

The University of North Carolina  
at Greensboro

JACKSON LIBRARY



CA

no. 1583

UNIVERSITY ARCHIVES

## ABSTRACT

DOUGLAS LANE MILLS. A Comparison of Multiple Regression Computer Programs and Their Usefulness in Analysis of Variance. (1977)

Directed by: Dr. William A. Powers. Pp. 79.

This research will examine the multiple regression programs from major statistical packages and document for each the statistical procedure used, the cost, the difficulty of use, and the output. Each program will be utilized for analyses of variance in balanced and unbalanced designs with an interest in determining the accuracy of and the statistical techniques employed by each package.

The regression procedures, while using dummy variables, are superior to the ANOVA subroutines for the solution to ANOVA. All four packages studied have regression programs that can be used in an unbalanced ANOVA. The REGR procedure in SAS is best for the standard form of regression, while SPSS and BMD are best for stepwise regression.

All packages have the limitation of not giving all types of correct analyses. Therefore, Appendix 4 gives an ANOVA program written for the Programming Language One Optimizing Compiler for a two factor design that will give five correct analyses.

A COMPARISON OF MULTIPLE REGRESSION  
COMPUTER PROGRAMS AND THEIR  
USEFULNESS IN ANALYSIS  
OF VARIANCE

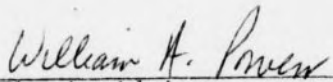
by

Douglas Lane Mills

A Thesis Submitted to  
the Faculty of the Graduate School at  
The University of North Carolina at Greensboro  
in Partial Fulfillment  
of the Requirements for the Degree  
Master of Arts

Greensboro  
1977

Approved by

  
Thesis Adviser

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Thesis Adviser William A. Powers

Committee Members Andrew E. Long, Jr.  
Karl Ray Gentry

May 10, 1977  
Date of Acceptance by Committee

## ACKNOWLEDGEMENT

The author expresses his appreciation to Dr. William A. Powers for his patience and encouragement in the preparation of this thesis.

TABLE OF CONTENTS

	Page
APPROVAL PAGE . . . . .	ii
ACKNOWLEDGEMENT . . . . .	iii
LIST OF TABLES . . . . .	vi
LIST OF FIGURES . . . . .	vii
CHAPTER 1	
Introduction . . . . .	1
Purpose . . . . .	2
CHAPTER 2	
Multiple Regression . . . . .	6
Stepwise Regression . . . . .	13
Choosing a Package . . . . .	15
Procedure for Evaluating the Packages . . . . .	16
CHAPTER 3	
Analysis of Variance . . . . .	25
Computer ANOVA Procedures . . . . .	25
Selecting an ANOVA Program . . . . .	26
CHAPTER 4	
Regression as ANOVA . . . . .	31
Generating Dummy Variables . . . . .	31
Using Dummy Variables . . . . .	38
Five Correct Analyses . . . . .	41
CHAPTER 5	
Conclusions on the Regression Programs . . . . .	45
Using the ANOVA Program . . . . .	45
BIBLIOGRAPHY . . . . .	49
APPENDIX 1	
Report and Proposal of the Committee on Evaluation of Program Packages . . . . .	51

	Page
APPENDIX 2	
Data Used in Testing the Four Packages . . . . .	62
APPENDIX 3	
Survey of Statistical Package Users . . . . .	65
APPENDIX 4	
Program Listing for the Analysis of Variance . .	67
APPENDIX 5	
Output of Computer Program for the Analysis of Variance . . . . .	74

LIST OF TABLES

TABLE		Page
1	Regression Output Available . . . . .	18
2	Stepwise Regression: Output for Final Equation . . . . .	19
3	Stepwise Regression: Output of Inter- mediate Steps . . . . .	20
4	Program Efficiency on Test Data Set . . . . .	23
5	Analysis of Variance Options Available . . . . .	27
6	Analysis of Variance Output Available . . . . .	28
7	Program Cost and Efficiency on Test Data ANOVA Procedures . . . . .	30
8	Values of Dummy Variables Using Zero- One Coding--Balanced Design . . . . .	34
9	Dummy Variables for Model After Reduction Zero-One-Negative One Coding--Balanced Design . . . . .	37
10	Results of Two By Three Data . . . . .	39
11	ANOVA Results of Two By Three Design . . . . .	40
12	Hypothesis for Each of the Five Correct Analyses . . . . .	42
13	Mean Rankings of Regression Procedures . . . . .	66



LIST OF FIGURES

FIGURE		Page
1	Diagram of General Linear Model . . . . .	12
2	Two By Three Analysis of Variance . . . . .	32
3	Arrangement of Data for Two By Three ANOVA . . . . .	38
4	Example of Entire Deck Set Up . . . . .	48

## CHAPTER 1

Introduction

In recent years the development of statistical computer package programs has revolutionized the analysis of data in the social and natural sciences as well as in mathematics. With these packages it is no longer necessary to write individual computer programs for each researcher. However, this methodology raises the fundamental problem: How does one go about choosing a statistical program?

With the wide range of statistical programs available to researchers, choosing the appropriate statistical program package for a given study has become an art. Frequently, the experienced programmer or statistician is unaware of the computational algorithms employed by each package. Other items of interest to the data analyst would be the capabilities (or liabilities), the cost, the difficulty of use, and the statistics that are available in the output generated by each package.

Recognizing the importance of this issue, the American Statistical Association formed the Committee on Evaluation of Programs Packages with the assignment of establishing criteria for selecting a computer package. In August 1974, the committee submitted a proposal (Francis, Heiberger, &

Velleman, 1974) on the standards that statisticians list as criteria to be used when selecting a computer package. To date such standards have yet to be systematically applied (see Appendix 1 for a summary of these standards).

Early studies concentrated on the efficiency of the computer itself and the effect the hardware had on program accuracy (Longley, 1967). As the interest changed to software, abstracts of the available packages became available to the user (Schucany, Minton, & Shannon, 1972). One of these brief surveys did follow up with analyses of the packaged programs (Slysz, 1974). Slysz concentrated on the cost, available procedures, and options within these subprograms. However, the first detailed analysis of particular program packages subroutines came when Ivor Francis (1973) studied the analysis of variance (ANOVA) programs.

This research will specifically examine packages with multiple regression capabilities with respect to their implementation in general multiple regression problems and for analysis of variance. Throughout it will be assumed that the reader has a basic knowledge of statistics and familiarity with ANOVA and multiple regression. Therefore, the reader will not receive a detailed discussion in these areas.

### Purpose

Multiple regression analysis is a general term which refers to a statistical method of investigating the mutual

and individual contributions of one or more independent variables to the variability of a dependent variable. Multiple regression is commonly considered a technique by which the linear dependence of a variable to one or more variables can be detected. This examination allows the researcher to infer relationships that may exist among the variables analyzed. In application, regression analysis is a procedure by which one can derive the least squares fit, and appraise the contribution of each variable or group of variables to this fit.

This study of regression analysis will examine the various algorithms by which a multiple regression analysis is commonly performed. Since multiple regression analysis is now done universally by statistical packaged programs, an examination will be made of how each of the selected packages approaches multiple regression.

Stepwise regression is a process in which the independent variables are taken into the regression that best improves the predictor equation. It should be noted that any subset of those independent variables specified can be chosen. The regular form of regression is derived if all the independent variables are forced into the analysis in a prescribed order. Both types of regression procedures will be examined to determine how well they detect singular and near singular matrices.

The relationship between regression and analysis of variance will be discussed. In regression the best linear

predictor equation is of primary interest, while in analysis of variance the concern is the proportion of variability in the dependent variable which can be attributed to each of the independent variables.

An ANOVA may be performed within the multiple regression framework, and, therefore, may be solved by manipulating regression procedures. This approach is noteworthy when dealing with an ANOVA for an unbalanced design. The unbalanced design, in ANOVA, arises when there is an unequal number of observations in each cell and is of particular interest because of the frequency of occurrence in applied research. The mathematician is concerned because of the various methods of computation, each theoretically sound, which can be implemented to give different numerical results and, therefore, different interpretations. The social or natural scientist is interested because the disciplines often yield data which fit these situations. Moreover, there is some difficulty in using traditional formulas since they do not give correct answers when there is an unequal number of observations in each cell.

The ANOVA procedures and the ANOVA options of regression programs will be examined to document their capabilities and limitations. Additionally, the program packages will be examined with respect to balanced and unbalanced designs to determine the accuracy of the solutions generated by the statistical techniques that the package implements.

The major criteria for choosing the statistical packages were their availability and popularity. With these standards in mind, the programs chosen were: (1) Biomedical Computer Programs (BMD) developed at the University of California at Los Angeles; (2) Statistical Analysis System (SAS) at North Carolina State University; (3) Statistical Package for the Social Sciences (SPSS) now located at the University of Chicago; and (4) Tele-Storage and Retrieval System (TSAR) developed at Duke University.

Finally, an original computer program will be presented that will give a generalized solution to both balanced and unbalanced designs in an ANOVA.

## CHAPTER 2

Multiple Regression

Multiple regression is a method of deriving a linear combination of independent variables  $X_1, X_2, \dots, X_n$  which give a prediction equation for some dependent variable  $Y$ . This technique finds the linear combination which minimizes the squared prediction error (residual).

A general form of the linear combination of the  $X$ 's may be denoted by

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n. \quad (1)$$

The prediction error,  $e$ , for a single observation is defined as

$$e = Y - \hat{Y}.$$

The coefficients in the linear combination are determined so that they will yield the "best" equation. For this research, the "best" equation is taken to be the "least-squares" equation which minimizes the total squared error in prediction over all subjects. This process can be demonstrated with one independent variable so that

$$\hat{Y} = b_0 + b_1 X.$$

Using subscript  $i=1, \dots, m$  to denote the  $m$  observations, the regression equation for the  $i$ th subject would be

$$Y_i^m = b_0 + b_1 X_i$$

and the error for the  $i$ th subject would be denoted by

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i.$$

Then the total square error is

$$\sum_{i=1}^m e_i^2 = \sum_{i=1}^m (Y_i - b_0 - b_1 X_i)^2.$$

Moreover,

$$\sum_{i=1}^m e_i^2 = \sum_{i=1}^m (Y_i - b_0 - b_1 X_i)^2 =$$

$$\sum_{i=1}^m (Y_i^2 - 2b_0 Y_i - 2b_1 Y_i X_i + 2b_0 b_1 X_i + b_0^2 + b_1^2 X_i^2).$$

The coefficients  $\hat{b}_0$  and  $\hat{b}_1$  which minimize the total squared error are the solutions to the two equations

$$\frac{\partial \sum e_i^2}{\partial b_0} = 0 \quad \text{and} \quad \frac{\partial \sum e_i^2}{\partial b_1} = 0.$$



$$\frac{\partial}{\partial b_0} \left( \sum_{i=1}^m e_i^2 \right) = -2 \sum_{i=1}^m Y_i + 2b_1 \sum_{i=1}^m X_i + 2mb_0 = 0 \text{ implies}$$

$$b_0 = \left( \sum_{i=1}^m Y_i - b_1 \sum_{i=1}^m X_i \right) / m = \bar{Y} - b_1 \bar{X}. \quad (2)$$

$$\frac{\partial}{\partial b_1} \left( \sum_{i=1}^m e_i^2 \right) = -2 \sum_{i=1}^m Y_i X_i + 2b_0 \sum_{i=1}^m X_i + 2b_1 \sum_{i=1}^m X_i^2 = 0 \text{ implies}$$

$$b_0 \sum_{i=1}^m X_i + b_1 \sum_{i=1}^m X_i^2 = \sum_{i=1}^m Y_i X_i. \quad (3)$$

By substitution of  $b_0$  from (2) into (3),

$$\bar{Y} \sum_{i=1}^m X_i - b_1 \bar{X} \sum_{i=1}^m X_i + b_1 \sum_{i=1}^m X_i^2 = \sum_{i=1}^m Y_i X_i.$$

$$b_1 \left( \sum_{i=1}^m X_i^2 - m\bar{X}^2 \right) = \sum_{i=1}^m Y_i X_i - m\bar{X}\bar{Y} \text{ implies}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^m X_i Y_i - m\bar{X}\bar{Y}}{m\sigma_X^2} = \frac{\text{cov}(X, Y)}{\sigma_X^2} \quad (4)$$

Now substituting this result into equation (2),

$$\hat{b}_0 = \bar{Y} - \bar{X} \left( \frac{\sum_{i=1}^m X_i Y_i - m\bar{X}\bar{Y}}{m\sigma_X^2} \right) = \frac{m\sigma_X^2 \bar{Y} + m\bar{X}^2 \bar{Y} - \left( \sum_{i=1}^m X_i Y_i \right) \bar{X}}{m\sigma_X^2} =$$

$$\frac{\bar{Y} \sum_{i=1}^m X^2 - \bar{X} \sum_{i=1}^m XY}{m\sigma_X^2}. \quad (5)$$

Thus, the error is at a minimum and the coefficients are "best" when

$$\hat{b}_0 = \frac{\bar{Y} \sum_{i=1}^m X^2 - \bar{X} \sum_{i=1}^m X_i Y_i}{m\sigma_X^2} \quad \text{and} \quad \hat{b}_1 = \frac{\sum_{i=1}^m X_i Y_i - m\bar{X}\bar{Y}}{m\sigma_X^2} .$$

The problem of finding  $b_0$  and  $b_1$  can also be approached from the matrix algebra viewpoint.

$$\text{Let } Y = \begin{bmatrix} Y_1 \\ \cdot \\ \cdot \\ Y_m \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_m \end{bmatrix}, \quad \beta = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}_{2 \times 1}, \quad \text{and } E = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_m \end{bmatrix}_{m \times 1} .$$

Then  $Y = X\beta + E$ , which corresponds to (1) and implies  $E = Y - X\beta$ .

So,  $E'E = (Y - X\beta)'(Y - X\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta$ .

Now, minimize  $E'E$  with respect to  $\beta$ .

$$\frac{\partial}{\partial \beta} E'E = \begin{bmatrix} \frac{\partial E'E}{\partial b_0} \\ \frac{\partial E'E}{\partial b_1} \end{bmatrix} = -2X'Y + 2X'X\beta \quad \text{which corresponds to (2) and (3)} .$$

Setting  $-2X'Y + 2X'X\beta = 0$  gives  $X'X\beta = X'Y$  or  $\hat{\beta} = (X'X)^{-1}X'Y$ .

$$\text{Thus, } \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X'X)^{-1}X'Y = \frac{1}{m\sigma^2 X^2} \begin{bmatrix} \frac{\Sigma X^2}{m} & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix} \begin{bmatrix} \Sigma Y \\ \Sigma XY \end{bmatrix} =$$

$$\frac{1}{m\sigma^2 X^2} \begin{bmatrix} \bar{Y}\Sigma X^2 - \bar{X}\Sigma XY \\ -\bar{X}\Sigma Y + \Sigma XY \end{bmatrix} =$$

$$\begin{bmatrix} \frac{\bar{Y}\Sigma X^2 - \bar{X}\Sigma XY}{m\sigma^2 X^2} \\ \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{m}}{m\sigma^2 X^2} \end{bmatrix}$$

which gives values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  equivalent to those in equations (4) and (5).

Note that, in general, with  $n$  independent variables the solution is also  $\hat{\beta} = (X'X)^{-1}X'Y$ . Since  $X'X$  is positive definite, the solution  $\hat{\beta} = (X'X)^{-1}X'Y$  minimizes the total squared error (Graybill, 1969).

Equivalently, by the geometric approach to least squares we wish to minimize

$$Q = \sum (Y_i - (b_0 + b_1 X_i))^2$$

with respect to  $b_0$  and  $b_1$  with  $E(e) = 0$  and  $E(ee') = \sigma^2 I$ .

Therefore, we show that

$$Q = \|Y - X\beta\|^2 \text{ where } X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{bmatrix} \text{ and } \beta = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} .$$

Since  $Q$  is the square distance from  $Y$  to  $X\beta$ , it is clear from Figure 1 that this distance is minimum when  $X\hat{\beta} = X\hat{\beta}$  and  $X\hat{\beta}$  is the perpendicular projection of  $Y$  on  $R_X$ , the space spanned by  $X\beta$ . That is,  $\|y - X\hat{\beta}\|^2 = \|y - Py\|^2$  where  $Py$  is the perpendicular projection of  $y$  on  $R_X$ .

Proof: Since  $y - Py \perp R_X$  and  $Py - X\beta \in R_X$ , then  $y - Py \perp Py - X\beta$ .

$$\begin{aligned} \|y - X\hat{\beta}\|^2 &= \|y - Py + Py - X\hat{\beta}\|^2 = \|(y - Py) + (Py - X\hat{\beta})\|^2 = \\ &= (y - Py)'(y - Py) + 2(y - Py)'(Py - X\hat{\beta}) + (Py - X\hat{\beta})'(Py - X\hat{\beta}). \end{aligned}$$

But  $(y - Py)'(Py - X\hat{\beta}) = 0$ , therefore  $\|y - X\hat{\beta}\|^2 = \|y - Py\|^2 + \|Py - X\hat{\beta}\|^2$ .

With  $Py \in R_X$ ,  $Py = X\hat{\beta}$  for some  $\hat{\beta}$ , then

$$\|y - X\hat{\beta}\|^2 \geq \|y - Py\|^2 \text{ for any } \hat{\beta}.$$

To find  $\hat{\beta}$ ,  $y = Py + (I - P)y = X\hat{\beta} + (I - P)y$ , so  $X'y = X'X\hat{\beta} + X'(I - P)y$ .

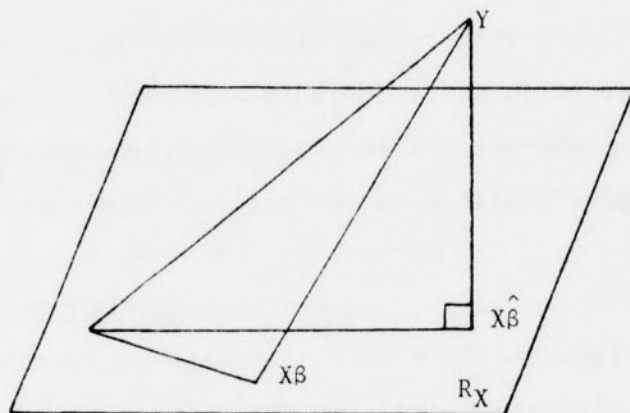
Since  $P = X(X'X)^{-1}X'$ , we have

$$X'(I - X(X'X)^{-1}X') = X' - X'X(X'X)^{-1}X' = X' - X' = 0$$

By substitution, we have  $X'y = X'X\hat{\beta}$  and thus  $\hat{\beta} = (X'X)^{-1}X'y$ .

Now that we have shown that if  $X\hat{\beta} = Py$  then  $X'y = X'X\hat{\beta}$ , we need to show that if  $X'y = X'X\hat{\beta}$ , then  $X\hat{\beta} = Py$  (i.e., show that  $y - X\hat{\beta} \perp R_X$ ).

FIGURE 1  
DIAGRAM OF GENERAL LINEAR MODEL



$$X'(y - X\hat{\beta}) = X'y - X'X\hat{\beta} = 0.$$

So,  $y - X\hat{\beta} \perp R_X$  and, therefore,  $X\hat{\beta} = Py$ .

We have shown that there is a minimum and that we can find it algebraically for all the X's. Consider the problem the best equation where not all the X's are used; this technique is called stepwise regression and will be discussed as it relates to the packaged programs.

### Stepwise Regression

Stepwise regression is a popular modification of multiple regression analysis which not only generates a least squares equation, but does this by allowing the user to choose the optimum independent variables while minimizing the error in prediction. The computations of stepwise regression are done by choosing at each step from the independent variables not in the equation the independent variable that has the largest partial correlation with the dependent variable. Partial correlations are correlations between the independent variables and dependent variable that are adjusted for those independent variables already entered into the equation. This iterative process ends when none of the remaining variables improves the prediction equation by a predetermined amount or all variables are entered into the equation.

An immediate problem is determining the point at which the variables should no longer be added to the equation.

An easy but non-statistical method is to specify a priori that only a certain number of variables will be included regardless of the contribution they make. The "best few" variables method limits the number of independent variables that can be entered into an equation. Although this procedure can easily be done, it is statistically weak. It is conceivable that some significantly influencing variable would be missing from the supposedly optimal equation.

A method of controlling admission of independent variables for which there is a statistical basis involves the use of the  $F$  statistic. The  $F$  value of the chosen variables are checked at each step for its significant contribution in the prediction of the independent variable. At each step, the variable that makes a significant contribution to those already chosen is selected for use in the equation. Further, at each step there is a test for those variables already chosen to insure that they have not become superfluous. If a variable has become unnecessary, it will be deleted from the selected list. For information on another less widely used technique for controlling the inclusion and deletion the reader is referred to Nie, 1975.

Regardless of the technique used for controlling the admission of the independent variables, both the information concerning each intermediate step and the results at the final step are printed.

### Choosing a Package

The first problem when beginning a computer analysis is the general one of choosing a package. For a good analysis, certain results are required to provide accurate representations and interpretations of the multiple regression analysis.

A good regression program should be able to give simple statistics (means, standard deviations, etc.) for each classification variable and correlation coefficients among the variables involved. Further essentials are an ANOVA table that has the sum of squares, degrees of freedom and mean square, along with an  $F$  test so that the statisticians can have a probability level for a test of possible significance. Finally, the package should have a multiple correlation coefficient  $R$  or  $R$  Square. Multiple  $R$  is the correlation between  $\underline{Y}$  and  $\hat{Y}$ . As an alternative, the package could have  $R$  Squared, that shows the quality of the least squares approximation by giving the amount of variance in the dependent variable that is accounted for by the independent variable.

The regression analysis must also include the regular coefficient of the regression equation as well as the standardized  $B$  coefficient. The standardized regression coefficients are important because they are uncontaminated by the magnitude of the corresponding independent variable; they weight each variable in common units (z-scores or standardized units). The programs must also give the standard



error of the coefficients and the standard error of the estimates. Also, there should be a test of hypothesis to check if the  $B$  values are significant. For the independent variables involved, the program should have partial and sequential sums of squares along with a test of significance for each variable. Partial sums of squares are sums of squares that each variable uniquely contributes to the total sums of squares. Sequential sums of squares are sums of squares that each variable contributes at the time that it is entered into the equation.

Some procedures which are only occasionally needed ought to be considered optional. Among them are sum of squares and cross products along with the covariance matrix. As for the residuals, there could be a method of retrieving the estimates of the  $\underline{Y}$  values and a method of plotting residuals.

For stepwise regression, the same rules hold for the final equation. However, at each iteration there should be an evaluation of the independent variable and the information it adds at that particular step.

#### Procedure for Evaluating the Packages

Each of the packages evaluated was run with specially prepared sets of test data (Appendix 2). These sets include data from actual studies as well as data to test the package's ability to detect and handle singular and near singular matrices. These test sets of data included independent

variables in which all but one of the observations were identical. This one observation differed by  $10^{-2}$  (the two decimal place test) and  $10^{-4}$  (the four decimal place test) in the two test situations. Also, tests were done for data that contained independent variables that were exact linear combinations of each other.

When making test runs, as many options and statistics were requested as possible. All options could not be used since some of these are mutually exclusive. The object of these tests was to determine the completeness of the output, the accuracy of the computations, ease of use of these packages, and the cost.

The four packages used in this study (BMD, SAS, SPSS, and TSAR) were selected on the basis of their availability and popularity at Triangle Universities Computation Center (TUCC). A careful examination of the regression capabilities was conducted to more objectively appraise each packaged program. Tables 1 through 3 indicate the output (some of which must be requested) that is available from the regression programs. These tables include the acquirable output of both regression (Table 1) and stepwise regression (Tables 2 and 3). Table 2 gives information on the final equation only and Table 3 shows what data is available for the intermediate steps.

When evaluating these four packages, it is noteworthy that only SAS has a separate procedure for both regression

## REGRESSION OUTPUT AVAILABLE

	BMD	SAS	SPSS	TSAR
Simple Statistics	*	*	*	*
Regression Coefficients	*	*	*	*
Correlation Matrix	*	*	*	*
Sum of Squares (Regression)	*	*	*	
Sum of Squares (Error)	*	*	*	
Sum of Squares (Total)		*		
Mean Square	*	*	*	
<u>F</u> Value for Regression	*	*	*	*
Covariance Matrix	*	*	*	
Multiple <u>R</u>	*	*	*	*
Adjusted <u>R</u>			*	
<u>R</u> Square	*	*	*	
Standard Error of Estimate	*	*	*	*
Summary Table	*		*	
List of Residuals	*	*	*	
Plot of Residuals			*	
Durbin-Watson <u>D</u>		*	*	
Coefficient of Variability		*		
Significance of <u>F</u> Value		*		
Sequential S.S. for each Ind. V.		*		
Partial S.S. for each Ind. V.		*		
Test for $H_0: \underline{B}=0$		*	*	
Zero Intercept Option	*	*		

## STEPWISE REGRESSION: OUTPUT FOR FINAL EQUATION

	BMD	SAS	SPSS	TSAR
Simple Statistics	*		*	*
Regression Coefficients	*	*	*	*
Correlation Matrix	*		*	*
Multiple <u>R</u>	*		*	*
<u>R</u> Square	*	*	*	
Standard Error of Estimate	*	*	*	*
Sum of Squares (Regression)	*	*	*	
Sum of Squares (Error)	*	*	*	
Sum of Squares (Total)		*		
Mean Square	*	*	*	
<u>F</u> Value for Regression	*	*	*	*
Covariance Matrix	*			
Summary Table	*	*	*	
List of Residuals	*		*	
Plot of Residuals			*	
Durbin-Watson <u>D</u>		*	*	
Coefficient of Variability		*		
Significance of <u>F</u> Value		*		
Sequential S.S. for each Ind. V.		*		
Partial S.S. for each Ind. V.		*		
Test for $H_0: \underline{B}=0$		*		
Zero Intercept Option	*	*		

## STEPWISE REGRESSION: OUTPUT OF INTERMEDIATE STEPS

	BMD	SAS	SPSS	TSAR
Sum of Squares (Regression)	*	*	*	
Sum of Squares (Error)	*	*	*	
Sum of Squares (Total)		*		
Mean Square	*	*	*	
F Value for Regression	*	*	*	*
Significance of F		*		
Multiple R ( <u>R</u> Square)	*	*	*	*
Sequential S.S. for each Ind. V.		*		
Partial S.S. for each Ind. V.		*		
<u>Variables in Equation</u>				
Regression Coefficients	*	*	*	*
F Value to Remove	*		*	*
Significance of F Value		*		
Standard B Value (Beta)		*	*	*
Standard Error Beta				*
Standard Error Estimate				*
Normalized <u>B</u>				*
Standard Error <u>B</u>	*	*	*	*
Test $H_0: \underline{B}=0$		*		
<u>Variables Not in Equation</u>				
Partial Correlation	*		*	
Tolerance	*		*	
F Value to Enter	*		*	
Beta In			*	

and stepwise regression. It is possible to use BMD, SPSS, and TSAR in regular regression even though they are written for stepwise regression by forcing the independent variables in a specified order.

When considering the output available from the programs for standard regression, SAS stands out. It is the only package written for this type of regression, and therefore SAS gives the most complete output for any regression analysis (Table 1). While SAS has the largest number of the suggested results previously mentioned, it does have some shortcomings. SAS does not have a method of plotting residuals directly in the regression procedure. Although SAS does have a separate plotting procedure, it is somewhat awkward to use since it is necessary to create a separate data set with only those variables and residuals to be plotted.

The stepwise regression procedure in SAS fares less well than the regression program. The output is less complete since it lacks a summary table as well as other essentials (Table 2). Further, SAS lacks information on the variables not in the equation at each iteration (Table 3).

BMD and SPSS are about equal with respect to options available. While SPSS does not have an option for a zero intercept (no constant in the regression equation), BMD

has this alternative. However, SPSS allows the plotting of residuals directly in the regression procedure, and it is the only package to have this capability. Finally, TSAR generates the least information. It does not give the sum of squares, a covariance matrix, or a list of residuals.

The accuracy involved in the computations is of importance in any research study. All packages provide accurate results for data with singular or near singular matrices. Both SPSS and SAS have the capability to detect singular and near singular matrices even when tested to four decimal places, an extreme in most applications. BMD was unable to detect a near singular matrix in the two decimal place test. However, it did detect the singular matrix. TSAR was able to recognize singular matrices and was superior to BMD in checking near singular matrices as it passed the two decimal place test. TSAR does not have the proficiency of SAS and SPSS with respect to the higher level of accuracy. Therefore, it was concluded on this basis that BMD02R and possibly TSAR are written in single precision while SAS and SPSS are written in double precision.

In addition to the problem of accuracy, an important but significant characteristic is its efficiency. TSAR is the most inexpensive package (Table 4) and has the quickest execution time. It is easy to conclude that while TSAR does not have as many capabilities as the other programs, its computations, however limited, are well done



when considering its accuracy with respect to its cost. BMD is relatively inexpensive, but it is the most difficult to use because of the precise card preparation required. SAS is the most expensive package and has the slowest execution time. Because of its excellent manual and easy deck preparation, SPSS stands out in its ease of use. This feature, along with a moderate cost, makes it an attractive package for stepwise regression. Appendix 3 contains subjective ratings of experienced users with respect to the manuals, ease of use, and completeness of output.

TABLE 4  
PROGRAM EFFICIENCY ON TEST DATA SET  
REGRESSION PROCEDURES  
30 Cases, 6 Independent Variables

	Cost	CPU Time	Priority	Region
BMD (BMD02R)	1.15	.9	0	114K
SAS (Regression)	1.69	1.5	0	108K
SAS (Stepwise)	.96	.9	0	108K
SPSS	.86	.6	0	184K
TSAR	.34	.2	0	100K



For standard regression, SAS has the necessary attributes in giving complete information that make it the best package for this type of analysis. For stepwise regression, SPSS is favored because of good output for each step and its summary table. Further, both packages are accurate (as shown by the test for near singular matrices) and are easy to use. Thus, on both objective and subjective bases, SAS and SPSS are the programs best suited for regression.

## CHAPTER 3

Analysis of Variance

An ANOVA is a statistical method by which it is possible to control separate factors which may affect a measured observation and to observe the measurable quantity at each level and, therefore, to evaluate the effect of the factors. Consequently, through experimental design, the researcher can infer the relationship between the measured observations and the separate factors.

Computer ANOVA Procedures

Each package has unique algorithms for the ANOVA calculations. These algorithms differ with respect to the design complexity (i.e., number of factors, number of levels, nesting) permitted. However, when used properly all packages should arrive at the same results. Many packages are so limited in the type of designs permitted that they are difficult to compare. Attempts to use a package unsuited for the analysis will usually give an appropriate diagnostic error message. Yet, in the recurring situation of unequal cell size, some of the packages produce erroneous results without warning the user in the manual or through messages in the output.

Even those programs which correctly perform an ANOVA for unequal cell sizes must be used with care since at least five different correct analyses are possible. The user must then decide which type of analysis is appropriate and select the package that has the corresponding algorithm; this is difficult, since not all packages identify the algorithm they use.

Most packages also have separate procedures for ANOVA and regression. These ANOVA subroutines do not necessarily give a correct analysis for an unbalanced ANOVA. The problem is that the procedures use formulae and algorithms for sums of squares dependent upon an equal number of observations in each cell. Had the developers of these packages used the regression viewpoint, which is a correct way to approach an unbalanced ANOVA (see Chapter 4), this problem could have been alleviated. Therefore, only data with an equal number of observations in each cell will be considered to evaluate the performance of each ANOVA package.

#### Selecting an ANOVA Program

For this ANOVA phase of research, the same four program packages were selected. The programs or subroutines used were the BMD08V program, the ANOVA procedure in SAS, the SPSS ANOVA, and the TFA0V3 procedure in TSAR.

As in regression, the selection of ANOVA procedure generates the question: What is available? The designs

TABLE 5  
ANALYSIS OF VARIANCE LIMITATIONS

	<u>BMD</u>	<u>SAS</u>	<u>SPSS</u>	<u>TSAR</u>
Maximum Number of Factors	10	unlimited	5	3
Maximum Number of Levels	100	unlimited	unlimited	12 columns, unlimited rows, layers
Nesting (Repeated Measures)	*	*		
Analysis of Covariance	*		*	
Mixed Models	*			
Regression Approach Option			*	

that these four programs can accommodate are given in Table 5. The output available with these ANOVA procedures is shown in Table 6. The data used in testing the procedures is in Appendix 2.

TABLE 6  
ANALYSIS OF VARIANCE OUTPUT AVAILABLE

	BMD	SAS	SPSS	TSAR
Simple Statistics	*	*		*
Sum of Squares (Effects)	*	*	*	*
Sum of Squares (Error)		*	*	*
Sum of Squares (Total)		*	*	
Mean Square	*	*	*	*
<u>F</u> Value		*	*	*
Multiple <u>R</u>			*	
Significance of <u>F</u> Value		*	*	
Coefficient of Variability		*		
Expected Mean Square	*			

The ANOVA procedures of the four packages fare less well than their regression counterparts. When considering both options and output available, the BMD package is best. BMD allows a reasonable number of factors and enough levels within each factor for most analyses. Further, it can

perform analysis of covariance and handle nested designs. BMD is the only package that can directly utilize a mixed model.

All other packages show severe limitations. SAS can handle neither an analysis of covariance nor a mixed model, although it gives the best output. SPSS cannot handle either repeated measures nor mixed models. Although the SPSS manual is generally good, the section on ANOVA is ambiguous and requires a great amount of statistical expertise to choose the correct options for an analysis. The problem has even caused confusion among the writers of SPSS. In their September 1976 newsletter, they admit to the problem and plan to remedy the problem in Version 7.0 of SPSS. Therefore, one should not use the SPSS ANOVA procedure unless he knows the option he needs and that this option is correct. For this reason, SPSS ANOVA was not used in the test on cost and efficiency (Table 7). TSAR has the most limitations. It does not include an analysis of covariance, repeated measures, nor mixed models and is limited to at most a three-factor design.

TABLE 7  
PROGRAM COST AND EFFICIENCY ON TEST DATA  
ANOVA PROCEDURES

18 Cases, 2 Independent Variables

	BMD08V	SAS	TSAR
Cost	.53	1.43	.54
CPU Time (Seconds)	.4	.8	.2
Priority	1	1	1
Region	140K	200K	200K

## CHAPTER 4

Regression as ANOVA

The most general way to think of an ANOVA is as a special form of regression. When using multiple regression for ANOVA, the independent variables will not be continuous variables but discrete dummy variables that indicate an observation's cell location within the design. The theory of dummy variables can best be shown by their generation and use.

Generating Dummy Variables

Consider the case of a two by three ANOVA (i.e., the first variable has two levels and the second variable has three levels). A convenient mathematical model for the observations in ANOVA is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad \text{where}$$

$$i = 1,2; j = 1,2,3; k = 1,2,\dots,n$$

which is denoted as the general linear model. The  $Y_{ijk}$  is the  $k$  th observation in the  $i$  th level (row) of the first effect and the  $j$  th level (column) of the second effect. The arrangement of rows and columns in a balanced design



with  $n$  observations per cell is shown in Figure 2. The  $\mu$  term is attributed to all observations. The  $\alpha$  term is the effect of the first variable and the  $\beta$  term is the effect of the second. The  $\gamma$  term is the effect of the interaction of the first and second variables. The  $\epsilon$  term is the error. It is usually assumed that

$$\sum_i \alpha_i = 0, \sum_j \beta_j = 0, \text{ and } \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \quad (6)$$

(see Searle, 1971, or Winer, 1971, for a more detailed discussion of the general linear model).

FIGURE 2  
TWO BY THREE ANALYSIS OF VARIANCE

	B1	B2	B3
A1	$Y_{111}$ $Y_{112}$ $\vdots$ $Y_{11n}$	$Y_{121}$ $Y_{122}$ $\vdots$ $Y_{12n}$	$Y_{131}$ $Y_{132}$ $\vdots$ $Y_{13n}$
A2	$Y_{211}$ $Y_{212}$ $\vdots$ $Y_{21n}$	$Y_{221}$ $Y_{222}$ $\vdots$ $Y_{22n}$	$Y_{231}$ $Y_{232}$ $\vdots$ $Y_{23n}$

In the two-way analysis with two levels of A and three levels of B, this model can be rewritten as

$$Y_{ijk} = \mu + \alpha_1 w_{11} + \alpha_2 w_{22} + \beta_1 x_{11} + \beta_2 x_{22} + \beta_3 x_{33} + \gamma_{11} w_{11} x_{11} + \gamma_{12} w_{11} x_{12} + \gamma_{13} w_{11} x_{13} + \gamma_{21} w_{22} x_{21} + \gamma_{22} w_{22} x_{22} + \gamma_{23} w_{22} x_{23} + \epsilon_{ijk} \quad (7)$$

where  $w_i$  is one if the observation is in row  $i$ , otherwise it is zero; similarly,  $x_j$  is one if the observation is in column  $j$ , otherwise it is zero. Therefore, by taking the information given by the  $x$ 's along with  $w$ 's the exact location of the observation can be determined. Consider the first observation of the first level of A and the third level of B, then

$$Y_{131} = \mu + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 1 + \gamma_{11} \cdot 0 + \gamma_{12} \cdot 0 + \gamma_{13} \cdot 1 + \gamma_{21} \cdot 0 + \gamma_{22} \cdot 0 + \gamma_{23} \cdot 0 + \epsilon_{131}$$

which becomes

$$Y_{131} = \mu + \alpha_1 + \beta_3 + \gamma_{13} + \epsilon_{131}$$

An entire analysis could be set up with dummy variables as in Table 8. This procedure is exactly the same for the unbalanced design.

TABLE 8  
VALUES OF DUMMY VARIABLES USING ZERO-ONE  
CODING--BALANCED DESIGN

row	column	replication	$w_1$	$w_2$	$x_1$	$x_2$	$x_3$	$w_1 x_1$	$w_1 x_2$	$w_1 x_3$	$w_2 x_1$	$w_2 x_2$	$w_2 x_3$
1	1	1	1	0	1	0	0	1	0	0	0	0	0
1	1	2	1	0	1	0	0	1	0	0	0	0	0
1	1	3	1	0	1	0	0	1	0	0	0	0	0
1	2	1	1	0	0	1	0	0	1	0	0	0	0
1	2	2	1	0	0	1	0	0	1	0	0	0	0
1	2	3	1	0	0	1	0	0	1	0	0	0	0
1	3	1	1	0	0	0	1	0	0	1	0	0	0
1	3	2	1	0	0	0	1	0	0	1	0	0	0
1	3	3	1	0	0	0	1	0	0	1	0	0	0
2	1	1	0	1	1	0	0	0	0	0	1	0	0
2	1	2	0	1	1	0	0	0	0	0	1	0	0
2	1	3	0	1	1	0	0	0	0	0	1	0	0
2	2	1	0	1	0	1	0	0	0	0	0	1	0
2	2	2	0	1	0	1	0	0	0	0	0	1	0
2	2	3	0	1	0	1	0	0	0	0	0	1	0
2	3	1	0	1	0	0	1	0	0	0	0	0	1
2	3	2	0	1	0	0	1	0	0	0	0	0	1
2	3	3	0	1	0	0	1	0	0	0	0	0	1

Observe that the arrangement in Table 8 carries more information than necessary. Certainly, if an observation is in the first level of A, then it cannot be in the second level. The same logic holds for the levels of B.

Recall the assumptions of the model from (6). For the example of the two by three analysis

$$\begin{aligned} \alpha_1 + \alpha_2 &= 0 && \text{implies} && \alpha_2 = -\alpha_1 \\ \beta_1 + \beta_2 + \beta_3 &= 0 && \text{implies} && \beta_3 = -\beta_1 - \beta_2 \\ \gamma_{11} + \gamma_{12} + \gamma_{13} &= 0 && \text{implies} && \gamma_{13} = -\gamma_{11} - \gamma_{12} \\ \gamma_{12} + \gamma_{22} &= 0 && \text{implies} && \gamma_{22} = -\gamma_{12} \\ \gamma_{13} + \gamma_{23} &= 0 && \text{implies} && \gamma_{23} = -\gamma_{13} \end{aligned}$$

Now, substitution of these results into (7) gives

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_1 w_1 - \alpha_2 w_2 + \beta_1 x_1 + \beta_2 x_2 + (-\beta_1 - \beta_2) x_3 + \gamma_{11} w_1 x_1 + \\ &\gamma_{12} w_1 x_2 + (-\gamma_{11} - \gamma_{12}) x_1 w_3 - \gamma_{11} w_2 x_1 - \gamma_{12} w_2 x_2 + (\gamma_{11} + \gamma_{12}) w_2 x_3 \\ &= \mu + \alpha_1 (w_1 - w_2) + \beta_1 (x_1 - x_3) + \beta_2 (x_2 - x_3) + \gamma_{11} (w_1 - w_2) (x_1 - x_3) \\ &\quad + \gamma_{12} (w_1 - w_2) (x_2 - x_3) + \epsilon. \end{aligned}$$

However, these dummy variables can be reduced in number by letting  $u_1 = w_1 - w_2$ ;  $u_2 = x_1 - x_3$ ;  $u_3 = x_2 - x_3$ ;  $u_4 = (w_1 - w_2)(x_1 - x_2)$ ;  $u_5 = (w_1 - w_2)(x_2 - x_3)$ . (8)

The equation then becomes

$$Y_{ijk} = \mu + \alpha_1 u_1 + \beta_1 u_2 + \beta_2 u_3 + \gamma_{11} u_4 + \gamma_{12} u_5 + \epsilon.$$

Therefore, the number of terms in the equation can be reduced by changing from the zero-one coding to zero-one-negative one coding (see Table 9). This reduced format can be obtained by using the five  $u_i$  substitutions into Table 7. In a design with a levels of factor A and b levels of factor B, this process reduces the number of dummy variables from  $2ab-1$  to  $ab-1$ , a decrease of  $ab$ . This reduction in the number of dummy variables is of great importance in the actual computational process. Recalling the idea of regression, one independent variable is equivalent to one dummy variable and the computer builds a matrix of size  $a+b+1 \times a+b+1$  rather than the unreduced size  $2ab \times 2ab$ . This smaller matrix, which is half of the original size in the two by three example, is much more efficient to use with respect to both core allocation and, more importantly, the time required for expensive matrix inversion computations necessary to solve the regression problem. Consequently, these two advantages cause important savings on extensive research projects.

TABLE 9

DUMMY VARIABLES FOR MODEL AFTER REDUCTION  
 ZERO-ONE-NEGATIVE ONE CODING--BALANCED DESIGN

			$u_1$	$u_2$	$u_3$	$u_4$	$u_5$
1	1	1	1	1	0	1	0
1	1	2	1	1	0	1	0
1	1	3	1	1	0	1	0
1	2	1	1	0	1	0	1
1	2	2	1	0	1	0	1
1	2	3	1	0	1	0	1
1	3	1	1	-1	-1	-1	-1
1	3	2	1	-1	-1	-1	-1
1	3	3	1	-1	-1	-1	-1
2	1	1	-1	1	0	-1	0
2	1	2	-1	1	0	-1	0
2	1	3	-1	1	0	-1	0
2	2	1	-1	0	1	0	-1
2	2	2	-1	0	1	0	-1
2	2	3	-1	0	1	0	-1
2	3	1	-1	-1	-1	1	1
2	3	2	-1	-1	-1	1	1
2	3	3	-1	-1	-1	1	1

Using Dummy Variables

FIGURE 3  
ARRANGEMENT OF DATA FOR TWO BY THREE ANOVA

	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
A <sub>1</sub>	8 1 0	10 8 6	8 6 4
A <sub>2</sub>	14 10 6	4 2 0	15 12 9

Consider the data collected for the two by three analysis (see Figure 3). The matrix generated for this data would be similar to Table 9 and thus the problem would become

$$\begin{bmatrix} 8 \\ 1 \\ 0 \\ 10 \\ 8 \\ 6 \\ 8 \\ 6 \\ 4 \\ 14 \\ 10 \\ 6 \\ 4 \\ 2 \\ 0 \\ 15 \\ 12 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{113} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{123} \\ \epsilon_{131} \\ \epsilon_{132} \\ \epsilon_{133} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{213} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{223} \\ \epsilon_{231} \\ \epsilon_{232} \\ \epsilon_{233} \end{bmatrix}$$

This analysis performed by a regression program will give the results shown in Table 10. Note that the sums of squares due to each  $u_i$  gives the variability in Y explained by that  $u_i$ .

TABLE 10  
RESULTS OF TWO BY THREE DATA

Source	DF	Sum of Squares	Mean Square	F Value
Total	5	316.000	63.200	
$u_1$	1	18.000	18.000	2.038
$u_2$	1	12.000	12.000	1.358
$u_3$	1	36.000	36.000	4.075
$u_4$	1	0.000	0.000	0.000
$u_5$	1	144.000	144.000	16.302
Error	12	106.000	8.333	

Therefore, the sum of squares for  $u_1$  is the measure of the variability due to the difference in  $\alpha_1$  and  $\alpha_2$ . Recalling the change of variables from (8) that gave the  $u$  terms, the researcher can infer the relationship among the variables. For example,  $u_1$  was the difference between  $w_1$  and  $w_2$  in Table 8. Further,  $w_1$  was associated with  $\alpha_1$  and  $w_2$  was associated with  $\alpha_2$ . Therefore, the researcher can assess the difference between  $\alpha_1$  and  $\alpha_2$ , which is equivalent to assessing the effect of factor A. Similarly, for the levels



within B, recall the change of variables from (8) that gave the  $u$  terms. Recall  $u_2 = x_1 - x_3$  and  $u_3 = x_2 - x_3$ . Also,  $x_1$  was associated with  $\beta_1$ ,  $x_2$  with  $\beta_2$ , and  $x_3$  with  $\beta_3$ . Therefore, the difference  $u_2$  is the contrast between  $\beta_1$  and  $\beta_3$  while the difference  $u_3$  is the contrast between  $\beta_2$  and  $\beta_3$ . When  $u_2$  and  $u_3$  are taken together, they give the statistician information desired on the effect of the levels within B. Finally,  $u_4$  and  $u_5$  together give the information on the possible interaction of the levels of A with those of B. To change this regression form of Table 9 to the ANOVA form,  $u_1$  is the sum of squares of A,  $u_2 + u_3$  is equal to the sum of squares due to factor B, and  $u_4 + u_5$  is equal to the sum of squares of the interaction of A and B. Hence, Table 10 would be rewritten into an ANOVA form as Table 11. Remember that if dummy variables are used, it makes no difference whether the data is balanced or not.

TABLE 11  
ANOVA RESULTS OF TWO BY THREE DESIGN

Source	DF	Sum of Squares	Mean Square	F Value
Total	5	316.000	63.200	
A	1	18.000	18.000	2.038
B	2	48.000	24.000	2.717
A*B	2	144.000	144.000	16.302
Error	12	106.000	8.833	

### Five Correct Analyses

When a research study calls for a two-way ANOVA, three hypotheses are commonly stated. These hypotheses are (1) no difference in the row effects, (2) no difference in the column effects, and (3) no interaction between the rows and columns.

In an unbalanced design, there are several ways of specifying the meaning of "no difference in row effects." Table 12 (Herr & Powers, 1976) gives the hypothesis for each of the corresponding correct analyses. In a balanced design, all such methods are equivalent; that is, they are reduced to the same parametric statement and they yield the same sums of squares due to row effect. However, in unbalanced designs, the various methods yield different analyses. Each analysis is a correct analysis within the context of the specified meaning of "no difference in row effect." But each analysis will result in a different sum of squares due to row effects and possibly a different interpretation of significant effects.

The five correct analyses for an unbalanced design are: (1) hierarchical--first A then B, (2) hierarchical--first B then A, (3) adjusting each main effect for each other, (4) a weighted means analysis, and (5) the standard parametric analysis (weighted squares of means). Each of these methods will give different sums of squares for one or more of the factors, therefore a different mean square and  $F$  value.

TABLE 12

## HYPOTHESIS FOR EACH OF THE FIVE CORRECT ANALYSES

<u>Subspaces</u>	<u>Identifying Phrase</u>	<u>Abbrevi- ation</u>	<u>Parametric Hypothesis Tested</u>	
			<u>Hypothesis (Rows)</u>	<u>Hypothesis (Columns)</u>
$G_r \quad G_c$				
$A J \quad B J$	Standard parametric	STP	$\mu_{1.} = \dots = \mu_{a.} (\alpha_p = 0)$	$\mu_{.1} = \dots = \mu_{.b} (\beta_q = 0)$
$\hat{A} \hat{B} \quad \hat{B} \hat{A}$	Each adjusted for the other	EAD	$\sum_{q=1}^b n_{kq} (\mu_{kq} - \mu_{*q}) = 0$ $k = 1, \dots, a-1$	$\sum_{p=1}^a n_{pk} (\mu_{pk} - \mu_{p*}) = 0$ $k = 1, \dots, b-1$
$\hat{A} J \quad \hat{B} \hat{A}$	Hierarchical--A first, then B adjusted for A	HAB	$\mu_{1*} = \dots = \mu_{a*}$	as above
$\hat{A} \hat{B} \quad \hat{B} J$	Hierarchical--B first, then A adjusted for B	HBA	$\sum_{q=1}^b n_{kq} (\mu_{kq} - \mu_{*q}) = 0$ $k = 1, \dots, a-1$	$\mu_{*1} = \dots = \mu_{*b}$
$\hat{A} J \quad \hat{B} J$	Weighted means	WTM	$\mu_{1*} = \dots = \mu_{a*}$	$\mu_{*1} = \dots = \mu_{*b}$

In the hierarchical method--first A then B--the computations are first done for the A effect and then for the B effect. Only that information not given by the A sum of squares will be given in the B effect (i.e., B is adjusted for A). With this, some variability is attributed to A (the rows) that is caused by column variability. This concept is similar to that of stepwise regression, when only unique information is entered in the equation by the process of partial correlation. However, that variability attributed to B (the columns) has no row variability. Similarly, in the hierarchical method--first B then A--the computations are done for B and then for A (i.e., A is adjusted for B).

There are two methods that are combinations of the two hierarchical approaches. When each main effect is adjusted for the other, the sums of squares are: A adjusted for B and B adjusted for A. The other "fusion" method is the weighted means analysis, in which neither main effect is adjusted for the other. Finally, the standard parametric analysis, using the weighted squares of means, gives an exact analysis for a balanced design. This approach tests the hypothesis of equal cell means.

There is a sixth method which is an approximation that has been almost outdated by digital computers. The unweighted means analysis (Winer, 1971, pp. 402-404) uses cell means in a balanced design with one observation per cell.

Since it is an approximation, this method's effectiveness is dependent upon the degree to which the data is unbalanced.

Rarely do the manuals or output of computer program packages identify which of the methods is used. All four regression programs use the hierarchical method when used with the user supplying dummy variables and specifying the order of adjustment.

In the ANOVA programs, BMD08V will not perform correctly for unbalanced designs, while BMD10V will perform the hierarchical method. SAS uses the weighted means analysis. SPSS has options that allow the user to choose the analysis he desires. The default is the weighted squares of means while OPTION 10 gives the hierarchical--first A then B--approach. As previously mentioned, the SPSS ANOVA options may give erroneous results for unbalanced data. Finally, TSAR uses the unweighted means analysis.

Unfortunately, the analyst usually is unaware of the approach of the individual procedures. However, an even greater problem is that the researcher has only one analysis for each computer run. It is quite possible that if he had a variety of analyses for his project, he would have a better understanding of the effects involved.

## CHAPTER 5

Conclusions on the Regression Programs

The regression procedures, while using dummy variables, are superior to the ANOVA subroutines for the solution to ANOVA. All four packages studied have regression programs that can be used in an unbalanced ANOVA. The REGR procedure in SAS is best for the standard form of regression, while SPSS and BMD are best for stepwise regression.

However, all packages have the limitation of not giving all types of correct analyses. Therefore, the analyst is confined when other analyses are needed. Appendix 4 gives an ANOVA program, written for the Programming Language One (PL/I) Optimizing Compiler, for a two factor design that will give all five correct analyses discussed in Chapter 4. Additionally, the program gives main effect and interaction means. Appendix 5 gives the sample output of the analyses.

Using the ANOVA Program

Using this ANOVA program requires little control card preparation. Only two cards are required before the observations; these include the Levels card and Input card.

The Levels card gives the number of levels for the row and column effects. This card is free field, that is, no special card columns are used. Only one space is needed



between the number of the rows and the number of columns in the design. For example, suppose a 4 X 3 design is desired; the Levels card would be:

Example of Levels Card	
4	3

This program has no limit to the number of rows and columns.

Next, the Input card gives the card columns in which the observations are punched. At the same time, the user must specify variable names (up to eight characters) for the rows, columns and dependent variables. (A requirement is that all variable names must be set off by apostrophes.) The user specifies his row variable name followed by the card column in which the row cell is punched. Next, the user specifies the variable name for the column and its card column. Finally, the user specifies observation name, the card column in which the observation starts, and the card column in which the observation terminates. After skipping another space, the user gives the number of places to the right of decimal for the observation. (Note: A zero is not assumed; you must punch the number even if the number is zero.)

An example of the Input card is:

Example of Input Card				
'A'	1	'B'	2	'OBS' 3 4 0

This example gives the row classification in card column 1, the column classification in card column 2, and the observation in columns 3 through 4 with zero places to the right of the decimal.

Data cards must include row and column classification codes preceding the actual observation. The decimal point for the observation may be either implicit or explicit in the data.

An example of an entire deck set up for the data suggested by Kutner (1974) (see Appendix 2) could give the control cards in Figure 4. This analysis will generate five pages of the printout found in Appendix 4.

It should be noted that there are limitations for any two-way ANOVA. There must be at least one observation in each cell (there is a check for empty cells). There can be at most 50 observations, unless the user changes the DECLARE statement. All the user needs to do is change the space allocation for variable Y in the DECLARE to

$$Y(II, JJ, n)$$

where  $n$  is the maximum number of observations in each cell.

This program shows that the regression approach can be used effectively for the solution to ANOVA.



## FIGURE 4

## EXAMPLE OF ENTIRE DECK SET UP

```
// JOB  
// EXEC PLOCG  
// } (Additional JCL if required system)  
//C.SYSIN DD *
```

Program Deck (Appendix 3)

```
//G.SYSIN DD *  
4 3  
'A' 1 'B' 3 'TREAT' 5 6 0  
1 1 42  
1 1 44  
1 1 36  
1 1 13  
1 1 19  
1 1 22  
1 2 33  
1 2 26  
1 2 33  
.  
.  
4 3 25  
4 3 5  
4 3 12  
//
```

## BIBLIOGRAPHY

- Adler, H. L., & Roessler, E. B. Introduction to probability and statistics. San Francisco: W. H. Freeman, 1972.
- Dixon, W. J. (Ed.). Biomedical computer programs. Berkeley: University of California Press, 1974.
- Draper, N. R., & Smith, H. Applied regression analysis. New York: Wiley, 1968.
- Francis, I. A comparison of several analysis of variance programs. Journal of the American Statistical Association, 1973, 68, 860-865.
- Francis, I., Heiberger, R. M., & Velleman, P. F. Report and proposal of the Committee on Evaluation of Program Packages to the Section on Statistical Computing. Washington, D.C.: American Statistical Association, August 1974.
- Frane, J. W. Comments on "A comparison of BMD, SAS, and SPSS" by Papaioannou, Styan, and Ward. SAS ONE: Proceedings of First International S.A.S. User's Conference. Raleigh, N. C.: S.A.S. Institute, Inc., 1976, 389.
- Graybill, F. Introduction to matrices with applications in statistics. Belmont, Calif.: Wadsworth, 1969.
- Herr, D. G., & Powers, W. A. Unbalanced two-way ANOVA and SAS. SAS ONE: Proceedings of First International S.A.S. User's Conference. Raleigh, N. C.: S.A.S. Institute, Inc., 1976, 309-317.
- IBM system/360 scientific subroutine package (PL/I). White Plains, N. Y., 1968. (Manual GH20-0586-0)
- Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston, 1973.

- Kutner, M. H. Hypothesis testing in linear models (Eisenhart Model I). The American Statistician, 1974, 28, 98-100.
- Longley, J. W. An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association, 1967, 62, 819-841.
- Nie, N., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical package for the social sciences. New York: McGraw-Hill, 1975.
- Papaioannou, T., Styban, G. P. H., & Ward, L. L. A comparison of BMD, SAS, and SPSS. SAS ONE: Proceedings of First International S.A.S. User's Conference. Raleigh, N. C.: S.A.S. Institute, Inc., 1976, 362-379.
- Schucauy, W. R., Minton, P. D., & Shannon, B. S. A survey of statistical packages. Computing Surveys, 1972, 4, 65-79.
- Searle, S. R. Linear models. New York: Wiley, 1971.
- Service, J. A user's guide to the statistical analysis system. Raleigh, N. C.: Sparks Press, 1972.
- Slysz, W. D. An evaluation of statistical software in the social sciences. Communications of American Computing Machinery, 1974, 17, 326-332.
- Snedecor, G. W., & Cochran, W. G. Statistical methods. Ames: Iowa State University Press, 1967.
- SPSS Newsletter. Chicago: SPSS, Inc. September 1976.
- Tele-storage and retrieval system: User's Manual. Durham, N. C.: Duke University Computation Center, 1974.
- Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

## APPENDIX 1

Report and Proposal  
of the  
Committee on Evaluation of Program Packages  
August 1974

by

Ivor Francis\* and Richard M. Heiberger,\*\* Co-Chairmen  
and Paul F. Velleman\*\*\*

Section 4

Criteria and Considerations

A program package consists of a front end plus a set of procedures. The front end is that part of the package used to prepare the data, and select a particular procedure to which it delegates specific tasks. The front end is what makes a package a package, as distinct from a set of subroutines. A stand-alone program is a degenerate package with only one procedure.

Package features fall into three broad categories: user interface, statistical effectiveness, and implementation. The first includes features of immediate interest to the user, such as form of control cards and manual comprehensibility. The second includes a package's statistical properties and capabilities, and the third set includes the computer-science features of a package.

Package designers have to balance many conflicting demands. In choosing a package, individual users must, in turn, personally determine the relative importance of available features. The major compromise to be made is between

---

\* Cornell University

\*\* University of Pennsylvania, support provided by a  
Faculty Summer Research Fellowship

\*\*\* Princeton University

handiness (ease of use, convenience, versatility, etc.) and cost (computing efficiency, size, etc.)

As responses from contributors made clear, there is no general agreement on the definition of ease of use. To achieve handiness designers must balance the versatility (and attendant complexity) wanted by regular users, the extensiveness desired by the statistically sophisticated, the mnemonic command structure needed by occasional users, and pedagogical extras suitable for the statistically naive.

#### 4.1 User Interface

##### 4.11 User's documentation

A key determinant of the usefulness of a statistical package is its user's documentation. The documentation will be most effective if it has been thoroughly edited for technical and literary accuracy.

The complete documentation which accompanies a package is required to meet several distinct needs, and includes:

- a) Novice's manual - a clear introduction to the package which would permit an inexperienced user to run simple jobs with minimum effort.
- b) User's reference manual - everything the user needs to know about the package.
- c) Printed output - partial documentation within each run.
- d) Installation instructions - detailed description of the steps a computer installation must follow to get the package running.
- e) Programmer's manual - description of the structure of the package.
- f) Source listing of the package.

The first three items constitute user's documentation and of these the printed output is discussed in Section 4.14. The remaining three items are discussed in Section 4.3, on implementation.

A novice's manual is judged by its conciseness and clarity. Elementary examples should be illustrated before complications are introduced, and each should include a statement of the problem, and complete input and output listings.

The reference manual is judged by its accuracy, completeness, and organization. It must document exactly what the package does. It should be organized for ease of use, and include a table of contents and an index. A high quality of printing and binding of the manual also adds to its usefulness and permanence under constant reference. Interactive packages can provide additional reference documentation on-line.

Language syntax conventions should be documented, and potential errors described. For each procedure, the reference manual must include: a complete and accurate description of the procedure used; references to the literature for both the statistical and numerical techniques employed; complete documentation of examples used in illustration; a specification of default values; and a list and explanation of error messages. Additional desirable features include: clear and accurate descriptions of the control statements, and of the available options; documentation of size limitations and estimates of execution time, and precision (single or double). A user who wishes to use only one procedure will prefer that the documentation of each procedure be self-contained. For interactive packages, a list of choices should be available on request at any decision point. If greater depth is desired, lengthy explanations, perhaps quoting directly from the reference manual, may be provided.

#### 4.12 User's Environment

All packages create a computing environment which can affect the user's approach to his analysis much as architecture can affect our moods and thoughts. The key determinant of this environment is the package's command language.

The most computationally efficient languages usually employ code numbers in fixed card columns to specify procedures, variables, and options. This can force the user to refer regularly to a code book both for issuing commands and for interpreting output. Confusion can occur if a variable is assigned different numbers in different runs. Nevertheless, when similar calculations are repeated, the simplicity of this format can make control card preparation and checking routine. Some packages allow the specification of variable labels and translate procedure and option numbers into descriptive terms to produce more readable output. Others relax the fixed column requirements.

With the added expense of a command translator, a mnemonic language can be used, and procedures named with terms descriptive of their function. Often options are also



given names. Provision may be made for assigning names to variables and variable categories to be used in commands as well as in output labeling. These names can be included in an on-line data archive. The resulting commands are easier to write and check since they are immediately understandable. Free format input, in which delimiting punctuation separates the terms, may be available, especially on interactive systems where column counting is awkward.

The specification of variable transformations can be made in any of the above ways. An algebraic syntax is concise, although it requires a language interpreter of greater complexity (and cost), since several operations can be specified in a single command and function hierarchy must be recognized. Some packages also allow user-defined macros of elementary operations.

The computer's operating system can significantly affect the user's environment. It should be easy for a user to gain access to the package. Whether it is in batch or interactive operating mode is also important.

The user's environment consists of not merely the vocabulary and syntax of the control language, but also the variety and completeness of the available procedures and the ease of access to them. Experience has shown that what a user can do with his data in a chosen package frequently limits what he considers doing in his analysis. (v. Section 4.21.)

Packages should provide some protection for the user. Violations of standard assumptions (e.g., singular covariance matrices) should be reported, as should any corrective action taken. In interactive packages syntax errors must not be fatal. In batch packages the control card scan should be completed before the job is abandoned if any error is encountered. All packages should generate readable error messages which are sufficiently specific (e.g., "COMMAND 'OPTIONS' NOT RECOGNIZED", is better than "OPTIONS MUST FOLLOW PROCEDURE SPECIFICATION").

#### 4.13 Missing Values

In the analysis of multivariate data sets collected in surveys or experiments, it is seldom true that values for all the variables have been recorded for every individual. If a package is to be useful for such analyses, it must be able to handle data with missing values.

The manual must state whether more than one missing-value code can be specified, and what the codes are; whether missing values can be selectively included or excluded

from specific analyses; whether they can be replaced (e.g. by the mean, median, or mode of that variable, or in more sophisticated ways), and whether transformations of [missing] produce [missing].

Missing values present thorny statistical problems, many of which remain unsolved, and so it is not surprising that the ad hoc methods used by some procedures in analyzing data with missing values are inadequate. The user must be told at least what individual routines do when they encounter missing values. Are missing values automatically excluded (e.g. in the computation of correlations) and do other routines allow for this (e.g., is the correlation matrix then used in a regression procedure with algorithms valid only for complete data)? Does the documentation make this clear? How are missing values reported in the output?

#### 4.14 Printed Output

The printed output is the visible end product of a package. It must provide the user's results in readable form and ought to include additional identification and documentation.

The visual impact of the printed output greatly affects the true usefulness of a package. While a package must print all necessary information, it should not burden the user with unwanted output. Some provisions to suppress standard output or request additional output are common and useful. Output should be succinct but not cluttered, complete but not voluminous.

The printed output is often the only permanent record of the run and as such it should include all information necessary for tracing both user and package bugs. A listing of the user's control cards (with translations of any non-mnemonic codes) will help to document what was actually done. Such common identifiers as the date, system, and package version help to identify package bugs.

Output from the individual statistical procedures can also function as documentation. Particularly with newer statistical methods it is advantageous to include references to the literature in the printed output.

#### 4.15 Graphics

Graphics are desirable both at the initial data preparation phase of analysis (e.g., histograms of the data), and at the conclusion of a procedure (e.g., plots of the residuals from a regression.) Graphics add an important dimension to data analyses. Such things as probability



plots, histograms, principal components plots for clusters, and contour plots have only poor non-graphical substitutes. We think that plotting capabilities are necessary for statistical completeness.

Any output device can be used for plotting. Full-page scatter-plots are commonly generated on non-graphical devices such as line printers and occasionally on time-sharing terminals. Recently more sophisticated techniques (e.g. Andrews-Tukey 6-line plots have been proposed for plotting on such devices.

The evaluation of graphical capabilities is complicated by the recent proliferation of graphics devices. Despite this diversity, there are several areas in which graphics can be evaluated: Are all plots clear and well labelled? How are the scaling decisions made? Are the residuals plotted on the same scale as the data or on an expanded scale? Will a few outliers force a "window" so large as to dwarf the important patterns? How are multiple points at the same coordinate handled? How are points falling outside the plotting "window" indicated? Can plotting information be stored for later use or use on other devices? Are special grids and axes (e.g. log scales, calendar axes) available? Can the user override the default decisions for the above features?

#### 4.16 Cost

The usefulness of a package also depends on the cost of running a job. A package that is most desirable in terms of its control language, capabilities, and accuracy, may be too expensive to use regularly. But is the program that costs least per run necessarily cheapest? Not if it takes more user time or additional runs to debug control cards.

Different computer centers' charging algorithms assign different relative weights to core storage, CPU time, and I/O. A package designed for an installation with cheap core and expensive I/O will run at a disadvantage at a center with different weights. Large or long-running programs are often disproportionately penalized by schedulers, resulting in a slow response or turnaround. Procedure documentation should include a formula for estimating components of running cost in terms of the dimensions of the problem.

#### 4.17 Audience

Usually a package is designed with a particular audience in mind (e.g. agricultural experimenters, economists, social scientists). The capabilities of the package and

even the vocabulary and form of the control language may be specialized. Some designers intentionally expect users to have considerable technical knowledge, while others assume the defaults commonly wanted by their audience. Packages designed primarily for teaching may prevent actions which would be errors for naive users but may be desired by the experienced. A package's reference manual should identify its intended audience and, recognizing that widely distributed packages are frequently used by others, provide warnings of idiosyncracies.

#### 4.18 Pedagogy

Many packages are intended to be used in teaching and all may be so used at times. Some aspects of package design may be of particular importance to students and to teachers selecting a package for their classes, and these ought to be noted by reviewers.

Even a clear and well-written user's manual may be unsuitable as a text for statistically naive users, although examples and explanations of the value of different options will help. Output designed for the student may be more verbose than that desired by an experienced researcher. (e.g. "CHI SQUARED = 98 ON 24 DF; SEX = MALE TENDS WITH INCOME = HIGH; SMALLEST CELL COUNT = 19")

#### 4.2 Statistical Effectiveness

In Section 4.1 we examined package features from the point of view of handiness. In this section we shall discuss generally those features not already mentioned in Section 4.1, that are of particular interest to a practicing statistician, namely the available statistical tools, capabilities, and accuracy. In the planned reviews these would be discussed in detail.

##### 4.21 Versatility

Seldom, if ever, does data analysis consist simply of choosing the one correct statistical technique to analyze some data. Statistical analysis, in practice, is a more continuous or dynamic process, consisting of a sequence of procedures. After each procedure one is likely to pause to review the results to that point and consider the next step. A package of statistical programs, either batch or interactive, should facilitate such analyses.

The first step in an analysis is the preparation of the data. This is facilitated by a convenient file system. The data might be screened for outliers, and missing values; plots might be made, frequency distributions, and

various univariate statistics computed; departures from normality, or other assumptions might be noted, and transformations applied.

When the data have been prepared, the user will choose some of the procedures. In many cases, the output from the first procedure will be needed as input to another. For example, the residuals from a regression program might be plotted, or used in another regression, or coefficients from a discriminant analysis might be used to classify new observations.

A package designed for a particular audience should be able to compute the statistics needed by that audience from the data commonly collected by that audience. It should contain a full range of procedures, and each procedure should compute and print all the necessary statistics, with sufficient description to make them understandable.

No package can remain up-to-date, nor incorporate every new statistical technique invented. If a package does not offer a particular statistical procedure, the package can do one of two things: it can either make available, in machine-readable form, the results to that point, or it can offer a front-end structure that makes it convenient for the user to add his own program into the system (v. Section 4.33).

#### 4.22 Accuracy

Accuracy, as applied to statistical programs, means statistical accuracy. Several kinds of accuracy are required in order to achieve statistical accuracy.

First, the literature source used by the designer of the program must correctly describe the statistical technique to be programmed. Second, the programmer must follow the directions correctly. Third, the algorithms used for potentially difficult computations, for example matrix inversion, must be accurate. Fourth, statistics to evaluate the accuracy of the data, the statistical method, and the numerical method, must be computed and printed. Finally, the user's documentation and the printed output should make it absolutely clear what has been done. Warnings and error messages should be clear, and any fix-ups documented clearly.

To document the first kind of accuracy, specific references to the statistical literature, perhaps to standard textbooks, should be made in the reference manual, if not in the output. Precise definition of terminology must be given. For the second, a source listing following standard stylistic conventions provides the ultimate documentation,

although a flow chart would help. For the third, published algorithms should be used, and appropriate references given. If a new algorithm is used it should be published. The manual should comment on any machine differences in accuracy (e.g. the effects of word length on a particular algorithm).

Examples of statistics required for the fourth kind of accuracy are standard errors, coefficients of variation, condition number of a matrix, multiple correlations of each variable with the other variables. Also, the robustness of a procedure to input errors or outliers could be judged by removing outliers. Robustness of a procedure to departures from common assumptions about the data could be evaluated. The sensitivity of the analysis to rounding in the original data should be evaluated.

### 4.3 Implementation

#### 4.31 Programmer's documentation

Programmer's documentation consists of the programmer's manual and the source listing of the package. The source listing is the primary documentation of what the package does and how it does it. As such it will be most useful if it follows standard stylistic conventions. While it is helpful for a source listing to be available for reference at all installations using the package, developers of proprietary packages may not be willing to distribute their code -- a potential disadvantage to the users of their packages.

The source listing is useful for two purposes. One is to determine exactly what the package does: new statistical techniques can often be better understood if an implementation guide is available for study along with an algebraic description of the calculations. The other is to locate and correct bugs in the package. A bug can sometimes be found more rapidly by a user bitten by it than by a package designer who is interested in the whole package. Designers of large packages understandably do not want bugs locally fixed at each installation. They do want to know of bugs and suggested fixes for them which can then be distributed in later releases of the package.

#### 4.32 Portability

Packages developed at one installation are often taken to others. When the second installation has the same computer and operating system, a copy of the executable load module can be transferred. In other cases the source code itself must be used for the transfer. For these transfers to be possible, the package must be written in transportable code, for example ANSI Standard FORTRAN. Even here there



are difficulties since the standard leaves some serious ambiguities which have been resolved differently by different computer manufacturers. This problem can be met, at some programming expense, by writing in a pure, unambiguous subset of ANSI FORTRAN, by isolating system dependencies in a few subroutines which must then be rewritten for each installation, or by preprocessing to select statements appropriate to the object system.

Few interactive packages are transferrable across systems since such things as program segmenting and terminal I/O conventions are highly system-dependent. Nevertheless, the growth of computer networks and the tendency of large time-sharing systems to reach wide areas with multiplexor lines has, to some extent, replaced problems of transportability with those of access. Several time-sharing environments now run at installations across the country, and thus access to them is growing easier.

#### 4.33 Front End

Some packages make provision for the addition of procedures which were not provided by the designer, but rather are supplied by the user. There are two ways in which this can be done. Data processed by the front end of the package can be made available to the user by outputting it, so that it can then be read into another program. This method is quick but forces the user out of the package's working environment. Alternatively the user can add a new procedure which appears to be part of the package, and can be called through the package control language in the same manner as any other procedure.

Packages which make this provision have the advantage of allowing the user to concentrate his efforts on the programming of the analysis technique, and relieving him of the tribulations of data handling.

The structure of the package determines whether it is feasible to add a special-purpose routine with low anticipated usage. If the package is a single executable load module, the new procedure must be written as a subroutine with access to the COMMON areas of the package. In addition, the main program in the package must be modified to include the new name. Even more important, in terms of cost of development, the user must pay to re-link-edit the entire package.

On the other hand, if the package consists of a collection of independent routines which are dynamically linked together by the computer's operating system, additions are more easily made with less user difficulty. The new

procedure is written as an independent routine which makes specified subroutines calls for control information. These subroutines (but not the whole package) are linked to the new procedure. There is no need to modify the main program since the new procedure is located by the operating system and not by the main program. This method is dependent on the operating system and therefore not portable to computer systems other than the one on which it was written.

#### 4.34 Source Language

The language in which the package itself is written affects the features made available in the package. For example, the high overhead cost of the interpretation of free format control languages or algebraic interpreters. The required string handling operations take significantly less time in languages such as PL/I, APL, Algol-68, and BASIC in which string handling is a primitive operation. Further, in languages such as Dartmouth BASIC, APL, or Algol-68, the package designer is given object-time access to the source language, and thus can harness existing code.

## APPENDIX 2

## DATA USED IN TESTING THE FOUR PACKAGES

ANOVA  
Data Set 1

8	10	8
4	8	6
0	6	4
14	4	15
10	2	12
6	0	9

ANOVA  
Data Set 2

Data Manufactured to Conform to That  
of Snedecor and Cochran (p. 472)

31		27	23
29		25	25
		24	26
		25	25
20	22	17	
18	23	13	
16	21		
23	17		

ANOVA  
Data Set 3

Kutner's Data For 4 X 3 Design (Kutner, 1974)

42 36 19	53 53	31 25 24
44 13 22	26 21	-3 25
28 34 13	34 36	3 32 26
23 42	33 31	4 28 16
1 19 29	11 7 -6	21 9 1
	9 1	3
24 22 15	27 12 16	22 25 12
9 -2	12 -5 15	7 5

REGRESSION  
Data Set 4

Data Used For Two and Four Decimal Place Tests

Data Point	X1	X2	X3	Y (Dependent Variable)
1	50.71	10.19	10.19	5.01
2	49.80	18.87	18.87	10.00
3	35.91	31.01	31.01	18.65
4	76.31	17.00	17.00	19.51
5	43.47	13.81	13.81	17.76
6	88.01	19.75	19.75	20.01
7	63.50	14.92	14.92	19.21
8	59.81	12.15	12.15	25.25
9	81.08	10.88	10.88	35.10
*10	53.27	19.00	19.01	16.10

\*Data point for 4 decimal place test

10	53.27	19.00	19.0001	16.10
----	-------	-------	---------	-------



REGRESSION  
Data Set 5

Data Suggested by Alder & Rossler (1972)

X1	X2	X3	X4	X5	X6	Y
124	83	13	12	15	13	4
115	109	7	26	27	13	4
108	111	6	17	21	20	3
140	79	8	28	23	11	3
111	115	10	20	26	4	8
109	120	5	26	19	10	5
119	101	6	24	14	20	4
119	87	10	8	14	17	5
108	106	7	10	16	5	7
131	100	6	21	20	8	5
118	99	4	26	16	10	4
96	119	7	20	17	12	7
148	66	10	16	11	9	2
132	77	5	27	19	21	3
118	104	9	18	23	21	4
123	89	6	13	19	13	5
106	126	11	29	30	18	11
109	83	6	16	23	15	3
150	64	6	11	7	7	1
113	96	4	25	26	19	4
117	91	6	30	19	2	4
114	118	9	24	19	10	3
106	140	13	40	31	10	9
114	120	12	17	21	18	9
134	93	9	17	24	12	3
115	116	15	41	33	22	5
109	123	10	34	24	23	4
109	105	11	22	25	13	5
111	111	8	25	19	12	4
114	105	13	44	20	24	3

## APPENDIX 3

## SURVEY OF STATISTICAL PACKAGE USERS

A survey of attitudes toward the four packaged programs was taken among users of these packages. The sample was limited to experienced users of program packages in the North Carolina Educational Computing Service (NCECS) community. The members of NCECS use the Triangle Universities Computational Center (TUCC) as their source of computer packages.

These users were asked to rank the regression procedures of BMD, SAS, SPSS, and TSAR on three bases--the program manual, output, and ease of use. The packages were ranked on a 4-3-2-1 basis where 4 was the top rating and 1 the lowest. The overall rankings and their mean ranking are given in Table 13.

The number of qualified respondents was disappointing. Since the major packages are constantly improving with respect to flexibility, the researcher is no longer required to be familiar with five to eight packages to analyze his data, but can now do so with two or three. Therefore, of those contacted only eight users felt they had the expertise to evaluate all four packages.

The SPSS manual was rated best by all respondents. The raters considered the SAS (1972) and TSAR manuals about equal in helpfulness, while BMD was considered the least helpful.

TABLE 13  
MEAN RANKINGS OF REGRESSION PROCEDURES

<u>Manual</u>		<u>Output</u>		<u>Ease of Use</u>	
Package	$\bar{X}$	Package	$\bar{X}$	Package	$\bar{X}$
SPSS	4.0	SAS72	3.4	SPSS	3.8
TSAR	2.3	SPSS	3.4	SAS72	2.6
SAS72	2.1	BMD	1.7	TSAR	2.6
BMD	1.6	TSAR	1.5	BMD	1.1

When considering the completeness of the output of the packages, SAS and SPSS appeared to be comparable since they had the same mean rating. BMD and TSAR were thought to be about equal but well below SPSS. Finally, BMD was perceived to be the most difficult to use by 85 percent of the respondents. A package's ease of use is apparently related to the amount of documentation available.

## APPENDIX 4

PROGRAM LISTING FOR THE  
ANALYSIS OF VARIANCE

```

ANOVA:  PROCEDURE OPTIONS(MAIN) REORDER;
/*THIS PROGRAM WILL COMPUTE AN ANALYSIS OF VARIANCE FOR AN
UNBALANCED DESIGN USING FIVE DIFFERENT METHODS OF ADJUSTING
THE DEPENDENT VARIABLES. IT ALLOWS LABELING OF VARIABLES.
THE CODE WILL CHECK FOR EMPTY CELLS. IF ANY CELL IS EMPTY
THE PROGRAM WILL TERMINATE.*/
DECLARE(A1,A2,A3)CHARACTER(8) VARYING,
        (F1,F2,F3,F4,F5,F6)FIXED BIN;
/*READ IN NUMBER OF LEVELS,LABELS,FORMATING*/
GET LIST(II,JJ);
GET SKIP LIST(A1,F1,A2,F2,A3,F3,F4,F5);
F6=F4-F3+1;
BEGIN;
BI:     DECLARE(N(II,JJ),(NID)(II),(NDJ)(JJ))FIXED BIN(31),
        ((YIJD,YIJD5)(II,JJ),(YIDD)(II),(SSAB,SSASB)
        FLOAT(16),(YDJD)(JJ)) FLOAT(16),(Y(II,JJ,50))FLOAT(16)
        ,(W,V,XBARI,XBARJ)(5)FLOAT(16),(SSA,SSAA,SSB)FLOAT(16)
        ,(T,X)(5,5),U(5),C2(5))FLOAT BIN(53),
        (C3,SPT1,SPT2,SPT3,XBARI,XBAR2,YDD,YSOR,Y2)FLOAT(16),
        (TITLE)CHAR(40)VARYING,(SSTGT,SSBSA,SSE)FLOAT(16),
        (D) FLOAT BIN(53),(OP)FIXED BIN,(SV1,SUJ)FLOAT(16);
        DECLARE MINV ENTRY(*,*) FLOAT BIN(53),FIXED BIN,
        FLOAT BIN(53),FLOAT BIN(53));
PRINTX: PROCEDURE(A1,A2,A3,TITLE,SSTGT,SS1,SS2,SSAB,SSE,K1,K2,NDD);
/*SUBPROGRAM WILL SET UP THE ANALYSIS OF VARIANCE TABLES*/
DECLARE(A1,A2,A3)CHARACTER(8)VARYING,
        (TITLE)CHARACTER(40)VARYING,
        (SS1,SS2,SSAB,SSTGT,SSE,MSTGT,MSA,MSB,MSAB,MSE)
        FLOAT(16);
FI:     FORMAT(COL(16),A,COL(35),F(2),COL(48),F(14,8),COL(71),
        F(12,6),COL(90),F(8,5));
        PUT PAGE;
        PUT EDIT('ANALYSIS OF VARIANCE -- FIVE',
        'METHODS')(COL(26),A,A)SKIP(3);
        PUT EDIT('ANALYSIS OF VARIANCE TABLE FOR DEPENDENT VARIABLE ',
        'A3, TWO WAY DESIGN')(COL(24),(3) A) SKIP(4);
        PUT EDIT(TITLE)(COL(40),A) SKIP(3);
        PUT EDIT('SOURCE','DF','SUM OF SQUARES','MEAN SQUARE',
        'F VALUE')(COL(16),A,COL(35),A,COL(48),A,COL(71),
        A,COL(90),A) SKIP(3);
K=NDD-1;

```

```

KA=K1-1;
KB=K2-1;
KAB=KA*KB;
KERR=NDD-(11*JJ);
MSTGT=SSTGT/K;
MSA=SS1/KA;
MSB=SS2/KB;
MSAB=SSAB/KAB;
MSE=SSE/KERR;
PUT EDIT('TOTAL',K,SSTGT,MSTGT)(COL(16),A,COL(35),F(2),
      COL(48),F(14,8),COL(71),F(12,6))SKIP(2);
PUT EDIT(A1,KA,SS1,MSA,MSA/MSE)(R(F1))SKIP(2);
PUT EDIT(A2,KB,SS2,MSB,MSB/MSE)(R(F1))SKIP(2);
PUT EDIT(A1,'*',A2,KAB,SSAB,MSAB,MSAB/MSE)(COL(16),(3) A,
      COL(35),F(2),COL(48),F(14,8),COL(71),F(12,6),
      COL(90),F(8,5))SKIP(2);
PUT EDIT('ERROR',KERR,SSE,MSE)(COL(16),A,COL(35),F(2),
      COL(48),F(14,8),COL(71),F(12,6))SKIP(2);
END PRINTX;
N=0;
Y=0;
ON ENDFILE(SYSIN) GO TO L2;
/*READ OBSERVATIONS AND ASSIGN TO PROPER CELL*/
J=1;
L1: DO WHILE (J<=JJ);
      GET EDIT(I,J,OBS)(COL(F1),F(1), COL(F2), F(1), COL(F3),
      F(F6,F5));
      N(I,J)=N(I,J)+1;
      Y(I,J,N(I,J))=OBS;
END L1;
L2: DO I=1 TO 11;
L3: DO J=1 TO JJ;
      IF N(I,J)=0 THEN
L4: DO;
L5: PUT SKIP(2) LIST('WARNING--EMPTY CELL RECHECK DATA');
      STOP;
      END D1;
      END L3;
END L2;
NDD=0;
NID=0;
NDJ=0;
/*COMPUTE ROW AND COLUMN COUNTS FOR OBSERVATIONS*/
L4: DO I=1 TO 11;
L5: DO J=1 TO JJ;
      NID(I)=NID(I)+N(I,J);
      END L5;
END L4;
L6: DO J=1 TO JJ;
L7: DO I=1 TO 11;
      NDJ(J)=NDJ(J)+N(I,J);

```

```

      END L7;
      END L6;
L8:   DO I=1 TO III;
      NDD=NDD+NID(I);
      END L8;
      YDDD,Y2,YSQR=0;
      YIJD, YIJD=0;
      YIDD=0;
      YDJD=0;
L9:   DO I=1 TO III;
L10:  DO J=1 TO JJ;
      L=N(I,J);
L11:  DO K=1 TO L;
      YDDD=YDDD+Y(I,J,K);
      Y2=Y(I,J,K)*Y(I,J,K);
      YSQR=YSQR+Y2;
      YIJD(I,J)=YIJD(I,J)+Y2;
      YIJD(I,J)=YIJD(I,J)+Y(I,J,K);
      END L11;
      YIDD(I)=YIDD(I)+YIJD(I,J);
      END L10;
      END L9;
L12:  DO J=1 TO JJ;
L13:  DO I=1 TO III;
      YDJD(J)=YDJD(J)+YIJD(I,J);
      END L13;
      END L12;
/*CALCULATE MEAN WITHIN EACH LEVEL OF ANALYSIS*/
      PUT EDIT('MEANS WITH EACH LEVEL OF THE ANALYSIS')
        (COL(42),A)SKIP(3);
      PUT EDIT('INDEPENDENT VARIABLE', 'LEVEL', 'MEAN')
        (COL(28),A,COL(58),A,COL(75),A)SKIP(2);
      PUT SKIP;
ML1:  DO I=1 TO III;
      PUT EDIT(A1,I,YIDD(I)/NID(I))(COL(35),A,COL(60),F(1),
        COL(73),F(10,4))SKIP;
      END ML1;
ML2:  DO J=1 TO JJ;
      PUT EDIT(A2,J,YDJD(J)/NDJ(J))(COL(35),A,COL(60),F(1),
        COL(73),F(10,4))SKIP;
      END ML2;
ML3:  DO I=1 TO III;
ML4:  DO J=1 TO JJ;
      PUT EDIT(A1,'*',A2,I,J,YIJD(I,J)/N(I,J))(COL(35),
        (3) A,COL(60),F(1),F(2),COL(73),F(10,4))SKIP;
      END ML4;
      END ML3;
/*COMPUTE SUM OF SQUARES FOR A*/
      TITLE='HIERARCHICAL, FIRST A THEN B';
      SPT1,SPT2,C3=0;
      T=0;

```

```

U,C2=0;
L14:  DO I=1 TO II;
      SPT1=SPT1+((YIDD(I)*YIDD(I))/NID(I));
      END L14;
      SSA=SPT1-((YDDD*YDDD)/NDD);
      /*COMPUTE SUM OF SQUARES FOR B ADJUSTED FOR A*/
      IF II<JJ THEN
D2:    DO;
L15:    DO J=1 TO JJ;
      SPT2=SPT2+((YDJD(J)*YDJD(J))/NDJ(J));
      END L15;
L16:    DO I=1 TO II-1;
      SPT3=0;
L17:    DO J=1 TO JJ;
      SPT3=SPT3+((N(I,J)*N(I,J))/NDJ(J));
      END L17;
      T(I,1)=NID(I)-SPT3;
      END L16;
L18:    DO I=1 TO II-1;
L19:    DO K=I+1 TO II-1;
L20:    DO J=1 TO JJ;
      T(I,K)=T(I,K)-(N(I,J)*N(K,J))/NDJ(J);
      END L20;
      T(K,1)=T(I,K);
      END L19;
      END L18;
      OR=II-1;
      CALL MINV(T,OR,D,0);
L21:    DO I=1 TO II-1;
      U(I)=YIDD(I);
L22:    DO J=1 TO JJ;
      U(I)=U(I)-N(I,J)*(YDJD(J)/NDJ(J));
      END L22;
      END L21;
L23:    DO J=1 TO II-1;
L24:    DO I=1 TO II-1;
      C2(J)=C2(J)+(U(I)*T(I,J));
      END L24;
      END L23;
L25:    DO I=1 TO II-1;
      C3=C3+(C2(I)*U(I));
      END L25;
      SSBSA=SPT2-SPT1+C3;
      END D2;
      ELSE
D3:    DO;
L26:    DO J=1 TO JJ-1;
      SPT3=0;
L27:    DO I=1 TO II;
      S=NID(I);

```



```

      SPT3=SPT3+((N(I,J)*N(I,J))/S);
      END L27;
      T(J,J)=NDJ(J)-SPT3;
      END L26;
L28:   DO J=1 TO JJ-1;
L29:     DO K=J+1 TO JJ-1;
L30:       DO I=1 TO II;
          S=NID(I);
          T(J,K)=T(J,K)-(N(I,J)*N(I,K))/S;
        END L30;
      T(K,J)=T(J,K);
      END L29;
      END L28;
      OR=JJ-1;
      CALL MINV(T,OR,D,0);
L31:   DO J=1 TO JJ-1;
          U(J)=YDJ(J);
L32:     DO I=1 TO II;
          U(J)=U(J)-(N(I,J)*YIDD(I))/NID(I);
        END L32;
      END L31;
L33:   DO I=1 TO JJ-1;
L34:     DO J=1 TO JJ-1;
          C2(I)=C2(I)+(U(J)*T(I,J));
        END L34;
      END L33;
      SSBSA=0;
L35:   DO J=1 TO JJ-1;
          SSBSA=SSBSA+(C2(J)*U(J));
        END L35;
      END D3;
L36:   DO I=1 TO II;
L37:     DO J=1 TO JJ;
          C3=C3+(YIJD(I,J)*YIJD(I,J))/N(I,J);
        END L37;
      END L36;
      SSE=YSQR-C3;
      SSTGT=YSQR-(YDDD*YDDD)/NDD;
      SSAB=SSTGT-SSE-SSA-SSBSA;
      CALL PRINTX(A1,A2,A3,TITLE,SSTGT,SSA,SSBSA,SSAB,SSE,II,JJ,
          NDD);
      TITLE='THE STANDARD PARAMETRIC ANALYSIS';
      W=0;
      V=0;
L38:   DO I=1 TO II;
L39:     DO J=1 TO JJ;
          W(I)=W(I)+(1/((JJ*JJ)*N(I,J)));
        END L39;
      W(I)=1/W(I);
      END L38;
L40:   DO J=1 TO JJ;

```

```

L41:      DO I=1 TO III
           V(J)=V(J)+(1/((III*III)*N(I,J)))
         END L41
           V(J)=1/V(J)
         END L40
           SWI,SVJ=0
L42:      DO I=1 TO III
           SWI=SWI+W(I)
         END L42
L43:      DO J=1 TO JJ
           SVJ=SVJ+V(J)
         END L43
           XBAR1=0
           XBARJ=0
L44:      DO I=1 TO III
L45:          DO J=1 TO JJ
               X(I,J)=YIJD(I,J)/N(I,J)
           END L45
         END L44
L46:      DO I=1 TO III
L47:          DO J=1 TO JJ
               XBAR1(I)=XBAR1(I)+X(I,J)/JJ
           END L47
         END L46
L48:      DO J=1 TO JJ
L49:          DO I=1 TO III
               XBARJ(J)=XBARJ(J)+X(I,J)/III
           END L49
         END L48
           XBAR1,XBAR2,SSAA,SSB=0
L50:      DO I=1 TO III
           XBAR1=XBAR1+W(I)*XBAR1(I)
         END L50
           XBAR1=XBAR1/SWI
L51:      DO J=1 TO JJ
           XBAR2=XBAR2+V(J)*XBARJ(J)
         END L51
           XBAR2=XBAR2/SVJ
L52:      DO I=1 TO III
           SSAA=SSAA+(V(I)*((XBAR1(I)-XBAR1)**2))
         END L52
L53:      DO J=1 TO JJ
           SSB=SSB+(V(J)*((XBARJ(J)-XBAR2)**2))
         END L53
           CALL PRINTX(A1,A2,A3,TITLE,SSTOT,SSAA,SSB,SSAB,SSE,III,
                    JJ,NDD)
           SSB=0
L54:      DO J=1 TO JJ
           SSB=SSB+(YDJD(J)*YDJD(J))/NDJ(J)
         END L54
           SSB=(YDDD=YDDD)/NDD

```

```
SSASB=SSTOT-SSAB-SSB-SSE;
TITLE='HIERARCHICAL, FIRST B THEN A';
CALL PRINTX(A2,A1,A3,TITLE,SSTOT,SSB,SSASB,SSAB,SSE,II,
            JJ,NDD);
TITLE='EACH MAIN EFFECT ADJUSTED FOR EACH OTHER';
CALL PRINTX(A1,A2,A3,TITLE,SSTOT,SSASB,SSBSA,SSAB,SSE,
            II,JJ,NDD);
TITLE='A WEIGHTED MEANS ANALYSIS';
CALL PRINTX(A1,A2,A3,TITLE,SSTOT,SSA,SSB,SSAB,SSE,II,JJ,
            NDD);
END B1;
END ANOVA;
```

## APPENDIX 5

OUTPUT OF COMPUTER PROGRAM FOR  
THE ANALYSIS OF VARIANCE

## MEANS WITH EACH LEVEL OF THE ANALYSIS

INDEPENDENT VARIABLE	LEVEL	MEAN
A	1	26.0667
A	2	25.5333
A	3	8.7500
A	4	13.5000
B	1	22.7895
B	2	18.2105
B	3	15.8000
A*B	1 1	29.3333
A*B	1 2	28.2500
A*B	1 3	20.4000
A*B	2 1	28.0000
A*B	2 2	33.5000
A*B	2 3	18.1667
A*B	3 1	16.3333
A*B	3 2	4.4000
A*B	3 3	8.5000
A*B	4 1	13.6000
A*B	4 2	12.8333
A*B	4 3	14.2000

## ANALYSIS OF VARIANCE -- FIVE METHODS

ANALYSIS OF VARIANCE TABLE FOR DEPENDENT VARIABLE TREAT TWO WAY DESIGN

HIERARCHICAL, FIRST A THEN B

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE
TOTAL	57	9340.15517241	163.862371	
A	3	3133.23850575	1044.412835	9.45576
B	2	418.83381901	209.416910	1.89599
A*B	6	707.26618099	117.877697	1.06722
ERROR	46	5080.81666667	110.452536	

## ANALYSIS OF VARIANCE -- FIVE METHODS

ANALYSIS OF VARIANCE TABLE FOR DEPENDENT VARIABLE TREAT TWO WAY DESIGN

## THE STANDARD PARAMETRIC ANALYSIS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE
TOTAL	57	9340.15517241	163.862371	
A	3	2997.47238751	999.157463	9.04603
B	2	415.87332193	207.936661	1.88259
A*B	6	707.26618099	117.877697	1.06722
ERROR	46	5080.81666667	110.452536	

## ANALYSIS OF VARIANCE -- FIVE METHODS

ANALYSIS OF VARIANCE TABLE FOR DEPENDENT VARIABLE TREAT TWO WAY DESIGN

HIERARCHICAL, FIRST B THEN A

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE
TOTAL	57	9340.15517241	163.862371	
B	2	488.63938294	244.319691	2.21199
A	3	3063.43294182	1021.144314	9.24510
B*A	6	707.26618099	117.877697	1.06722
ERROR	46	5080.81666667	110.452536	



## ANALYSIS OF VARIANCE -- FIVE METHODS

ANALYSIS OF VARIANCE TABLE FOR DEPENDENT VARIABLE TREAT TWO WAY DESIGN

EACH MAIN EFFECT ADJUSTED FOR EACH OTHER

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE
TOTAL	57	9340.15517241	163.862371	
A	3	3063.43294182	1021.144314	9.24510
B	2	418.83381901	209.416910	1.89599
A*B	6	707.26618099	117.877697	1.06722
ERROR	46	5080.81666667	110.452536	

## ANALYSIS OF VARIANCE -- FIVE METHODS

ANALYSIS OF VARIANCE TABLE FOR DEPENDENT VARIABLE TREAT TWO WAY DESIGN

## A WEIGHTED MEANS ANALYSIS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE
TOTAL	57	9340.15517241	163.862371	
A	3	3133.23850575	1044.412835	9.45576
B	2	488.63938294	244.319691	2.21199
A*B	6	707.26618099	117.877697	1.06722
ERROR	46	5080.81666667	110.452536	