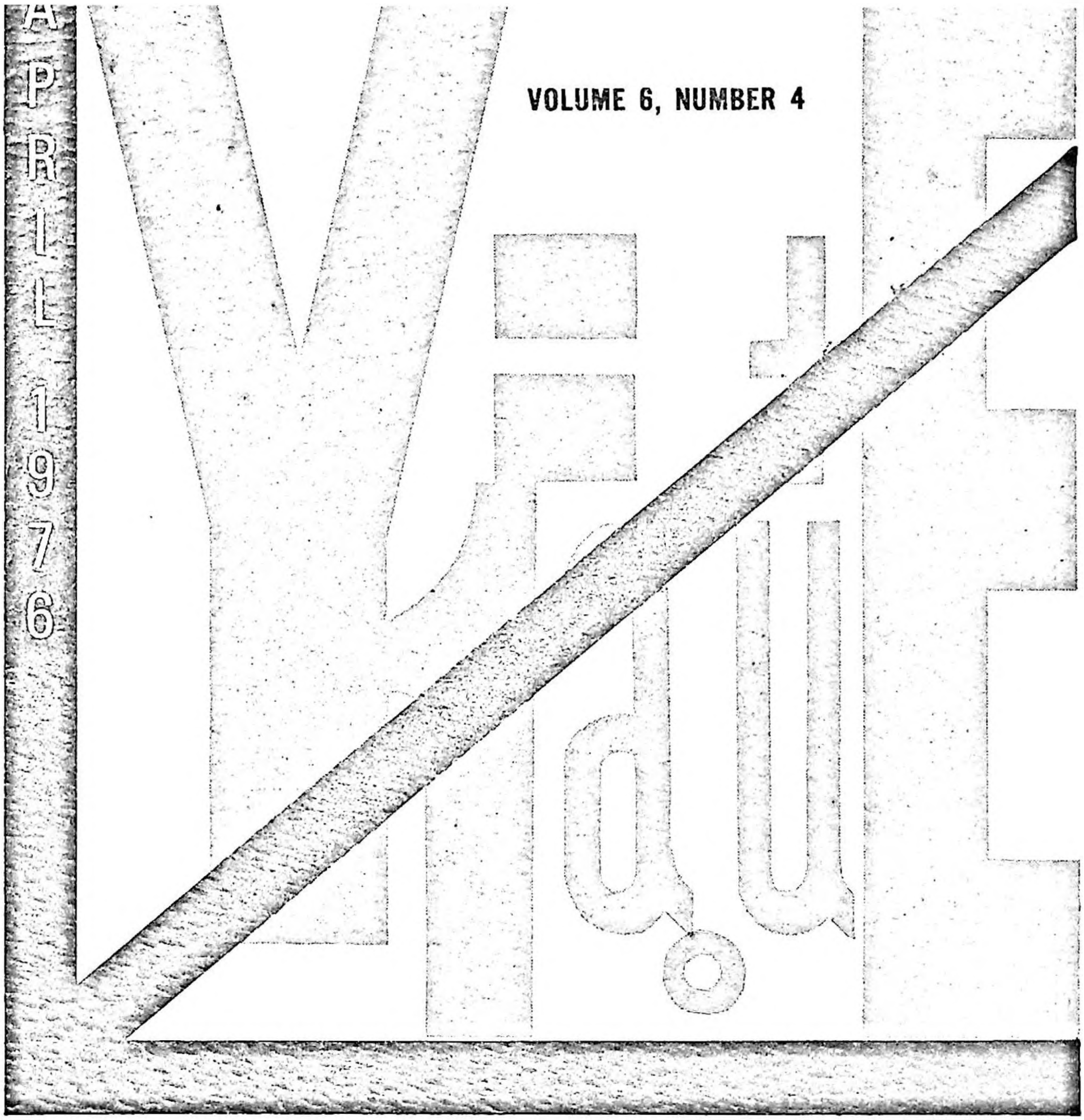


MULTIPLE LINEAR REGRESSION VIEWPOINTS

A publication of the Special Interest Group
on Multiple Linear Regression

MULTIPLE LINEAR REGRESSION VIEWPOINTS

TABLE OF CONTENTS

Multiple Linear Regression Viewpoints
Vol. 6, No. 4, 1976

TYPE VI ERROR: INCONSISTENCY BETWEEN THE STATISTICAL PROCEDURE AND THE RESEARCH QUESTION*

ISADORE NEWMAN, ROBERT DEITCHMAN, JOEL BURKHOLDER
RAYMOND SANDERS, The University of Akron
LEROY ERVIN, JR., Oberlin College

While a great deal of money and energy is currently being directed toward research, there also seems to be a general lack of acceptance of the relevance of research findings. This skepticism is somewhat justifiable, and we will attempt to discuss what we feel is the major cause for this state of affairs. These issues may explain to some extent why less and less monies are being made available for certain types of research.

One reason people ignore research findings has been that the statistical models used have frequently been unrelated or tangentally related to the research question of interest. There is a variety of reasons for this, some of which will be discussed in this paper.

---

*Paper presented to the World Population Society, Western Regional Meeting; January 24, 1976, aboard the Queen Mary, Long Beach, California.

We would like to greatfully acknowledge the assistance of Dr. Ralph Blackwood, Dr. Keith McNeil, and Dr. Charles Dye for their editorial comments.

1. The courses that teach research methods generally emphasize data analysis, rather than practicing appropriate methods and procedures for asking and developing research questions. These courses do not adequately develop the skills of evaluating the research question and the statistical models that are most capable of reflecting the research question.

Quite often a student coming out of these courses tends to select a familiar, "canned" standard statistical design, or package (cookbook approach) such as a 2 x 3, or 2 x 2 x 3, because he has not been taught to develop his own models to reflect their research question. Therefore, he uses these standard models which dictate the question being investigated. Sometimes a researcher is aware that these models do not completely represent his true research question. When this happens, he may then make inferential jumps from his data. These influences may well be inappropriate. In some cases, the researcher is unaware that his models are nct really reflective of his research questions; sometimes the unsophisticated researcher allows the statistical model to totally dictate his research question. Under these conditions, we find research that is technically correct but is not relevant because it is not related in a pragmatic way to a specific problem.

2. People often misinterpret and misapply research findings because they confuse statistical analysis with experimental design. Traditional analysis of variance, in which the design and statistical procedures are totally related, has fostered this confusion.

One sometimes finds even his most respected colleagus making statements such as:

a.) We can only infer causation when the research is analyzed through traditional analysis of variance.

b.) If we use regression or correlation, we cannot infer causation.

Rembember, causal relationship can only be inferred through experimental design. No statistical technique allows us to infer or assume causal relationships.

In a recent issue of the Harvard Review Luecke & McGinn (1975) wrote a paper entitled, "Regression Analyses And Educational Production Functions: Can They Be Trusted?" Their conclusion was that we cannot appropriately infer causation from regression techniques. This is an obvious statement since causation cannot be inferred from any statistical technique, whether it be regression, analysis of variance, t-test, etc. However, their conclusion has been inappropriately generalized beyond this, in some quarters it has been accepted to the extent that some believe causation cannot be inferred anytime regression is used regardless of the research design. This stems from the earlier problem of confusing statistics with design.

Regression is a statistical procedure and should be used in relationship with some design in the Campbell & Stanley Tradition (1969). A researcher should begin by asking a research question that reflects an area of interest. He should then design his study so that it can answer that research question. His research design should include how data will be collected, how subjects will be selected; how treatments will be administered, etc. (See Fig. 1). To the extent that this design has internal validity, he can infer causal relationships

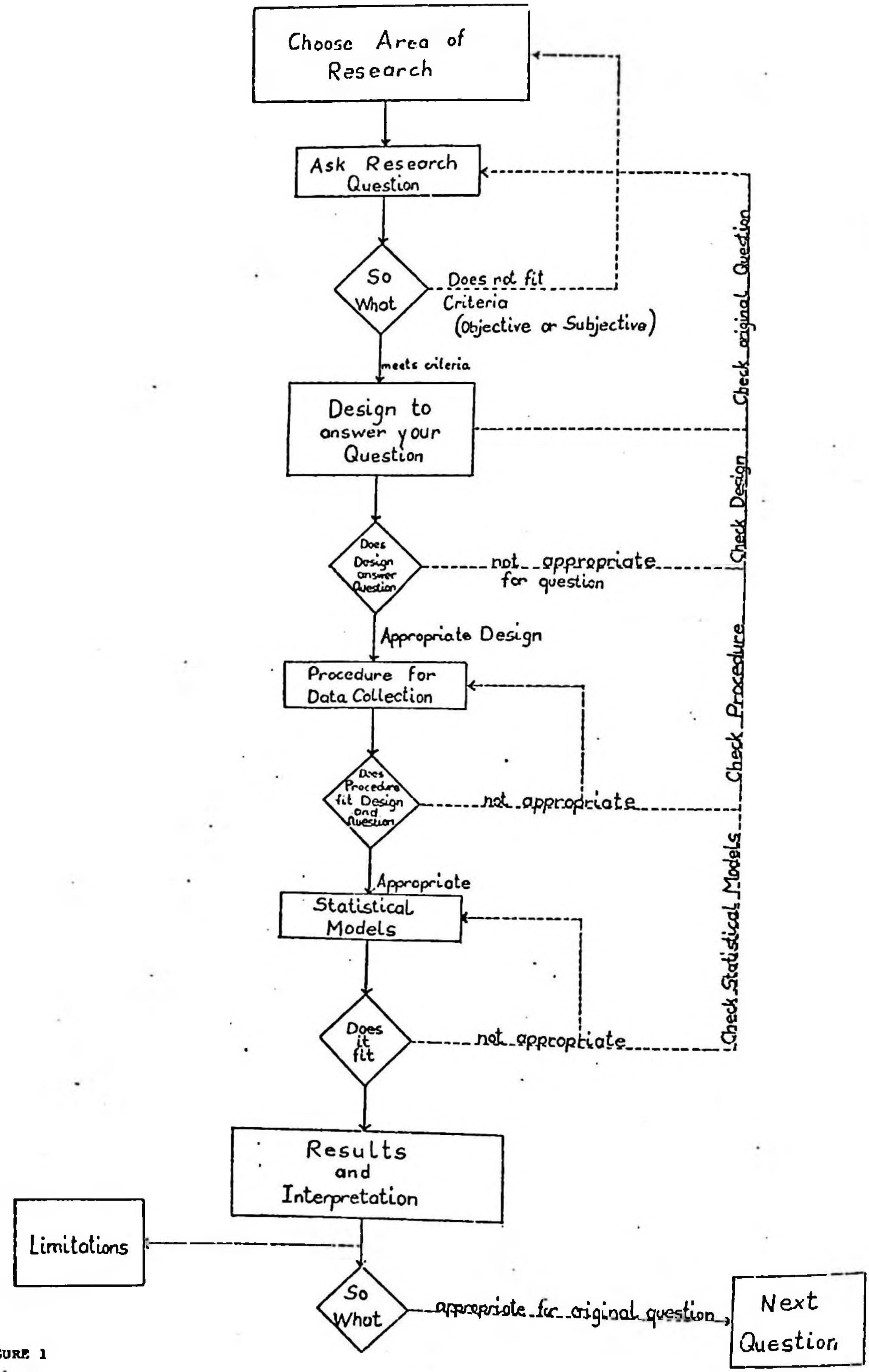

Choose Area of Research
Ask Research Question
So What
Does not fit Criteria (Objective or Subjective)
meets criteria
Design to answer your Question
Check original Question
Does Design answer Question
not appropriate for question
Check Design
Appropriate Design
Procedure for Data Collection
Does Procedure fit Design and Question
not appropriate
Check Procedure
Appropriate
Statistical Models
Does it fit
not appropriate
Check Statistical Models
Results and Interpretation
Limitations
So What
appropriate for original question
Next Question

FIGURE 1

between the independent and dependent variables. If this design is internally valid, no matter what statistical technique he uses to test it, such as regression, correlation, etc., he can legitimately assume causal relationships.

If his research design is ex post facto, where the independent variable is not under the control of the researchers, no matter what techniques he uses, he cannot infer causation. All the studies on desegregation and busing such as the Coleman Report (1966); Jensen, Jencks, et al. (1972), and Gerald & Miller (in press) are of this nature. It is not technically legitimate to infer causation when the design is ex post facto. Even though a variety of statistical techniques such as path analysis as developed by Blalock (1962, 1964, 1970, & 1972) and more recently component analysis developed by Mood (1971), have attempted to get at causal relationships of ex post facto data, through the manipulation of regression techniques one still cannot technically infer causation (Newman & Newman, 1975). Causation cannot be inferred from an ex post facto design even if you analyze data with ANOVA. These techniques tend to eliminate some rival hypotheses, making the study somewhat stronger. However, they cannot eliminate all of the rival hypotheses so the study cannot be totally internally valid and causation cannot be inferred. This is not to say that these studies are not very valuable. This simply says one cannot infer causation between the independent and dependent variables.

A necessary but not sufficient prerequisite for an experimental design is that at least one independent variable is active (under the control of the researcher). If there are no active variables in a research study, the study by definition would be ex post facto, even if the analyses is ANOVA. For instance, assure one study with a 2 x 2 factorial design looked at Black Students in integrated schools versus Blacks who were not, and sex. If the subjects were not under

	Male	Female
Integrated		
non-Integrated		

the researchers' control, this study would then be an ex post facto design, not a true experimental design, since random assignment could not be employed. In effect, this is a straight correlational study even though one is getting A-main effects, B-main effects and AB-interaction. It is not a true experimental design and causation cannot be inferred. The statistical question being asked in such a design is, whether there is a significant relationship between these independent variables and the criterion. Any inference about causal relationships would be inappropriate.

3. When the research question of interest is one of trends or functional relationships, one often finds the use of inappropriate statistical models which cannot accurately reflect the research questions (Newman, 1974).

When researching developmental questions, one is often more interested in functional relationships than mean differences. There is generally a continuous variable that is of interest, such as time, age, population sizes, I.Q., SES. When traditional analysis of variance is employed, continuous variables are forced into categorizations. This causes the researcher to lose degrees of freedom, and there is a potential loss of information. This loss is contingent upon how representative the categories are of the inflections in the naturally occuring continuous variable.

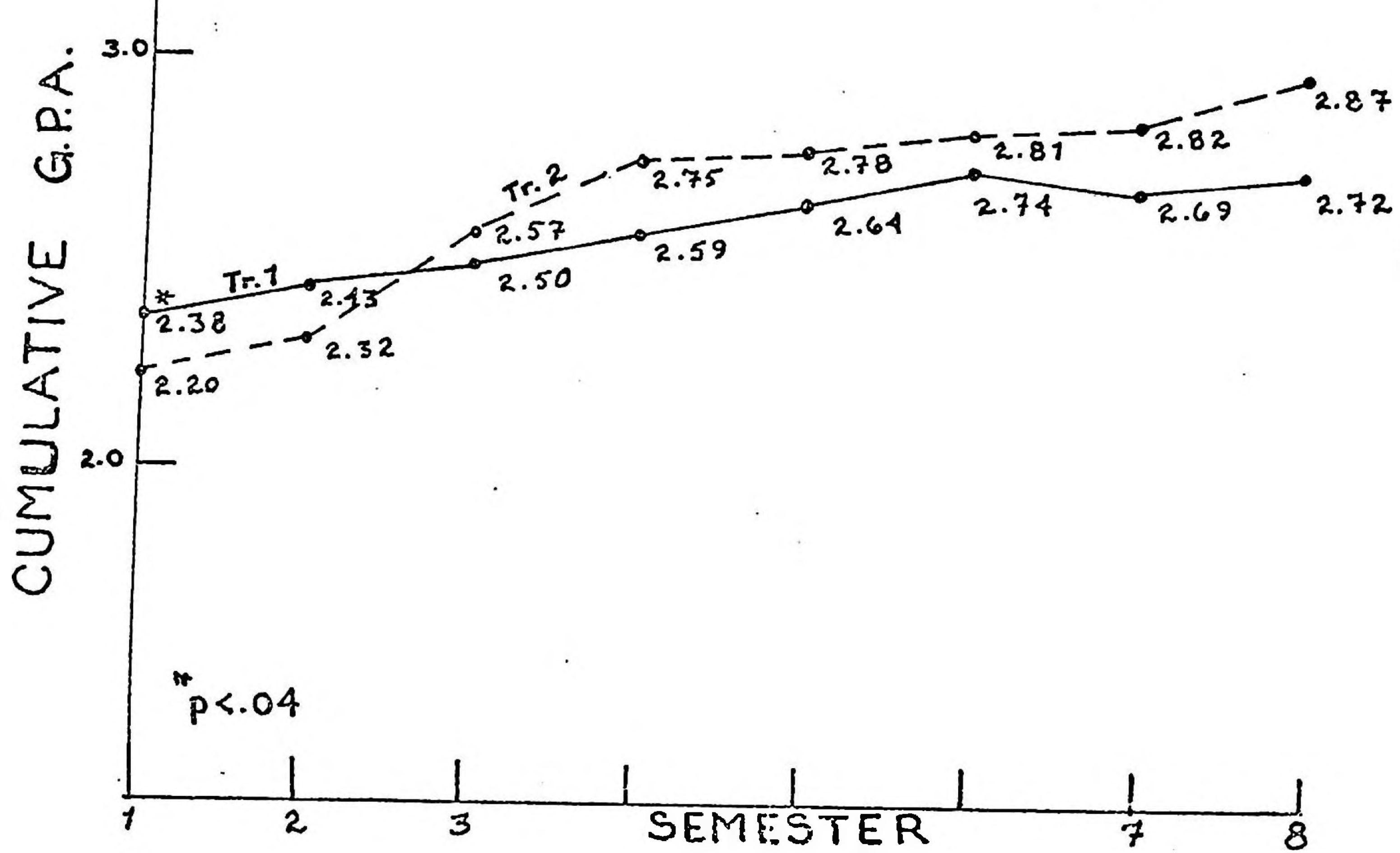
Since continuous variables are frequently artifically categorized, the analysis produced by such a procedure may not really reflect the researcher's queston or interest. The most efficient method for writing statistical models that reflect trend or curve fitting questions, is the general case of the least squares solution, linear model (Multiple Linear Regression Procedures, Newman (1974), McNeil, Kelly, McNeil (1975), Draper & Smith (1966), Kelly, Newman, and McNeil (1973)). This procedure allows one to write linear models, which specifically reflect the research question.

Linear Regression is an excellent statistical tool for looking at a population trend or comparing multiple trends over time (Newman 1974, Ervin 1975). The following examples are taken from a study comparing performance trends of underprepared students receiving special treatment, with the performance of regularly admitted students requiring special assistance over four years (Ervin, 1975). Although the example provided is concerned with an educational issue at a single institution, regression is not limited to such situations. In fact, it has more flexibility than any other single statistical tool and can be used with large populations incorporating a sizable number of independent variable in a single model.

In Figure 2, a graph is presented that reflects the researcher's interest in learning if there are significant differences in trends (in this case slope differences) between Black subjects who received the Developmental Program (X_2) and Black students who did not receive

FIGURE 2

TREATMENTS AND CUMULATIVE G.P.A. OF BLACK SUBJECTS



the program (X_1), as it relates to their cumulative G.P.A.

This research question was then restated as a specific hypothesis.

Hypothesis I: Are these significant differences in slopes for X_1 vs. X_2 in predicting cumulative G.P.A. for Black students at Oberlin.

The models needed to test this hypothesis are as follows:

Model 1: $X_{14} = a_0U+a_1X_1+a_2X_2+a_3X_3+a_4X_4+E$

Model 2: $X_{14} = a_0U+a_1X_1+a_2X_2+a_5X_5+E_2$

where: X_{14} = cumulative G.P.A.

X_1 = 1 if student had program, 0 otherwise

X_2 = 1 if student did not have program, 0 otherwise

X_3 = number of the semester for the subjects who had the program, 0 otherwise

X_4 = number of the semester for the subjects who did not have the program, 0 otherwise

X_5 = X_3+X_4 = number of semesters for all subjects

U = unit vector, 1 if subject is in the sample, 0 otherwise

a_0-a_5 = partial regression weight

E = error $Y - \hat{Y}$

If Model 1 is found to be significantly different from Model 2, this would indicate that there is a significantly different functional relationship (at some specific α level) between Black students who took the program and Black students who did not take the program in terms of their cumulative G.P.A.

There are an infinite number of other questions that could be asked, such as: Are there significantly different 2nd degree or 3rd degree functional relationships (curvilinear relationships), and are the means of the groups significantly different over all semesters or over specific semesters? etc. Then, regression models could be written that specifically reflect the questions of interest.

4. The problem of unequal N's is really an intriguing one. It is of importance because it is a common problem for applied researchers and depending upon the models employed it can produce widely different answers.

There are a variety of solutions to the unequal N's problem, which can be divided into two major categories--approximate and exact. Examples of approximate solutions are: randomly eliminating data and running the analysis on just group means, therefore, decreasing the number and power. A researcher using any of these solutions is generally aware of the limitations and problems.

What may be more misleading are the exact solutions which are all technically correct but which, like the mean, median, and mode, are answering different questions. The three exact solutions are the hierarchial model, the unadjusted main effects method, and the fitting constants method.

1. Hierarchial Model (Cohen, 1968; Williams, 1974). This requires apriori knowledge of which variables in a solution are more important or of greatest interest. With this approach, each effect is only adjusted for the effects that are of greatest interst.

2. Unadjusted Main Effects Methods (Williams, 1974). This solution requires that the main effects may not be adjusted while the

interaction effects would be adjusted for the variance accounted for by the main effects.

3. Method of Fitting Constants. This method as a solution for disproportionality has been suggested and described most frequently (Winter, 1971; Schagger, 1959; and others). In this procedure, each main effect is adjusted for each other main effect,and the interactions are adjusted for all main effects. This procedure is similar to the analysis of covariance procedure suggested by Overall and Spiegel (1969).

By examining Table 1, one can see that the hierarchial method is most sensitive to detecting questions of major interest. It accounts for the total variance of the variance of the main effect, even though it overlaps with the variance of each main effect minus the variance of the other main effects. The third method of fitting constants may be most conservative of the three. It determines the unique variance accounted for by each main effect and the intereaction. As one can see from Table 1, these three methods, which are all statistically accurate, are answering three different questions.

It is important to determine which of these methods are reflecting the question that we are interested in answering. One can only do this by being sensitive to one's research question and by being aware of the different statistical techniques which are more appropriate than others.

5. Still another area of concern for applied researchers is the relationship between statistical significance and N size. When the N's are very large, any difference is likely to be statistically significant at extremely conservative alpha levels. It is the

TABLE 1

CONCEPTUAL PRESENTATION OF THE THREE METHODS FOR ADJUSTING UNEQUAL N's

	A - Main Effect	B - Main Effect	A B - Interaction
Hierarchial Method	Total proportion of variance accounted for by the effect of major interest. Example: (R^2 for A - main effect)	Proportion of variance accounted for by the effect of the second most important variable holding constant the first most important variable. Example: (R^2 for B holding A	Proportion of variance accounted for by the last important variable holding the first and second most important variables constant. Example: (R^2 for AB holding A, & B constant)
Unadjusted Main Effects Method	Proportion of variance for one main effect holding the other(s) main effect(s) constant. Example: (R^2 for A-main effect holding B main effect constant)	Proportion of Variance accounted for by the second main effect holding constant the other(s) main effect(s). Example: (R^2 for B-main effect holding constant A-main effect (s)	Proportion of Variance accounted for by the interaction while holding constant all the main effects. Example: (R^2 for AB holding constant A & B main effects)
Fitting Constant Method	Unique variance accounted for by the first main effect holding constant the other main effect(s) and interaction(s). Example: (R^2 for A-hold constant B & AB)	Unique variance accounted for by the second main effect while holding constant the other main effect(s) and interaction(s). Example: (R^2 for B-effect holding constant A & AB)	Unique variance accounted for by the interaction holding constant all the main effects. Example: (R^2 AB hold constant A & B main effects)

author's position that it is to the advantage of the researchers and the implementers of research, to always look at the R^2 (proportion of variance accounted for). This is a very important index to estimate the pragmatic importance of the study. This becomes especially important when the N's are large and one finds significance at $p <.0001$. It is possible that the R^2 is accounting for 1% of the variability and 99% is unaccounted. It is important that the researcher be aware of the magnitude of effect as well as the statistical significance. Without this awareness, one increases the likelihood of inappropriately estimating the usefulness of the results.

6. Another error that sometimes occurs due to inappropriate statistical procedures, is the loss of power (increase in Type II Error, failure to reject a false null hypothesis). This may be due to not using a directional hypothesis when one should, especially when the N is small.

A relatively frequent error found in the literature is the statement that, "Significance was found, but in the wrong direction." This is incorrect because if it was in the wrong direction, one would have to fail to reject the null hypothesis, and therefore conclude the findings were non significant.

A possibly more dangerous error, because of its subtlety, is when a researcher makes a non directional test to "play it safe" and because most statistical text books suggest this procedure. He then might find statistical significance in the opposite direction from which the question was initially generated. (In other words, the researcher should have used a directional test, but did not.) Some researchers would then state that they have found statistical sense, but it is a conceptual error. In this case, statistical significance is likely to mean that there was something wrong with the initial logic or theory from which the hypothesis was derived, and a rethinking of the initial thinking and procedures is called for. Therefore, this aspect must be carefully looked at before one can determine appropriate estimates of the findings applicability. (See McNeil *et al*. (1975), for an excellent discussion on the importance of directionality in hypotheses testing.)

Suggested Approach for Conducting Research

It has been suggested that the sequential steps for conducting research, which is presented in Figure 1, be adopted as a guide (Deitchman, Newman, Burkholder, and Sander (1975)).

At best, one should always generalize with caution. The magnitude of the relationship that one finds in a sample is quite often an overestimate of the magnitude of the effect in the population. Therefore, one should try to replicate whenever possible.

If one does not wish to or cannot replicate, the next best procedure of estimating the effect size in the population from a sample, is by using shrinkage estimates. Shrinkage formulas are most appropriate as unbiased estimates of R^2 when the predictor variables are not solely selected on the basis of their initial correlation with the dependent variable (Newman, 1973).

One of the best estimates of shrinkage is a cross validation procedure. This procedure is described in detail in McNeil *et al* (1975). Three easily computed mathematical estimates of shrinkage (Wherry, McNemar, and Lord) are discussed and evaluated by Newman (1973) and Klein & Newman (1974).

In addition, anyone who is conducting or reading research should keep in mind the important difference between statistical methods and research design. As depicted in Figure 1, the first important step is to ask a relevant research question and then to decide on procedures and methods (design) which will lead to an answer for that question. This design should then be evaluated for limitations in terms of its internal and external validity. Once one establishes the basic research question and a design that is acceptable, specific hypotheses can be derived which can be statistically tested (keeping in mind that every test of significance is a test of a specific hypothesis.) Linear models (regression models) can then be developed to reflect each specific hypothesis and to test each for significance. In this manner, when one gets statistical significance from a test, they know specifically the hypothesis being tested. They also know how accurately the results can be interpreted in terms of causation and its general ability, based upon the internal and external validity of the research design.

In Retrospect

We have attempted to convince trainers of researchers and the researchers themselves to beware of the Type VI Error. We believe that the examples will help the researcher to become sensitive to these types of problems. Bad research not only drives out good money but creates a large credibility gap where there should be none. From the authors perspective, applied and theoretical research really only differ in terms of face validity which implies an attempt to answer the "So What" question of the research. It is to some extent an artificial dichotomy that applied researchers are dealing with the "So What" and the theoretical researchers are not. In reality, this is not true. Asking the "So What" question may be one of the discriminating features between good and bad research, regardless of orientation.

The authors of this paper feel that good research can benefit everyone, but poor research may harm everyone. We also believe that the suggestions presented in this paper, if followed, are likely to facilitate good research.

In conclusion: a Type VI Error was defined as the inconsistency between the researcher's question of interest and the statistical procedures he employs to analyze his data. The reasons for these commonplace errors vary. They may be the result of courses which fail to emphasize logical analysis as a prerequisite to the statistical analysis, confusion caused by teaching traditional analysis of variance without differentiating between the statistic and the design, the misconception that causation can only be inferred from traditional analysis of variance, and forgetting that causation is a design question, not a statistical one.

A major problem discussed was disproportionality (unequal N's). The three exact solutions hierarchial models, unadjusted main effects, and fitting constants, were discussed in terms of the conceptual questions the answer. Stress was placed on understanding that each solution is answering different questions. Therefore, one should not concern himself with which model is more statistically accurate, but with which one is more reflective of the researcher's question.

In dealing with a research question, one must consider practical as well as statistical significance. This means that N size, and effect size as well as probability level and power, must be considered.

Finally, a procedure is suggested for conducting research along with a flow chart. We feel that following the procedures, outlined in the flow chart, will help one avoid some of the pitfalls that lead to Type VI errors.

References

1. Applebaum, M. I. and Carmer, E. M. Some problems in the analysis of variance. Psychological Bulletin, 1974, 81, 335-343.

2. Blalock, H. M. Four-variable causal models and partial correlation. American Journal of Psychology, 1962, 68, 182-194.

3. Blalock, H. M. Causal inferences in non-experimental research. Chapel Hill: University of North Carolina Press, 1964.

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

5. Blalock, H. M. (Ed.). Causal models in the social sciences. Chicago: Aldine, 1971.

6. Campbell & Stanley. Experimental and Quasi-Experimental, Designs for Research, Rand McNally, 1969, N.Y.

7. Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. Equality of educational opportunity. Washington, D.C.: U.S. Department of Health, Education and Welfare, 1966.

8. Deitchman, R., Newman, I., Burkholder, Joel, Sander, R. E. "This is the answer. What is the Question?" Paper presented at APA, Chicago, August 1975.

9. Draper, N. R., and Smith, H. Applied Regression Analysis. John Wiley and Sons, Inc., New York, 1966.

10. Ervin, L. A Multivariable Approach To Evaluation of A Special Educational Program In Higher Education. Unpublished Dissertation, Ohio: Akron University, 1975.

11. Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D. K., Gintis, H., Heyns, B., & Michelson, S. Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basic Books, 1972.

12. Jencks, C., & Brown, M. Effects of high schools on their students. Harvard Educational Review, 1975, 45, 273-324.

13. Kelly, F., Newman, I., and McNeil, K. "Suggested Inferential Statistical Models for Research In Behavior Modification. The Journal of Experimental Education, Vol. 4, no. 4, Summer, 1973.

14. Klein, M., Newman, I. "Estimated Parameters of Three Shrinkage Formuli." Multiple Linear Regression Viewpoints, March 1974, Publication of the Special Interest Group on Multiple Linear Regression of AERA.

15. Lyecke, D. F. and McGinn, N. F. "Regression Analyses and Educational Production Functions: Can They be Trusted?" Harvard Educational Review Vol. 45 #3, August 1975.

16. McNeil, K., Kelly, F., and McNeil, J. Testing Research Hypotheses Using Multiple Linear Regression. Southern Illinois University Press: Carbondale, Illinois, 1975.

17. Mood, A. M. Macro-Analysis of the American Educational System. Operations Research, 1969, 17, 770-784.

18. Mood, A. M. Partitioning Variance in Multiple Regression Analyses as a Tool for Developing Learning Models. A.E.R.J., March, 1971, 191-202.

19. Newman, I. "A Demonstration of Multiple Regression Models that will Facilitate the Investigation of Trends and Functional Relationships. A paper presented to the Psychology Division at the Ohio Academy of Science, April, 1974.

20. Newman, I. Some Further Considerations of Using Factor Regression Analysis. Multiple Linear Regression Viewpoints, October, 1972, 39-41.

21. Newman, I. and Newman, C. A discussion of Component Analysis: Its intended purpose, strengths and weaknesses. A paper presented the American Educational Research Association, 1975; Multiple Linear Regression Special interest group.

22. Newman, I. "Variations Between Shrinkage Estimation Formula and The Appropriatness of Their Interpretation," Multiple Linear Regression Viewpoints. Vol. 4, 2, 45-48, August 1973.

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

24. Scheffe, H. The analysis of variance. New York: John Wiley & Sons, 1959.

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

26. Williams, J. D. Four-way disproportionate hierarchial models. Multiple Linear Regression Viewpoints, 1974, 5, #2.

27. Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Multiple Linear Regression Viewpoints
Vol. 6, No. 4, 1976

CANONICAL ANALYSIS AS A GENERALIZED REGRESSION TECHNIQUE FOR MULTIVARIATE ANALYSIS

JOHN D. WILLIAMS
The University of North Dakota

The use of characteristic coding (dummy coding) is made in showing solutions to four multivariate problems using canonical analysis. The canonical variates can be analyzed by the use of multiple linear regression. When the canonical variates are used as criteria in a multiple linear regression, the R^2 values are equal to θ, where θ is the squared canonical correlation coefficient. Several different methods exist for testing multivariate hypotheses. Where the interest is in a two-way disproportionate multivariate analysis of variance, the trace criterion ($\Sigma\theta_i$) seems particularly applicable.

Characteristic (dummy) coding has been used in multiple linear regression to analyze univariate analysis of variance problems; the same coding scheme can be extended to multiple criteria. While the resulting data are analyzed through canonical analysis, the design matrix conforms to the usual multiple linear regression design matrices. Thus, the utilization of multiple criteria can be pursued in a logical sequence without necessitating continuously changing the entire terminology.

In the present paper, four multivariate research designs are examined in a canonical analysis framework: a multivariate two-group situation, sometimes referred to as Hotelling's T^2 test; a multivariate multiple group situation; a multivariate two-way analysis; and a multivariate two-way analysis with disproportionate cell frequencies.

Tests of Significance in Canonical Analysis

In multivariate analysis, several different tests of significance are used. Typically, the multivariate analysis of variance has focused on solving the following equation for λ_i:

$$\left| S_{12}S_{22}^{-1}S_{21} - \lambda(S_{11} - S_{22}^{-1}S_{21}) \right| = 0,$$

whereas canonical analysis has focused on solving

$$\left|S_{12}S^{-1}_{22}S_{21} - \theta S_{11}\right| = 0$$

where $\theta = R^2_{c_i}$ and $\theta_i = \frac{\lambda_i}{1 + \lambda_i}$; also,

S_{11}, S_{12}, S_{21} and S_{22} are variance - covariance matrices.

Roy's (1957) largest root criterion tests the significance of the largest characteristic root. Hotelling's (1951) trace criterion tests the overall multivariate hypotheses for all dimensions simultaneously and is given by trace = $\Sigma\lambda_i$. Tables for testing either of these two tests have been given by Pillai (1960). A trace criterion using Pillai's tables wherein the trace is $\Sigma\theta_i$ is useful in testing the significance of the overall set of canonical correlations and is analogous to Hotelling's trace criterion.

Wilks Λ also provides a test of the overall hypothesis. All necessary tables for testing these hypotheses can be found in Timm (1975). Typically, in canonical analysis simultaneous tests of each characteristic root using Roy's approach is used. On the other hand, multivariate analysis of variance programs usually employ an overall test (either the trace criterion or Λ). Harris (1975) has argued that the use of the largest root criterion is more sensible in that if either the trace criterion or Λ shows significance but the largest root criterion does not, then the differences among the groups cannot be pinpointed by any single linear combination of variables. Harris would see the use of the overall hypotheses as being more useful in only those cases where λ_1 and λ_2 are close to the same value.

A Multivariate Two'Group Situation (Hotelling's T^2 Test)

The simplest multivariate analysis of variance situation is the multivariate analog to the usual t test; here, several criteria are observed for two groups and an overall test for group differences can be made. As an example, suppose four criteria (Y_1, Y_2, Y_3 and Y_4) are observed for two groups as indicated in Table 1.

Table 1

Four Criteria Y_1, Y_2, Y_3, and Y_4

for the Multivariate Two-Group Situation

Group 1				Group 2			
Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
20	17	17	25	15	26	13	16
22	19	16	28	19	25	15	15
24	14	18	23	23	21	17	17
26	16	17	17	24	17	22	18
28	18	16	29	25	19	16	22
30	20	15	32	26	14	8	23
32	22	14	35	27	22	14	24
34	16	16	42	28	20	18	20
36	9	18	38	30	17	21	18

To accomplish a canonical analysis with the data in Table 1, it is necessary to define a <u>first</u>-set and a <u>second</u>-set. For convenience, the criteria will constitute the first set and the predictors (group membership variables) will constitute the second set. Actually, only a single group membership variable is necessary:

X_1 = 1 if a member of Group One; 0 otherwise.

Table 2 contains the criteria and design matrix necessary to accomplish this analysis.

Table 2

Criteria and Design Matrix
for a Two-Group Multivariate Analysis

Y_1	Y_2	Y_3	Y_4	X_1
20	17	17	25	1
22	19	16	28	1
24	14	18	23	1
26	16	17	17	1
28	18	16	29	1
30	20	15	32	1
32	22	14	35	1
34	16	16	42	1
36	9	18	38	1
15	26	13	16	0
19	25	15	15	0
23	21	17	17	0
24	17	22	18	0
25	19	16	22	0
26	14	8	23	0
27	22	14	24	0
28	20	18	20	0
30	17	21	18	0

Using canonical analysis to find the relationship between the left set and the right set, the following results are obtained:

$\theta = .60066$; canonical $R_c = \sqrt{\theta} = .77502$;

Wilk's $\Lambda = .39934$, with $p < .01$. Also, the trace = .60066; in every case, $p < .01$.

The coefficients for the first set are

	First root
Y_1	-.65403
Y_2	-.32522
Y_3	.24849
Y_4	1.26372.

The coefficient for the X_1 variable is of course 1.000.

Also,

$\bar{Y}_1 = 26.0556$, $s_{Y_1} = 5.3189$;

$\bar{Y}_2 = 18.4444$, $s_{Y_2} = 4.0905$;

$\overline{Y}_3 = 16.1667$, $s_{Y_3} = 3.0534$; and

$\overline{Y}_4 = 24.5556$, $s_{Y_4} = 7.9648$.

If a new variable, Y_5, is formed as

$$Y_5 = -.65403(Y_1 - \overline{Y}_1)/s_{Y_1} - .32522(Y_2 - \overline{Y}_2)/s_{Y_2} + .24849(Y_3 - \overline{Y}_3)/s_{Y_3} + 1.26372(Y_4 - \overline{Y}_4)/s_{Y_4},$$

then a regression equation can be formed as

$$Y_5 = b_0 + b_1X_1 + e_1. \qquad (1)$$

If a regression is completed with the formulation in equation 1, then

$R = .77502$,

$R^2 = .60066$, and

$1-R^2 = .39934$.

This information is identical to that found in the use of the canonical analysis; the relationship is, for the two-group situation:

$R = \sqrt{\theta} = R_c$;

$R^2 = \theta$ and $\Lambda = 1-R^2$.

The use of equation 1 shows that a composite variable, Y_5, is a linear composite of variables Y_1, Y_2, Y_3 and Y_4 such that the relationship with X_1 remains maximized.

Because there is only one group membership variable involved, an interesting reversal of the roles of the criteria and predictor can be made:

$$X_1 = b_0 + b_1Y_1 + b_2Y_2 + b_3Y_3 + b_4Y_4 + e_2. \qquad (2)$$

If equation 2 is utilized in a multiple regression framework,

$R = .77502$, $R^2 = .60066$, $1\text{-}R^2 = .39934$ and $F = 4.888$.

The first three results were obtained in the prior two analyses.

There are some differences between the two analyses, however. For the use of equation 2, $b_1 - b_4$ are different from the coefficients for the first set given previously; this, of course, was to be accepted. The beta coefficients also differ from the coefficients given earlier;

$\beta_1 = -.50688$,

$\beta_2 = -.25206$,

$\beta_3 = .19259$,

$\beta_4 = .97940$.

If some thought is given to it, this difference comes as no surprise either. In a canonical analysis, each canonical variate has a mean of zero and standard deviation of one. In a regression analysis, the beta coefficients are such that for every predictor variable, there is a mean of zero and standard deviation of one. The difference is that in canonical analysis, the new variate is created with mean zero and standard deviation of one.

Finally if a multivariate analysis of variance program is executed, $\Lambda = .39934$ and $F = 4.888$, results that were obtained earlier. Thus, if the interest is in comparing two groups on several criteria simultaneously, several different strategies allow equivalent solutions. In this special case, the execution of equation 2 (using the group membership variable as the criterion and the Y variables as predictors) is perhaps the easiest solution to employ. The use of canonical analysis and subsequent formation of a composite variable would also seem to be of some value.

A Multivariate Multiple Group Situation

If several groups are involved in the analysis with multiple criteria, then the usual one-way multivariate analysis of variance is often employed. As an example of such a situation, suppose three criteria are available for three groups of subjects. Such a situation is encountered in Table 3.

Table 3

Criteria and Design Matrix for Multivariate Analysis of Variance Through Regression

Y_1	Y_2	Y_3	X_1	X_2
17	23	1	1	0
22	28	2	1	0
14	22	3	1	0
18	27	4	1	0
29	25	5	1	0
22	32	6	0	1
24	34	8	0	1
26	36	10	0	1
28	42	12	0	1
25	31	14	0	1
26	23	15	0	0
29	32	16	0	0
32	29	17	0	0
35	42	18	0	0
33	23	19	0	0

Two group membership variables are used:

X_1 = 1 if a member of Group 1; 0 otherwise, and

X_2 = 1 if a member of Group 2; 0 otherwise.

Using the data in Table 3, a canonical analysis is performed with the Y scores (criteria) as the first set and the X variables (predictors) as the second set. Several useful items are typically available from a canonical analysis. Either the canonical roots or canonical correlations (or both) will be available. For the data in Table 3, θ_1 = .89286, and θ_2 = .43602. The first canonical correlation is $\sqrt{\theta_1}$ = .94491, and the second canonical correlation is $\sqrt{\theta_2}$ = .66032. The weights for the Y side and X side are:

Y side weights	1	2	3
1	.09556	-.01807	-1.07322
2	-.79902	1.10739	.27948

X side weights	1	2
1	1.15457	.56235
2	.01725	1.00851

It is interesting to form variables to correspond to those suggested by the Y side and X side weights and investigate these transformed variables using ordinary multiple regression. To utilize the weights, it is first necessary to transform all the data in Table 1 into z scores. As $\overline{X}_1 = .3333$, $\overline{X}_2 = .3333$, $\overline{Y}_1 = 25.3333$, $\overline{Y}_2 = 29.9333$, $\overline{Y}_3 = 10.0$, $s_{x_1} = .4880$, $s_{x_2} = .4880$, $s_{y_1} = 6.0198$, $s_{y_2} = 6.5407$ and $s_{y_3} = 6.2678$, the transformation equations are:

$$Z_1 = \left[.09556(Y_1-25.3333)/6.0198\right] + \left[-.01807(Y_2-29.9333)/6.5407\right] + \left[-1.07322(Y_3-10.)/6.2678\right]$$

$$Z_2 = \left[-.79902(Y_1-25.3333)/6.0198\right] + \left[1.10739(Y_2-29.9333)/6.5407\right] + \left[.27938(Y_3-10.)/6.2678\right]$$

$$Z_3 = \left[1.15457(X_1-.3333)/.4880\right] + \left[.56235(X_2-.3333)/.4880\right]$$

$$Z_4 = \left[.01725(X_1-.3333)/.4880\right] + \left[1.00851(X_2-.3333)/.4880\right]$$

Using Z_1 as the criterion and X_1 and X_2 as predictors, $R^2 = .89286$, $R = .94491$, identically the same results as found for the first canonical root. Similarly, using Z_2 as the criterion and X_1 and X_2 as predictors, $R^2 = .43602$ and $R = .66032$. If Y_1, Y_2, and Y_3 are used as predictors of Z_3 and then Z_4, again the canonical correlations appear as multiple correlations.

Also the following correlations are of interest:

$r_{Z_1Z_2} = 0$ $\qquad$ $r_{Z_2Z_3} = 0$

$r_{Z_1Z_3} = .94491$ $\qquad$ $r_{Z_2Z_4} = .66032$

$r_{Z_1Z_4} = 0$ $\qquad$ $r_{Z_3Z_4} = 0$

If a traditional multivariate analysis of variance is performed, the test for H_2 (overall difference among all groups) yields $\Lambda = .06042$. While some canonical printouts (such as Cooley and Lohnes, 1971) include this value as part of the output, Λ can be found as $\Pi_i^k\ (1-\theta_i)$ where the θ_i are the canonical roots. For this particular data, $\Lambda = (1-.89286)(1-.43602) = .06042$. Because $\theta_i = R_i^2$, this result can be written as $\Lambda = \Pi_i^k(1-R_i^2)$. The trace criterion yields Tr = .89286 + .43602 = 1.2888.

Also, from the multivariate analysis printout, an F = 10.227 ($p < .01$) tests the overall group differences among the three groups.

The use of the canonical variates found through the use of the canonical vectors should present an attractive alternative to those researchers who wish to complete multiple comparisons after the rejection of the overall null hypothesis. One suggestion, made by Hummel and Sligo (1971) is to compare the groups on univariate tests on each variable after the rejection of the overall null hypothesis. An alternative is to use the first canonical variate for the criteria set and run an analysis among the groups with this (and subsequent) canonical variates from the criteria set. Scheffé's test would seem appropriate as a multiple comparison method.

A Two-Way Multivariate Situation

The two-way multivariate analysis of variance is not quite as available as the one-way multivariate analysis of variance, but several programs are available, including Cramer (1974), Finn (1974), Ondrack (1974), and Cooley and Lohnes (1971). Suppose three criteria are measured in a 2X3 design. The design matrix is given in Table 4.

Table 4

Criteria and Design Matrix For Two-Way
Analysis of Variance Through Regression

Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
13	18	16	1	0	0	0	0	1	1	0
14	19	17	1	0	0	0	0	1	1	0
15	18	22	1	0	0	0	0	1	1	0
16	15	27	1	0	0	0	0	1	1	0
14	19	20	1	0	0	0	0	1	1	0
17	18	23	1	0	0	0	0	1	1	0
15	9	13	0	1	0	0	0	1	0	1
17	13	21	0	1	0	0	0	1	0	1
16	16	12	0	1	0	0	0	1	0	1
17	18	22	0	1	0	0	0	1	0	1
14	13	17	0	1	0	0	0	1	0	1
15	11	18	0	1	0	0	0	1	0	1
16	17	27	0	0	1	0	0	1	0	0
16	14	21	0	0	1	0	0	1	0	0
17	16	23	0	0	1	0	0	1	0	0
18	15	24	0	0	1	0	0	1	0	0
20	13	22	0	0	1	0	0	1	0	0
22	9	18	0	0	1	0	0	1	0	0
17	15	16	0	0	0	1	0	0	1	0
18	17	15	0	0	0	1	0	0	1	0
19	19	14	0	0	0	1	0	0	1	0
20	21	14	0	0	0	1	0	0	1	0
21	23	18	0	0	0	1	0	0	1	0
22	25	19	0	0	0	1	0	0	1	0
13	17	22	0	0	0	0	1	0	0	1
15	13	22	0	0	0	0	1	0	0	1
13	16	21	0	0	0	0	1	0	0	1
12	12	17	0	0	0	0	1	0	0	1
11	11	19	0	0	0	0	1	0	0	1
11	17	18	0	0	0	0	1	0	0	1
17	15	9	0	0	0	0	0	0	0	0
18	17	8	0	0	0	0	0	0	0	0
20	15	7	0	0	0	0	0	0	0	0
19	17	6	0	0	0	0	0	0	0	0
17	19	5	0	0	0	0	0	0	0	0
16	22	4	0	0	0	0	0	0	0	0

where

X_1 = 1 for a member of Row 1 and Column 1, 0 otherwise;

X_2 = 1 for a member of Row 1 and Column 2, 0 otherwise;

X_3 = 1 for a member of Row 1 and Column 3, 0 otherwise;

X_4 = 1 for a member of Row 2 and Column 1, 0 otherwise;

X_5 = 1 for a member of Row 2 and Column 2, 0 otherwise;

X_6 = 1 for a member of Row 1, 0 otherwise;

X_7 = 1 for a member of Column 1, 0 otherwise and

X_8 = 1 for a member of Column 2, 0 otherwise.

An analysis of the data in Table 5 is necessarily complex. Four sets of canonical relationships are possible; the Y_1, Y_2 and Y_3 variables can be related to X_1 - X_5, then X_6 and X_7, then X_8 and finally X_6, X_7 and X_8. In the univariate situation these relationships correspond to the full model, the rows, the columns and the rows and columns as predictors. Table 5 contains the various canonical relationships from these four different sets of predictors.

Table 5

Eigenvalues, Canonical Correlations and Wilks Lambda For Two-Way Multivariate Analysis of Variance by Regression

Preditors: X_1, X_2, X_3, X_4, X_5 (Full Model)

	$R^2_{c_i}$	Canonical R	Λ
First Root	.86535	.93024	.03594 **
Second Root	.58716	.76626	.26692 **
Third Root	.35345	.59452	.64655 **

Predictors: X_6 (Rows)

First Root	.32400	.56921	.65600 **

Predictors: X_7, X_8 (Columns)

First Root	.45924	.67767	.40033 **
Second Root	.25968	.50959	.74032

Predictors: X_6, X_7, X_8 (Rows and Columns)

First Root	.54326	.73706	.25558 **
Second Root	.30626	.55341	.55957 **
Third Root	.19340	.43977	.80660 **

** significant at .01 level

Corresponding to each canonical root are the weights to create the canonical variables; they can be found (for the Y side only) in Table 6.

Table 6

Canonical Weights for Canonical Variates from Table 5

Predictors: X_1 - X_5

	Y-Weights		
Variate	1	2	3
Z_1	-.44831	-.20691	.73004
Z_2	-.85876	-.05983	-.68227
Z_3	.40964	-1.02427	-.23006

Predictor: X_6

Variate	1	2	3
Z_4	.16443	-.45391	.85726

Predictors: X_7, X_8

Z_5	.62240	.65303	.10464
Z_6	-.62870	.69649	.59305

Predictors: X_6, X_7, X_8

Z_7	-.46410	-.62886	.32592
Z_8	.00110	-.58279	-.91247
Z_9	.94386	-.60035	.33548

If Z_1 - Z_9 are used as criteria in a multiple linear regression layout, then when X_1 - X_5 are used as predictors of Z_1, $R^2 = .86535 = \theta_1$; similar findings will occur with the Z variates that correspond to the X predictors used in the original canonical analysis. If the trace (sum of the squared canonical correlations) of the four models are found

$$Tr\ (Full) = .86535 + .58716 + .35345 = 1.80596$$

$$Tr\ (Rows) = .32400$$

$$Tr\ (Columns) = .44924 + .25968 = .71892$$

Tr (Rows and Columns) = .54326 + .30626 + .19340 = 1.04292

also,

Tr (Rows) + Tr (Columns) = Tr (Rows & Columns) .

The interaction can be defined as

Tr (Full) - Tr (Rows & Columns) = 1.80596 - 1.04292 = .76304.

The sum of the squared canonical correlations can be broken down into the separate R_c^2 values through the use of orthogonal coefficients, as there are an equal number of entries in each of the six cells. If five new variables are defined as follows

X_9 = 1 if a member of Row 1, -1 if a member of Row 2;

X_{10} = 1 if a member of Column 1, 0 if a member of Column 2, -1 if a member of Column 3;

X_{11} = 1 if a member of either Column 1 or Column 3, -2 if a member of Column 2;

$X_{12} = X_9 \cdot X_{10}$ and

$X_{13} = X_9 \cdot X_{11}$.

Using X_9 as the predictor of the three criteria, $R_c^2 = .32400$, the same result as found in Table 6 for rows. When X_{10} and X_{11} are used as predictors, $R_{c_1}^2 = .45924$ and $R_{c_2}^2 = .25968$, the same result as found in Table 5 for columns. If X_9, X_{10}, X_{11}, X_{12} and X_{13} are used as predictors, $R_{c_1}^2 = .86535$, $R_{c_2}^2 = .58716$ and $R_{c_3}^2 = .35345$, the same results as found for the full model. If X_9, X_{10} and X_{11} were used as predictors, the results would duplicate those found by using X_6, X_7 and X_8 as predictors. If X_{12} and X_{13} are used as predictors, the following results are found:

$R_{c_1}^2 = .42305 \quad R_{c_1} = .65042 \quad \Lambda_1 = .38079, \quad p < .01;$ and

$R_{c_2}^2 = .33999 \quad R_{c_2} = .58309 \quad \Lambda_2 = .66001, \quad p < .01.$

Finding the interaction directly through the use of orthogonal polynomials appears to be limited to those cases in which the cell entries are either equal or proportional. The last problem to be discussed considers the multivariate disproportional case.

A Two-Way Disproportionate Multivariate Analysis

An analysis similar to the one employed for the two-way equal cell case shown in Table 4 can be considered. In fact, the same 36 "subjects" are reconsidered, after deliberately creating a disproportionate situation. The first 3 subjects are, for the disproportionate case, in cell 1 (Row 1 and Column 1); the next four subjects are in Cell 2 (Row 1 and Column 2); the next 10 scores are in Cell 3 (Row 1 and Column 3; the next 9 subjects are in Cell 4 (Row 2 and Column 1); the next 7 subjects are in Cell 5 (Row 2 and Column 2); finally, the last 3 subjects are in Cell 6 (Row 2 and Column 3). The number of entries in each cell for the 2X3 layout is as given in Table 7.

Table 7

Frequencies for 2X3 Multivariate Analysis with Disproportionate Cells

	Column 1	Column 2	Column 3
Row 1	3	4	10
Row 2	9	7	3

The design matrix is as before with X_1 - X_8 having the same meaning. The results of the canonical analysis are found in Table 8.

Table 8

Eigenvalues, Canonical Correlations and Wilks Lambda for Two-Way Disproportionate Cell Frequencies

Predictors: X_1, X_2, X_3, X_4, X_5 (Full Model)	$R^2_{c_i}$	Canonical R	Λ
First Root	.61960	.78715	.22565**
Second Root	.31954	.56528	.59320*
Third Root	.12824	.35811	.87176
Predictors: X_6 (Rows)			
First Root	.27262	.52213	.72738*
Predictors: X_7, X_8 (Columns)			
First Root	.21675	.46557	.76287
Second Root	.02601	.16127	.97399
Predictors: X_6, X_7, X_8 (Rows and Columns)			
First Root	.34802	.58993	.49760**
Second Root	.21674	.46555	.76322
Third Root	.02559	.15998	.97441

*significant at .05 level
**significant at .01 level

Interpretation of the data in Table 8 may be made, but the lack of consensus on interpreting univariate disproportionate situations will only be increased as the situation becomes multivariate. Many authors prefer the "fitting constants" solution (see Anderson & Bancroft, 1952, Overall and Spiegel, 1969, and Rao, 1965). Cohen (1968) describes a partioning solution called the hierarchical model. An unadjusted main effects solution is shown in Williams (1972). Searle (1971) and Appelbaum and Cramer (1974) prefer a multiple step decision making process that combines the fitting constants solution and the unadjusted main effects solution. The multivariate situation is complicated by the existence of several criteria for judging the significance of an experiment. The approach taken here is to describe both the fitting constants solution and the unadjusted main effects solution; those who prefer the decision rules given in Searle could easily employ them with the information given.

The Unadjusted Main Effects Solution

The unadjusted main effects solution follows in a manner very similar to the one presented in regard to the equal cell case. In fact, the data in Table 8 can be interpreted (except for the interaction) as being an unadjusted main effects solution. The interaction can be found as the difference between the trace of the full model and the trace of the rows and columns model:

$$(.61960 + .31954 + .12824) - (.34802 + .21674 + .02559) = .47703.$$

The only available method to test the interaction hypothesis is Pillai's trace criterion; $p < .05$.

A Multivariate Analog to the Fitting Constants Solution

Because of the disproportionality of the data, the direct calculation of the R_c^2 terms is precluded; the traces can be found in a manner similar to finding the trace for the interaction, however. The trace for rows (after removing the effect for columns) can be found as the trace for rows and columns minus the trace for columns:

$$(.34802 + .21674 + .02559) - (.21675 + .02601) = .34759, \ p < .01.$$

The trace for columns can be found as the trace for rows and columns minus the trace for rows:

$$(.34802 + .21674 + .02559) - (.27262) = .31773, \ p > .05.$$

The interaction is the same as given for the unadjusted main effects model.

Discussion

While Harris's argument for the use of the greatest characteristic root criterion as a measure for multivariate analysis is noteworthy, the trace criterion ($\Sigma\theta_i$) is particularly useful with disproportionate cell frequencies. If the focus of the greatest characteristic root and the corresponding canonical variates are made on the full model (cell model, or full rank model) then either criterion is applicable, and perhaps Harris's suggestion is appropriate. If the intent is on producing a two-way MANOVA with disproportionate cells and there is interest in the row, column and interactions effects, then Pillai's trace criterion is most appropriate. Even where there is interest in the usual effects, the most likely canonical variate to be of interest is the variate associate with the greatest characteristic root from the full model.

Four different multivariate applications have been shown herein. Other applications (multivariate trend analysis, multivariate analysis of covariance and other analogs to univariate designs) are possible through a canonical approach. Also, the univariate analyses that can be performed by multiple linear regression can be conceptualized as a canonical problem. That the canonical analyses and multiple regression analyses would yield quite similar results is not quite the same as saying that the analyses are identical for the univariate situation. As was shown in Hotelling's T^2 test, some differences in weighting coefficients occur. The overall results (R^2's) are identical, however.

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 nonorthogonal 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.

Cooley, W.W. and Lohnes, P.R. Multivariate data analysis. New York: Wiley, 1971.

Cramer, E.M. Revised MANOVA program. Chapel Hill, N.C.: The L.L. Thurstone Psychometric Laboratory, The University of North Carolina, 1974.

Finn, J.D. A general model for multivariate analysis. New York: Holt, Rinehart and Winston, 1974.

Harris, R.J. A primer of multivariate statistics. New York: Academic Press, 1975.

Hotelling, H. A generalized T-test and measure of multivariate dispersion. Proceedings of the second Berkeley symposium on mathematics and statistics, 23-41, 1951.

Hummel, T.J. and Sligo, J.R. Empirical comparison of univariate and multivariate analysis of variance procedures. Psychological Bulletin, 1971, 76, 49-57.

Ondrack, M. Analysis of variance designs: A manual for the computer analysis of selected models. Toronto: Institute for Behavioral Research, York University, 1974.

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

Pillai, K.C.S. Statistical tables for tests of multivariate hypotheses. Manila: Statistical Center, 1960.

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

Roy, S.N. Some aspects of multivariate analysis. New York: Wiley, 1957.

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

Timm, N.H. Multivariate analysis with applications in education and psychology. Belmont, Calif: Wadsworth, 1975.

Williams, J.D. Two way fixed effects analysis of variance with disproportionate cell frequencies. Multivariate Behavioral Research, 1972, 7, 66-83.

Multiple Linear Regression Viewpoints
Vol. 6, No. 4, 1976

CAUSAL INFERENCE: MULTIPLE LINEAR REGRESSION vs. ANALYSIS OF VARIANCE ORTHOGONAL AND NON-ORTHOGONAL DESIGNS

KEITH MCNEIL
Educational Monitoring Systems

JUDY PLATT
Graduate Student, University of Michigan

Viewpoint: Causal statements about continuous predictors (derived from correlational designs) which are correlated may be more fruitful than those about categorical predictors (derived from analysis of variance designs) which are orthagonal.

Since ANOVA designs are a subset of MLR, all of the causal statements permissable with ANOVA are permissable in MLR, if two requirements hold: random assignment and orthogonality hold. Random assignment means that the researcher manipulated the predictor variables. Orthogonality implies that the predictors are uncorrelated. When these conditions hold, sample results are generalizable to the population. Unfortunately, few researchers are able to totally manipulate their predictors. If manipulation is obtained, it is usually in an environment which is somewhat different from the real world to which a generalization is desired.

When one forces predictors to be orthogonal, (as in ANOVA designs), one may be tampering with how those variables are related in the real world. Even if Ss are randomly assigned to the various levels in, say, a 2 x 3 orthogonal design, it seems that a logical falacy exists which has not previously been emphasized. Let us assume in the 2 x 3 design that there is neither an interaction effect nor a B main effect, although the A main effect is highly significant. One would identify A as a cause of the criterion variable. Upon subsequent manipulation of A, certain changes in the criterion would presumably be observed. But because the B variable in the real world is correlated with the A variable, manipulation of A could well change (manipulate) B, and hence the observed change in the criterion would be caused by both A and B, not just caused by A as inferred from the 2 x 3 design. In short, analyses of orthogonal data leads to erroneous conclusions, because the orthogonal variables are not representative of the real world.

What results from an analysis wherein the predictors have been artificially orthogonalized is a permissable causal inference to an artificial population. One might say the causal inference has internal validity but lacks to some extent the external validity for the population about which one is really concerned. [It should be noted that there have been recent attempts to handle the problem of non-orthogonal ANOVA designs. These attempts have been in response to missing data, and have relied heavily on regression models.]

Another inferential weakness of ANOVA concerns the analysis of predictors which are continuous. The MLR approach is uniquely applicable as the ANOVA designs can only handle categorical predictors.

At this point in methodological development, precise causal inferences with correlated predictors are not possible. The work in path analysis (Blalock, 1964) and commonality analysis (Newman and Newman, 1975) may eventually lead to a solution, but as yet only tentative guesses based on theory lead to causal inferences. But we would argue that since continuous variables reflect the world better than do categorical variables, that eventually the "causal payoff" will be with continuous variables.
We definitely feel that if one wants to describe or to predict, that continuous variables should be utilized. And even though the ultimate goal of scientific investigation should be ascertaining causal functions, we probably need to do a lot more in the area of describing and predicting.

One procedure which some investigators follow is to assess the effects of a predictor "over and above" the effects of other predictors. One does this by eliminating that predictor from the full model and measuring the loss in R^2. The problem here, of course, is to decide which variables to include in the restricted model. Some reasons for including variables would be: 1) theoretical - theory says they are dominant, 2) temporal - an "ultimate" cause must preceed an "intermediate" cause, or 3) confounding - a variable which may have an effect, but you were not able to stratify upon. One might argue, though, that extracting or focusing on one of a set of causers is not a beneficial enterprise. That is, the criterion might best be understood and dealt with in the presence of all the necessary causal predictors.

In conclusion, artificial categorization of continuous predictors, and artificial orthogonalizations of correlated variables most likely leads one to make incorrect causal interpretations. Dealing with correlated observations in a MLR framework may eventually lead to valid causal conclusions.

REFERENCES

Blalock, H. M. Causal Inferences in Nonexperimental Research. Chapel Hill: University of North Carolina Press, 1964.

Newman, I. and Newman, C. A discussion of component analysis: its intended purpose, strengths, and weaknesses. Multiple Linear Regression Viewpoints, 1975, 5 (4), 27-34.

ERRATUM

All SIG members should have received by now a copy of the text: TESTING RESEARCH HYPOTHESES USING MULTIPLE LINEAR REGRESSION by Keith McNeil, Francis Kelly, and Judy McNeil. The text was published by Southern Illinois University Press in 1975. There is a typographical error on page 514, as the line starting with "43" should read:

43 21 17 34 19 10 89 87 1001 98 17 13

NOTE: The data listing in the accompanying Manual is correct.

Multiple Linear Regression Viewpoints
Vol. 6, No. 4, 1976

SHOULD A FIRST COURSE IN ANOVA BE THROUGH MLR?

JOHN D. WILLIAMS
The University of North Dakota

While many adherents to the MLR approach might be oriented toward answering the question given in the title to this paper with a "yes", several points of view should be examined. Given that it makes pedagogic sense (which traditionalists may dispute) to use a linear models approach, practical considerations may temper the decision to implement MLR in the teaching of ANOVA.

First, a most important consideration is an adequate computer together with the necessary software. As most colleges and universities will have the computer and necessary software, the concern changes to accessibility to students. Are key-punch machines readily available? Are other computer services cost-free to the student? Is the software easily accessible? If the answers to any of these questions is no, then implementing a MLR approach can be extremely frustrating.

Second, it must be remembered that most students in an applied course are going to eventually work on a research project of some sort and will, of course, have to be able to communicate with their major advisor. In general, most of our non-statistician colleagues are familiar with traditional ANOVA designs and terminology. Thus, to help the student and his advisor to communicate, some degree of dependence on traditional ANOVA terminology is necessary.

Implementing a Regression Approach in a First Course in ANOVA

If the concerns just given are reasonably satisfied (computer availability, and the expectations of the major departments) then a definite yes can be given as an answer to the question posed in the title to this paper. Still to be resolved, though, is the question, "Should the course, if taught by MLR, use traditional ANOVA terminology?" Also, "Should such a course be oriented toward a direct translation of ANOVA type questions to MLR solutions?"

Remembering that students will eventually have to communicate with their major advisors and also will be reading journals using a traditional ANOVA format, the most judicious choice seems to me to use ANOVA terminology, but also point out the conveniences provided by the MLR approach. Perhaps this process can be illustrated through the use of the analysis of covariance (at least as I do this on a personal basis).

First, before any discussion is made in class of the linear models involved in a solution, the students are required to read either Lindquist's (1953), Edward's (1968), or Winer's (1971) presentation on the analysis of covariance. They are specifically instructed to follow the numerical example to completion and then reflect upon their cognitive understanding of the process. Then, in the ensuing class period, the students consider the same problems, but through multiple regression. For example*, if there are three groups and one covariate, then a full model can be defined as

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3 + e_1, \qquad (1)$$

where

Y = the score,

X_1 = the covariate,

X_2 = 1 if a member of Group 1, 0 otherwise,

X_3 = 1 if a member of Group 2, 0 otherwise,

b_o - b_3 are regression coefficients and e_1 is the error in prediction with this model.

*under the assumption of homogeniety of regression

Then a restricted model can be formed as

$$Y = b_0 + b_1X_1 + e_2. \qquad (2)$$

Equation 1, the full model, contained all the information (both covariate and group membership variables) whereas equation 2 included only the covariate. The "difference" between these models would then tell us what group differences there are that are independent of the covariate and are tested by

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

where

R^2_{FM} = the multiple correlation squared value from the full model;

R^2_{RM} = the multiple correlation squared value from the restricted model;

k = the number of groups, and

N = the number of subjects.

Several students often then bring up the question, "If it's this easy, then why did you make us read these other books?" Some ask, "Why would anyone ever use the methods described in those other books, if we can get the same answers through two simple linear models?"

While I have no answer to the last posed question, I would note that those students who ask either of the last two questions have become "converts" to the use of multiple linear regression for solving ANOVA problems.

As to whether the course should be oriented toward a direct translation of ANOVA designs to MLR, the overriding concern (at least to me) regards the original data. If the data is associated with natural mutually exclusive groups (male-female, Protestant-Catholic-Jew-other) then a direct ANOVA type solution is appropriate; if the ANOVA design is arrived at through "cutting up" the subjects into groups such as high-middle-low on a continuous variable, it makes more sense to me to use the original continuous variable.

References

Edwards, A.L. Experimental design in psychological research. New York: Holt, Rinehart and Winston, 1968.

Lindquist, E.F. Design and analysis of experiments in psychology and education. Boston: Houghton-Mifflin, 1953.

Winer, B.J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

Multiple Linear Regression Viewpoints
Vol. 6, No. 4, 1976

POSITION STATEMENT ON THE ROLES AND RELATIONSHIPS BETWEEN STEPWISE REGRESSION AND HYPOTHESIS TESTING REGRESSION

KEITH MCNEIL
Educational Monitoring Systems

The two statistical procedures, which are the same in their underlying mathematical development, are quite different in their applications and interpretations. Therefore we need to be careful in identifying the procedure that we are using. Following is a description of my position on these topics. Further discussion by SIG members will be sought at AERA.

HYPOTHESIS TESTING REGRESSION- a full and restricted regression model are specified and tested. This test of significance answers the one research hypothesis that was stated.

STEPWISE REGRESSION- a whole sequence of models and tests resulting in a model which includes predictors which meet certain conditions for inclusion and which omits others on the basis of certain conditions for exclusion.

ROLES

STEPWISE- to determine a set of variables which "best" account for the criterion variance. The technique selects a parsimonious set of predictors, which are a function of that sample. No a priori hypotheses (other than selection of variables) are stated. The results are valuable, though, for future use in constructing hypotheses.

HYPOTHESIS TESTING REGRESSION- tests one a priori hypothesis in the presence of none or a set of variables. The full regression model contains the variable(s) to be tested and the covariable(s) (if any). The restricted model contains only the covariables.

One could use the hypothesis testing procedure to obtain the results of the stepwise procedure, but to do so would be laborious to say the least.

It might be instructive to do this, though, in order to discover just how many tests of significance are calculated in one stepwise analysis.

On the other hand, the stepwise procedure might not test the hypothesis of concern which is tested in the hypothesis testing procedure. We are thus reminded that there are an infinite number of hypotheses one can test against a particular full model.

The two roles are thus quite different, with neither procedure being able to efficiently reproduce the desired results of the other procedure. The two roles are necessary in empirical work, the stepwise procedure providing information for the development of hypotheses, and the hypothesis testing procedure testing those hypotheses.

RELATIONSHIPS BETWEEN STEPWISE AND HYPOTHESIS TESTING REGRESSION

1. Kinds of data analyzed. Most stepwise applications have involved only continuous variables. Few applications have included categorical variables or interactions of categorical and continuous variables. One reason is that only one out of a set of mutually exclusive categorical variables may end up in the model. How would one interpret such a result? Williams (1973) has developed a solution to the problem, but few users have turned to the technique. Additionally, almost all stepwise analyses have been performed on "static" data- data over which the researcher has not had any manipulative control. Many would propose that one necessary condition for causal inference is that the causer has been manipulated. (Unfortunately few hypothesis testing applications have manipulated the data either.)

In the recent past, members of this Special Interest Group have been over-zealous in defending the MLR procedure. We have published paper after paper illustrating how an ANOVA design (using categorical predictors) can be performed in MLR. I must admit that I am as guilty as anyone (although maybe not as sophisticated-who really is interested in Chi Square types of hypotheses [McNeil, 1974]???). It is time for us all to return to stating hypotheses and employing regression models which test those hypotheses.

2. Shrinkage estimates. Sample R^2 values are indeed overestimates of the population R^2, therefore one of the various shrinkage formulae should be used (see Newman, 1973; Klein and Newman, 1974). Since the sample R^2 reduction is a function of the total number of predictors, it would seem that stepwise applications ought to be reduced on the basis of the number of variables in the entire pool, not only on those in the resulting model. I am not aware of any discussion on this point other than that by Newman(1973). The usual dismal results in any cross-validation supports the view that stepwise results, in general, give an exaggerated picture of the true magnitude of predictibility.

3. Non-linear terms. Only a few applications of the stepwise procedure have included non-linear terms. In our recent text(McNeil, Kelly, and McNeil, 1975), we have emphasized two reasons for including non-linear terms- either the expected functional relationship is non-linear, or the way the construct has been originally measured needs to be modified.

One advantage of the stepwise procedure in this area is the possibility of, say, a second degree term appearing in the final model, but not the original linear term. Some traditional thinking suggests that the second degree term can appear in the model only if the linear term is also in the model. One disadvantage of the hypothesis testing procedure is that we too soon come to closure on what the expected functional relationship is, or how the construct should be measured. In a sense, the stepwise procedure (if non-linear terms are included) calculates the empirical answer to those questions.

4. Causal inferences. A detailed comparison of the two procedures on this point was presented by Jack Byrne and myself in Viewpoints in 1975. To summarize, most stepwise applications are based on one-shot studies that are not based on a priori hypotheses. In such instances, causal inferences are questionable. On the other hand, hypothesis testing applications have, in general, overemphasized the plausible causal inferences. These statements are based on the proposition that causal inferences should be made only after repetitive findings on manipulated findings.

REFERENCES

Byrne, J. F. and McNeil, K. A. A further discussion of issues related to causal inference. Multiple Linear Regression Viewpoints, 1975, 5(4), 1-10.

Klein, M. and Newman, I. Estimated parameters of three shrinkage estimate formuli. Multiple Linear Regression Viewpoints, 1974, 4(4), 6-11.

McNeil, K. A. The multiple linear regression approach to "Chi Square" hypotheses. The Journal of Experimental Education, 1974, 43, 53-55.

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

Newman, I. Variations between shrinkage estimation formulas and the appropriateness of their interpretation. Multiple Linear Regression Viewpoints, 1973, 4(2), 45-48.

Williams, J. D. Applications of setwise regression analysis. Multiple Linear Regression Viewpoints, 1973, 4(2), 1-7.

Multiple Linear Regression Viewpoints
Vol. 6, No. 4, 1976

BRIEF NOTE ON THE JUSTIFICATION FOR USING MULTIPLE LINEAR REGRESSION

ISADORE NEWMAN
The University of Akron

People involved in research should be able to justify the statistical procedure they select. This is especially necessary for students who must be able to explain their choice to advisors who may be unfamiliar with the chosen techniques. Therefore, with the increasing use of multiple linear regression, which is unfamiliar and/ or misunderstood by many, it is important to clarify reasons why multiple linear regression is an appropriate and in many cases a preferable procedure. This note will present some of the general arguments which can support the use of regression.

The F-test, which is the analysis of variance, is a statistical technique (a test of significance) and is calculated on the basis of a least squares solution. It has unfortunately been confused with what has become known as traditional, one way or factorial analysis of variance. Traditional analyses of variance tend to confound the statistical procedures with the research design. However, if one separates the two, some of the advantages of the regression hypotheses testing procedures become more obvious. A few of these advantages are:

a.) Multiple Linear Regression (MLR) is the general case of the least sum of squares solution. Chi squares, t and F tests are special cases of the least squares solution. Therefore anytime anyone could use any of the special cases of the least squares solution, the more general case would also be appropriate.

b.) A significant F on a factorial design is more difficult to interpret. It may not reflect your specific hypothesis. With the regression procedure one states the hypothesis and then writes the regression model to test that hypothesis. So every test of significance is a test of a specific hypothesis.

c.) Regression is more flexible in being able to write the models that specifically reflect the research hypotheses.

d.) With traditional analysis of variance one can only ask interaction questions that have catagorical variables interacting with catagorical variables. With regression one can ask interaction questions between catagorical variables, between catagorical and continuous variables or between continuous variables. Since regression can deal with catagorical and continuous variables, it is more flexible in its ability to reflect real-world problems. With regression, there is no need to catagorize variables that are continuous in nature as required by traditional ANOVA;therefore,one would not lose degrees of freedom and power. (McNeil, Kelly, McNeil, 1975, Kerlinger, 1973).

e.) All analyses of covariance procedures are really regression procedures because the covariate(s) are always held constant by regressing it on the criterion. The multiple linear regression procedure makes the covariance procedure easier to calculate and interpret (Kerlinger, 1973).

f.) Regression also facilitates the calculation and interpretation of trends (functional relationship). Trends which are usually continuous in nature must be catagorized when traditional analysis of variance is used. Since regression can deal with continuous variables, no artificial categories must be imposed.

g.) The applied researcher must often deal with unequal N's and nonorthogonal designs. When these problems occur and one is using traditional analyses of variance, a correction is required. All of the corrections that produce the three exact solutions are regression procedures (Newman, Deitchman, Burkholder, Sanders, Ervin, 1976). In other words, if one has unequal N's and they are using traditional analysis of variance, regression will have to be used regardless of whether the research is aware of it or not. Once again regression is more flexible.

In conclusion, whatever can be done with traditional analysis of variance you can do with MLR, and more.

REFERENCES

Kerlinger, Fred. Foundations of Behavioral Research, Holt, Rinehart and Winston, New York, 1973.

McNeil, K., Kelly, F., and McNeil, J. Testing Research Hypotheses Using Multiple Linear Regression, Southern Illinois University Press: Carbondale, Illinois, 1975.

Newman, I., Deitchman, R., Burkholder, J., Sanders, R., and Ervin, L. "Type VI Error: Inconsistency between the Statistical Procedure and the Research Question." Paper presented to the World Population Society, Western Regional Meeting, Long Beach, California; Jan. 1976.

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

Title
Author and affiliation
Single-space indented abstract (the rest of the manuscript should be double-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, double-spaced typed copy. A cost of $1 per page should be sent with the submitted paper. Reprints are available to the authors from the editor. Reprints should be ordered at the time the paper is submitted and 20 reprints will cost $.50 per page of manuscript. Prices may be adjusted as necessary in the future.

"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.

THE UNIVERSITY OF AKRON
AKRON, OHIO 44325

TABLE OF CONTENTS