*Multiple Linear Regression Viewpoints*
Volume 2, Number 2
October, 1971

A publication of the Special Interest Group on Multiple Linear Regression

of the American Educational Research Association



Table of Contents

Empirical Exercises for the Study of Multiple Regression - J. T. Bolding.....21

A Note on Multiple Comparisons - William Connett.....23

Setwise Regression Analysis-A New Data-Analytic Tool - John D. Williams and Al C. Lindem.....25

Updated 1971 Membership List - Bill Connett.....28

# EMPIRICAL EXERCISES FOR THE STUDY OF MULTIPLE REGRESSION

J. T. Bolding

Theory takes on real value when it can be demonstrated to the student with known results. The following is a descriptive list of several exercises which demonstrate theory and raise questions.

Exercise 1: Use a table of random numbers or a random number generator to assign ten predictor scores to each individual in a generated sample, say N = 100. Let the ten scores be designated by $X_1, \ldots, X_{10}$. Use as variable Y the sum of the ten random numbers. $Y = X_1 + \ldots + X_{10}$. Construct a regression model using the ten predictor scores for each individual to predict their sum Y. The raw weight coefficients should all be equal to 1.00 and $R^2$ should be 1.00. Deviations from the expected results are due to the procedure (or computer program) used to obtain them.

Exercise 2: Generate samples as in exercise 1, but construct models using a restricted set of the ten predictors. If all ten times N of the predictor scores are randomly selected from the same population, then the ten predictor variables are pair wise independent and should account for 10 percent of the variance of the criterion. Using only four predictors in a model where the criterion is the sum of all ten, one would expect all raw weight coefficients to be equal to 1.00 but $R^2$ should only be 0.40. Using eight predictors $R^2$ should be 0.80. Deviations from the expected results are a result of sampling error when a good procedure is used. To improve the results one may increase the sample size.

Exercise 3: Observe the effects of including in a model a predictor which was not used to construct the criterion. Such a variable should not contribute to $R^2$ and should have a regression coefficient of zero. This is indeed the case when $R^2$ is 1.00 before including the extra predictor. As an exercise, generate eleven predictor scores for each individual in the sample. Let the criterion score be the sum of the first ten, but use all eleven as predictors.

Exercise 4: As in exercise 3 generate eleven predictor scores and let the criterion score be the sum of the first ten. Construct the model which has ten predictors, one of which did not contribute to the criterion and nine which did contribute. The expected value of $R^2$ is 0.90.

Exercise 5: Generate samples as in exercise 1, but include as predictors only five of the contributing scores. Also include power terms and interaction terms for those five.

Exercise 6: First generate five random variables, $X_1$, ..., $X_5$, for each individual in the sample. Then determine generated variables according to the rules:

$$W = X_1 * X_2$$

$$U = X_3 * X_3$$

$$V = X_4 + X_5 \quad .$$

Construct the criterion $Y = W + U + V + 7$. Finally use regression techniques to determine the contribution of each $W_i$ to Y.

# A NOTE ON MULTIPLE COMPARISONS

William Connett
University of Northern Colorado

Multiple regression is often used to compare a multiplicity of models based on a single set of data. Checks for interaction, main effects, simple effects, curvilinearity, etc. are made, while the effect of these comparisons on the overall alpha level is ignored. It is safe to say that regression analysis as commonly used lends itself very well to the commission of a sort of Type I error due to the modification of the overall alpha level through multiple comparisons on the same set of data. This need not be the case if the researcher is aware of the problem and controls the overall alpha level through some adjustment technique.

Two general methods are available for controlling the overall alpha level. One method is to adjust the F value required for significance. This method in relation to multiple regression was recently discussed in Viewpoints (Williams, 1970). The second way to maintain an overall alpha level is by proper choice of the alpha levels for the individual comparisons. The purpose of this note is to review a method for determining the individual alpha levels necessary to maintain some selected overall alpha level.

Kimball (1951) has shown that the overall alpha $\alpha_o$ of any number of comparisons within the same set of data can be determined by:

$$1. \quad \alpha_o < 1-(1-\alpha_1)(1-\alpha_2) \ . \ . \ . \ (1-\alpha_N)$$

If the same alpha level is chosen for each comparison the equation may be simplified to:

$$2. \quad \alpha_o < 1-(1-\alpha_S)^N$$

It is then possible to solve equation 2 for $\alpha_S$ which is the alpha necessary for the specific contrasts if some overall alpha $\alpha_o$ is to be maintained.

$$3. \quad \alpha_S > 1-\sqrt[N]{1-\alpha_o}$$

From formula 3 it follows, for example, if an overall alpha of .05 is required, and five comparisons are to be made, the alpha level for each comparison would be set at .0032.

Table 1 indicates some specific alpha levels necessary to maintain overall alpha levels for a selected number of comparisons.

TABLE 1

Specific alpha levels required to maintain overall alpha levels for selected numbers of comparisons

| | | Number of Comparisons | | | | |
|---|---|---|---|---|---|---|
| | | 2 | 4 | 6 | 8 | 10 |
| Overall $\alpha$ Level | .10 | .0513 | .0131 | .0033 | .0008 | .0002 |
| | .05 | .0253 | .0064 | .0016 | .0004 | .0001 |
| | .02 | .0101 | .0025 | .0006 | .00016 | .00004 |
| | .01 | .0050 | .0013 | .0003 | .00008 | .00002 |

The inequalities indicate that this method will result in a conservative overall alpha level. It is indeed a "quick and dirty" method which should prove of some value to the applied researcher. While the first method of controlling overall alpha level is perhaps preferred, certainly, the method presented here is in most cases preferable to simply ignoring the multiple comparison problem.

Bibliography

Kimball, A.W. On dependent tests of significance in the analysis of variance. Annals of Mathematical Statistics, 1951, 22, 600-602.

Williams, J.D. Multiple comparisons in a regression framework. Multiple Linear Regression Viewpoints, 1970, 1 (2), 26-39.

# SETWISE REGRESSION ANALYSIS-A NEW DATA-ANALYTIC TOOL

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

Most researchers have a familiarity with a common technique known as stepwise regression analysis. The stepwise technique has proven to be a useful data analytic method which allows a computing procedure for selecting variables to be dropped from a predictive system (in the backward stepwise regression procedure), or to be added to a system (in the forward stepwise procedure); in either case, the technique proceeds one variable at a time. Also, both quantitative and binary coded (i.e., a 1 if a trait is present, 0 if the trait is absent) variables can be included in the usual stepwise regression analysis.

It is precisely with the binary coded variables that the usual stepwise regression analysis begins to become less useful. If there are more than two categories involved in the binary coding, then the usual stepwise regression procedure becomes almost meaningless.

Consider the following situation. A researcher is interested in finding the relationships of several variables to alienation in undergraduate students. The variables selected as predictor variables are the following:

1. socioeconomic status of parents
2. religion: 1 if Catholic, 0 otherwise;
3. religion: 1 if Protestant, 0 otherwise;
4. religion: 1 if Jewish, 0 otherwise;
5. sex: 1 if male, 0 otherwise;
6. class: 1 if freshman, 0 otherwise;
7. class: 1 if sophomore, 0 otherwise;
8. class: 1 if junior, 0 otherwise;
9. grade point average overall;
10. grade point average in major field;

11. a measure of interpersonal aggression;

12. a measure of anomie;

13. a measure of need for people; and

14. a measure of need to help people.

Two variables that seem to be missing are a religion variable and a class variable. Actually, the remaining categories are simply zero-coded. Thus the possibility of "other" or "none" for religion is zero-coded, as is being a senior. It is not important that these two particular variables were zero-coded. For example, the sophomore group could be zero-coded, with the other three classes being binary coded. The results would be unchanged.

With the previous 14 variables, the researcher might have defined seven sets as follows:

| Set | Variables |
|---|---|
| 1 | 1 socioeconomic status |
| 2 | 2,3,4 religion |
| 3 | 5 sex |
| 4 | 6,7,8 class |
| 5 | 9,10 grade point average |
| 6 | 11,12 interpersonal aggression and anomie |
| 7 | 13,14 need for people and need to help people |

The setwise procedure drops one <u>set</u> at a time in a stepwise fashion. There will be as many steps as there are sets; for the illustrative example, seven steps would be present. Statistically, the steps are accomplished by an iterative procedure that allows the $R^2$ (multiple correlation coefficient squared) term to be maximized at each stage in a backward stepwise procedure. Once a set is discarded, the set is no longer considered at later stages.

As an example of the procedure, the computer internally computes the $R^2$ value for the seven sets. Then the $R^2$ values are found for all possible combinations of 6 sets of variables. Only the highest $R^2$ value is retained. This process automatically eliminates one set. The procedure is then continued but with only the remaining six sets until only one set remains.

## Difficulties Involved with Setwise Regression

While the difficulty regarding the use of binary coded predictors has been at least partially solved, other difficulties in regard to the stepwise procedure are also involved in the setwise procedure; additionally, the setwise procedure has a new problem unique to itself.

It has been pointed out several times that probability levels in the stepwise procedure are usually violated. Further, when k of the N variables have been dropped, the N - k remaining variables are not necessarily the set of N - k variables that would yield the highest $R^2$ value. These criticisms would also be valid in regard to the setwise procedure. Additionally, the differences in the number of variables in a set will have some effect upon when that set of variables would be dropped. Other things being equal, a set with 6 variables will be retained longer than a set with 3 variables. Notwithstanding these difficulties, if the setwise procedure is judiciously employed by researchers, then addition data analysis power can be obtained.

The program and sample printout are available on request.