropriate.)

Let's consider the correlation between $X_2$ and $X_3$ in Figure 5. Reading back from $X_2$ to $X_3$, one possible traverse is the path, $p_{23}$. Another is formed by the product, $p_{24}r_{34}$. These exhaust the traverses from $X_2$ to $X_3$; therefore

$$r_{23} = p_{23} + p_{24}r_{34} \qquad (16).$$

Except for the change in subscripts, this equation is identical to equation (5).

Now let's consider the decomposition of the correlation between $X_1$ and $X_3$. The resulting expression would be:

$$r_{13} = p_{13} + p_{12}p_{23} + p_{12}p_{24}r_{34} + p_{14}r_{34} \quad (17).$$

The advantage of these decompositions is that we may now attach *causal* interpretations to the various traverses. For example, the path, $p_{13}$, is the measure of a *direct effect* of $X_3$ on $X_1$. The product, $p_{12}p_{23}$, is a measure of an *indirect causal effect* of $X_3$ on $X_1$ through $X_2$. That is, not only does $X_3$ effect $X_1$ directly, but $X_3$ also has an effect on $X_2$, which in turn effects $X_1$. Let's give names to these variables for illustration; let $X_1$ = grade-point average, $X_2$ = attendance, $X_3$ = a measure of motivation, and $X_4$ = IQ. Assuming $p_{13}$, $p_{12}$, and $p_{23}$ are all positive, we would conclude that of the total relationship between motivation and grades, $r_{13}$, a portion, $p_{13}$, is a direct causal effect--the higher the motivation, the higher the grades. The indirect causal effect, $p_{12}p_{23}$, arises from the fact that the higher the motivation, the greater the attendance, *and* the greater the attendance, the higher the grades.

That leaves two parts of the total relationship between $X_1$ and $X_3$ yet to be discussed. The two traverses, $p_{12}p_{24}r_{34}$ and $p_{14}r_{34}$, both include the correlation $r_{34}$, to which we have previously decided *not* to attach any causal interpretations. We should be correspondingly hesitant to attach any causal interpretation to a traverse which includes such a term. There is no single label we may call this part of the decomposition, but perhaps a *joint association* is as good as any. In terms of the names we have attached to these variables, we would say, after direct and indirect effects that a portion of the relationship of motivation

and grades is due to the fact that motivation and IQ are correlated for unknown reasons, and that IQ also has effects, both direct and indirect, on grades.

Finally, let's examine the decomposition of the correlation of $X_1$ and $X_2$. The correlation may be decomposed into the following set of traverses:

$$r_{12} = p_{12} + p_{13}p_{23} + p_{14}p_{24} + p_{13}r_{34}p_{24} + p_{14}r_{34}p_{23} \qquad (18).$$

The path, $p_{12}$, is the usual direct effect; and two traverses include the correlation between exogenous variables, to which we attach no *causal* interpretation. But consider the traverse, $p_{13}p_{23}$; this part of the total relationship arises because $X_1$ and $X_2$ are both caused by $X_3$. That is, $X_3$ is an antecedent cause of both $X_1$ and $X_2$, and *that* part of the total relationship we call a *spurious effect* due to $X_3$. Similarly, there is a spurious effect due to $X_4$, namely $p_{14}p_{24}$. In substantively interpreting these variables, we would say that part of the association between attendance and grades is a spurious effect of motivation and IQ. That is, people with higher rates of attendance will receive higher average grades, but part of this association is due to the fact that highly motivated people have both greater attendance and higher grades, and part is due to the fact that more intelligent people also have greater attendance and higher grades.

Total relationships may, therefore, be decomposed through path analysis into direct effects, indirect effects, spurious effects, and joint associations. Direct effects are partial derivatives; indirect effects occur only through intervening

variables; spurious effects occur from joint antecedent variables; joint associations involve as one of their components a correlation between variables to which no causal interpretation may be attached. These decomposed effects may be expressed as proportions of the total relationship. For example, the proportion of the total relationship between $X_1$ and $X_2$ in Figure 5 which may be considered a direct causal effect is $p_{12}/r_{12}$. I mentioned above that regression coefficients in their metric form may be substituted for the path coefficients; indeed they are preferred. If regression coefficients are used, the proportion of the total relationship, $b_{12}$, which may be considered a direct causal effect is given by $b_{12.34}/b_{12}$, that is, the ratio of the partial regression coefficient controlling for $X_3$ and $X_4$, to the zero-order regression slope, $b_{12}$. These ratios, whether one uses standardized or metric coefficients, are equal. That is,

$$\frac{p_{ij.klm...}}{r_{ij}} = \frac{b_{ij.klm...}}{b_{ij}} \qquad (19).$$

The other components of the decomposition are also equivalent, regardless of whether path coefficients or regression coefficients are used.

Alwin and Hauser (1975) have explicated the decomposition of effects through the use of ordinary least squares regression. Good examples of substantive applications may be found in Alexander and McDill (1976), Alwin (1976), Duncan, Featherman, and

Duncan (1972), Featherman and Hauser (1976), Hauser, Sewell, and Alwin (1976), and Wolfle (1977).

Using the rules for reading path diagrams helps in appreciating the properties of the causal relationships in the model, but for algebraic manipulation and calculation, it is more convenient to use the fundamental theorem of path analysis, which may be written (Duncan, 1966: 5):

$$r_{ij} = \sum_q p_{iq} r_{jq} \qquad (20),$$

where $i$ and $j$ denote two variables in the model, and the index $q$ runs over all variables from which direct paths lead to $X_i$.

For example, the correlation between $X_1$ and $X_2$ in Figure 5 may be written in terms of equation (20):

$$r_{12} = p_{12}r_{22} + p_{13}r_{23} + p_{14}r_{24} \qquad (21).$$

Similarly,

$$r_{23} = p_{23}r_{33} + p_{24}r_{34} \qquad (22),$$

and

$$r_{24} = p_{24}r_{44} + p_{23}r_{34} \qquad (23).$$

Any variable correlated with itself is equal to 1. And substituting equations (22) and (23) into (21), we have

$$r_{12} = p_{12} + p_{13}(p_{23} + p_{24}r_{34}) + p_{14}(p_{24} + p_{23}r_{34}) \qquad (24).$$

Expanding this equation, we find

$$r_{12} = p_{12} + p_{13}p_{23} + p_{14}p_{24} + p_{13}p_{24}r_{34} + p_{14}p_{23}r_{34} \qquad (25),$$

which is equivalent to equation (18) which we obtained from decomposing the model by "reading" the appropriate traverses from $X_1$ to $X_2$.

## STRATEGIES OF ANALYSIS

### Recursive Equation Models

The models we have been considering so far have all been recursive. The label applies to models in which there are no feedback loops; that is, the causal flow in the model is unidirectional. Stated differently, it means that a variable cannot be both a cause and an effect of another variable, either directly or indirectly. Models that include reciprocal effects are called nonrecursive; estimates of coefficients for such models may not be obtained by ordinary least squares, and associations are not decomposable, as I have explicated above. Erlanger and Winsborough (1976) provide an example of how to solve a simple nonrecursive model with the two-stage least squares procedure.

The strategy of analysis for recursive models is twofold. First, one will want to obtain estimates of the extent to which intervening variables account for relationships among variables. These are the indirect effects discussed above. For example, in industrial societies much of the effect a father's socioeconomic status has on his son's socioeconomic achievement is not a direct effect, but occurs through the son's educational attainment (Blau and Duncan, 1967). The second strategy of analysis for recursive models is to obtain estimates of the extent to which antecedent

variables account for relationships between other variables. These are the spurious effects discussed above. For example, Duncan (1968) found that about 56% of the association (r = .59) between educational attainment and adult intelligence was a spurious effect of "early" intelligence.

Block Equation Models

A block equation model is one in which each of a set of dependent variables is regressed on the same set of independent variables. For example,

$$Y_1 = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + e_1$$

$$Y_2 = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + e_2$$

$$Y_3 = \alpha_3 + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_3 + e_3$$

$$Y_4 = \alpha_4 + \beta_{41}X_1 + \beta_{42}X_2 + \beta_{43}X_3 + e_4 \qquad (26).$$

The set of equations (26) may be diagrammed as shown in Figure 6.

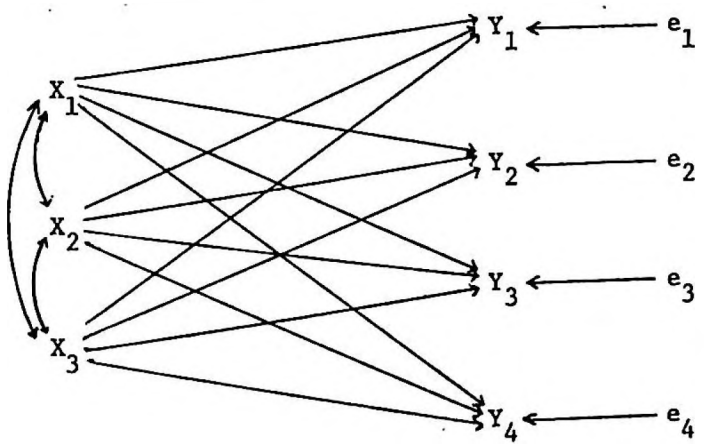



Figure 6.

The analysis goals for this kind of model are (1) to compare the zero-order coefficients with their corresponding partial coefficients in order to determine how much of the zero-order association may be considered a direct effect, and how much a joint association; and (2) to examine the residuals for correlated errors. That is, we want to determine how good a job the independent variables, the $X_i$'s do of accounting for the correlations among the $Y_i$'s. To effect this comparison, we may decompose the correlation between two $Y_i$, allowing for the correlation of their residuals. The lengthy equation would include only one unknown, the correlation of the residuals. The resulting correlation of the residuals may be interpreted as that portion of the correlation between the two $Y_i$ which is left unexplained by the $X_i$. In other words, it is the higher-order partial correlation coefficient, controlling for the $X_i$.

To illustrate this, consider the block equation model

$$x_1 = p_{13}x_3 + p_{1e_1}e_1$$

$$x_2 = p_{23}x_3 + p_{2e_2}e_2 \qquad (27),$$

which has already been diagrammed in Figure 4. Solving for $p_{13}$ and $p_{23}$ would result in

$$p_{13} = r_{13} \qquad (28)$$

$$p_{23} = r_{23} \qquad (29),$$

and

$$p_{1e_1} = \sqrt{1 - r_{13}^2} \qquad (30)$$

$$p_{2e_2} = \sqrt{1 - r_{23}^2} \qquad (31).$$

Permitting the residuals to be correlated, we would trace all possible paths from $x_2$ to $x_1$ to obtain the correlation coefficient:

$$r_{12} = p_{13}p_{23} + p_{1e_1}r_{e_1e_2}p_{2e_2} \qquad (32).$$

By substitution from equations (28) through (31) we obtain

$$r_{12} = r_{13}r_{23} + (\sqrt{1 - r_{13}^2})\, r_{e_1e_2}\, (\sqrt{1 - r_{23}^2}) \qquad (33),$$

which means

$$r_{e_1e_2} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = r_{12.3} \qquad (34);$$

that is, the correlation of the residuals is the partial correlation of $x_1$ and $x_2$, controlling for $x_3$.

Block-Recursive Equation Models

It is entirely possible to have a model which is nothing more than a combination of a block system and a recursive system. For example, consider the model in Figure 7.

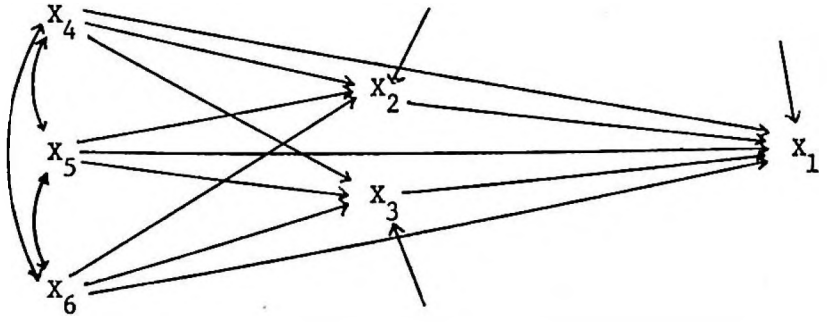



Figure 7.

We assume $r_{e_1e_2} = r_{e_1e_3} = 0$, but it is unnecessary to assume $r_{e_2e_3} = 0$.

The analysis goals are defined for us by what we want, and

by what we can do with block and recursive equation models. In this example, we might want to:

1) Determine the extent to which $X_2$ and $X_3$ mediate the effects of the three exogenous variables. For example, determine the indirect effects of $X_4$ on $X_1$ through $X_2$ and $X_3$.

2) Determine the extent to which $X_4$, $X_5$, and $X_6$ account for the effects of $X_2$ and $X_3$ on $X_1$. That is, calculate the amount of the association of $X_1$ and $X_2$ (and $X_3$) which can be considered to arise spuriously from joint, antecedent, exogenous causes.

3) We can examine the possibility of correlated errors between $X_2$ and $X_3$. We do so by determining the higher-order partial correlation coefficient of $X_2$ and $X_3$, controlling for $X_4$, $X_5$, and $X_6$.

4) Finally, focus on $X_2$, and determine the extent to which the zero-order relationship between $X_1$ and $X_2$ is accounted for by the contemporaneous variable $X_3$; and vice versa. That is, once we have discovered that $X_2$ and $X_3$ have correlated errors, it is possible, indeed probable, that part of the zero-order association between $X_1$ and $X_2$ (and $X_3$) includes an effect attributable to the correlated errors of $X_2$ and $X_3$. For example, part of the decomposition of $r_{12}$ would include the product, $p_{13}p_{3e_3}r_{e_2e_3}p_{2e_2}$.

## APPLICATIONS TO EDUCATIONAL TOPICS

Wright (1925) developed the technique of path analysis to explain the causes of observed correlations between corn prices and hog production. In recent applications, path analysis has been applied most usefully to understand the process of socio-economic achievement (for example, Blau and Duncan, 1967; Duncan, Featherman and Duncan, 1972). Practically all of these models use educational attainment as a mediating variable between background variables and measures of socioeconomic achievement.

A number of reports exist that use path analysis to explain educational performance and attainment. Among the best are Hauser's (1971) examination of the educational performance in Nashville. Duncan, Haller, and Portes (1968) employed a nonrecursive structural equation model which addressed the effect that a friends' educational plans had on ego's plans, and vice versa. Hauser (1973) used the panel data of 1957 Wisconsin high school graduates to develop an elaborate model of the process by which background characteristics, ability, academic performance, and the influence of significant others effect one's socioeconomic plans and educational attainment. There has been a subsequent wave to the panel survey, and Sewell and Hauser (1975) have reported these new results, which now include measures of the effects these variables and educational attainment are having on early career achievements. Recent studies of educational attainment disaggregate social processes that occur within

secondary schools; for example, Alexander and McDill (1976) measured the influence of curriculum tracking on subsequent levels of educational attainment.

In Jencks, *et al.*'s (1972) examination of the effects of schooling, he addressed the question of the inheritability of intelligence, using structural equation models to disaggregate the effects of genotype and environmental influences; Loehlin, Lindzey, and Spuhler (1975) have suggested some revisions to Jencks' conclusions.

Sewell, Hauser and Featherman (1976) recently published a collection of papers on schooling and achievement processes. These papers represent a variety of styles, but nearly all are rigorous, and many include a structural equation approach to their topics. These few examples do not by any means exhaust the possible applications of path analysis; more than anything else, they merely represent the contents of my bookshelf. These sources are, however, excellent samples of the rigor that is currently being brought to bear on substantive topics in education.

## FURTHER CONSIDERATIONS

Let me finish by mentioning two more advantages of path analysis, and acknowledging a topic which I have not explicitly addressed above, but will confront anyone who "does" path analysis. One of the advantages of path analysis is that it does not depend on a complete knowledge of all intercorrelations between the variables in the model; by specifying some causal assumptions it

is sometimes possible to derive estimates from piecemeal data. Examples include the works by Duncan (1968) and Jencks *et al.* (1972: Appendix A).

Another advantage of path analysis is its ability to evaluate in certain circumstances the effects of unmeasured variables. Hauser and Goldberger (1971) discuss methods for measuring effects of such variables, and examples using a cannonical correlation approach and a principal components analysis may be found in Hauser (1973) and Hauser and Featherman (1977: 39), respectively. The latter example may more accurately be considered an example of measurement error, a topic that may also be addressed through causal models (Siegel and Hodge, 1968; Namboodiri, Carter, and Blalock, 1975).

Identification becomes a problem when more unknown coefficients exist in a model than do known intercorrelations. This is particularly a problem in nonrecursive models. Path analysis does not provide a solution to this generic problem, but it does help make explicit the assumptions made in its resolution. These issues have been of most concern to econometricians (for example, Goldberger, 1964; Johnston, 1972), but are beginning to receive explicit attention in sociological multivariate texts (for example, Namboodiri, Carter, and Blalock, 1975). The other side of the same coin is overidentification; that is, there are more intercorrelations than paths to estimate. As a result, there is no unique solution to the normal equations. In recursive models the solution is to simply use the ordinary least squares estimates (Duncan, 1975: 46).

For those interested in pursuing these topics further, your first step should be Duncan's (1975) excellent text on structural equation models.

REFERENCES

Alexander, K. L. and E. L. McDill. Selection and allocation within schools: Some causes and consequences of curriculum placement. *American Sociological Review*, 1976, *41*, 963-980.

Alwin, D. F. Socioeconomic background, colleges, and post-collegiate achievements. Pp. 343-372 in W. H. Sewell, R. M. Hauser, and D. L. Featherman (eds.), *Schooling and Achievement in American society*. New York: Academic Press, 1976.

Alwin, D. F., and R. M. Hauser. The decomposition of effects in path analysis. *American Sociological Review*, 1975, *40*, 37-47.

Anderson, J. G., and F. B. Evans. Causal models in educational research: Recursive models. *American Educational Research Journal*, 1974, *11*, 29-39.

Blalock, H. M., Jr. Four-variable causal models and partial correlations. *American Journal of Sociology*, 1962, *68*, 182-194.

Blalock, H. M., Jr. *Causal inferences in nonexperimental research*. Chapel Hill: University of North Carolina Press, 1964.

Blalock, H. M., Jr. Path coefficients versus regression coefficients. *American Journal of Sociology*, 1967, *72*, 675-676.

Blalock, H. M., Jr. (ed.). *Causal models in the social sciences*. Chicago: Aldine Press, 1971.

Blalock, H. M., Jr. *Social Statistics*, second edition. New York: McGraw Hill, 1972.

Blau, P. M., and O. D. Duncan. *The American occupational structure*. New York: John Wiley, 1967.

Boudon, R. A method of linear causal analysis: Dependence analysis. *American Sociological Review*, 1965, *30*, 365-374.

Boudon, R. A new look at correlation analysis. Pp. 199-235 in H. M. Blalock, Jr. and A. B. Blalock (eds.), *Methodology in social research*. New York: McGraw-Hill, 1968.

Duncan, O. D. Path analysis: Sociological examples. *American Journal of Sociology*, 1966, *72*, 1-16.

Duncan, O. D. Ability and achievement. *Eugenics Quarterly*, 1968, *15*, 1-11.

Duncan, O. D. Inheritance of poverty or inheritance of race? Pp. 85-110 in D. P. Moynihan (ed.), *On understanding poverty*. New York: Basic Books, 1969. (a)

Duncan, O. D. Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*, 1969, *72*, 177-182. (b)

Duncan, O. D. Partials, partitions, and paths. Pp. 38-47 in E. F. Borgatta and G. W. Bohrnstedt (eds.), *Sociological methodology 1970*. San Francisco: Jossey-Bass, 1970.

Duncan, O. D. *Introduction to structural equation models*. New York: Academic Press, 1975.

Duncan, O. D., D. L. Featherman, and B. Duncan. *Socioeconomic background and achievement*. New York: Seminar Press, 1972.

Duncan, O. D., A. O. Haller, and A. Portes. Peer influences on aspirations: A reinterpretation. *American Journal of Sociology*, 1968, *74*, 119-137.

Erlanger, H. S. and H. H. Winsborough. The subculture of violence thesis: An example of a simultaneous equation model in sociology. *Sociological Methods and Research*, 1976, *5*, 231-246.

Featherman, D. L. and R. M. Hauser. Changes in the socioeconomic stratification of the races, 1962-1973. *American Journal of Sociology*, 1976, *82*, 621-651.

Finney, J. M. Indirect effects in path analysis. *Sociological Methods and Research*, 1972, *1*, 175-186.

Goldberger, A. S. *Econometric theory*. New York: John Wiley, 1964.

Goldberger, A. S. Structural equation methods in the social sciences. *Econometrica*, 1972, *40*, 979-1001.

Goldberger, A. S., and O. D. Duncan (eds.). *Structural equation models in the social sciences*. New York: Seminar Press, 1973.

Hauser, R. M. *Socioeconomic background and educational performance*. Rose Monograph Series. Washington, DC: American Sociological Association, 1971.

Hauser, R. M. Disaggregating a social-psychological model of educational attainment. Pp. 255-284 in A. S. Goldberger and O. D. Duncan (eds.), *Structural equation models in the social sciences*. New York: Seminar Press, 1973.

Hauser, R. M. and D. L. Featherman. *The process of stratification: Trends and analyses.* New York: Academic Press, 1977.

Hauser, R. M. and A. S. Goldberger. The treatment of unobservable variables in path analysis. Pp. 81-117 in H. L. Costner (ed.), *Sociological methodology 1971.* San Francisco: Jossey-Bass, 1971.

Hauser, R. M., W. H. Sewell, and D. F. Alwin. High school effects on achievement. Pp. 309-341 in W. H. Sewell, R. M. Hauser, and D. L. Featherman, (eds.), *Schooling and achievement in American society.* New York: Academic Press, 1976.

Heise, D. R. Problems in path analysis and causal inference. Pp. 38-73 in E. F. Borgatta (ed.), *Sociological methodology 1969.* San Francisco: Jossey-Bass, 1969.

Heise, D. R. Employing nominal variables, induced variables, and block variables in path analysis. *Sociological Methods and Research,* 1972, *1*, 147-174.

Heise, D. R. *Causal analysis.* New York: John Wiley, 1975.

Jencks, C., M. Smith, H. Acland, M. J. Bane, D. Cohen, H. Gintis, B. Heyns, and S. Michelson. *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books, 1972.

Johnston, J. *Econometric methods,* second edition. New York: McGraw-Hill, 1972.

Kim, J. and C. W. Mueller. Standardized and unstandardized coefficients in causal analysis: An expository note. *Sociological Methods and Research,* 1976, *4*, 423-438.

Land, K. C. Principals of path analysis. Pp. 3-37 in E. F. Borgatta (ed.), *Sociological Methodology 1969.* San Francisco: Jossey-Bass, 1969.

Loehlin, J. C., G. Lindzey, and J. N. Spuhler. *Race differences in intelligence.* San Francisco: W. H. Freeman, 1975.

Namboodiri, N. K., L. F. Carter, and H. M. Blalock, Jr. *Applied multivariate analysis and experimental designs.* New York: McGraw-Hill, 1975.

Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Brent. *Statistical package for the social sciences,* second edition. New York: McGraw-Hill, 1975.

Sewell, W. H. and R. M. Hauser. *Education, occupation, and earnings: Achievement in the early career.* New York: Academic Press, 1975.

Sewell, W. H., R. M. Hauser, and D. L. Featherman. *Schooling and achievement in American society.* New York: Academic Press, 1976.

Siegel, P. M. and R. W. Hodge. A causal approach to the study of measurement error. Pp. 28-59 in H. M. Blalock, Jr., and A. B. Blalock (eds.), *Methodology in social research.* New York: McGraw-Hill, 1968.

Williams, J. D. and R. J. Klimpel. Path analysis and causal models as regression techniques. *Multiple Linear Regression Viewpoints,* 1975, *5*, 1-20.

Wolfle, L. M. Path analysis and causal models as regression techniques: A comment. *Multiple Linear Regression Viewpoints,* 1977, *7*, 33-40.

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

Wright, S. *Corn and Hog Correlations.* Washington, DC: U.S. Department of Agriculture Bulletin 1300, January, 1925.

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

UNIVERSITY OF MISSOURI-ST. LOUIS

School of Education

Department of Behavioral Studies

8001 Natural Bridge Road
St. Louis, Missouri 63121
Telephone: (314) 453-5782

June 30, 1977

MEMO

TO: MLR/SIG Members

FROM: Steve Spaner, MLR/SIG 1978 Program Chairman

SUBJECT: Call for papers for the 1978 paper session.

The title of our 1978 SIG paper session will be "Multiple linear regression substitutions for traditional and other forms of analysis." I am soliciting a variety of applied analysis papers to demonstrate the breadth and power of MLR: ANOVA, ANCOVA, repeated measures, trend analysis, path analysis, validity studies, and any others you have worked through MLR. I would like to limit the submissions to data based, applied studies even though theory may be an integral part. I feel the audience we attract are primarily interested in the how to use MLR and less interested in the why use MLR.

I have requested our chair-elect, John Williams, to conduct a second MLR/SIG session on path analysis and its relationship and overlap with MLR. John has written a couple of articles in the Viewpoints on path analysis and, I feel, knows the topic better than most. John's session could be considered a mini-workshop or mini-training session since it will be a one man show (except for his own assistants).

I hope we will have many interested members in both sessions, both as presenters and participants. Please submit your paper proposals in the standard AERA-APA form by August 15, 1977 to me:

Steve Spaner
MLR/SIG Program Chairman
Behavioral Studies Department
University of Missouri - St. Louis
St. Louis, Mo. 63121

Please plan to attend the paper session and training session at the AERA Convention in Toronto, Can next March 27-31, 1978. See you there.

Report from the Executive Secretary

# MINUTES OF THE 1977 ANNUAL MEETING OF MLR/SIG

Approximately 20 persons were in attendance at the business meeting.

The chairman, Mike McShane, brought the meeting to order.

1) The first item of business was a report from the MLR Viewpoints editor, Izzy Newman, indicating that the cost of the Viewpoints production has risen to the point that we must increase revenue, somehow. The two methods available are increased page costs to contributors and/or increased membership-subscription fees. The sentiment of the group was that to raise page costs to contributors would have a discouraging effect on article submission while raising membership-subscription fees a modest amount would be more palatable.

It was moved and seconded to increase membership-subscription fees from $2.00 to $3.00 per year for individual membership and to leave the library/institutional agency membership fee at $12.00. The motion was unanimously approved. Comment was made that we (MLR/SIG) should strictly enforce the library/institutional agencies fee which we have not done in the past and the Secretary-Treasurer, Steve Spaner, was so instructed. Additionally, members were urged to request their respective campus or institutional libraries to become MLR Viewpoints subscribers.

2) The next item of business was a motion to establish the position of Executive Secretary and redefine the duties of the Chair, Chair-Elect, Editor and Editorial Board to more appropriately and efficiently conduct the business of the MLR/SIG.

The motion was seconded and discussion led off with the explanation that the yearly change of the membership files and other MLR/SIG materials to the new chair-elect/secretary-treasurer has resulted in delays as long as half-year in getting the next issue of the Viewpoints out. The proposed establishment of an Executive Secretary post, with a 3-year tenure, was viewed as a solution to this problem. Further discussion resulted in various friendly amendments to more specifically define the tenure and function of the MLR/SIG offices and policy boards. Following is the structure as amended and unanimously approved:

| Position/Tenure | Function |
|---|---|
| Chair/1 year<br>1977-78 | -Organize the Convention program<br>-Liason with and official representative to AERA and other organizations<br>-Chair of the Executive Board |
| Chair-Elect/1 year<br>1977-78 | -Assist in organizing the Convention program<br>-Member of the Executive Board |
| Executive Secretary/3 years<br>1977-80 | -Maintain membership files and address labels<br>-Conduct the ongoing business of the MLR/SIG<br>-Financial Officer<br>-Secretary for the Annual Business meeting<br>-Member of and Secretary for the Executive Board meetings |
| Editor/until resignation | -Receive, review or send out for review, and assemble articles for the MLR/SIG journal, Viewpoints. |

| | |
|---|---|
| Executive/Editorial Board/ rotation of the two most senior members each year | -Review articles for Viewpoints<br>-Set publication policies for the Viewpoints<br>-Set MLR/SIG policy<br>-Act as a nominating committee for replacement of position holders |

Note should be made of the fact that members of the Editorial Board are now members of the Executive Board as well. The feeling was that a policy committee was needed and the Editorial Board was an existing and capable group of members for such decisions (should they arise).

3) The meeting then turned to the election of officers. Mike McShane called for nominations for the position of Chair-Elect (as newly defined).

John Williams of the University of North Dakota was nominated and elected by acclamation.

Next the position of Executive Secretary was opened for nomination (as newly created and defined).

Izzy Newman nominated Steve Spaner with the justification that the membership files and treasury were currently in his possession and his election would eliminate the transfer lag. Steve Spaner raised the issue and ethics of holding two positions in this coming year: chair and , if elected, executive secretary. The members present expressed no feelings of conflict of interest so long as Steve Spaner felt he could carry-out both functions. Steve indicated that several of his graduate students were MLR/SIG members who he was "sure" would be willing to help him if need be.

Steve Spaner of the University of Missouri at St. Louis was elected by acclamation.

The nominations of Executive/Editorial Board replacement (as newly defined) was opened next. The matter of tenure seniority was eliminated by the resignation from the Board of Thomas Jordan of the University of Missouri at St. Louis and Keith McNeil of Durhan, North Carolina.

Lee Wolfle of Virginia Polytechnic Institute and Mike McShane of the Association of American Medical Colleges were nominated (volunteered) and elected by acclamation.

In anticipation of other openings or resignations on the Editorial Board, the Chair requested any interested members to give their name to the Secretary. Volunteers were: Jack McArdle and James Maxwell. Other interested members may send their name to Steve Spaner, the Executive Secretary.

The 1977-78 Executive/Editorial Board in order of decreasing tenure (as assigned by the 1977-78 Chairperson)is as follows:

1. John Pohlman, Southern Illinois University at Carbondale
2. Earl Jennings, University of Texas
3. William Connett, Montant State Department of Education
4. Joe H. Ward, Jr., Lackland Air Force Base
5. Robert Deitchman, University of Akron
6. Samuel Houston, University of Northern Colorado
7. Lee Wolfle, Virginia Polytechnic Institute
8. John Williams, University of North Dakota
9. Mike McShane, AAMC, Wash, D.C.

The meeting was adjourned until next year's AERA convention.

Respectfully submitted,

Steven D. Spaner
Executive Secretary of the
MLR/SIG

UNIVERSITY OF MISSOURI-ST. LOUIS

School of Education

Department of Behavioral Studies

8001 Natural Bridge Road
St. Louis, Missouri 63121
Telephone: (314) 453-5782

June 27, 1977

MEMO:

TO: All past and current Multiple Linear Regression/Special Interest Group members

FROM: Steve Spaner, Executive Secretary

SUBJECT: First dues notice for MLR/SIG 1977-78 membership year.

At the 1977 MLR/SIG business meeting (minutes will appear in the next Viewpoints) dues were increased from $2.00 a year to $3.00 a year for individual memberships subscriptions. Agency or institutional subscriptions were left at $12.00 per year but enforcement of this rate will now be exercised. Members were encouraged to request their agency or institutional libraries to become MLR Viewpoints subscribers since the cost of the journal is rapidly increasing.

Make your check or money order out to MLR/SIG or Multiple Linear Regression/ Special Interest Group and send it to:

Steve Spaner
MLR/SIG Executive Secretary
Behavioral Studies Department
University of Missouri-St. Louis
St. Louis, Mo. 63121

(NOTE: If your address label has a code in the first line of a number followed by PD 77 you owe $3.00 for the 1977-78 membership year. If your code is a number followed by PD 78, you only owe $1.00 since you have already paid $2.00 for the 1977-78 year. If your code is only a number, you were not a paid member during the 1976-77 membership year and we sure would like to have you rejoin us.)

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

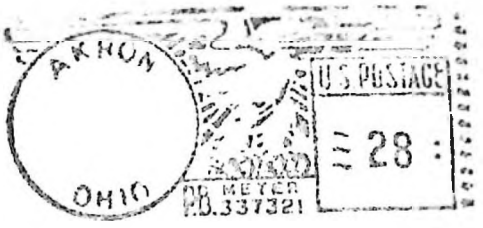
All manuscripts should be sent to the editor at the above address. (All manuscripts should be camera-ready copy.)

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, single-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.

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

"Membership in the *Multiple Linear Regression Special Interest Group* is renewed yearly at the time of the American Educational Research Association Convention. Membership dues pay for a subscription to the *Viewpoints* and are divided into two categories: individual=$3.00; and institutional (libraries and other agencies)=$12.50. Membership dues and subscription requests should be sent to the Executive Secretary of the MLRSIG."

**THE UNIVERSITY OF AKRON**
**AKRON, OHIO 44325**






---

# TABLE OF CONTENTS