REGRESSION ANALYSIS OF FULL-RANK EXPERIMENTAL DESIGN MODELS

A Thesis

Presented to

the Faculty of the Department of

Industrial and Systems Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

Sheridan J. Berthiaume

December 1971

REGRESSION ANALYSIS OF FULL-RANK EXPERIMENTAL DESIGN MODELS

An Abstract of a Thesis

Presented to

the Faculty of the Department of

Industrial and Systems Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

Sheridan J. Berthiaume

December 1971

ABSTRACT

Regression analysis is a powerful and general solution method for the analysis of variance of experimental design problems. However, when the traditional experimental design model is expressed in the matrix form, $Y = Xb + e$, the X matrix will always be singular. Since $X'X$ will also be singular, the normal equations, $X'X\hat{b} = X'Y$, will have no unique solution. This means that standard regression techniques cannot be used for an analysis of variance without reparameterizing the model into a full-rank form.

In this study, a new method of formulating experimental design models is developed that leads directly to a full-rank system of normal equations without reparameterization. The full-rank model bases the expected value of the response variable on a standard cell of the experiment, rather than the overall mean of the experiment.

The technique is demonstrated for several example problems. It is concluded that the combination of full-rank model formulation and regression analysis is a very useful tool for the analysis of designed experiments. This is especially true for nonorthogonal design that are difficult or impossible to handle by the traditional sum-of-squares method.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

It is known that regression analysis can be used to find the variance estimates required for the analysis of variance of experimental design problems. In fact, regression is the most general solution method available since it solves problems with missing data, incomplete blocks, or unequal group sizes as easily as it solves problems with complete data and equal group sizes. However, in spite of its generality, the application of regression has been limited.

One of the reasons for this is that it has computational disadvantages. The heart of the regression technique is the solution of a system of simultaneous linear equations called the normal equations. This is tedious work for even fairly small problems since these systems of equations tend to become large very fast. For example, a two treatment, five levels per treatment, factorial experiment with one observation per cell would call for a solution to a system of twenty-five equations with thirty-six unknowns to find the error sum of squares. Additionally, three smaller systems must be solved to find the sums of squares associated with the main effects and the interaction effect. Obviously regression is not a hand or desk calculator technique except for the very smallest problems.

In this era of digital computers, these computational difficulties would not be sufficient to hold back the application of regression analysis if there were no other disadvantages. Unfortunately there are. Returning to the example, notice the excess of the number of unknowns over the number of normal equations. This means there are an infinite number of solutions to the normal equations. Increasing the number of replications per cell will produce more normal equations but since the new equations are not independent of the original twenty-five, there is basically no change in the system. The standard regression techniques, by hand or computer subroutines, are designed to solve systems of N normal equations with M unknowns where N is equal to M. A system of normal equations from an experimental design problem where N is less than M cannot be solved without modifications. These modifications could be any one of the following which are listed on the next page.

1. Add K more independent equations to the system so that $N + K = M$.

2. Combine or reparameterize the M unknowns such that there are L less of them so that $N = M - L$.

3. Change the normal equation solution method so that it will find a feasible solution when N is less than M.

4. Change the experimental design model so that when the normal equations are formed, $N = M$.

It appears that the main effort to tailor regression analysis to experimental design problems has been by methods 1 and 2. The difficulty is that the application of these two methods seems to be almost unique for every type of problem. That is, no general method of adding independent equations or reparameterizing the unknowns can be applied to all problems. Each problem entails considerable effort on the part of the experimenter to fit the problem to a regression routine.

This difficulty is reflected in the lack of application of regression analysis in experimental design textbooks. These books usually stress the traditional sum of squares approach to analysis of variance problems. This method is popular since it is amenable to hand or desk calculator solution of fair-sized problems as long as they have equal group sizes. If regression analysis is mentioned at all, it is usually in the context of being only an interesting fact that analysis of variance problems can be solved by regression. For example, in Hicks (7), the use of regres is demonstrated only for single-factor problems where ha solution of the normal equations is feasible. The book never demonstrates how to set up simple factorial models for regression solution. In Draper and Smith (4), a regression textbook, it reads:

> "We are not recommending that fixed-effects analysis of variance problems be handled by general regression methods. We are pointing out that they can be, if the correct steps are taken in handling the problem and that it is valuable to realize thi is possible."

In Cooley and Lohnes (2), after describing their analysi of variance computer program, they state:

> "The multiple-regression approach to analysis variance allows greater flexibility than the appro used here, but the preparation for execution of th programs is more complicated."

In summary, the difficulty of adding more independ equations or reparameterizing the unknowns seems to over the generality advantage of the regression technique.

Method 3, where the normal equations are solved fo feasible solution with N less than M can be handled by either linear programming or generalized inverse techniques. The linear programming approach as presented by Cashler (1), has all the advantages of regression analysis with respect to solving large unbalanced problems but it also has two unique disadvantages. The first one is that the number of unknown variables in the model must be doubled when the problem is formulated to overcome the linear programming non-negativity constraint. The second disadvantage is one of higher computer processing time for linear programming routines as compared with regression routines.

This brings us to method 4, which is the topic of this paper. Is there a way to write experimental design models that leads directly to a full-rank system of normal equations? If there is, then the application of regression analysis to experimental design problems will be greatly simplified and advantage can be taken of its generality.

A restriction on the new model will be that it also has physical significance to the experimenter rather than being an abstract combination of parameters. If this is true, the experimenter who knows the technique of formulating the model, which will apply to all problems, can feed problems directly into regression routines and determine the various sums of squares required for an analysis of variance.

Chapter II contains a brief overview of some of the background material for experimental design problems. In Chapter III, the new model is developed. Chapter IV contains examples showing the application of the technique to various types of problems. The advantages and disadvantages of the technique are summarized in Chapter V along with the conclusions about its application to analysis of variance problems.

CHAPTER II

ANALYSIS OF VARIANCE AND REGRESSION ANALYSIS

Analysis of Variance

The analysis of variance is a statistical technique introduced by R. A. Fisher about 1923 in connection with experimental design applications in biological research. It is a method of dividing the variation observed in experimental data into different parts, each part assignable to a known source, cause, or factor. It allows the assessment of the relative magnitude of variation resulting from different sources and the determination whether a particular part of its variation is greater than expected under a null hypothesis.

Normally the analysis of variance is used to test the significance of the differences between the means of the observed dependent variables in different groups where each group has received a different treatment. The purpose being to see if the treatment has a significant effect on the dependent variable or if the deviations in the group means are due to random error.

The analysis of variance makes two basic assumptions about the distribution of the dependent variable within each group. These are listed on the following page.

1. The dependent variable in each of the treatment groups is normally distributed.

2. The variance of the dependent variable in each of the treatment groups is equal.

Assume an experiment is performed to determine the effect of a factor that has been set or measured at k different levels. A measurement of the dependent variable, Y from one of k treatment groups is considered to be composed of three quantities:

u - the overall expected value of the dependent variable

t_i - the deviation from the expected value of the dependent variable due to the effect of the i^{th} treatment

e - a deviation from the expected value due to the fact that measurements of the dependent variable are normally distributed with a mean of zero and a variance of σ_e^2

To represent the j^{th} observation from the i^{th} treatment group the model is written as

$$Y_{ij} = u + t_i + e_{ij}$$
$$i = 1, 2, \ldots, k$$

The null hypothesis is that all the treatment effects are equal to zero.

$$t_i = 0$$
$$i = 1, 2, . . . , k$$

This hypothesis is tested by first partioning the total sum of squares of the deviation of the measurements from the overall mean, $\overline{Y}$, into two additive and independent parts. These are called the within groups sum of squares and the between groups sum of squares. To show this is possible let n_i be the number of observations in the i^{th} group and let $\overline{Y}_i$ be the mean of the i^{th} group. We begin by writing the identity

$$(Y_{ij} - \overline{Y}) = (Y_{ij} - \overline{Y}_i) + (\overline{Y}_i - \overline{Y})$$

Squaring this identity and summing over the n_i cases in the i^{th} group yields

$$\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 + \sum_{j=1}^{n_i} (\overline{Y}_i - \overline{Y})^2$$
$$+ 2(\overline{Y}_i - \overline{Y}) \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)$$

The last term on the right disappears since the sum of deviations of group observations from the group mean is zero. Therefore

$$\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = \sum_{j=1}^{n_i} (Y_{ij} - Y_i)^2 + n_i (Y_i - \bar{Y})^2$$

We now sum over the k groups to obtain

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k} n_i (\bar{Y}_i - Y)^2$$

Thus the total sum of squares is partioned into two additive groups, a sum of squares within groups and a sum of squares between groups. Of the three terms, any two are independent and could be used to estimate the common group variance, σ_e^2.

To do this we must know the degrees of freedom associated with each sum of squares since the estimate, S^2, of a variance is

$$S^2 = \frac{(\text{deviations from mean of distribution})^2}{\text{degrees of freedom of estimate}}$$

If the total number of observations, $\sum_{i=1}^{k} n_i$, is equal to N, then the total sum of squares has N-1 degrees of freedom. One degree of freedom is lost due to the mean of the distribution being estimated. The within groups degrees of freedom can be found by knowing that in each group there are $n_i - 1$ degrees of freedom. One degree of freedom is lost in each group to estimate the group mean. Summing over the k groups yields

$$\sum_{i=1}^{k} (n_i - 1) = \sum_{i=1}^{k} n_i - k = N - k$$

For the between groups sum of squares there are k means and one degree of freedom is lost by expressing the group mean as deviations from the grand mean so there are k - 1 degrees of freedom. Notice that the degrees of freedom are additive.

$$\underset{(N - 1)}{\text{Total}} = \underset{(N - k)}{\text{Within}} + \underset{(k - 1)}{\text{Between}}$$

With the sums of squares and the degrees of freedom we can now estimate the within and between groups variances. These variance estimates are also called the mean squares.

$$S_W^{\ 2} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2}{N - k}$$

$$S_B^{\ 2} = \frac{\sum_{i=1}^{k} n_i \ (\overline{Y}_i - \overline{Y}\ )^2}{k - 1}$$

While the sums of squares and the degrees of freedom are additive, the variance estimates are not.

$$S_T^2 \neq S_W^2 + S_B^2$$

Going back to the second basic assumption of the analysis of variance, remember that the within groups variance is the same for all groups. This means the expected value of the within groups variance is σ_e^2, the population variance.

$$E(S_W^2) = \sigma_e^2$$

The expected value of S_B^2 may be shown to be

$$E(S_B^2) = \sigma_e^2 + \frac{\sum_{i=1}^{k} (u_i - u)^2}{k - 1} \quad \frac{(N - \sum_{i=1}^{k} n_i^2/N)}{k - 1}$$

Where u_i, and u are population means. When the null hypothesis is true, the term on the right is equal to zero since the mean of each group is equal to the overall mean. Therefore the expected value of S_B^2 reduces to σ_e^2 and

$$E(S_B^2) = E(S_W^2)$$

When the null hypothesis is false and the means of the groups differ from u,

$$E(S_B^2) = \sigma_e^2 + \text{measure of the variation of } u_i \text{ from } u$$

To test the null hypothesis, the ratio of S_B^2/S_W^2 is examined. If the population means differ from each other $E(S_B^2/S_W^2)$ will be greater than unity. Therefore if this ratio is significantly greater than unity this is evidence for the rejection of the null hypothesis and for the acceptance of the alternative hypothesis that a significant difference exists between the treatment group means. The significance of the deviation from unity may be assessed by reference to a table of F values with k - 1 degrees of freedom associated with the numerator and N - k degrees of freedom associated with the denominator. The quantities involved in the preceeding discussion are usually displayed in an analysis of variance table which follows on the next page.

AOV TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio
Between Groups SS_B	$k - 1$	$\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2$	$SS_B/(k - 1)$	$\frac{SS_B(N - k)}{SS_W(k - 1)}$
Within Groups SS_W	$N - k$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$	$SS_W/(N - k)$	
Total	$N - 1$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2$		

Regression Analysis

Regression analysis is a statistical technique for extracting the main features of the relationships hidden or implied in tabulated figures. Even if no sensible physical relationship exists between the variables, we may wish to relate them by some sort of mathematical equation. While the equation might be physically meaningless it may nevertheless be extremely valuable for predicting the values of some variables from knowledge of other variables. In this paper we will be concerned with only linear regressi analysis which assumes that the relationship is linear in unknown parameters.

The variables involved can be classified as either independent or dependent variables. The dependent vari is also called the response variable. The independent variables are those which can be set to a desired value or else the values can be observed but not controlled. As a result of changes in the independent variables, an effec is reflected in the dependent variables. In general, we shall be interested in finding out how changes in the independent variables affect the response variables. Howev the end result is a mathematical formula that describes the relationship between the independent and dependent variables.

The simplest example of this is the case with only one independent variable, x, and one dependent variable, y. The problem is to find an equation that will predict the expected value of y given the value of x.

$$E(y|x = X) = f(X)$$

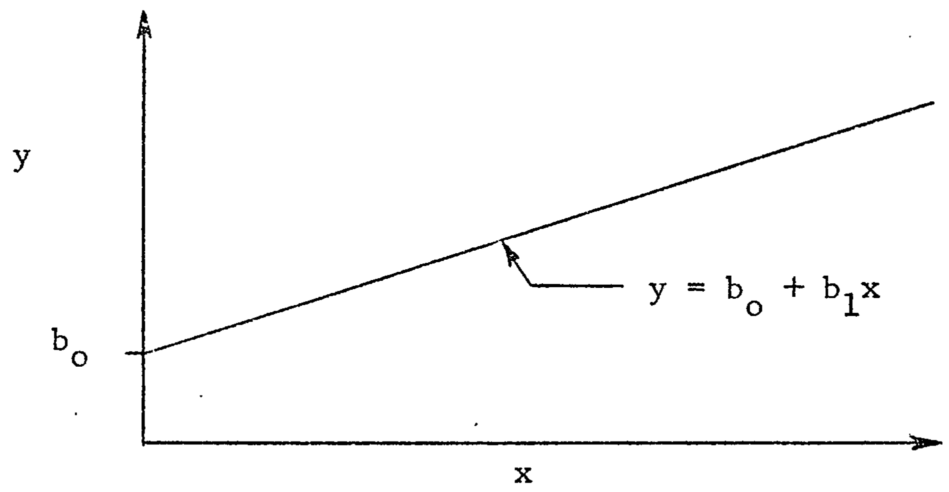
where f(X) is the regression equation. The highest power of x found in f(x) is called the order of the regression equation so that

$$f(x) = b_o + b_1x + b_2x^2 + b_3x^3$$

would be a linear third order model with constant coefficients b_i. Examining the first order equation

$$E(y|x = X) = b_o + b_1X$$

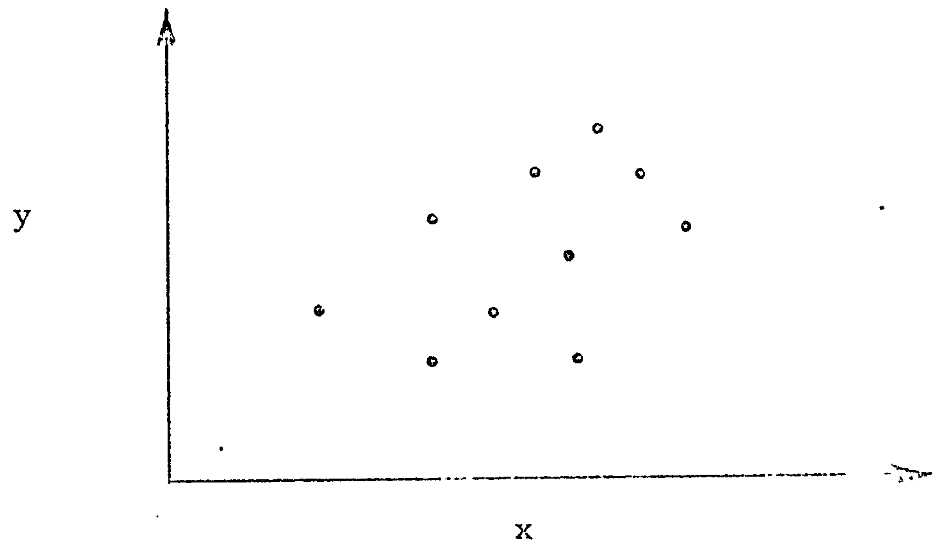
we see that this equation describes a straight line on the plot of y versus x.



So to develop this regression equation, the straight line relationship between y and x must be determined. The task here, of course, is to find the value of b_o and b_1 so that they do the best job of describing the relationship between y and x. The estimation of these coefficients is the essence of regression analysis. To perform the estimation of b_o and b_1 it is required to obtain some empirical data consisting of pairs of observations of y and x where x was set or measured and the response of y was simultaneously observed.

$$(y_1, x_1)$$
$$(y_2, x_2)$$
$$\vdots$$
$$(y_N, x_N)$$

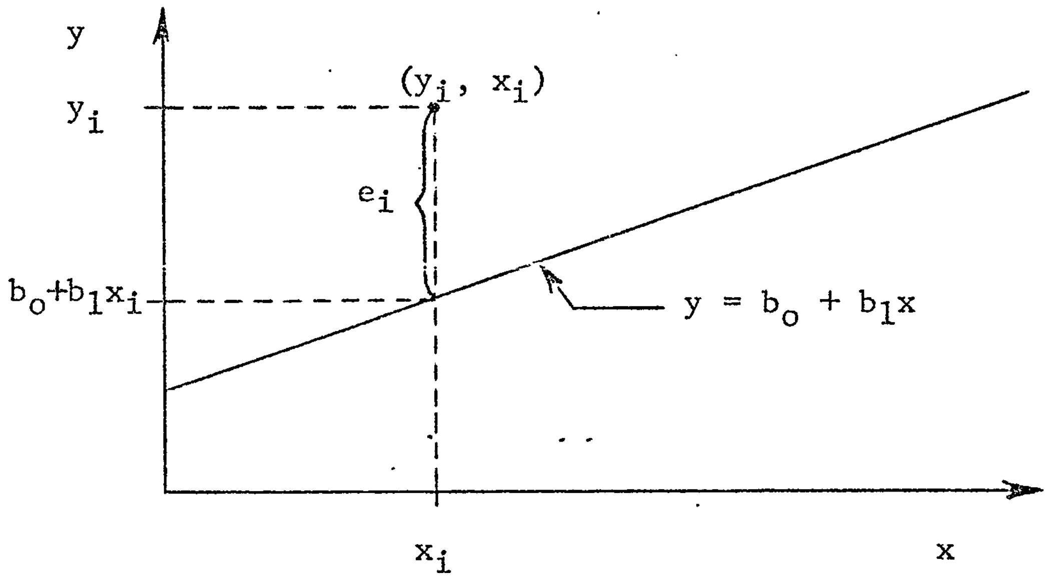
Plotting the observations might yield



Obviously no straight line can pass through all the N points so some method must be adopted to "fit" the line to the points. Since the points will not all lie on the regression line, we can express each point y_i by the model

$$y_i = b_o + b_1 x_i + e_i$$
$$i = 1, 2, \ldots, N$$

or graphically



when e_i is the deviation from the regression line. To find the best regression line we will estimate b_o and b_1 by the method of least squares. This method finds the b_o and b_1 that minimizes the sum of the squares of the e_i.

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b_o} = 0 \qquad \frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b_1} = 0$$

To do this the equations

$$y_i = b_o + b_1 x_i$$
$$i = 1, 2, \ldots, N$$

are expressed in the matrix form

$$Y = Xb + e$$

$$\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_N \end{pmatrix} \begin{pmatrix} b_o \\ b_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ e_N \end{pmatrix}$$

Then from Theorem 6.2 in Graybill (6), it is shown that the best (minimum variance) linear unbiased estimate of b is given by least squares. That is, the $\hat{b}$ that is the solution to the normal equations

$$\hat{b} = (X'X)^{-1} X'Y$$

is the best linear unbiased estimate of b.

If X'X is nonsingular, the estimates of b_o and b_1 are given by

$$\begin{pmatrix} \hat{b}_o \\ \hat{b}_1 \end{pmatrix} = \hat{b} = (X'X)^{-1} X'Y$$

The regression equation can then be written as

$$E(y|x = X) = b_o + b_1X$$

knowing that it is the best estimate of the expected value of y based on the method of least squares. This equation can now be used to predict the value of y given the values of x, or in other words, it describes a relationship between the independent and dependent variables.

The question may be asked "How well does the regression equation fit the data?" This is answered by partioning the total sum of squares about the line $y = 0$ (SS_T) into two categories, the sum of squares due to regression and the sum of squares about regression.

The sum of squares due to regression is the portion of the total sum of squares that is explained or accounted for by the regression equation. The larger the sum of squares due to regression, the better the fit of the regression equation to the data.

The sum of squares about regression is the sum of squares of the deviations of the data points from the regression line.

If the regression line passed through every data point, the sum of squares about regression would be zero and it would be apparent that the regression was perfectly fitted to the data. If the sum of squares about regression is large, it shows that there are significant deviations of the data from the regression line. This means that there is some lack of fit present. Since the sum of squares about regression is a measure of the fit error of the regression line, it will hereafter be referred to as the error sum of squares (SS_E). The sum of squares due to regression will be referred to as the regression sum of squares (SS_R). Therefore, we have

$$SS_T = SS_R + SS_E$$

To illustrate these quantities, suppose we were asked to find the first order regression of y on x given the following quantities.

x	y
1	1
1	3
2	4
2	6

Expressing the data in matrix form yields

$$\begin{pmatrix} 1 \\ 3 \\ 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} b_o \\ b_1 \end{pmatrix} + (e)$$

The normal equations $X'X\hat{b} = X'Y$ are developed as follows.

$$X'X = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \end{vmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} = \begin{vmatrix} 4 & 6 \\ 6 & 10 \end{vmatrix}$$

$$X'Y = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \end{vmatrix} \begin{pmatrix} 1 \\ 3 \\ 4 \\ 6 \end{pmatrix} = \begin{vmatrix} 14 \\ 24 \end{vmatrix}$$

$$\underset{X'X}{\begin{pmatrix} 4 & 6 \\ 6 & 10 \end{pmatrix}} \underset{\hat{b}}{\begin{pmatrix} \hat{b}_o \\ \hat{b}_1 \end{pmatrix}} = \underset{X'Y}{\begin{pmatrix} 14 \\ 24 \end{pmatrix}}$$

Solving for $\hat{b}$

$$\begin{pmatrix} \hat{b}_o \\ \hat{b}_1 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 6 & 10 \end{pmatrix}^{-1} \begin{pmatrix} 14 \\ 24 \end{pmatrix}$$

$$\begin{pmatrix} \hat{b}_o \\ \hat{b}_1 \end{pmatrix} = \begin{pmatrix} 5/2 & -3/2 \\ -3/2 & 1 \end{pmatrix} \begin{pmatrix} 14 \\ 24 \end{pmatrix}$$

$$\begin{pmatrix} \hat{b}_o \\ \hat{b}_1 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$$

Therefore

$$E(y) = -1 + 3x$$

The total sum of squares about y = 0 is

$$Y'Y = (1 \quad 3 \quad 4 \quad 6) \begin{pmatrix} 1 \\ 3 \\ 4 \\ 6 \end{pmatrix} = 62$$

The degrees of freedom associated with the total sum of squares is equal to the number of observations.

The regression sum of squares is equal to the sum of the squares of the distances of the regression line from y = 0 at each observation of x.

x	y	E(y)
1	1	3
1	3	3
2	4	5
2	6	5

For our example $SS_R = 3^2 + 3^2 + 5^2 + 5^2 = 58$. Equivalently, by matrix analysis

$$SS_R = \hat{b}'X'Y$$

$$SS_R = (-1 \quad 3) \begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \end{vmatrix} \begin{pmatrix} 1 \\ 3 \\ 4 \\ 6 \end{pmatrix}$$

$$SS_R = 58$$

The degrees of freedom of the regression sum of squares is equal to the number of coefficients in the regression equation, p.

$$DF_R = p = 2$$

The error sum of squares is the sum of squares of deviations of the observed points from the regression line. In our example, since each point has a deviation of 1 from the regression line

$$SS_E = 1^2 + 1^2 + 1^2 + 1^2 = 4$$

Normally the error sum of squares is determined by

$$SS_E = Y'Y - \hat{b}'X'Y$$

$$SS_E = 62 - 58$$

$$SS_E = 4$$

As mentioned earlier, the regression equation with the best fit to the data is the one with the lowest value for the error sum of squares. This means it has a minimum of variation of data points from the regression line. The degrees of freedom associated with the error sum of squares is equal to the number of data points, N, minus the number of regression coefficients to be estimated, p. In our example

$$DF_E = N - p = 4 - 2 = 2$$

Summarizing the results yields

Source	DF	Sum of Squares
Regression	$p = 2$	$\hat{b}'X'Y = 58$
Error	$N - p = 2$	$Y'Y - \hat{b}'X'Y = 4$
Total	$N = 4$	$Y'Y = 62$

Relation between Regression Analysis and Traditional Analysis of Variance Techniques

In this section the methods of the two previous sections will be drawn together to show how regression can be used to find the sums of squares and their associated degrees of freedom required for analysis of variance problems. This will be done by an example rather than a theoretical development. Using an example means that a loss of generality will be inevitable, but this is accepted with the hope of increasing the visibility of the relationship.

For our example we will take results from a hypothetical two-level single-factor experiment and demonstrate how the analysis of variance would be performed by the traditional method and by regression analysis. The computations shown here are designed to emphasize the similarities between the two methods and not to demonstrate exactly how the methods would be used to solve the problem.

The data for the problem is as follows:

Treatment	t_1	t_2
Results	3 5	7 9

The model for this experiment is

$$y_{ij} = u + t_i + e_{ij}$$

$$i = 1,\ 2$$
$$j = 1,\ 2$$

$$N = 4$$
$$k = 2$$

In equation form the experiment is written

$$\begin{aligned}
3 &= u + t_1 & &+ e_{11} \\
5 &= u + t_1 & &+ e_{12} \\
7 &= u & + t_2 &+ e_{21} \\
9 &= u & + t_2 &+ e_{22}
\end{aligned}$$

or expressed in the matrix notation

$$Y = Xb + e$$

$$\begin{pmatrix} 3 \\ 5 \\ 7 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} u \\ t_1 \\ t_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \end{pmatrix}$$

In Table (2.1), which follows, the calculations required to form an analysis of variance table are shown. In the left-hand column is the traditional sum of squares method and the right-hand column shows the associated regression calculations.

TABLE 2.1

Sum of Squares	Regression
1. No reparameterization necessary	1. Before beginning the regression analysis, the problem of the singularity of the X matrix must be overcome. For this example the problem will be reparameterized so that the X matrix is of full-rank. The matrix equation Y = Xb + e is written: $\begin{pmatrix} 3 \\ 5 \\ 7 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u + t_1 \\ t_2 - t_1 \end{pmatrix} + (e)$ Notice that this yields the same four equations as the previous matrix expression of the experiment.
2. Find the total sum of squares about $\overline{Y}$. $\overline{Y} = \frac{3 + 5 + 7 + 9}{4} = 6$ $SS_T = (3 - 6)^2 + (5 - 6)^2 + (7 - 6)^2 + (9 - 6)^2$ $SS_T = 20$	2. Find the total sum of squares about Y = 0. $SS_T = Y'Y = (3\ 5\ 7\ 9) \begin{pmatrix} 3 \\ 5 \\ 7 \\ 9 \end{pmatrix} = 164$ The total sum of squares for the two methods has a different reference point so the numerical results will be different.

TABLE 2.1 (2)

Sum of Squares	Regression
3. Find the total degrees of freedom. $DF_T = N - 1 = 4 - 1 = 3$	3. Find the total degrees of freedom. $DF_T = N = 4$
4. Find the within groups sum of squares about the group means. $\overline{Y}_1 = \frac{3 + 5}{2} = 4$ $\overline{Y}_2 = \frac{7 + 9}{2} = 8$ $SS_W = \sum_{i=1}^{2} \sum_{j=1}^{2} (Y_{ij} - \overline{Y}_i)^2$ $SS_W = (3 - 4)^2 + (5 - 4)^2 + (7 - 8)^2 + (9 - 8)^2$ $SS_W = 1 + 1 + 1 + 1 = 4$	4. Find the error sum of squares. $SS_E = SS_T - SS_R$ $SS_E = Y'Y - \hat{b}'X'Y$ $\hat{b} = (X'X)^{-1} X'Y = \begin{vmatrix} 4 & 2 \\ 2 & 2 \end{vmatrix}^{-1} \begin{vmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{vmatrix} \begin{vmatrix} 3 \\ 5 \\ 7 \\ 9 \end{vmatrix}$ $\hat{b} = \begin{vmatrix} 4 \\ 4 \end{vmatrix}$ $E(y) = 4 + 4x_1$

TABLE 2.1 (3)

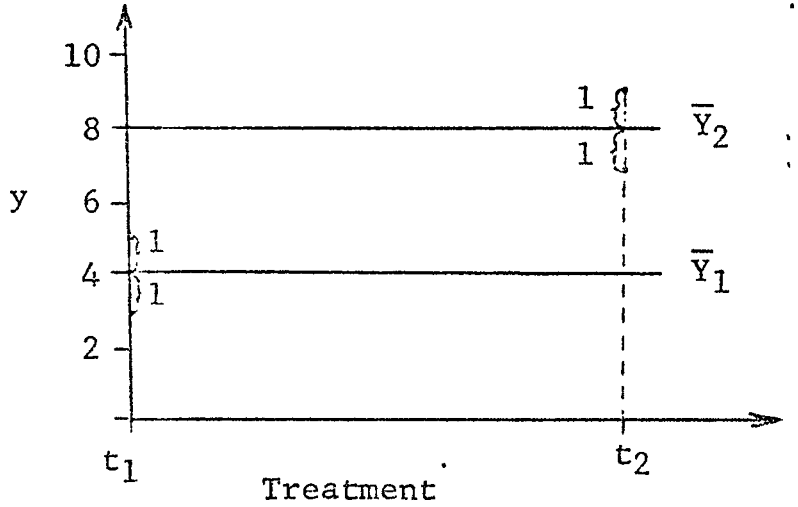
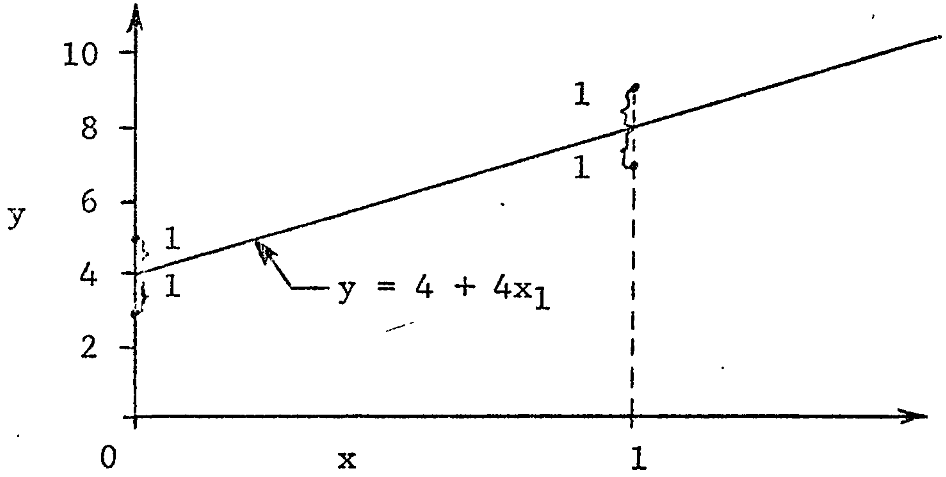
Sum of Squares	Regression
4. (continued)	4. (continued) $SS_R = (4\ 4) \begin{vmatrix} 24 \\ 16 \end{vmatrix} = 160$ $SS_E = 164 - 160 = 4$
	
5. Find the degrees of freedom for the within groups sum of squares. $DF_W = N - k = 4 - 2 = 2$	5. Find the degrees of freedom for the error sum of squares. $DF_E = N - \text{Rank}(X) = 4 - 2 = 2$

TABLE 2.1 (4)

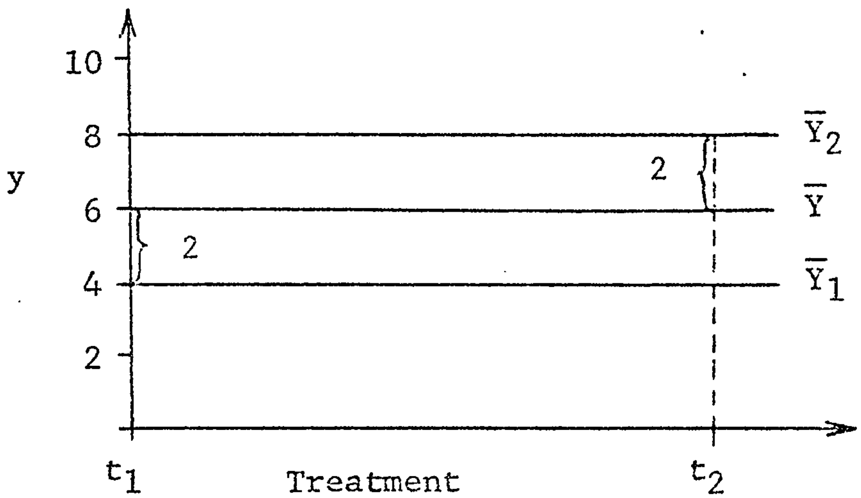
Sum of Squares	Regression
6. Find the between groups sum of squares. $SS_B = \sum_{i=1}^{2} n_i(\overline{Y}_i - \overline{Y})^2$ $SS_B = 2(2)^2 + 2(2)^2 = 16$	6. With the hypothesis H_0: $t_1 = t_2 = 0$, rewrite the matrix equation in the form, $Y = Za + e$ $Y = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u - 0 \\ 0 \end{pmatrix} + (e)$ which reduces to $Y = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (u) + (e)$ Find the reduction in the regression sum of squares if the hypothesis is true. $SS_R - SS_{R(a)} = \hat{b}'X'Y - \hat{a}'Z'Y$ $\hat{a} = (Z'Z)^{-1} Z'Y = (4)^{-1}(1\ 1\ 1\ 1) \begin{pmatrix} 3 \\ 5 \\ 7 \\ 9 \end{pmatrix}$ $\hat{a} = 6$

TABLE 2.1 (5)

Sum of Squares	Regression
6. (continued)	6. (continued)

Sum of Squares

6. (continued)



Regression

6. (continued)

$$SS_{R(a)} = \hat{a}'Z'Y$$

$$SS_{R(a)} = (6)(1\ 1\ 1\ 1)\begin{pmatrix}3\\5\\7\\9\end{pmatrix} = 144$$

$$SS_R - SS_{R(a)} = 160 - 144 = 16$$

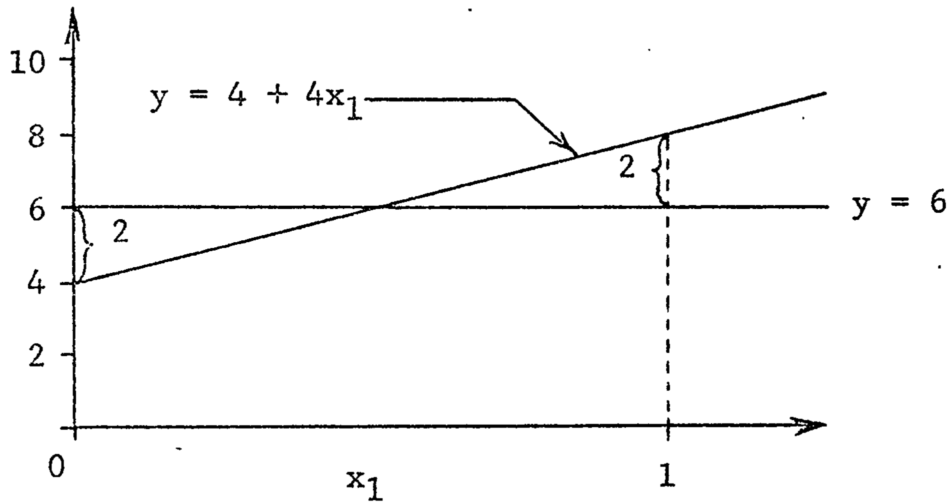


TABLE 2.1 (6)

Sum of Squares	Regression
7. Find the degrees of freedom for the between groups sum of squares. $$DF_B = k - 1 = 2 - 1 = 1$$	7. Find the difference in the degrees of freedom for the regression sum of squares with H_0 false and H_0 true. $$DF = \text{rank}(X) - \text{rank}(Z) = 2 - 1 = 1$$
8. Set up the Source, Degrees of Freedom, and Sum of Squares columns for the AOV Table.	8. Set up the Source, Degrees of Freedom, and Sum of Squares columns for the AOV Table.

Source	DF	SS
Between Groups	$k - 1 = 1$	$\sum_{i=1}^{k} n_i(Y_i - \bar{Y})^2 = 16$
Within Groups	$N - k = 2$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = 4$
Total (About $Y = \bar{Y}$)	$N - 1 = 3$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = 20$

Source	DF	SS
Regression (H_0 False)	$R(X) = 2$	$\hat{b}'X'Y = 160$
Regression (H_0 True)	$R(Z) = 1$	$\hat{a}'Z'Y = 144$
Reduction in SS_R if H_0 is true	$R(X) - R(Z) = 1$	$\hat{b}'X'Y - \hat{a}'Z'Y = 16$
Error SS (H_0 False)	$N - R(X) = 2$	$Y'Y - \hat{b}'X'Y = 4$
Total (About $Y = 0$)	$N = 4$	$Y'Y = 164$

The regression analysis of experimental design models is summarized in the following steps.

1. Write the experiment model in terms of a full-rank matrix equation.

$$Y = Xb + e$$

2. Find the total sum of squares about Y = 0 and its degrees of freedom.

$$SS_T = Y'Y$$

$$DF_T = N$$

3. Find the regression sum of squares and its degrees of freedom.

$$SS_R = \hat{b}'\ X'Y$$

$$SS_R = Y'X\ (X'X)^{-1}\ X'Y$$

$$DF_R = \text{rank}(X)$$

4. Find the error sum of squares and its degrees of freedom.

$$SS_E = SS_T - SS_R$$

$$SS_E = Y'Y - \hat{b}'X'Y$$

$$SS_E = Y'Y - Y'X\ (X'X)^{-1}\ X'Y$$

$$DF_E = N - \text{rank}(X)$$

5. Form a hypothesis that states that the effect of each of the experimental factors is zero. For each hypothesis, write a reduced model of the experiment that assumes the hypothesis is true. For a fixed factor, t, the t_j are assumed to be fixed constants and the hypothesis would be

$$H_0: \quad t_j = 0 \quad \text{for all } j$$

If t is a random factor, the t_j are assumed to be normally distributed random variables with a mean of zero and a variance of σ_t^2. The hypothesis to test the effect of t in this case would be

$$H_0: \quad \sigma_t^2 = 0$$

In either case, a model is formed by setting all terms in the original model that contain a t to zero which will yield a reduced model.

$$Y = Z_i a_i + e$$

$$i = 1, 2, \ldots, N_H$$

where $\quad N_H$ = number of hypotheses

6. Find the regression sum of squares and its degrees of freedom for each model.

$$SS_{Ri} = \hat{a}_i{}' Z_i{}' Y = Y' Z_i (Z_i{}' Z_i)^{-1} Z_i{}' Y$$

$$DF_{Ri} = \text{rank}(Z_i)$$

7. Find the difference in the regression sum of squares and the degrees of freedom for this model (i^{th} hypothesis is true) and the original model (i^{th} hypothesis is false.)

$$SS_{R-Ri} = SS_R - SS_{Ri} = \hat{b}'X'Y - \hat{a}_i'Z_i'Y$$

$$DF_{R-Ri} = \text{rank}(X) - \text{rank}(Z_i)$$

8. Form the analysis of variance table based on the following quantities.

AOV TABLE

Source	Degrees of Freedom	Sum of Squares
Regression	rank(X)	$\hat{b}'X'Y$
Factor 1	rank(X) - rank(Z_i)	$\hat{b}'X'Y - \hat{a}_i'Z_i'Y$
.	.	.
.	.	.
.	.	.
.	.	.
Factor n	rank(X) - rank(Z_n)	$\hat{b}'X'Y - \hat{a}_n'Z_n'Y$
Error	N - rank(X)	$Y'Y - \hat{b}'X'Y$
Total	N	$Y'Y$

From the previous table it is clear that regression is a fairly straight-forward, although computationally tedious, method of determining the sum of squares. The real problem as stated earlier, is in step 1. That is, the reparameterization of the model to a full-rank model. The remainder of this paper will be concerned with the method and examples of writing experimental design models that make this step of reparameterization unnecessary.

CHAPTER III

FULL-RANK EXPERIMENTAL DESIGN MODELS

The first part of this chapter will demonstrate why the traditional experimental design models always lead to an indeterminant system of normal equations with an excess of unknowns over independent equations.

Consider a single-factor experiment with r levels of the factor to be investigated as to their effect on a response variable. Also assume there are n_i replications for each level i. The model for this experiment is expressed by the following:

$$y_{ik} = u + t_i + e_{ik}$$
$$i = 1, 2, \ldots, r$$
$$k = 1, 2, \ldots, n_i$$

where

y_{ik} is the k^{th} observation of the response variable under the experimental condition of level i of the treatment,

u is the overall expected value of the response variable for the entire experiment.

t_i is the deviation from u caused by the effect of level i of the treatment t.

e_{ik} is the random error in the experiment which is normally distributed with a mean of 0 and a variance of σ_e^2.

In matrix form (y = Xb + e), the experiment is expressed by the following:

$$
\begin{vmatrix} y_{11} \\ y_{12} \\ \cdot \\ \cdot \\ \cdot \\ y_{1n} \\ y_{21} \\ \cdot \\ \cdot \\ \cdot \\ y_{2n_2} \\ y_{31} \\ \cdot \\ \cdot \\ \cdot \\ y_{rn_r} \end{vmatrix}
=
\begin{vmatrix}
1 & 1 & 0 & 0 & . & 0 & 0 \\
1 & 1 & 0 & 0 & . & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & 1 & 0 & 0 & . & 0 & 0 \\
1 & 0 & 1 & 0 & . & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & 0 & 1 & 0 & . & 0 & 0 \\
1 & 0 & 0 & 1 & . & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & 0 & 0 & 0 & . & 0 & 1
\end{vmatrix}
\begin{pmatrix} u \\ t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_r \end{pmatrix}
+
\begin{vmatrix} e_{11} \\ e_{12} \\ \cdot \\ \cdot \\ \cdot \\ e_{1n_1} \\ e_{21} \\ \cdot \\ \cdot \\ \cdot \\ e_{2n_2} \\ e_{31} \\ \cdot \\ \cdot \\ \cdot \\ e_{rn_r} \end{vmatrix}
$$

where the matrices have the following dimensions

Matrix	Dimension	
	Row	Column
Y	$\sum_{i=1}^{r} n_i$	1
X	$\sum_{i=1}^{r} n_i$	r + 1
b	r + 1	1
e	$\sum_{i=1}^{r} n_i$	1

When the normal equations

$$X'X\hat{b} = X'Y$$

are formed, the matrix X'X will be a square (r + 1) by (r + 1) matrix. From Theorem 1.20 in Graybill (6), the rank of X'X will be equal to the rank of X which will be equal to the number of independent rows in the matrix. In our experiment there are r different experimental conditions, one for each level of t, and for each condition there is an equation expressing the expected value of y for the condition.

$$\begin{array}{ll}
\text{Level 1:} & E(Y) = u + t_1 \\
\text{Level 2:} & E(Y) = u + t_2 \\
\vdots & \vdots \\
\text{Level r:} & E(Y) = u + t_r
\end{array}$$

Without adding any supplemental conditions on the experiment, it is clear that there is only one independent equation for each different condition in the experiment. This means, for our example, there are only r independent rows in the X matrix. Since X is of rank r, X'X is a singular (r + 1) by (r + 1) matrix of rank r. So there is no unique solution

$$\hat{b} = (X'X)^{-1} X'Y$$

to the normal equations.

Regardless of the number of factors or type of experiment, there will be no more independent rows in the X matrix than there are different experimental conditions. For a two-factor experiemnt with r and s levels per factor there will be rs different conditions so the rank of the X matrix will be rs. So, in general, rank(X) equals the product of the number of levels for each factor of the experiment.

The column dimension of X, designated p, will be equal to the number of unknowns in the experimental design model. For the single-factor example, p was equal to r + 1. For a two-factor experiment with r and s levels per factor, we have:

Factor	Number of Terms
u	1
1	r
2	s
1 X 2 interaction	rs
	Total = 1 + r + s + rs = p

It is apparent that p is much greater than the rank of X which is rs. This means, of course, that X'X is again singular. As the number of factors in an experiment is increased, p will always be greater than the rank of X. This is obvious since the rank of X will always be equal to the number of the highest level interaction terms in the b matrix. The dimension p will be equal to this, plus all the lower level terms in the b matrix. So, in summary, regardless of the experiment, the model will always lead to a set of indeterminant normal equations since X'X will be singular.

The preceeding discussion also leads us to the fact that the rank of X'X will be equal to the number of experimental conditions, or cells, in the experiment. Therefore, if the number of unknowns in the b matrix, p, can be reduced to the number of cells, X'X will be a p x p matrix of rank p. Under these conditions we can find $\hat{b}$ by

$$\hat{b} = (X'X)^{-1} X'Y$$

and proceed to find the sums of squares by the method of the preceeding chapter. We already know that this can be done by reparameterizing the model after it is written. However, this is extra work that must usually be done manually before the regression solution method begins.

The problem now, is how can we write the model directly so that the number of unknowns is equal to the number of experimental cells which leads directly to a full-rank X'X. It appears that the key to this is getting away from expressing the response variable as being equal to an overall mean, u, plus deviations caused by the experimental factors.

$$E(Y) = u + \text{deviations}$$

To reduce the number of unknowns, it is possible to select one of the cells of the experiment as standard from which the expected response of all other cells deviates.

$$E(Y) = \text{standard cell} + \text{deviations}$$

We will denote the expected response of the standard cell as S and will choose the cell where all factors are at level 1 as the standard cell. For a 2^2 factorial experiment (factors a and t) with 2 replications, the model would be developed as follows:

Cell a_1, t_1 is the standard cell so the expected response for this cell is simply S.

	t_1	t_2
a_1	E(Y) = S	
a_2		

For the a_1, t_2 cell the only deviation from S would be caused by the change in treatment t from level 1 to level 2. Therefore,

	t_1	t_2
a_1	E(Y) = S	E(Y) = S + t_2
a_2		

Likewise for a_2, t_1

	t_1	t_2
a_1	E(Y) = S	E(Y) = S + t_2
a_2	E(Y) = S + a_2	

For the a_2, t_2 there are two sources of deviations from S so there will also be an interaction term between the two factors.

	t_1	t_2
a_1	$E(Y) = S$	$E(Y) = S + t_2$
a_2	$e(Y) = S + a_2$	$E(Y) = S + a_2 + t_2 + at_{22}$

Notice that there are no terms in the equations with a subscript containing a 1. This is caused by the definition of cell 1 to be the standard from which deviations are measured. Writing the new model for the experiment yields

$$Y_{ijk} = S + a_i + t_j + at_{ij} + e_{ijk}$$

$$i = 1,\ 2$$

$$j = 1,\ 2$$

$$k = 1,\ 2$$

$$a_1 = t_1 = at_{11} = at_{12} = at_{21} = 0$$

Listing the number of unknowns in the model

$$S$$

$$a_2$$

$$t_2$$

$$at_{22}$$

shows that there are only four of them which equals the number of cells in the experiment. Therefore, we have succeeded for this case in expressing the experiment with the same number of unknowns as experimental conditions. Although not proved for the general case, it should be apparent that each cell introduces only one new term which is the highest level interaction possible between the single-factor terms in the cell. Since each cell introduces one new unknown and one more independent equation, this insures the fact that the number of unknowns and cells will be equal.

Writing the model of our 2^2 example in matrix form,

$$Y = Xb + e$$

yields

$$\begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} S \\ t_2 \\ a_2 \\ at_{22} \end{pmatrix} + \left| e \right|$$

To examine the rank of X it is rewritten as

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Notice that the top four rows form a diagonal of 1's with only 0's above the diagonal. Forming a determinant of the top four rows, it appears as

$$\begin{vmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{vmatrix} = 1$$

Since this 4 x 4 determinant has a nonzero value and the column dimension is 4, the rank of X is equal to four. This means X'X will be a (4 x 4) with a rank of 4 which means our system of normal equations will have a unique solution.

To summarize the method of expressing models that lead to a full rank X'X matrix:

1. Choose a cell of the experiment as the standard, S, from which the deviations in the expected value of Y for all other cells will be based on. In this paper, this will always be the level 1 cell for all factors.

2. Write the expected value for all other cells as Y = S + deviations from standard cell. Or, in equation form it may be written

$$Y_{rijk\ldots} = S + a_i + t_j + \ldots. + at_{ij} + \ldots. + e_{rijk\ldots}$$

where all experimental factors containing a subscript of one are zero.

The next chapter will contain examples, primarily from Hicks (7) and show how they can be solved with the combination of expressing the model in full-rank form and using regression analysis.

CHAPTER IV

EXAMPLE PROBLEMS

This chapter demonstrates how some representative experimental design problems can be systematically solved using the combination of full-rank model formulation and regression analysis. The method is applied to four different types of problems. They are as follows:

1. A completely randomized single-factor experiment with unequal group sizes.

2. A single-factor experiment with an incomplete block design.

3. A 2 x 2 factorial experiment with missing data.

4. A nested-factorial experiment with fixed and random factors.

The four problems are solved using a standard stepwise regression routine designed for regression analysis rather than analysis of variance problems. The routine is the BMD02R Stepwise Regression program which is one of the UCLA Biomedical Computer Programs described in Dixon (3). This widely-used package is available in many large-scale computing centers. The stepwise feature of the routine is not required but it is mandatory that the user be able to easily control which variables enter the regression equations. The BMDO2R routine accomplishes this by the Control-Delete commands which force variables into, or keep variables out of, the regression calculations. The routine also automatically provides the regression and error sums of squares and degrees of freedom for each regression as part of the output. This is a great advantage over a routine that only provides the regression coefficients and leaves the user to calculate

$$SS_R = \hat{b}'X'Y$$

and

$$SS_E = Y'Y - \hat{b}'X'Y$$

A limitation of BMDO2R for analysis of variance work is that it can handle no more than 80 variables in its regression calculations. The user must, therefore, insure that when the full-rank model is formulated, it contains no more than 80 different terms. If necessary, the number of terms in the model can be reduced by assuming certain factors or interactions have no effect and deleting the terms associated with those factors. BMDO2R can process up to 9999 observations which should be sufficient to handle most experiments.

On the following pages, the four example problems are solved with the BMD02R program and a step-by-step description of the solution is presented for each problem. A listing of each problem's input data for BMD02R is provided in Appendix A.

EXAMPLE NO. 1

Type:

A single fixed-factor experiment with unequal group sizes.

Source:

Hicks (7), page 42.

Problem:

In this experiment, a single factor, t, is set to five different levels and the number of measurements of the response variable in each group is different. The data for the experiment is the following:

Treatment	t_1	t_2	t_3	t_4	t_5
Response	83	84	86	89	90
	85	85	87	90	92
		85	87	90	
		86	87	91	
		86	88		
		87	88		
			88		
			88		
			88		
			89		
			90		

Solution:

1. Express model in terms of a full-rank matrix equation.

The full-rank model as developed in Chapter III for this experiment is:

$$y_{ik} = S + t_i + e_{ik}$$

$$i = 1, 2, 3, 4, 5$$

$$k = 1, \ldots, n_i \qquad \text{where} \quad n_1 = 2,\ n_2 = 6,\ n_3 = 11,\ n_4 = 4,\ n_5 = 2$$

$$t_1 = 0$$

The matrix representation of this model is the following:

$$\begin{vmatrix} 83. \\ 85. \\ 84. \\ 85. \\ 85. \\ 86. \\ 86. \\ 87. \\ 86. \\ 87. \\ 87. \\ 87. \\ 88. \\ 88. \\ 88. \\ 88. \\ 88. \\ 89. \\ 90. \\ 89. \\ 90. \\ 90. \\ 91. \\ 90. \\ 92. \end{vmatrix} = \begin{vmatrix} 10000 \\ 10000 \\ 11000 \\ 11000 \\ 11000 \\ 11000 \\ 11000 \\ 11000 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10100 \\ 10010 \\ 10010 \\ 10010 \\ 10010 \\ 10001 \\ 10001 \end{vmatrix} \begin{pmatrix} S \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{pmatrix} + (e)$$

2. <u>Find the total, error, and regression sums of squares and degrees of freedom.</u>

The regression routine is used to generate a regression equation which includes all five variables of the b matrix. The following output of the routine shows the regression and error (labeled residual) sum of squares and degrees of freedom.

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	5	191767.857	38353.571
RESIDUAL	20	23.140	1.157

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		[illegible]	)	
S	2	8.[illegible]+01	7.61-01	12196.9545
T2	3	1.5[illegible]22[illegible]	3.78-01	2.9171
T3	4	3.8182029+[illegible]0	8.27-01	21.3236
T4	5	6.[illegible]3221+00	9.32-01	41.4866
T5	6	7.[illegible]3219+00	[illegible]	42.3508

Using these to find the total sum of squares and degrees of freedom yields,

$$SS_{R(S,t)} = 191767.857 \qquad DF_{R(S,t)} = 5$$

$$SS_E = \underline{\quad 23.140} \qquad DF_E = \underline{20}$$

$$SS_T = 191790.997 \qquad DF_T = 25$$

3. <u>Form the appropriate hypotheses to test the significance of the experimental factors</u>.

For this example there is only one factor, t, which is fixed, so the only hypothesis to be tested is

$$H_0: \; t_2 = t_3 = t_4 = t_5 = 0$$

This null hypothesis states that treatment levels 2, 3, 4, and 5 cause no significant deviation from the standard response which is defined to be level 1 of the factor t.

4. <u>For each hypothesis, find the regression sum of squares and degrees of freedom for the reduced model that assumes the hypothesis to be true</u>.

This is accomplished by removing from the b matrix those variables assumed to be zero and finding a new regression equation for the reduced model. For this example, the new regression equation will include only the variable S since t_2, t_3, t_4, and t_5 are set to zero. The regression results for the reduced model are

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	1	171668.832	171668.832
RESIDUAL	24	122.164	5.090

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		[illegible]	)	
S	2	8.7559999+01	4.51-01	37459.6787

which show that

$$SS_{R(S)} = 191668.832 \qquad DF_{R(S)} = 1$$

5. Find the regression sum of squares and degrees of freedom associated with the factors tested in each hypothesis.

This is done by subtracting the regression quantities of the reduced model from the regression quantities of the full model. For this example, the sum of squares calculations are

$$SS_t = SS_{R(S,t)} - SS_{R(S)}$$

$$SS_t = 191767.857 - 191668.832$$

$$SS_t = 99.025$$

The degrees of freedom calculations are

$$DF_t = DF_{R(S,t)} - DF_{R(S)}$$

$$DF_t = 5 - 1$$

$$DF_t = 4$$

6. Form the analysis of variance table and make the appropriate F tests.

AOV

Source	DF	Sum of Squares	Mean Squares	F
Factor t	4	99.025	24.756	21.397
Error	20	23.140	1.157	
Total	25	191790.997		

The factor t is found to be significant at the 99% level of confidence.

EXAMPLE NO. 2

Type:

A single fixed-factor experiment with an incomplete block design.

Source:

Hicks (7), page 57.

Problem:

In this example, the factor t is set to four different levels and only three levels can be run in a block. There are four blocks of data as follows.

Treatment		t_1	t_2	t_3	t_4
Response	Block 1	2	-	20	7
	Block 2	-	32	14	3
	Block 3	4	13	31	-
	Block 4	0	23	-	11

Solution:

1. Express model in terms of a full-rank matrix equation.

The full-rank model as developed in Chapter III for this experiment is shown on the following page.

$$y_{ijk} = S + t_i + b_j + e_{ijk}$$

$$i = 1, 2, 3, 4$$

$$j = 1, 2, 3, 4$$

$$k = 1$$

$$t_1 = b_1 = 0$$

The matrix representation of the model is

$$\begin{vmatrix} 2. \\ 20. \\ 7. \\ 32. \\ 14. \\ 3. \\ 4. \\ 13. \\ 31. \\ 0. \\ 23. \\ 11. \end{vmatrix} = \begin{vmatrix} 1000000 \\ 1010000 \\ 1001000 \\ 1100100 \\ 1010100 \\ 1001100 \\ 1000010 \\ 1100010 \\ 1010010 \\ 1000001 \\ 1100001 \\ 1001001 \end{vmatrix} \begin{pmatrix} S \\ t_2 \\ t_3 \\ t_4 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} + (e)$$

2. <u>Find the total, error, and regression sums of squares and degrees of freedom.</u>

The regression results for the full model are

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	7	3114.833	444.976
RESIDUAL	5	363.167	72.633

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000	)	
S	2	[illegible]	[illegible]	[illegible]
T2	3	2.0749999+01	7.38+00	7.9139
T3	4	2.0249998+01	7.38+00	7.5276
T4	5	5.4999990+00	7.38+00	.5553
B2	6	-2.4999950-01	7.38+00	.0011
B3	7	1.2500002+00	7.38+00	.0287
B4	8	1.5000002+00	7.38+00	.0413

To find the total sum of squares and degrees of freedom

$$
\begin{array}{llll}
SS_{R(S,t,b)} = 3114.833 & & DF_{R(S,t,b)} = 7 \\
SS_E = \underline{363.167} & & DF_E = \underline{5} \\
SS_T = 3478.000 & & DF_T = 12
\end{array}
$$

3. Form the appropriate hypotheses to test the significance of the experimental factors.

For this example there are two fixed factors, t and b, to be investigated. Therefore, two hypotheses are formed. To test the significance of the factor t, the hypothesis is

$$H_0(t): \quad t_2 = t_3 = t_4 = 0$$

To test the significance of the blocks, b, the hypothesis is

$$H_0(b): \quad b_2 = b_3 = b_4 = 0$$

4. For each hypothesis find the regression sum of squares and the degrees of freedom for the reduced model that assumes the hypothesis to be true.

For $H_0(t)$ the reduced model contains the variables S, b_2, b_3, and b_4. The regression results for this model are shown on the following page.

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	4	2234.000	558.500
RESIDUAL	8	1244.000	155.500

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000	)	
S	2	9.6666669+00	7.20+00	1.8028
B2	6	6.6666663+00	1.02+01	.4287
B3	7	6.3333331+00	1.02+01	.3869
B4	8	1.6666664+00	1.02+01	.0268

which yield

$$SS_{R(S,b)} = 2234.000 \qquad DF_{R(S,b)} = 4$$

For $H_0(b)$ the reduced model contains the variables S, t_2, t_3, and t_4. The regression results for this model are

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	4	3108.667	777.167
RESIDUAL	8	369.333	46.167

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000	)	
S	2	2.0000007+00	3.92+00	.2599
T2	3	2.0666665+01	5.55+00	13.8773
T3	4	1.9666665+01	5.55+00	12.5668
T4	5	4.9999991+00	5.55+00	.8123

which yield

$$SS_{R(S,t)} = 3108.667 \qquad DF_{R(S,t)} = 4$$

5. <u>Find the regression sum of squares and degrees of freedom associated with the factors tested in each hypothesis</u>.

For the factor t, the calculations are

$$SS_t = SS_{R(S,t,b)} - SS_{R(S,b)}$$

$$SS_t = 3114.833 - 2234.000$$

$$SS_t = 880.833$$

$$DF_t = DF_{R(S,t,b)} - DF_{R(S,b)}$$

$$DF_t = 7 - 4$$

$$DF_t = 3$$

For the blocks b, the calculations are

$$SS_b = SS_{R(S,t,b)} - SS_{R(S,t)}$$

$$SS_b = 3114.833 - 3108.667$$

$$SS_b = 6.166$$

$$DF_b = DF_{R(S,t,b)} - DF_{R(S,t)}$$

$$DF_b = 7 - 4$$

$$DF_b = 3$$

6. <u>Form the analysis of variance table and make the appropriate F tests</u>.

AOV

Source	DF	Sum of Squares	Mean Squares	F
Factor t	3	880.833	293.611	4.042
Factor b	3	6.166	2.055	.028
Error	5	363.167	72.633	
Total	12	3478.000		

Neither factor is significant at the 95% level of confidence.

EXAMPLE NO. 3

Type:

A 2 x 2 fixed-effect factorial design with three replications per cell and two missing values.

Source:

Dixon (3), page 550.

Problem:

In this example, two factors, a and b, are each set to two different levels and three response measurements are made in each cell. Two measurements are missing. The data for the experiment is

Treatment	b_1	b_2
a_1	5 3 -	6 5 7
a_2	13 14 15	12 10 -

Solution:

1. Express the model in terms of a full-rank matrix equation.

The full-rank model for the experiment is

$$y_{ijk} = S + a_i + b_j + ab_{ij} + e_{ijk}$$

$i = 1, 2$

$j = 1, 2$

$k = 1, 2, \ldots, n_{ij}$ where $n_{11} = 2$

$n_{12} = 3$

$n_{21} = 3$

$n_{22} = 2$

$$a_1 = b_1 = ab_{11} = ab_{12} = ab_{21} = 0$$

The matrix representation of this model is

$$\begin{pmatrix} 5. \\ 3. \\ 13. \\ 14. \\ 15. \\ 6. \\ 5. \\ 7. \\ 12. \\ 10. \end{pmatrix} = \begin{pmatrix} 1000 \\ 1000 \\ 1100 \\ 1100 \\ 1100 \\ 1010 \\ 1010 \\ 1010 \\ 1111 \\ 1111 \end{pmatrix} \begin{pmatrix} S \\ a_2 \\ b_2 \\ ab_{22} \end{pmatrix} + (e)$$

2. Find the total, error, and regression sums of squares, and degrees of freedom.

The regression results for the full model are

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	4	970.000	242.500
RESIDUAL	6	8.000	1.333

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.000000	)	
S	2	4.000002+00	8.16-01	24.0000
A2	3	9.999994+00	1.05+00	89.9998
B2	4	1.999995+00	1.05+00	3.6000
AB22	5	-4.999993+00	1.49+00	11.2500

To find the total sum of squares and degrees of freedom

$$SS_{R(S,a,b,ab)} = 970.000 \qquad DF_{R(S,a,b,ab)} = 4$$

$$SS_E = \underline{\quad 8.000} \qquad DF_E = \underline{\ 6}$$

$$SS_T = 978.000 \qquad DF_T = 10$$

3. <u>Form the appropriate hypotheses to test the significance of the experimental factors</u>.

For this example there are three fixed factors, a, b, and ab to be investigated. Therefore, three hypotheses are formed. For the interaction effect, ab, the hypothesis is

$$H_0(ab): \quad ab_{22} = 0$$

For the factor a, the hypothesis is

$$H_0(a): \quad a_2 = 0$$

Notice that this hypothesis also implicitly states that the interaction effect is removed. Whenever a factor is removed from the model, it also implies that all higher level interaction terms containing that factor are removed.

For the factor b, the hypothesis is

$$H_0(b): \quad b_2 = 0$$

4. <u>For each hypothesis, find the regression sum of squares and degrees of freedom for the reduced model that assumes the hypothesis to be true</u>.

For $H_0(ab)$, the results are

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	3	955.000	318.333
RESIDUAL	7	23.000	3.286

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.000000	)	
S	2	5.4999993+00	1.07+00	26.3043
A2	3	7.5000003+00	1.17+00	41.0869
B2	4	-5.0000002-01	1.17+00	.1826

which yield

$$SS_{R(S,a,b)} = 955.000 \qquad DF_{R(S,a,b)} = 3$$

For $H_0(a)$, the results are

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	2	820.000	410.000
RESIDUAL	8	158.000	19.750

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000)		
S	2	9.9999999+00	1.99+00	25.3165
B2	4	-1.9999999+00	2.81+00	.5063

which yield

$$SS_{R(S,b)} = 820.000 \qquad DF_{R(S,b)} = 2$$

For $H_0(b)$, the results are

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	2	954.400	477.200
RESIDUAL	8	23.600	2.950

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000)		
S	2	5.1999999+00	7.68-01	45.6305
A2	3	7.5999999+00	1.09+00	46.9491

which yield

$$SS_{R(S,a)} = 954.000 \qquad DF_{R(S,a)} = 2$$

5. <u>Find the regression sum of squares and degrees of freedom associated with the factors tested in each hypothesis</u>.

For the interaction effect, ab, the calculations are

$$SS_{ab} = SS_{R(S,a,b,ab)} - SS_{R(S,a,b)}$$
$$SS_{ab} = 970.0 - 955.0$$
$$SS_{ab} = 15.0$$
$$DF_{ab} = DF_{R(S,a,b,ab)} - DF_{R(S,a,b)}$$
$$DF_{ab} = 4 - 3$$
$$DF_{ab} = 1$$

For the factor a, the calculations are

$$SS_{a} = SS_{R(S,a,b,ab)} - SS_{R(S,b)} - SS_{ab}$$
$$SS_{a} = 970.0 - 820.0 - 15.0$$
$$SS_{a} = 135.0$$

Notice that the difference in the sum of squares between the full and reduced models yields the sum of squares due to factor a plus the sum of squares due to the interaction effect, ab. This is caused by the fact that the hypothesis $H_0(a)$ implicitly includes the assumption that all interaction effects with the factor a are also removed from the model.

For the factor b, the calculations are

$$SS_b = SS_{R(S,a,b,ab)} - SS_{R(S,a)} - SS_{ab}$$

$$SS_b = 970.0 - 954.4 - 15.0$$

$$SS_b = 0.6$$

$$DF_b = DF_{R(S,a,b,ab)} - DF_{R(S,a)} - DF_{ab}$$

$$DF_b = 4 - 2 - 1$$

$$DF_b = 1$$

6. <u>Form the analysis of variance table and make the appropriate F tests</u>.

AOV

Source	DF	Sum of Squares	Mean Squares	F
Factor a	1	135.00	135.00	101.25
Factor b	1	0.60	0.60	.45
Factor ab	1	15.00	15.00	11.25
Error	6	8.00	1.33	
Total	10	970.00		

. The factors a and ab are significant at the 99% level of confidence.

EXAMPLE NO. 4

Type:

A three-factor, nested-factorial experiment with fixed and random effects.

Source:

Hicks (7), page 172.

Problem:

In this experiment, three factors, methods (m), groups (g), and teams (t) are investigated to find their effect on the number of rounds of ammunition per minute that can be loaded into a gun. The factors m and g are fixed and t is a random factor which is nested within g. The data for the experiment is:

Groups (g)		1			2			3		
	Teams (t)	1	2	3	4	5	6	7	8	
Methods (m)	1	20.2	26.2	23.8	22.0	22.6	22.9	23.1	22.9	21.8
		24.1	26.9	24.9	23.5	24.6	25.0	22.9	23.7	23.5
	2	14.2	18.0	12.5	14.1	14.0	13.7	14.1	12.2	12.7
		16.2	19.1	15.4	16.1	18.1	16.0	16.1	13.8	15.1

Solution:

1. Express model in terms of a full-rank matrix equation.

The model for the experiment is written as a full-factorial model. After the sum of squares are determined, some of the interactions will be combined to account for the fact that the t is nested within g.

The full model is

$$y_{ijkl} = S + m_i + g_j + t_k + mg_{ij} + mt_{ik} + gt_{jk} + mgt_{ijk} + e_{ijkl}$$

$i = 1, 2$

$j = 1, 2, 3$

$k = 1, 2, 3$

$l = 1, 2$

$m_1 = g_1 = t_1 = 0$

$mg_{ij} = mt_{ik} = gt_{jk} = mgt_{ijk} = 0$ when $i = 1$ or $j = 1$ or $k = 1$

The matrix representation of the model is shown on the following page.

$$
\left|\begin{array}{c}
20.2\\ 24.1\\ 26.2\\ 26.9\\ 23.8\\ 24.9\\ 14.2\\ 16.2\\ 18.0\\ 19.1\\ 12.5\\ 15.4\\ 22.0\\ 23.5\\ 22.6\\ 24.6\\ 22.9\\ 25.0\\ 14.1\\ 16.1\\ 14.0\\ 18.1\\ 13.7\\ 16.0\\ 23.1\\ 22.9\\ 22.9\\ 23.7\\ 21.8\\ 23.5\\ 14.1\\ 16.1\\ 12.2\\ 13.8\\ 12.7\\ 15.1
\end{array}\right|
=
\left|\begin{array}{c}
100000000000000000\\
100000000000000000\\
110000000000000000\\
110000000000000000\\
101000000000000000\\
101000000000000000\\
100100000000000000\\
100100000000000000\\
110100100000000000\\
110100100000000000\\
101100010000000000\\
101100010000000000\\
100010000000000000\\
100010000000000000\\
110010000010000000\\
110010000010000000\\
101010000000100000\\
101010000000100000\\
100110001000000000\\
100110001000000000\\
110110101010001000\\
110110101010001000\\
101110011001000100\\
101110011001000100\\
100001000000000000\\
100001000000000000\\
110001000000100000\\
110001000000100000\\
101001000000010000\\
101001000000010000\\
100101000100000000\\
100101000100000000\\
110101100100100010\\
110101100100100010\\
101101010100010001\\
101101010100010001
\end{array}\right|
\left|\begin{array}{c}
S\\ t_2\\ t_3\\ m_2\\ g_2\\ g_3\\ mt_{22}\\ mt_{23}\\ mg_{22}\\ mg_{23}\\ gt_{22}\\ gt_{23}\\ gt_{32}\\ gt_{33}\\ mgt_{222}\\ mgt_{223}\\ mgt_{232}\\ mgt_{233}
\end{array}\right|
+ (e)
$$

2. <u>Find the total, error, and regression sums of squares and degrees of freedom.</u>

The regression output for the full model is

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	18	14175.167	787.509
RESIDUAL	18	41.591	2.311

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000	)	
S	2	2.2149967+01	1.07+00	424.6674
T2	3	4.4000165+00	1.52+00	6.3788
T3	4	2.2000130+00	1.52+00	2.0947
M2	5	-6.9499941+00	1.52+00	2[illegible].9046
G2	6	6.0001171-01	1.52+00	.1558
G3	7	[illegible]773-01	1.52+00	.3127
MT22	8	-1.0500031+00	2.15+00	.2386
MT23	9	-3.45[illegible]056+00	2.15+00	2.5756
MG22	10	-7.[illegible]11-01	2.15+00	.1060
MG23	11	-9.5000521-01	2.15+00	.1953
GT22	12	-3.5500091+00	2.15+00	2.7271
GT23	13	-1.0000125+00	2.15+00	.2164
GT32	14	-4.1[illegible]0060+00	2.15+00	3.6376
GT33	15	-2.5500097+00	2.15+00	1.4071
MGT222	16	1.15[illegible]010+00	3.04+00	.1431
MGT223	17	2.0000044+00	3.04+00	.4328
MGT232	18	-1.350000[illegible]+00	3.04+00	.1972
MGT233	19	2.[illegible]02+00	3.04+00	.7314

which shows that

$$SS_{R(C,m,g,t,mg,mt,gt,mgt)} = 14175.167$$

$$SS_E = \underline{\quad 41.591}$$

$$SS_T = 14216.758$$

and

$$DF_{R(S,m,g,t,mg,mt,gt,mgt)} = 18$$
$$DF_E = \underline{18}$$
$$DF_T = 36$$

3. <u>Form the appropriate hypotheses to test the significance of the experimental factors</u>.

Still considering the problem as a full-crossed factorial model, the following hypotheses are used to test the significance of the three factors and their interactions.

The mgt interaction includes the random factor t, so

$$H_0(mgt):\quad \sigma^2_{mgt} = 0$$

This random-factor hypothesis is different from a fixed-factor hypothesis but since the random factor is assumed to be $N(0,\sigma^2_{mgt})$ the reduced model is still developed by setting all the terms containing mgt to zero.

The mg interaction term contains only fixed effects so

$$H_0(mg):\quad mg_{22} = mg_{23} = 0$$

The mt and gt terms include the random factor t, so

$$H_0(mt):\quad \sigma^2_{mt} = 0$$
$$H_0(gt):\quad \sigma^2_{gt} = 0$$

The m and g factors are fixed so

$$H_0(m):\quad m_2 = 0$$
$$H_0(g):\quad g_2 = g_3 = 0$$

The t factor is random so

$$H_0(t): \quad \sigma^2_t = 0$$

4. <u>For each hypothesis, find the regression sum of squares and degrees of freedom for the reduced model that assumes the hypothesis to be true.</u>

For $H_0(mgt)$ the result is

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQ.
REGRESSION	14	14170.006	1012.143
RESIDUAL	22	46.752	2.125

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.000000	)	
S	2	2.2394432+01	9.09-01	606.8409
T2	3	4.4333300+00	1.19+00	13.8731
T3	4	1.4333452+00	1.19+00	1.4532
M2	5	-7.4368834+00	1.09+00	46.8715
G2	6	7.5010567-02	1.19+00	.0040
G3	7	6.4167592-01	1.19+00	.2936
MT22	8	-1.1106696+00	1.19+00	.8632
MT23	9	-1.9156700+00	1.19+00	2.5930
MG22	10	3.4999751-01	1.19+00	.0865
MG23	11	-5.3333563-01	1.19+00	.2008
GT22	12	-2.9750069+00	1.46+00	4.1648
GT23	13	-9.7338196-06	1.46+00	.0000
GT32	14	-4.7750066+00	1.46+00	10.7292
GT33	15	-1.2500001+00	1.46+00	.7353

which shows that

$$SS_{R(S,m,g,t,mg,mt,gt)} = 14170.006$$

$$DF_{R(S,m,g,t,mg,mt,gt)} = 14$$

For $H_0(mg)$ the result is

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	12	14168.819	1180.735
RESIDUAL	24	47.939	1.997

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		[illegible]	)	
S	2	2.2424950+01	8.16-01	755.2727
T2	3	4.4333437+00	1.15+00	14.7595
T3	4	1.4333452+00	1.15+00	1.5428
M2	5	-7.4999961+00	8.16-01	84.4615
G2	6	2.5000964-01	9.99-01	.0626
G3	7	3.7500007-01	9.99-01	.1408
MT22	8	-1.1106096+00	1.15+00	.9364
MT23	9	-1.9166750+00	1.15+00	2.7587
GT22	12	-2.9750065+00	1.41+00	4.4309
GT23	13	-9.7234092-06	1.41+00	.0000
GT32	14	-4.7750066+00	1.41+00	11.4147
GT33	15	-1.2500081+00	1.41+00	.7022

which shows that

$$SS_{R(S,m,g,t,mt,gt)} = 14168.819$$

$$DF_{R(S,m,g,t,mt,gt)} = 12$$

For $H_0(mt)$ the result is

ANALYSIS OF VARIANCE

	DF.	SUM OF SQUARES	MEAN SQUARE
REGRESSION	12	14164.446	1180.370
RESIDUAL	24	52.313	2.180

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT)		0.0000000	)	
S	2	2.2899968+01	3.52-01	721.7635
T2	3	3.0750091+00	1.04+00	13.7778
T3	4	4.7591017-01	1.04+00	.2079
M2	5	-6.4499762+00	6.52-01	98.2737
G2	6	7.5011018-02	1.21+00	.0039
G3	7	6.4167065-01	1.21+00	.2834
MG22	10	3.4999733-01	1.21+00	.0843
MG23	11	-5.3333582-01	1.21+00	.1957
GT22	12	-2.9750065+00	1.48+00	4.0605
GT23	13	-9.7392254-06	1.48+00	.0000
GT32	14	-4.7750067+00	1.48+00	10.4605
GT33	15	-1.2500081+00	1.48+00	.7169

which shows that

$$SS_{R(S,m,g,t,mg,gt)} = 14164.446$$

$$DF_{R(S,m,g,t,mg,gt)} = 12$$

For $H_0(gt)$ the result is

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	10	14143.520	1414.352
RESIDUAL	26	73.239	2.817

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		[illegible]	)	
S	2	2.3394435+01	8.85-01	697.4536
T2	3	1.8500050+00	9.69-01	3.6450
T3	4	1.0166720+00	9.69-01	1.1006
M2	5	-7.4382034+00	1.25+00	35.3006
G2	6	-9.1666179-01	9.69-01	.8949
G3	7	-1.3666623+00	9.69-01	1.9892
MT22	8	-1.1166696+00	1.37+00	.6640
MT23	9	-1.9166701+00	1.37+00	1.9562
MG22	10	3.4999747-01	1.37+00	.0652
MG23	11	-5.3333560-01	1.37+00	.1515

which show that

$$SS_{R(S,m,g,t,mg,mt)} = 14143.520$$

$$DF_{R(S,m,g,t,mg,mt)} = 10$$

For $H_0(m)$ the result is

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	9	13511.308	1501.256
RESIDUAL	27	705.451	26.128

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000	)	
S	2	1.8674991+01	2.56+00	53.3722
T2	3	3.8750061+00	3.61+00	1.1494
T3	4	4.7500920-01	3.61+00	.0173
G2	6	2.5000871-01	3.61+00	.0048
G3	7	3.7500717-01	3.61+00	.0108
GT22	12	-2.9750074+00	5.11+00	.3387
GT23	13	-8.6337768-06	5.11+00	.0000
GT32	14	-4.7750056+00	5.11+00	.8727
GT33	15	-1.2500037+00	5.11+00	.0598

which shows that

$$SS_{R(S,g,t,gt)} = 13511.308$$

$$DF_{R(S,g,t,gt)} = 9$$

For $H_0(g)$ the result is

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	6	14126.281	2354.380
RESIDUAL	30	90.478	3.016

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.0000000	)	
S	2	2.2633327+01	7.09-01	1019.1249
T2	3	1.8500054+00	1.00+00	3.4044
T3	4	1.0166726+00	1.00+00	1.0282
M2	5	-7.4999761+00	1.00+00	55.9529
MT22	8	-1.1166695+00	1.42+00	.6202
MT23	9	-1.9166700+00	1.42+00	1.8271

which shows that

$$SS_{R(S,m,t,mt)} = 14126.281$$

$$DF_{R(S,m,t,mt)} = 6$$

And finally, for $H_0(t)$, the result is

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	6	14125.187	2354.198
RESIDUAL	30	91.571	3.052

VARIABLES IN EQUATION

VARIABLE		COEFFICIENT	STD. ERROR	F TO REMOVE
(CONSTANT		0.000000	)	
S	2	2.4349794+01	7.13-01	1165.4970
M2	5	-8.4499762+00	1.01+00	70.1773
G2	6	-9.1666165-01	1.01+00	.8259
G3	7	-1.3666521+00	1.01+00	1.8357
MG22	10	3.4999726-01	1.43+00	.0602
MG23	11	-5.3333569-01	1.43+00	.1396

which shows that

$$SS_{R(S,m,g,mg)} = 14125.187$$

$$DF_{R(S,m,g,mg)} = 6$$

5. Find the regression sum of squares and degrees of freedom associated with the factors tested in each hypothesis.

For mgt

$SS_{mgt} = SS_{R(S,m,g,t,mg,mt,gt,mgt)} - SS_{R(S,m,g,t,mg,mt,gt}$

$SS_{mgt} = 14175.167 - 14170.006$

$SS_{mgt} = 5.161$

$DF_{mgt} = DF_{R(S,m,g,t,mg,mt,gt,mgt)} - DF_{R(S,m,g,t,mg,mt,gt}$

$DF_{mgt} = 18 - 14$

$DF_{mgt} = 4$

For mg

$SS_{mg} = SS_{R(S,m,g,t,mg,mt,gt,mgt)} - SS_{R(S,m,g,t,m}$

$- SS_{mgt}$

$SS_{mg} = 14175.167 - 14168.819 - 5.161$

$SS_{mg} = 1.187$

$DF_{mg} = DF_{R(S,m,g,t,mg,mt,gt,mgt)} - DF_{R(S,m,g,t,mt,}$

$- DF_{mgt}$

$DF_{mg} = 18 - 12 - 4$

$DF_{mg} = 2$

Similar calculations for mt and gt yield

$$SS_{mt} = 5.560$$

$$DF_{mt} = 2$$

$$SS_{gt} = 26.486$$

$$DF_{gt} = 4$$

For m

$$SS_m = SS_{R(S,m,g,t,mg,mt,gt,mgt)} - SS_{R(S,g,t,gt)} - SS_{mgt} - SS_{mg} - SS_{mt}$$

$$SS_m = 14175.167 - 13511.308 - 5.161 - 1.187 - 5.560$$

$$SS_m = 651.951$$

$$DF_m = DF_{R(S,m,g,t,mg,mt,gt,mgt)} - DF_{R(S),g,t,gt)} - DF_{mgt} - DF_{mg} - DF_{mt}$$

$$DF_m = 18 - 10 - 4 - 2 - 2$$

$$DF_m = 1$$

Similar calculations for g and t yield

$$SS_g = 16.052$$

$$DF_g = 2$$

$$SS_t = 12.773$$

$$DF_t = 2$$

Up to this point the problem has been treated as a fully-crossed factorial experiment. To correct for the fact that t is nested within g, the following terms are adjusted to include the interaction terms.

For the factor t

$$SS_{tk(j)} = SS_t + SS_{gt}$$

$$SS_{tk(j)} = 12.773 + 26.486$$

$$SS_{tk(j)} = 39.259$$

$$DF_{tk(j)} = DF_t + DF_{gt}$$

$$DF_{tk(j)} = 2 + 4$$

$$DF_{tk(j)} = 6$$

For the factor mt

$$SS_{mt_{ik(j)}} = SS_{mt} + SS_{mgt}$$

$$SS_{mt_{ik(j)}} = 5.560 + 5.161$$

$$SS_{mt_{ik(j)}} = 10.721$$

$$DF_{mt_{ik(j)}} = DF_{mt} + DF_{mgt}$$

$$DF_{mt_{ik(j)}} = 2 + 4$$

$$DF_{mt_{ik(j)}} = 6$$

6. <u>Form the analysis of variance table and make the appropriate F tests</u>.

When the analysis of variance table for this problem is formed it will include an expected mean squares (EMS) column. Since this problem has both fixed and random factors, the appropriate F tests are determined from the EMS quantities.

AOV

Source	DF	Sum of Squares	Mean Squares	EMS	F
m_i	1	651.951	651.951	$\sigma_e^2 + 2\sigma_{mt}^2 + 18\sigma_m^2$	364.830
g_j	2	16.052	8.026	$\sigma_e^2 + 4\sigma_t^2 + 12\sigma_g^2$	1.227
$t_{k(j)}$	6	39.259	6.543	$\sigma_e^2 + 4\sigma_t^2$	2.831
mg_{ij}	2	1.187	0.594	$\sigma_e^2 + 2\sigma_{mt}^2 + 6\sigma_{mg}^2$	0.332
$mt_{ik(j)}$	6	10.721	1.787	$\sigma_e^2 + 2\sigma_{mt}^2$	0.775
Error	18	41.591	2.311	σ_e^2	
Total	36	14216.758			

The factor m is significant at the 99% level of confidence.

The four preceeding examples demonstrate that widely different types of problems can be solved by the one consolidated method of regression analysis of full-rank models. The fact that the solution method is the same, regardless of the orthogonality of the problem, has both advantages and disadvantages. For nonorthogonal problems, it is a great advantage since the experimenter need only have one regression routine to solve any type of analysis of variance problems. However, if the problem to be solved is orthogonal it can usually be solved in a short time with only a desk calculator by the traditional sum of squares method. Therefore, the main benefits of the full-rank model and regression technique are realized when solving nonorthogonal problems.

Most of the work involved in using a standard regression package for experimental design problems is concerned with the following four items.

1. Writing the full-rank X matrix for the model.

2. Generating the commands to include or delete variables for the regression calculations.

3. The addition and subtraction of regression quantities to find the sums of squares and degrees of freedom associated with the experimental factors.

4. The division required to compute the mean squares and F ratios to complete the analysis of variance table.

To demonstrate how a regression routine might be modified to more efficiently handle analysis of variance problems, the program, ANOVA, was written. It consists of a regression routine with a front end that converts traditional experimental design data to the full-rank form and a back end that outputs an analysis of variance table. ANOVA is described in Appendix B where the four example problems of this chapter are solved with the ANOVA routine to demonstrate how it simplifies the regression procedure.

CHAPTER V

CONCLUSIONS

The advantages of the full-rank model formulation and regression analysis of experimental design problems are as follows.

1. The approach is completely general since any design model, regardless of orthogonality, can be written as a full-rank model and solved by regression analysis.

2. The full-rank model is easily formulated since the terms of the model have physical significance to the experimenter.

3. The method eliminates the task of reparameterization since the full-rank model always leads to a system of normal equations that have a unique solution.

4. The analyst needs only one computer program, a regression routine, for all his analysis of variance work.

5. Regression analysis codes are available at almost all computing facilities.

The disadvantages of the technique are as follows.

1. Orthogonal problems are more easily solved using a desk calculator and the traditional sum of squares method.

2. The number of variables that a regression code can handle may limit the number of factors that can be tested for their effect on the response variable.

3. The standard regression codes leave the analyst with several computations to make, manually or with another computer run, prior to the construction of an analysis of variance table.

4. The regression calculations cannot be done manually except for small problems that could be easily handled by the traditional methods.

The first disadvantage leads to the conclusion that the regression technique is profitable in terms of time and effort only for nonorthogonal problems. The second and third disadvantages could be overcome by a specialized computer code such as ANOVA, to facilitate the solution of analysis of variance problems. The fourth disadvantage is lessened by the fact that the analyst should use regression only for nonorthogonal problems which are difficult to solve manually by any method.

In summary it appears that regression analysis is well known to be a powerful and general solution method for experimental design problems but its application has been retarded by the additional work of preparing the problem for the regression calculations. The full-rank formulation of experimental design models eliminates this task and makes regression a much more desireable solution method for analysis of variance work.

BIBLIOGRAPHY

1. Cashler, F. L. Linear Programming Applied to the Analysis of Variance of Designed Experiments. Unpublished Masters Thesis, University of Houston, 1970.

2. Cooley, W. W. and Lohnes, P. R. Multivariate Procedures for the Behavioral Sciences. New York: John Wiley & Sons, 1962.

3. Dixon, W. J. BMD Biomedical Computer Programs. Berkeley: University of California Press, 1970.

4. Draper, N. R. and Smith, H. Applied Regression Analysis. New York: John Wiley & Sons, 1966.

5. Ferguson, G. A. Statistical Analysis in Psychology and Education. New York: McGraw-Hill, 1966.

6. Graybill, F. A. An Introduction to Linear Statistical Models. New York: McGraw-Hill, 1961.

7. Hicks, C. R. Fundamental Concepts in the Design of Experiments. New York: Holt, Rinehart, and Winston, 1964.

APPENDICES

APPENDIX A

BMDO2R Input Data For Examples of Chapter IV

The following pages show the listings of the BMDO2R input cards for the examples in Chapter IV.

```
PROBLM   UNEQAL   25    6                       2    5                 YES
LABELS   2S        3T2        4T3        5T4        6T5
(F10.4,70F1.0)
     83.    10000
     85.    10000
     84.    11000
     85.    11000
     85.    11000
     86.    11000
     86.    11000
     87.    11000
     86.    10100
     87.    10100
     87.    10100
     87.    10100
     88.    10100
     88.    10100
     88.    10100
     88.    10100
     88.    10100
     89.    10100
     90.    10100
     89.    10010
     90.    10010
     90.    10010
     91.    10010
     90.    10001
     92.    10001
SUBPRO    1                                              YES
CONDEL 33333
SUBPRO    1                                              YES
CONDEL 31111
FINISH
```

Input for Example 1

```
PROBLM    IBLOCK    12     8                       3     7                  YES
LABELS    2S        3T2         4T3         5T4         6B2         7B3         8E
(F10.4,70F1.0)
     2.      1000000
    20.      1010000
     7.      1001000
    32.      1100100
    14.      1010100
     3.      1001100
     4.      1000010
    13.      1100010
    31.      1010010
     0.      1000001
    23.      1100001
    11.      1001001
SUBPRO    1                                               YES
CONDEL 3333333
SUBPRO    1                                               YES
CONDEL 3333111
SUBPRO    1                                               YES
CONDEL 3111333
FINISH
```

Input for Example 2

```
PROBLM    2X2MVL    10      5                        4    4                YES
LABELS    2S         3A2          4B2        5AB22
(F10.4,70F1.0)
      5.    1000
      3.    1000
     13.    1100
     14.    1100
     15.    1100
      6.    1010
      5.    1010
      7.    1010
     12.    1111
     10.    1111
SUBPRO    1                                               YES
CONDEL 3333
SUBPRO    1                                               YES
CONDEL 3331
SUBPRO    1                                               YES
CONDEL 3311
SUBPRO    1                                               YES
CONDEL 3131
FINISH
```

Input for Example 3

```
PROBLM    NESFAC    36.  19                          8  18               YES      1
LABELS    2S        3T2        4T3        5M2        6G2        7G3        8MT
LABELS    9MT23     10MG22     11MG23     12GT22     13GT23     14GT32     15GT
LABELS    16MGT222  17MGT223   18MGT232   19MGT233
(F10.4,70F1.0)
     20.2    1000000000000000000
     24.1    1000000000000000000
     26.2    1100000000000000000
     26.9    1100000000000000000
     23.8    1010000000000000000
     24.9    1010000000000000000
     14.2    1001000000000000000
     16.2    1001000000000000000
     18.0    1101001000000000000
     19.1    1101001000000000000
     12.5    1011000100000000000
     15.4    1011000100000000000
     22.0    1000100000000000000
     23.5    1000100000000000000
     22.6    1100100000100000000
     24.6    1100100000100000000
     22.9    1010100000010000000
     25.0    1010100000010000000
     14.1    1001100010000000000
     16.1    1001100010000000000
     14.0    1101101010100001000
     18.1    1101101010100001000
     13.7    1011100110010000100
     16.0    1011100110010000100
     23.1    1000010000000000000
     22.9    1000010000000000000
     22.9    1100010000000100000
     23.7    1100010000000100000
     21.8    1010010000000010000
     23.5    1010010000000010000
     14.1    1001010001000000000
     16.1    1001010001000000000
     12.2    1101011001001000010
     13.8    1101011001001000010
     12.7    1011010101000010001
     15.1    1011010101000010001
SUBPRO    1                                            YES
CONDEL 3333333333333333333
SUBPRO    1                                            YES
CONDEL 3333333333333331111
SUBPRO    1                                            YES
CONDEL 3333333333111111111
SUBPRO    1                                            YES
CONDEL 3333331133333331111
SUBPRO    1                                            YES
CONDEL 3333333311333331111
SUBPRO    1                                            YES
CONDEL 3113331133111111111
SUBPRO    1                                            YES
CONDEL 3331331111333331111
SUBPRO    1                                            YES
CONDEL 3333113311111111111
FINISH
```

Input for Example 4

APPENDIX B

ANOVA Description

The routine ANOVA was written to demonstrate how the analysis of variance calculations might be performed automatically as part of a specialized regression routine. The ANOVA user provides as input the following:

1. Number of factors.

2. Number of observations.

3. Number and identification of factors that are blocks and have no interaction with other factors.

4. Data for the problem consisting of a response measurement and the levels of the factors associated with the response.

The program then does the following:

1. Builds a full-rank model of the experiment as described in Chapter III.

2. Finds the total, error, and regression sums of squares and degrees of freedom for the full model.

3. Forms a full-rank reduced model for each possible factor to be tested (up to three-level interactions).

4. Finds the regression sum of squares and degrees of freedom for each reduced model.

5. Finds the sum of squares and degrees of freedom associated with each factor.

6. Computes and outputs an analysis of variance table.

There are several limitations to the program that could be eliminated by additional programming effort.

First of all, the program is limited to 150 observations and a combined total of 100 single, two-level interaction and three-level interaction terms. It is reasonable to assume that this problem could be overcome by transfers between core storage and disk or drum storage units for the manipulation of larger matrices.

Secondly, the analysis of variance table generated by ANOVA assumes that all factors are fixed. Therefore, the last column of the table provides the F ratio between the mean squares of the factor and the error mean squares. To be complete, ANOVA should include an algorithm that computes the correct F ratio for fixed or random factors.

The third limitation is that the program treats all problems as fully-crossed factorial problems. Therefore, for problems with nested factors, some of the sums of squares and degrees of freedom must be manually combined to obtain the proper results for nested terms. An algorithm to combine the appropriate interaction terms prior to the printing of the analysis of variance table should be included in a program of this type.

In spite of the previously described shortcomings, the program appears to be a useful tool for analysis of variance problems, especially ones with nonorthogonal designs.

The following pages contain a listing of ANOVA and its subroutine HYPOTH, the input data for the examples in Chapter IV, and the ANOVA results for the examples in Chapter IV. The results agree with the BMDO2R solutions except for Example Number 4. ANOVA treated it as a fully-crossed, fixed-effect, factorial design, so the sums of squares and degrees of freedom must be appropriately combined to account for the nested factor, t. Once these quantities are computed, the analysis of variance table would have to be manually completed.

```
      FOR ANOVA
        COMMON Y(150,1),X(150,100),XH(150,100),XTX(100,100),BT(1,100),
       1LEVEL(10),NLEVEL(10),NCOL(100),SS(7,7,7),NDF(7,7,7),ICOL(100),
       1IBLOCK(6),ICHK(100)
        NPROB=1
C
C           READ SIZE OF PROBLEM
C
      1 READ(5,100,END=47 )NFAC,NOBS,NBLOCK
    100 FORMAT(3I2)
        IF(NBLOCK .EQ. 0)GO TO 2
        READ 101,(IBLOCK(I),I=1,NBLOCK)
    101 FORMAT(40I2)
      2 NFAC1=NFAC-1
        NFAC2=NFAC-2
        N=1
        M=2
        DO 3 I=1,NFAC
      3 NLEVEL(I)=2
        DO 4 I=1,NOBS
      4 X(I,1)=1.
C
C           READ DATA
C
      5 READ(5,102)Y(N,1),(LEVEL(J),J=1,NFAC)
    102 FORMAT(F10.0,35I2)
C
C           ENTER SINGLE-FACTOR TERMS INTO X AND B MATRICES
C
        DO 8 I=1,NFAC
        IF (LEVEL(I) .GT. NLEVEL(I))NLEVEL(I)=LEVEL(I)
        IF(LEVEL(I) .EQ. 1)GO TO 8
        ICOL(M)=I*1000 +LEVEL(I)*100
        M1=M-1
        DO 6 ICHK=1,M1
      6 IF(ICOL(M) .EQ. ICOL(ICHK+1))GO TO 7
        X(N,M)=1.
        M=M+1
        GO TO 8
      7 X(N,ICHK    )=1.
      8 CONTINUE
        IF((NFAC .EQ. 1).OR.(NBLOCK .EQ. NFAC1))GO TO 19
C
C           ENTER TWO-FACTOR INTERACTION TERMS INTO X AND B MATRICES
C
        DO 13 I=1,NFAC1
        IJ=I+1
        DO 13 J=IJ,NFAC
        IF(NBLOCK .EQ. 0)GO TO 10
        DO 9 IB=1,NBLOCK
      9 IF((I.EQ.IBLOCK(IB)).OR.(J.EQ.IBLOCK(IB)))GO TO 13
     10 IF((LEVEL(I) .EQ. 1).OR.(LEVEL(J) .EQ. 1))GO TO 13
        ICOL(M)=I*100000+J*10000+LEVEL(I)*100+LEVEL(J)*10
        DO 11 ICHK=3,M
     11 IF(ICOL(M) .EQ. ICOL(ICHK-1))GO TO 12
        X(N,M)=1.
        M=M+1
        GO TO 13
     12 X(N,ICHK-1)=1.
```

```
   13 CONTINUE
      IF(NFAC .EQ. 2)GO TO 19
C
C         ENTER THREE-FACTOR INTERACTION TERMS INTO X AND B MATRICES
C
      DO 18 I=1,NFAC2
      IJ=I+1
      DO 18 J=IJ,NFAC1
      JK=J+1
      DO 18 K=JK,NFAC
      IF(NBLOCK .EQ. 0)GO TO 15
      DO 14 IB=1,NBLOCK
   14 IF((I.EQ.IBLOCK(IB)).OR.(J.EQ.IBLOCK(IB)).OR.(K.EQ.IBLOCK(IB)))
     1GO TO 18
   15 IF((LEVEL(I).EQ.1).OR.(LEVEL(J).EQ.1).OR.(LEVEL(K).EQ.1))GO TO 1
      ICOL(M)=I*1[illegible]+J*10000+K*1000+LEVEL(I)*100+LEVEL(J)*10+LEVEL(K
      DO 16 ICHK=3,M
   16 IF(ICOL(M).EQ.ICOL(ICHK-1))GO TO 17
      X(N,M)=1.
      M=M+1
      GO TO 18
   17 X(N,ICHK-1)=1.
   18 CONTINUE
   19 CONTINUE
C
C         DETERMINE IF X AND B MATRICES ARE COMPLETE
C
      M1=M-1
      IF(N .EQ. NOBS)GO TO 20
      N=N+1
      GO TO 5
C
C         FIND TOTAL SUM OF SQUARES
C
   20 SSTOT=0.
      M=M-1
      V=4.
      DO 21 I=1,N
   21 SSTOT=SSTOT+Y(I,1)**2
C
C         FORM NORMAL EQUATIONS
C
      DO 22 I=1,M
      DO 22 J=1,M
      XTX(I,J)=0.
      DO 22 IROW=1,N
   22 XTX(I,J)=XTX(I,J)+X(IROW,I)*X(IROW,J)
      DO 23 I=1,M
      XTX(I,M+1)=0.
      DO 23 J=1,N
   23 XTX(I,M+1)=XTX(I,M+1)+X(J,I)*Y(J,1)
C
C         SOLVE NORMAL EQUATIONS
C
      [illegible] (XTX,1[illegible],1[illegible],M,M+1[illegible],JC,V)
      D[illegible] 24 I=1,M
   24 BT(1,I)=XTX(I,M+1)
C
C         FIND REGRESSION SUM OF SQUARES
```

```
C
      SSREG=0.
      DO 26 I=1,M
      XTY=0.
      DO 25 J=1,N
   25 XTY=XTY+X(J,I)*Y(J,1)
   26 SSREG=SSREG+BT(1,I)*XTY
C
C        FIND TOTAL, REGRESSION AND ERROR DEGREES OF FREEDOM
C
      IDFTOT=N
      IDFREG=M
      IDFERR=N-M
C
C        FIND ERROR SUM OF SQUARES AND MEAN SQUARES
C
      SSERR=SSTOT-SSREG
      VARERR=SSERR/FLOAT(N-M)
C
C        DETERMINE IF MODEL HAS THREE-LEVEL INTERACTION TERMS
C
      IF(NFAC .EQ. 1)GO TO 35
      IF(NFAC .EQ. 2)GO TO 29
      IF(NBLOCK .EQ. NFAC2)GO TO 29
C
C        FIND SUMS OF SQUARES AND DEGREES OF FREEDOM FOR THREE-LEVEL
C        INTERACTION TERMS
C
      DO 28 I=1,NFAC2
      IJ=I+1
      DO 28 J=IJ,NFAC1
      JK=J+1
      DO 28 K=JK,NFAC
      ITEST=I*100+J*10+K
      NDELET=0
      DO 27 IM=2,M
      MTEST=ICOL(IM)/1000
      IF(ITEST .NE. MTEST)GO TO 27
      NDELET=NDELET+1
      NCOL(NDELET)=IM
   27 CONTINUE
      CALL HYPOTH(NDELET,IDELET,M,N,SSHYPO)
      SS(I,J,K)=SSREG-SSHYPO
      NDF(I,J,K)=NDELET
   28 CONTINUE
C
C        DETERMINE IF MODEL HAS TWO-LEVEL INTERACTION TERMS
C
   29 IF(NBLOCK .EQ. NFAC1)GO TO 35
C
C        FIND SUMS OF SQUARES AND DEGREES OF FREEDOM FOR TWO-LEVEL
C        INTERACTION TERMS
C
      DO 34 I=1,NFAC1
      IJ=I+1
      DO 34 J=IJ,NFAC
      NDELET=0
      IDELET=0
      SSCOR=0.
```

```
      NCOR=0
      DO 33 IM=2,M
      IFAC=ICOL(IM)/100000
      JFAC=ICOL(IM)/10000-(ICOL(IM)/100000)*10
      KFAC=ICOL(IM)/1000-(ICOL(IM)/10000)*10
      IF((I .NE. IFAC).AND.(I .NE. JFAC))GO TO 33
      IF((J .NE. JFAC).AND.(J .NE. KFAC))GO TO 33
      IF(KFAC .EQ. 0)KFAC=1
      IF(KFAC .EQ. 1)GO TO 32
      IDELET=IDELET+1
      IF(NCOR  .EQ. 0)GO TO 31
      DO 30 ICHK=1,NCOR
   30 IF(ICOL(IM)/1000 .EQ. ICOR(ICHK))GO TO 32
   31 NCOR=NCOR+1
      ICOR(NCOR)=ICOL(IM)/1000
      SSCOR=SSCOR+SS(IFAC,JFAC,KFAC)
   32 NDELET=NDELET+1
      NCOL(NDELET)=IM
   33 CONTINUE
      CALL HYPOTH(NDELET,IDELET,M,N,SSHYPO)
      SS(I,J,1)=SSREG-SSHYPO-SSCOR
      NDF(I,J,1)=NDELET-IDELET
   34 CONTINUE
C
C        FIND SUMS OF SQUARES AND DEGREES OF FREEDOM FOR SINGLE-FACTOR
C        TERMS
C
   35 DO 40 I=1,NFAC
      NDELET=0
      IDELET=0
      SSCOR=0.
      NCOR=0
      DO 39 IM=2,M
      IFAC=ICOL(IM)/100000
      JFAC=ICOL(IM)/10000-(ICOL(IM)/100000)*10
      KFAC=ICOL(IM)/1000-(ICOL(IM)/10000)*10
      IF((I.NE.IFAC).AND.(I.NE.JFAC).AND.(I.NE.KFAC))GO TO 39
      IF(JFAC .EQ. 0)JFAC=1
      IF(KFAC .EQ. 0)KFAC=1
      IF((JFAC .EQ. 1).AND.(KFAC .EQ. 1))GO TO 38
      IDELET=IDELET+1
      IF(NCOR .EQ. 0)GO TO 37
      DO 36 ICHK=1,NCOR
   36 IF(ICOL(IM)/1000 .EQ. ICOR(ICHK))GO TO 38
   37 NCOR=NCOR+1
      ICOR(NCOR)=ICOL(IM)/1000
      SSCOR=SSCOR+SS(IFAC,JFAC,KFAC)
   38 NDELET=NDELET+1
      NCOL(NDELET)=IM
   39 CONTINUE
      CALL HYPOTH(NDELET,IDELET,M,N,SSHYPO)
      SS(I,1,1)=SSREG-SSHYPO-SSCOR
      NDF(I,1,1)=NDELET-IDELET
   40 CONTINUE
C
C        PRINT ANOV TABLE HEADING AND REGRESSION DATA
C
      PRINT 103,NPROB,NOBS,(I,NLEVEL(I),I=1,NFAC)
  103 FORMAT(1H1,30X,'EXAMPLE NUMBER',I3//
```

```
     15X,'NUMBER OF OBSERVATIONS'',I4//
     15X,'FACTOR'/
     15X,'NUMBER',4X,'LEVELS'//
     11O(7X,I2,7X,I2/))
      PRINT 1(4,IDFREG,SSREG
  1O4 FORMAT(/3CX,'ANALYSIS OF VARIANCE'/
     163X,'MS RATIO TO'/
     18X,'SOURCE',8X,'DF',3X,'SUM OF SQUARES',3X,'MEAN SQUARES',8X,'ER
     1R MS'//
     16X,'REGRESSION',3X,I2,3X,E14.8//
     18X,'FACTOR'/)
C
C         PRINT DATA FOR SINGLE-FACTOR TERMS
C
      DO 41 I=1,NFAC
      VAR=SS(I,1,1)/FLOAT(NDF(I,1,1))
      F=VAR/VARERR
   41 PRINT 1C5,I,NDF(I,1,1),SS(I,1,1),VAR,F
  1O5 FORMAT(7X,I1,11X,I2,3X,E14.8,4X,E14.8,6X,F11.2)
C
C         PRINT DATA FOR TWO-FACTOR INTERACTION TERMS
C
      DO 42 I=1,NFAC1
      IJ=I+1
      DO 42 J=IJ,NFAC
      IF(NDF(I,J,1).EQ.0)GO TO 42
      VAR=SS(I,J,1)/FLOAT(NDF(I,J,1))
      F=VAR/VARERR
      PRINT 1O6,I,J,NDF(I,J,1),SS(I,J,1),VAR,F
  1O6 FORMAT(7X,I1,' X ',I1,7X,I2,3X,E14.8,4X,E14.8,6X,F11.2)
   42 CONTINUE
C
C         PRINT DATA FOR THREE-FACTOR INTERACTION TERMS
C
      DO 43 I=1,NFAC2
      IJ=I+1
      DO 43 J=IJ,NFAC1
      JK=J+1
      DO 43 K=JK,NFAC
      IF(NDF(I,J,K).EQ.0)GO TO 43
      VAR=SS(I,J,K)/FLOAT(NDF(I,J,K))
      F=VAR/VARERR
      PRINT 1O7,I,J,K,NDF(I,J,K),SS(I,J,K),VAR,F
  1O7 FORMAT(7X,I1,' X ',I1,' X ',I1,3X,I2,3X,E14.8,4X,E14.8,6X,F11
   43 CONTINUE
      PRINT 1O8,IDFERR,SSERR,VARERR,IDFTOT,SSTOT
  1O8 FORMAT(/6X,'ERROR',6X,I2,3X,E14.8,4X,E14.8//
     16X,'TOTAL',8X,I2,3X,E14.8)
C
C         CLEAR FOR NEXT PROBLEM
C
      NPROB=NPROB+1
      DO 44 I=1,N
      DO 44 J=1,M
   44 X(I,J)=0.
      DO 45 I=1,7
      DO 45 J=1,7
      DO 45 K=1,7
      NDF(I,J,K)=0
```

```
   45 SS(I,J,K)=0.
      GO TO 1
C
C          ERROR MESSAGE IF SOLUTION TO NORMAL EQUATIONS NOT FOUND
C
   46 PRINT 109,
  109 FORMAT(' GJR FOR X')
   47 STOP
      END
```

```
      FOR HYPOTH
          SUBROUTINE HYPOTH(NDELET,IDELET,M,N,SSHYPO)
          COMMON Y(150,1),X(150,100),XH(150,100),XTX(100,100),BT(1,100),
         1LEVEL(10),NLEVEL(10),NCOL(100),SS(7,7,7),NDF(7,7,7),ICOL(100),
         1IBLOCK(6),ICOR(100)
          V=4.
    C
    C         FORM X MATRIX FOR REDUCED MODEL
    C
          MH=M-NDELET
          IF(NDELET .EQ. 0)GO TO 5
          J=1
          K=1
        1 DO 2 I=1,NDELET
        2 IF(J .EQ. NCOL(I))GO TO 4
          DO 3 I=1,N
        3 XH(I,K)=X(I,J)
          IF(J .EQ. M)GO TO 7
          J=J+1
          K=K+1
          GO TO 1
        4 IF(J .EQ. M)GO TO 7
          J=J+1
          GO TO 1
        5 DO 6 I=1,N
          DO 6 J=1,M
        6 XH(I,J)=X(I,J)
    C
    C         FORM X'X MATRIX FOR REDUCED MODEL
    C
        7 DO 8 I=1,MH
          DO 8 J=1,MH
          XTX(I,J)=0.
          DO 8 IROW=1,N
        8 XTX(I,J)=XTX(I,J)+XH(IROW,I)*XH(IROW,J)
    C
    C         FORM NORMAL EQUATIONS
    C
          DO 9 I=1,MH
          XTX(I,MH+1)=0.
          DO 9 J=1,N
        9 XTX(I,MH+1)=XTX(I,MH+1)+XH(J,I)*Y(J,1)
    C
    C         SOLVE NORMAL EQUATIONS
    C
          CALL GJR(XTX,100,100,MH,MH+1,$13,JC,V)
          DO 10 I=1,MH
       10 BT(1,I)=XTX(I,MH+1)
    C
    C         RETURN REGRESSION SUM OF SQUARES FOR REDUCED MODEL
    C
          SSHYPO=0.
          DO 12 I=1,MH
          XTY=0.
          DO 11 J=1,N
       11 XTY=XTY+XH(J,I)*Y(J,1)
       12 SSHYPO=SSHYPO+BT(1,I)*XTY
          RETURN
    C
```

```
C        ERROR MESSAGE IF SOLUTION TO NORMAL EQUATIONS NOT FOUND
C
   13 PRINT 100,
  100 FORMAT(' GJR XH')
      STOP
      END
```

The following are listings of the ANOVA input data for the examples of Chapter IV.

Example 1:

```
  125
          83. 1
          85. 1
          84. 2
          85. 2
          85. 2
          86. 2
          86. 2
          87. 2
          86. 3
          87. 3
          87. 3
          87. 3
          88. 3
          88. 3
          88. 3
          88. 3
          88. 3
          89. 3
          90. 3
          89. 4
          90. 4
          90. 4
          91. 4
          90. 5
          92. 5
```

Example 2:

```
  212 1
  2
           2. 1 1
          20. 3 1
           7. 4 1
          32. 2 2
          14. 3 2
           3. 4 2
           4. [illegible]
          13. 2 3
          31. 3 3
           0. 1 4
          23. 2 4
          11. 4 4
```

Example 3:

210

5.	1	1
3.	1	1
13.	2	1
14.	2	1
15.	2	1
[illegible].	1	2
5.	1	2
7.	1	2
12.	2	2
10.	2	2

Example 4:

336

20.2	1	1	1
24.1	1	1	1
26.2	1	1	2
26.9	1	1	2
23.8	1	1	3
24.9	1	1	3
22.0	1	2	1
23.5	1	2	1
22.6	1	2	2
24.6	1	2	2
22.9	1	2	3
25.[illegible]	1	2	3
23.1	1	3	1
22.9	1	3	1
22.[illegible]	1	3	2
23.7	1	3	2
21.8	1	3	3
23.5	1	3	3
14.2	2	1	1
16.2	2	1	1
13.[illegible]	2	1	2
19.1	2	1	2
12.5	2	1	3
15.4	2	1	3
14.1	2	2	1
16.1	2	2	1
14.0	2	2	2
18.1	2	2	2
13.7	2	2	3
16.[illegible]	2	2	3
14.1	2	3	1
15.1	2	3	1
17.2	2	3	2
[illegible]	[illegible]	[illegible]	[illegible]
12.7	2	3	3
15.1	2	3	3

The following are the outputs from the ANOVA program for the example problems of Chapter IV.

EXAMPLE NUMBER 1

NUMBER OF OBSERVATIONS: 25

FACTOR NUMBER	LEVELS
1	5

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARES	MS RATIO TO ERROR MS
REGRESSION	5	.19176786+06		
FACTOR				
1	4	.99019531+02	.24754883+02	[illegible]
ERROR	20	.23142578+02	.11571289+01	
TOTAL	25	.19179100+06		

EXAMPLE NUMBER 2

NUMBER OF OBSERVATIONS: 12

FACTOR NUMBER	LEVELS
1	4
2	4

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARES	MS RATIO TO ERROR MS
REGRESSION	7	.31148333+04		
FACTOR				
1	3	.88083337+03	.29361112+03	[illegible]
2	3	[illegible]	.20555522+01	[illegible]
ERROR	5	.36316669+03	.72633337+02	
TOTAL	12	.34780000+04		

EXAMPLE NUMBER 3

NUMBER OF OBSERVATIONS: 10

FACTOR NUMBER	LEVELS
1	2
2	2

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARES	MS RATIO TO ERROR MS
REGRESSION	4	.97000000+03		
FACTOR				
1	1	.13500000+03	.13500000+03	101.[illegible]
2	1	.59999847+00	.59999847+00	.4[illegible]
1 X 2	1	.15000000+02	.15000000+02	11.2[illegible]
ERROR	6	.80000000+01	.13333333+01	
TOTAL	10	.97800000+03		

EXAMPLE NUMBER 4

NUMBER OF OBSERVATIONS: 36

FACTOR NUMBER	LEVELS
1	2
2	3
3	3

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARES	MS RATIO TO ERROR MS
REGRESSION	18	.14175168+05		
FACTOR				
1	1	.65195068+03	.65195068+03	282.16
2	2	.16051514+02	.80257568+01	3.47
3	2	.12771464+02	.63857422+01	2.76
1 X 2	2	[illegible]	[illegible]	[illegible]
1 X 3	2	[illegible]	.27803955+01	1.2[illegible]
2 X 3	4	.26487061+02	.66217651+01	2.87
1 X 2 X 3	4	.51607666+01	.12901917+01	.56
ERROR	18	.41590576+02	.23105676+01	
TOTAL	36	[illegible]		