

ANALYSIS OF COVARIANCE,  
A METHOD FOR HANDLING  
CONCOMITANT VARIABLES

by

Daniel Garber Brooks

ARTHUR LAKES LIBRARY  
COLORADO SCHOOL OF MINES  
GOLDEN, COLORADO

ProQuest Number: 10781859

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10781859

Published by ProQuest LLC (2018). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 – 1346

A Thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science in Mathematics.

Signed: Daniel G. Brooks  
Daniel G. Brooks

Golden, Colorado

Date: 6/21, 1973

ARTHUR LAKES LIBRATT  
COLORADO SCHOOL OF MINES  
GOLDEN, COLORADO

Approved: William R. Astle  
Prof. William Astle  
Thesis Advisor

W.D. Cypeland  
for Dean of the Graduate School  
Dr. Joseph R. Lee (Deceased)  
Head, Department of  
Mathematics

Golden, Colorado

Date: July 6, 1973

ABSTRACT

A brief review of linear regression analysis and analysis of variance is first presented. The rest of the thesis deals with the combining of these two techniques to form an analysis of covariance model which can be used to identify and separate from the data of interest variability which is due to variation in a concomitant variable upon which the data is dependent.

A computer program is finally included along with a discussion of the use of this program and interpretation of the output.

TABLE OF CONTENTS

INTRODUCTION.....	1
LINEAR REGRESSION ANALYSIS.....	4
Introduction.....	4
The Regression Model.....	5
Determining the Regression Relationship.....	9
Estimation of Parameters $\alpha$ and $\beta$ .....	11
Assumptions.....	11
Least Squares Estimation.....	12
Residual Sum of Squares Form.....	14
The Regression Adjustment in Error Sum of Squares.....	16
ANALYSIS OF VARIANCE.....	20
Introduction.....	20
Classes of ANOVA Problems.....	23
Assumptions.....	24
Fixed-Effects Model.....	25
Random-Effects Model.....	30
Results of Unsatisfied Assumptions.....	31
Some Terminology.....	31
Preparation.....	32
Setting up the Experiment.....	32
Running the Experiment.....	33

Analysis of Variance Model.....34

    One-Way Classification.....34

    Two-Way Classification.....39

    Example.....42

ANALYSIS OF COVARIANCE.....45

    Introduction.....45

    Simple Covariance.....47

        Introduction.....47

        Example.....47

            ANOVA Model.....47

                The Problem.....49

                Approach to the Problem.....50

            Regression Model.....50

    Analysis of Covariance Model.....50

    Application of the Model.....52

        Example Application.....53

        Regression Slopes Used in Adjustment.....55

    General Computational Procedure.....57

    Specific Results.....64

        Reduction in Variance (Error Control).....64

        Adjustment of Means.....64

    Assumptions.....65

    Tests of Assumptions.....66

Multiple Covariance.....69

    Multiple Regression Model.....69

    Adjustment for Regression.....72

Multiple Covariance Model.....	72
General Example--2 Factors, 3 Covariates.....	73
General Computational Approach.....	74
Uses.....	79
Direct Applications.....	79
Indirect Applications.....	83
Estimation of Missing Data.....	83
Regression in Multiple Classifications.....	84
Outlier Points.....	84
Regression Adjustment.....	85
Pooling Data.....	87
Individual Yield Estimation.....	88
Applications in Special Experimental Design Situations.....	88
Pitfalls in Applications.....	89
USE OF THE COMPUTER PROGRAM.....	92
Purpose.....	92
Documentation.....	92
Input.....	97
Execution.....	101
Output.....	102
APPENDICES.....	109
A. Consequences of Assumptions being Unsatisfied for ANOVA.....	110
B. Estimation of $\frac{\sigma^2}{n}$ .....	114
C. Decomposition Proof for Total Sum of Squares....	115

D. Mutually Distinct Partitioning.....	116
E. Two-factor Decomposition.....	117
F. Solution of Normal Equations for Multiple Regression.....	118
G. Residual Sum of Squares Deviation.....	121
H. Derivation for $R^2$ .....	122
I. Computer Program Listing.....	123
LITERATURE CITED.....	151

ACKNOWLEDGMENTS

I wish to express my thanks to Prof. William Astle, my thesis advisor, for reading and rereading the earlier drafts, for his suggestions contributing to the final form of the thesis, and direction in its completion. Also to Dr. John Kork and Dr. Raymond Mueller, the thesis committee members, I extend my thanks. I also wish to thank Prof. Baer for his assistance in using the computer and especially I wish to acknowledge Dr. Raymond Mueller for his patience and instruction during my graduate study.

I also gratefully acknowledge the financial aid given by the Colorado School of Mines in the form of teaching assistantships and tuition waivers.

INTRODUCTION

The primary purpose in collecting data is usually to be able to interpret the data as saying something about, or in some way describing, the processes or things being studied. Many difficulties can arise which make the interpretation of the data difficult. One of these difficulties is a lot of variation in the data. Unless this variation can be explained or reduced it is hard for the experimenter to draw conclusions with much precision or definiteness. There are several methods used in attempting to reduce or control variation among the data. Some of these methods are:

- (i) Exercise control over the homogeneity of the material being tested, and over the experimental environment, thus helping to create uniformity of conditions under which the experiment is performed.
- (ii) Group the material and the environment so that (i) holds for the subgroups, i.e., divide the experiment to achieve homogeneity within subgroups.
- (iii) Refine the experimental techniques, so that they are consistent throughout all phases of the experiment and don't contribute to variation in results.

If the experimental techniques are stable and it is not possible to subdivide the elements of the experiment into homogeneous subdivisions to control experimental variation, there is a fourth

technique available:

- (iv) Measure the variables related to the variable of interest, and use analysis of covariance.

The analysis of covariance is a statistical tool which can be used to identify in a variable of interest,  $Y$  say, the amount of variation which is a result of variation in another variable, say  $X$ , upon which  $Y$  is dependent. Analysis of covariance can then be applied to remove this variation from the variable of interest. The related variates are referred to as concomitant variables.

The presentation of analysis of covariance which follows is on a level which can be understood (hopefully) by a person who has had little background in statistics. A review of regression analysis and analysis of variance, which are vital to analysis of covariance, is presented in the first two chapters to help introduce notation and make the whole of the presentation as self-contained as possible.

The last section of the chapter on analysis of covariance presents several different applications. Although it is not feasible to present each of them in computational detail, references for further investigation are given if the reader is interested. The objective is to show the versatility of the technique and the information which is available for what is usually only a small investment of additional computational effort. There are many instances, in fact, where the measurements of what would be the concomitant variable are already known, or are readily obtainable, but because the experimenter is not aware of analysis of covariance, this additional data is not used.

The last chapter shows how to use the computer program which is available, and how to read the output.

LINEAR REGRESSION ANALYSISIntroduction

In working with statistical problems, many times there is an association between two or more of the variables being measured, and it would be helpful to establish a relationship, from the data, which would make it possible to estimate, or "predict", one or more variables in terms of other variables. One of the oldest, and probably one of the most useful, types of relationships is what is called a linear model. Such a model, or relationship, enables the experimenter to use additional, or known, information to help describe the behavior of the variable of primary interest. Such a relationship might make possible a prediction of the amount of sales of a new product from its price, or a student's future grade average from his I.Q. rating or entrance exam score.

Although it would be nice to be able to predict one quantity exactly in terms of others, it is hardly ever possible, so estimation of an "average" value in terms of others usually has to suffice. For example, it is not possible to predict the exact number of sales of a product from its price, but it may be possible to estimate the "average" number of expected sales based on information about past performances of like products.

The 19th century English mathematician Francis Galton developed the idea of "regression" in his studies of heredity. Speaking of the "law of universal regression" he said that "each peculiar

ilarity in a man is shared by his kinsman, but on the average in a less degree." To back this up, he collected data on the heights of fathers and their sons. It was found that although tall fathers tend to have tall sons, the average height of the sons of a group of tall fathers is less than the average height of their fathers. So we say there is a regression, or going back, of sons' heights toward the average height of all men.

In mathematical jargon, we would probably say that the predicted variable, say  $Y$ , is a function of the variable upon which the prediction was based, call it  $X$ , but in statistics Galton's term "regression" is usually used so that the relationship found between  $Y$  and  $X$  is called the "regression of  $Y$  on  $X$ ."

### The Regression Model

As an example, suppose we were interested in studying the weights of a certain population of men, and the relationship between the weights and the heights of these men. To study this the men are subdivided into groups according to height so that the men in any one group are all very nearly the same height, and the relationship between weight and height is examined by looking at the various subdivisions and their weights.

It is obvious that for any particular height there will be a whole range of weights. Not every man that is six feet tall will weigh 180 pounds. There will be light ones and heavy ones. This distribution of weights for a particular height has a mean value. In statistical language this would be called the expected value for the weights. This distribution would also have a var-

iance, the variance of weights of all men who have this height. The "regression of weight on height" in this case would be the relationship between the heights and the means of the distributions of weights for each of the heights. The regression relationship doesn't allow a prediction of a man's weight, given his height, but the average weight of all men who have that height can be predicted.

It can be seen here that the distribution of weights depends on the height chosen, and so weight would be referred to as the dependent variable, and height as the independent variable.

Introducing some common notation, the dependent random variable, which is usually the variable of interest, is denoted by  $Y$  and the independent variable by  $X$ , although any symbols may be used. The mean value for the distribution of  $Y$ , called the expected value, is denoted by  $E(Y)$ . The expected value of  $Y$  given the variable  $X$  has taken on the value  $X$ , denoted  $E(Y|X)$ , is called the conditional expected value of  $Y$ . It could represent the expected value for the weights given the height chosen was  $X$ . If we write  $E(Y|X)$ , this represents a whole set of conditional expected values of  $Y$  as  $X$  takes on all values in its domain.

In regression analysis, we want to determine what the relationship between  $X$  and  $E(Y|X)$  is, so that if we know the independent variable  $X$  has taken the value  $X$ , it is possible to predict the mean value of the  $Y$  for that  $X$ ,  $E(Y|X)$ . A first approximation to this relationship, and usually a good approximation, especially over short intervals, is a straight line. This could be written

as

$$E(Y|X) \doteq \alpha + \beta X .$$

Read the symbol " $\doteq$ " as "is approximated by." This says that the relationship between the  $X$  values and  $E(Y|X)$  is linear. So if we graphed the coordinates  $(X, E(Y|X))$  as  $X$  took on all the values of  $X$ , we assume this graph is a straight line. In the assumed linear case, the line  $\alpha + \beta X$  is by definition the "true" regression line, and deviations are measured with respect to this line. The linear assumption is shown graphically in the diagram below.

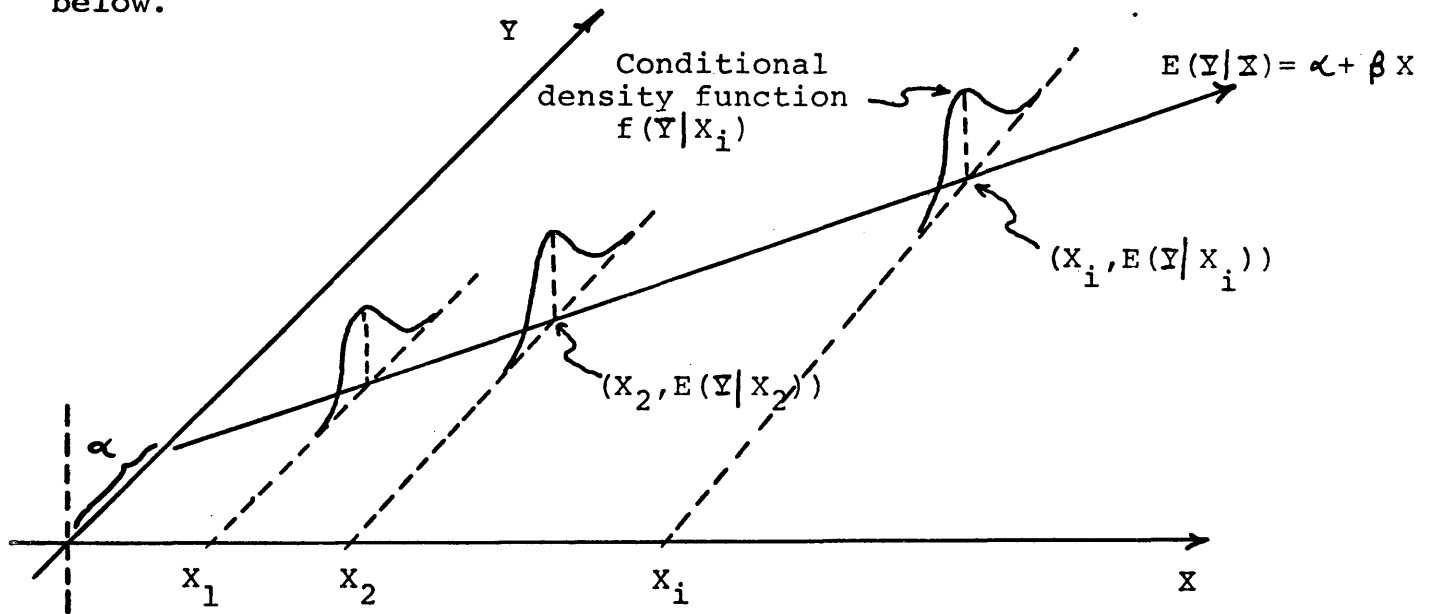


Figure 1. True linear regression relationship.

From the diagram, it can be seen that we are assuming that changes in the value of  $X$  change the value of the mean of the  $Y$ , but it is assumed that changes in  $X$  do not affect the shape of the density function of  $Y$ . More detail as to assumptions will be

given later.

It should be understood that in general the straight line is an approximation of the actual relationship between  $Y$  and  $X$ . Also, because  $Y$  is a random variable, in general, observed or measured values of  $Y$  will not be equal to the predicted value. That is, the observed values of  $Y$  will not all fall on the line  $E(Y|X)$ .

The observed value of  $Y$  can then be written as

$$Y = E(Y|X) + \epsilon ,$$

where  $\epsilon$  represents some error, the amount of  $Y$  not accounted for by the regression line of  $Y$  on  $X$ . This is logical, since for each  $X$  there is a whole population of  $Y$ 's, of which we have observed one. We would expect, in general, that this one we observed will not always be the mean value,  $E(Y|X)$ , of the population of  $Y$ 's for that  $X$ . The amount that the observed  $Y$  differs from the mean value is represented by  $\epsilon$ .

Since, in the linear case,  $E(Y|X) = \alpha + \beta X$ , we can rewrite this relation as

$$Y = \alpha + \beta X + \epsilon ,$$

where  $Y$  is a measured or observed value of  $Y$  for  $X$  taking on the value  $X$ .

It is assumed here that the only error involved is in observing the  $Y$ , that the  $X$  is measured without error, so that the error is entirely included in  $\epsilon$ .

The relationship above is for  $Y$  dependent on only one variable,  $X$ . The situation may arise where  $Y$  is influenced by  $k$  fixed variates,  $X_1, X_2, \dots, X_k$ . This is covered in chapter 3.

If our data consists of  $n$  observations of  $Y$  and  $X$ , simultaneously, then we have  $n$  points,  $\{(X_i, Y_i), i=1, 2, \dots, n\}$ , and each observed  $Y$  could be modeled

$$Y_i = \alpha + \beta X_i + \epsilon_i .$$

It remains now to find out what  $\alpha$  and  $\beta$  are so the regression model is usable.

### Determining The Regression Relationship

To be able to use, in the linear case, what we define as the true regression relation,

$$Y = E(Y|X) + \epsilon = \alpha + \beta X + \epsilon ,$$

we have to know the joint probability density function for  $X$  and  $Y$  to compute the true conditional expectations of  $Y$ . Since with fresh data this information is rarely available, the relationship between  $X$  and  $Y$  has to be estimated from the data on hand. If we have  $n$  observations on  $X$  and  $Y$ , we hope to use these to estimate the expected value of  $Y$  for each  $X$  value, i.e.,  $E(Y|X_i)$ ,  $i=1, \dots, n$ . If  $\hat{Y}_i$  is used to denote an estimate (however obtained) of the expected value value of  $Y$ , given  $X$  is  $X_i$ ,  $E(Y|X_i)$ , then each of the observed values,  $Y_i$ , can be represented by

$$Y_i = \hat{Y}_i + e_i \quad , \quad i=1, 2, \dots, n \quad ;$$

where  $\hat{Y}_i$  is the estimate of  $E(Y|X_i)$ , and  $e_i$  is what is called the residual.  $e_i$  is the amount by which the estimated value,  $\hat{Y}_i$ , missed the observed  $Y_i$ .  $\epsilon_i$  is the deviation of the observed value,  $Y_i$ , from the true regression line,  $E(Y|X)$ , and  $e_i$  is the deviation of the observed value,  $Y_i$ , from the estimated regression line,  $\hat{Y}$ .

Hence, for the linear case, the true regression relation

$$Y_i = E(Y|X_i) + \epsilon_i = \alpha + \beta X_i + \epsilon_i$$

can be approximated by

$$Y_i = \hat{Y}_i + e_i = a + bX_i + e_i \quad ,$$

where  $a$  is an estimate of  $\alpha$ ,  $b$  is an estimate of  $\beta$ , and  $\{e_i\}$  are the residuals.

The actual (assumed linear) regression curve is found by joining the expected values of  $Y$  given different values of  $X$ . This relation is compared to the estimated regression line below.

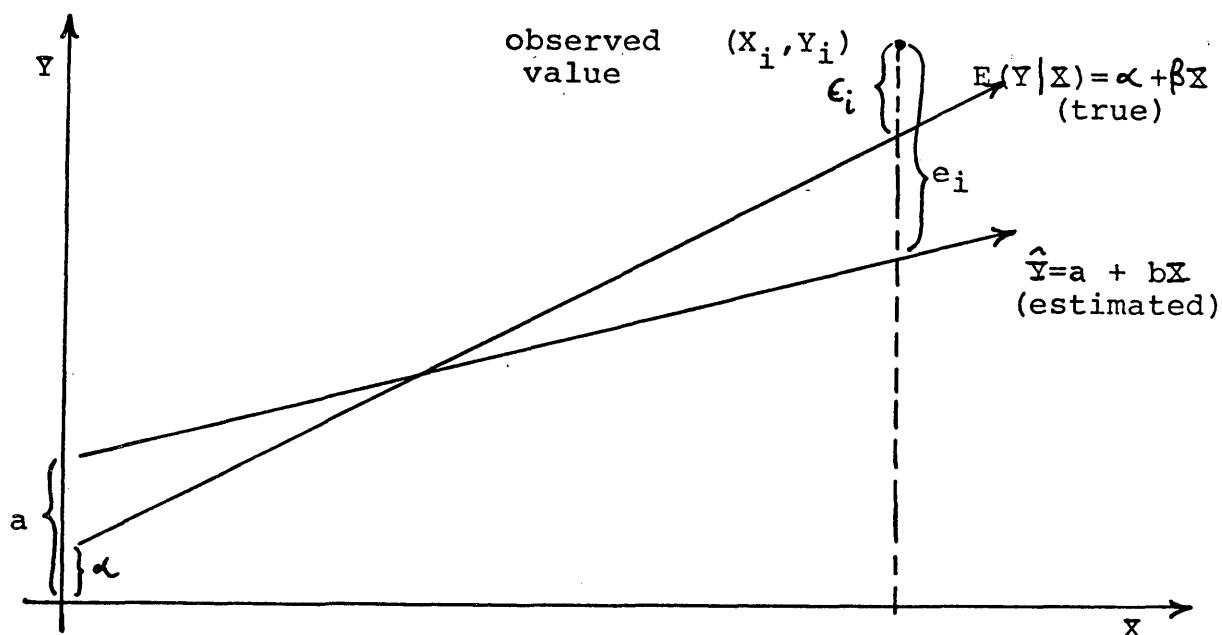


Figure 2. Comparison of true and estimated regression lines.

For a given observation  $(X_i, Y_i)$ , the true error is given by

$$\epsilon_i = Y_i - E(Y|X_i) = Y_i - (\alpha + \beta X_i) \quad (1)$$

and the estimated error, or residual, by

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i) \quad .$$

### Estimation of Parameters $\alpha$ and $\beta$

There are many methods for determining values for  $a$  and  $b$ , the estimates of  $\alpha$  and  $\beta$ , from the data to obtain a "best" fit of the line to the observed points. No matter how it is done, it seems reasonable to try and make the residuals,  $\{e_i\}$ , as small as possible. The problem, then, is how to go about making them small. Among possible approaches are:

- (i) Minimize the sum of the absolute values of the  $e_i$ , i.e., minimize  $\sum |e_i|$ .
- (ii) Minimize the greatest of the absolute residuals, i.e., minimize  $\left[ \max_i \{|e_i|\} \right]$ .
- (iii) Minimize the sum of the squares of the residuals, i.e., minimize  $\sum e_i^2$ .

Method (iii), called the "method of least squares", is one of the easier methods to apply, and under certain assumptions, it provides estimates of  $\alpha$  and  $\beta$  which are unique, and the best (linear) unbiased estimators.

Before showing how this method can be applied, some of the assumptions which are made for the linear regression model least squares estimation should be noticed.

Assumptions. The basic assumption, of course, is (i) that the conditional expected value of  $Y$ , given  $X$ ,  $E(Y|X)$ , is a linear function of  $X$ . It is also assumed (ii) that the conditional densities for  $Y$  are uncorrelated for different values of  $X$ . This is usually reasonable to assume, and just means that the value  $Y$  takes on for one value of  $X$  does not affect the value  $Y$  takes

on for another value of  $X$ . Thirdly, we assume (iii) that the  $X$  values are measured without error, so that all the error in the model is represented in the error term,  $\epsilon$ . We also assume that the conditional variances of the  $Y$  populations are independent of  $X$ . That is, the variance of  $Y$  is the same no matter what value  $X$  takes on. These assumptions are enough to be able to apply the method of least squares with assurance that the estimates of  $\alpha$  and  $\beta$  are "unbiased" (see [22], pg.147) and have the smallest variance of any estimators. If however, it is desired to make the usual tests of significance (such as the  $t$ - and  $F$ -tests) of our estimates (to see how sure we are about how good we think the estimates are), it must also be assumed that  $Y$  is distributed normally (or that  $X$  and  $Y$ , jointly, are distributed as a bivariate normal). This means that the  $Y$ 's are not just uncorrelated, but independent, and that we can say that the error terms,  $\{\epsilon\}$ , are independently and normally distributed with mean zero (see equation (1)) and a common variance,  $\sigma^2$ .

Least Squares Estimation. Let us consider the problem of estimating the best linear approximation of the relationship between  $Y$  and a single fixed variable,  $X$ , using the method of least squares, so that observed values of  $Y$  are given by

$$Y = \alpha + \beta X + \epsilon = a + bX + e$$

where  $\alpha$  and  $\beta$  are unknown parameters and  $a$  and  $b$  are their respective estimates,  $\epsilon$  is the true error, and  $e$  the residual ( $e = Y - \hat{Y}$ , where  $\hat{Y} = a + bX$ ). We now assume that a sample of  $n$   $X$ 's are selected (without error) and corresponding  $Y$ 's are measured. Using the method of least squares, we form what is called

the error sum of squares, usually represented by SSE,

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - (a+bX_i))^2 \\ &= \sum_{i=1}^n (Y_i - a - bX_i)^2 . \end{aligned}$$

We want to determine a and b so as to minimize SSE. This can be done by taking partial derivatives of SSE with respect to a and b, setting the two resulting equations equal to zero, and solving this system of equations for a and b.

Taking derivatives, first with respect to a, then with respect to b, equating the derivatives to zero, and rearranging them into what are called the "normal"\* equations, we obtain

$$\begin{aligned} \sum_{i=1}^n Y_i &= na + b \cdot \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i Y_i &= a \cdot \sum_{i=1}^n X_i + b \cdot \sum_{i=1}^n X_i^2 , \text{ for } i=1,2,\dots,n. \end{aligned}$$

We then solve these two equations simultaneously for a and b.

An example follows.

Suppose we are given the set of paired data below (data from Miller and Freund [32]), where the X's represent baking time, in minutes, of a mineral specimen, and the Y's are the oxide thicknesses on the specimens in Angstrom units resulting from the baking. We examine the thicknesses resulting from 10 different baking times.

Time	20	30	40	60	70	90	100	120	150	180
Thickness	3.5	7.4	7.1	15.6	11.1	14.9	23.5	27.1	22.1	32.9

$$\sum_{i=1}^n X_i = 860$$

$$\sum_{i=1}^n Y_i = 165.2$$

$$\sum_{i=1}^n X_i^2 = 98,860$$

$$\sum_{i=1}^n X_i Y_i = 18,469.0$$

---

\*Normal here does not refer to the Normal distribution.

The normal equations are

$$165.2 = 10a + 860b$$

$$18,469.0 = 860a + 98,800b .$$

Solving these two equations we obtain

$$a = 1.90$$

$$b = 0.17 .$$

So the equation of the straight line which provides the best fit in the sense of least squares is

$$Y = 1.90 + 0.17X .$$

The results are diagrammed below.

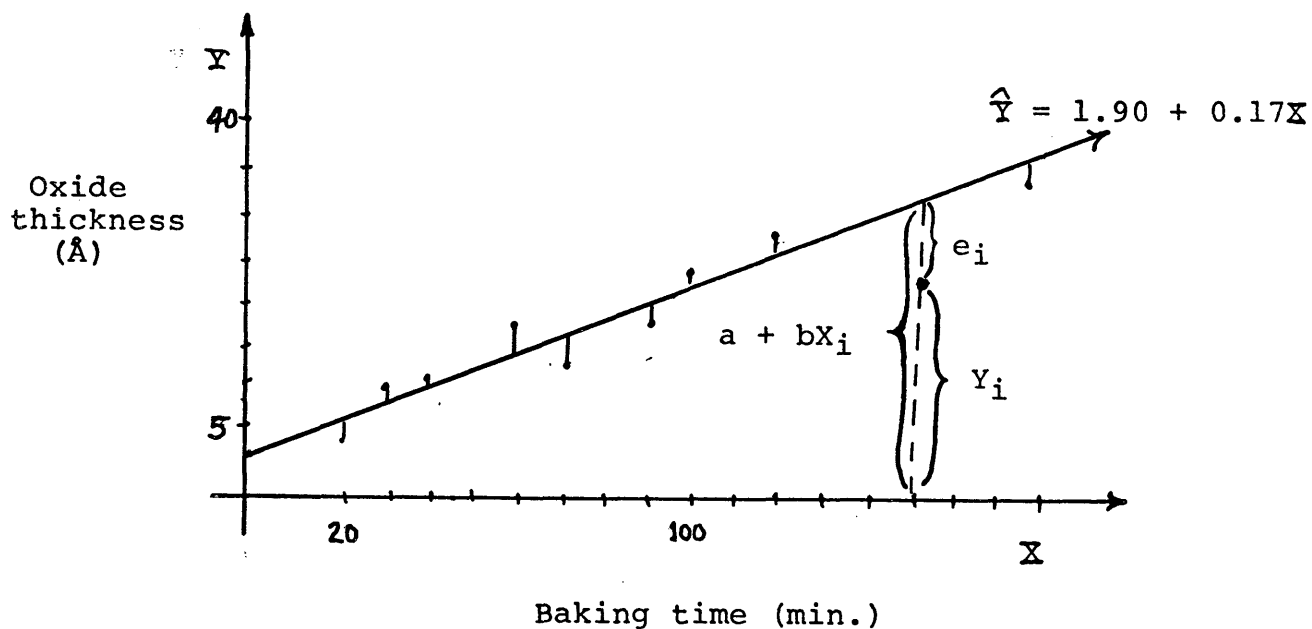


Figure 3. Regression of Oxide thickness (Y) on baking time (X).

#### Residual Sum of Squares Form

There is a second approach we might use in solving the normal equations for  $a$  and  $b$ , giving a different form for the regression model and resulting sometimes in easier calculations.

Taking the partial derivative of the error sum of squares with respect to  $a$ , equating it to zero, and solving for  $a$ , we have

$$\frac{\partial \text{SSE}}{\partial a} = 0 \quad \Rightarrow \quad \sum Y_i = na + b \cdot \sum X_i ;$$

hence,

$$a = \sum Y_i / n - (b \cdot \sum X_i) / n = \bar{Y} - b\bar{X} .$$

We know this solution for  $a$  results in a minimum sum of squares, rather than a maximum, since no finite maximum exists.

If we insert this value for  $a$  into the original SSE equation, we get that

$$\begin{aligned} \text{SSE} &= \sum_i (y_i - a - bx_i)^2 \\ &= \sum [(Y_i - \bar{Y}) - b(X_i - \bar{X})]^2 \\ &= \sum (y_i - bx_i)^2 \end{aligned}$$

where  $y_i = Y_i - \bar{Y}$  and  $x_i = X_i - \bar{X}$ . Using this value for  $a$  we could rewrite the regression model as

$$\hat{Y} = a + bX = \bar{Y} + bx .$$

Taking the partial derivative of this new form with respect to  $b$ , equating that to zero, and solving, we obtain

$$\frac{\partial [\sum (y - bx_i)^2]}{\partial b} = 0 \quad \Rightarrow \quad b = \frac{\sum x_i y_i}{\sum x_i^2} \quad (2)$$

where  $\sum x_i y_i = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - [(\sum X_i)(\sum Y_i)]/n$

$$\sum x_i^2 = \sum (X_i - \bar{X})(X_i - \bar{X}) = \sum X_i^2 - (\sum X_i)^2/n$$

We can use this new information to find the amount of the total error sum of squares ( $\sum y_i^2$ ) which is actually due to variation among the  $X$ , causing variation in  $Y$  by means of the dependency of  $Y$  on  $X$ , which might not otherwise have been there.

Using the new model

$$\begin{aligned} Y_i &= \hat{Y}_i + e_i \\ &= \bar{Y} + bx_i + e_i \end{aligned}$$

and rearranging to solve for  $e_i$ ,

$$\begin{aligned} e_i &= (Y_i - \bar{Y}) - bx_i \\ &= y_i - bx_i \end{aligned}$$

Using this deviation form to reconsider the residual sum of squares we obtain

$$\begin{aligned} \sum e_i^2 &= \sum (y_i - bx_i)^2 \\ &= \sum y_i^2 - 2b \cdot \sum x_i y_i + b^2 \cdot \sum x_i^2 \end{aligned}$$

Substituting for  $b$  the least squares estimate obtained above in equation (2), we have that

$$\begin{aligned} \sum e_i^2 &= \sum y_i^2 - 2 \frac{\sum x_i y_i}{\sum x_i^2} \cdot \sum x_i y_i + \frac{\sum x_i y_i}{\sum x_i^2} \cdot \sum x_i^2 \\ &= \sum y_i^2 - \left[ \frac{(\sum x_i y_i)^2}{\sum x_i^2} \right] \end{aligned} \quad (3)$$

### The Regression Adjustment in the Error Sum of Squares

Looking at this new result, equation (3), for the residual sum of squares we see we are reducing the original error sum of squares, prior to regression, which is

$$\sum_i (y_i - \bar{Y})^2 = \sum y_i^2,$$

by the amount

$$\frac{(\sum x_i y_i)^2}{\sum x_i^2}.$$

This is the amount of the error sum of squares that we can "explain" as variation among the  $Y$  values due to variation in the  $X$  values,

transferred through the dependence of  $Y$  on  $X$ .

The new adjusted sum of error squares can be viewed in a different way which may make clearer what this sum represents.

$\sum(Y_i - \bar{Y})^2$  gives an estimate of the variation among the variable of interest,  $Y$ , before we use any information concerning the dependence of  $Y$  on a concomitant variable  $X$ . Using this information as we did above, we obtained the new resultant estimate of variation. Rearranging equation (3), we get that

$$\sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - \left[ \frac{(\sum x_i y_i)^2}{\sum x_i^2} \right] + \left[ \begin{array}{l} \text{residual} \\ \text{sum of} \\ \text{squares} \end{array} \right]$$

The residual sum of squares is found by subtraction, so that

$$\text{Residual S.S.} = \sum e_i^2 = \sum Y_i^2 - \frac{(\sum x_i y_i)^2}{\sum x_i^2}$$

The total sum of squares of  $Y$ ,  $\sum_1 Y_i^2$ , has been partitioned into two parts:

- (i) A sum of squares attributable to variation among the  $X$ 's, said to be "attributable to regression."
- (ii) An unexplained portion, the residual sum of squares.

It may be of some assistance to show graphically what has been done. This is done in the diagram on the next page (Figure 4). From the diagram it can be seen that the two forms of the regression model are equivalent. That is, that

$$Y_i = a + bX_i + e_i$$

and

$$y_i = bx_i + e_i$$

are equivalent.

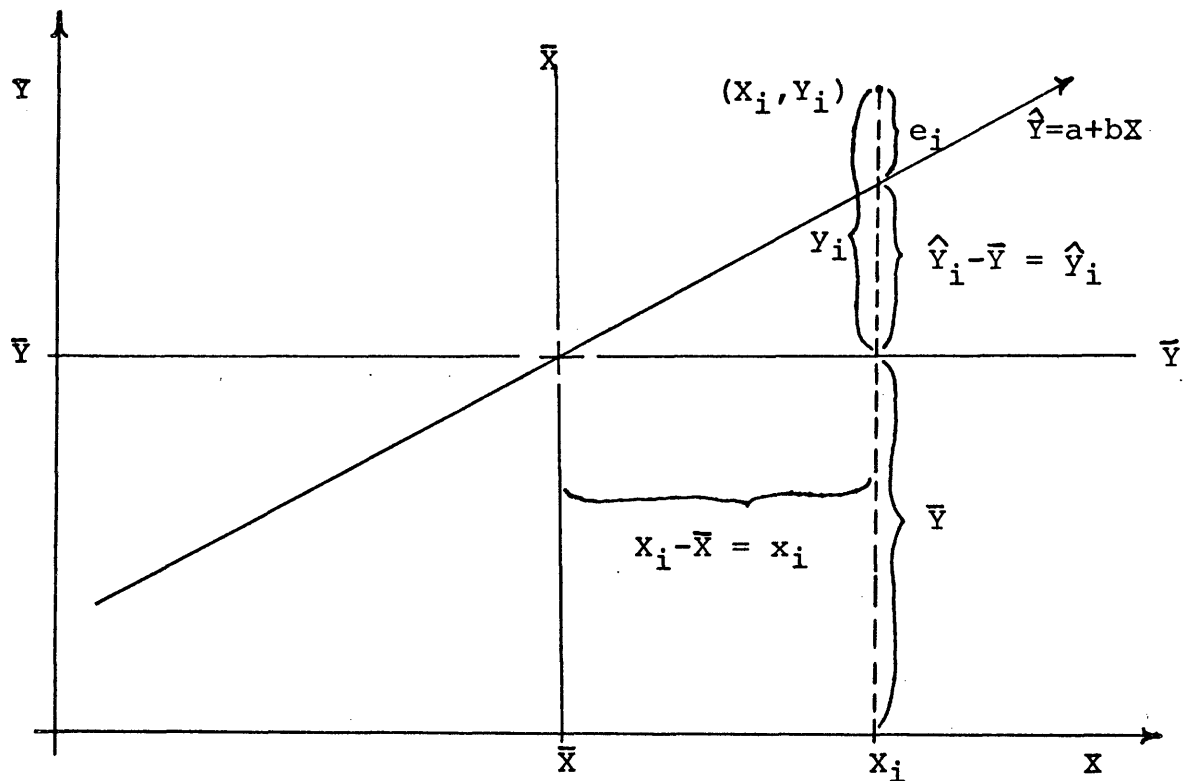


Figure 4. Deviation form of the regression model.

As a further explanatory note, we may notice something more from Figure 4. We know that the reduction in the sum of squares of the error term  $e$  is given by

$$(\sum x_i y_i)^2 / (\sum x_i^2) = b \cdot \sum x_i y_i \quad .$$

Reducing this, we obtain

$$\begin{aligned} b \cdot \sum x_i y_i &= [(\sum x_i y_i)^2 / (\sum x_i^2)^2] \cdot (\sum x_i^2) \\ &= b^2 \cdot \sum x_i^2 \\ &= \sum b^2 x_i^2 \\ &= \sum \hat{Y}_i^2 \\ &= \sum (\hat{Y}_i - \bar{Y})^2 \quad . \end{aligned}$$

So the sum of squares attributable to regression turns out to be the sum of squares of the  $n$  deviations of the estimate,  $\hat{Y}$ , from its mean  $\bar{Y}$ . What we have actually done by subtracting this amount

out is to measure the variation among the Y's as if they had all been measured for a standard value of X, namely  $\bar{X}$ . Since it would be tedious to calculate the sum of squared deviations of estimates from their mean, we use the shortcut form given in equation (3).

As a final note, we can now estimate the variance of the random variable Y which is not attributable to variation in X. The residual sum of squares, the unexplained part of the variation in the Y's, divided by (n-2), its degrees of freedom, gives us the residual mean square and is an (unbiased) estimate of the variance of Y. It is usually denoted by  $s_{yx}^2$ , and is defined by

$$s_{yx}^2 = \frac{\text{SSE(adjusted)}}{(n-2)} = \frac{\sum y_i^2 - (\sum x_i y_i)^2 / \sum x_i^2}{n - 2}$$

$$= \frac{\sum e_i^2}{n-2} .$$

It measures the amount of variation in Y not associated with, explained by or dependent upon changing values of the fixed variate X.

## ANALYSIS OF VARIANCE

### Introduction

First introduced by Sir R. A. Fisher of England in the early 1930's, the term "analysis of variance" has come to represent the use of a number of statistical techniques whereby the experimenter is able to examine the variability occurring in a group of data, and separate the variance ascribable to one group of causes from the variance due to other groups of causes. By separating, and identifying, different sources of variation in the data and the amount of variation each contributes, the experimenter is aided in making judgements about the populations from which the data has been drawn.

One important use of some of these techniques is as an aid in determining if the differences between means of different samples can be attributed to chance variation or if these differences indicate actual differences between the true means of the corresponding populations from which the samples were taken. Therefore, we want to analyze the variability that occurs in the entire group of data and determine its sources.

To illustrate the basic idea with an example, suppose a man can drive from his home to work along any one of three different routes, and he would like to know which of the three routes is the fastest. He records the time it takes him along each route on five different days, shown below in minutes.

Route 1:	22	26	25	25	31
Route 2:	25	27	28	26	29
Route 3:	26	28	27	30	30

The means of these three samples are 25.8, 27.0, and 28.2. Since the size of the samples is so small, we would like to know whether the differences among these means is because the routes really do, on the average, take different amounts of time to travel, or if the differences are just due to chance variations along the routes during the five days on each route.

To treat this kind of problem in general, suppose we have  $k$  independent samples (random) of size  $n^*$ , each from one of  $k$  populations, and let  $X_{ij}$  represent the  $j$ th observation from the  $i$ th population. We can express each of the observations, as well as the samples, in table form as shown below.

						Sample Means	
Sample 1:	$X_{11}$	$X_{12}$	$\cdots$	$X_{1i}$	$\cdots$	$X_{1n}$	$\bar{X}_1.$
Sample 2:	$X_{21}$	$X_{22}$	$\cdots$	$X_{2i}$	$\cdots$	$X_{2n}$	$\bar{X}_2.$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Sample $k$ :	$X_{k1}$	$X_{k2}$	$\cdots$	$X_{ki}$	$\cdots$	$X_{kn}$	$\bar{X}_k.$
							<hr/>
						grand mean	$\bar{X}_{..}$

Table 1.

---

\*Although computationally easier, it is not a requirement that the samples be of equal size. If the samples are of different sizes, we calculate the overall mean by weighting sample means according to the size of the sample.

These sample groups are often referred to as "treatments", owing to the origin of this method in the field of agriculture.

The dot as a subscript means that the variable has been summed for the subscript which has been replaced by the dot. For example,

$$x_{1.} = \sum_{j=1}^n x_{1j} ,$$

and in general

$$x_{i.} = \sum_j x_{ij} .$$

The means are calculated as

$$\bar{x}_{i.} = (\sum_j x_{ij})/n ,$$

$$\bar{x}_{..} = (\sum_{i=1}^k \sum_{j=1}^n x_{ij})/nk .$$

Each of the  $k$  samples comes from a population with a true mean,  $m_i$ . say, which has been estimated by  $\bar{x}_{i.}$  . We are interested in learning something about the relationship between the means of the populations. More specifically, we are concerned with testing the fact of whether the means of the populations are equal or not. Mathematically, we want to test the hypothesis that all  $k$  populations have the same (true) mean. Symbolically,

$$H_0: m_1 = m_2 = \dots = m_k = m..$$

is the hypothesis that is to be tested.

Each of the observations can be expressed now as

$$x_{ij} = m_{i.} + \epsilon_{ij} , \quad i=1,2,\dots,k ; j=1,2,\dots,n,$$

where the  $\epsilon_{ij}$  represent (random) "error" deviations of the observations,  $x_{ij}$ , from their respective sample means. This model can be further generalized to

$$x_{ij} = m.. + \alpha_i + \epsilon_{ij} , \quad i=1,2,\dots,k ; j=1,2,\dots,n.$$

where  $\alpha_i$  is the deviation of the  $i$ th sample mean,  $m_i$ , from the grand overall mean, and  $\epsilon_{ij}$  again is the deviation of the observation from its sample mean.

The hypothesis of equal population means which is to be tested could also be expressed as

$$H_0: \alpha_i = 0 \text{ for every } i.$$

The alternative to this hypothesis is

$$H_1: \alpha_i \neq 0 \text{ for some } i,$$

or in other words, at least one of the population means differs from the grand mean.

Before continuing to see how this hypothesis could be tested, it is important to consider the assumptions underlying the analysis of variance model and the practical importance of each.

### Classes of Analysis of Variance Problems

Two distinct classes of problems are solvable by analysis of variance (ANOVA). Although the calculus of the analysis in either case is the same, there are two major ways to interpret the results. If we use the analysis as a "fixed-effects model", we mean that we are comparing these  $k$  sample means and will draw conclusions about only the  $k$  population means involved in the analysis. If, as a result of the analysis, it is decided that there is a difference among the  $k$  means, we interpret this as meaning that at least one of the  $k$  population means differs from the others. In other words, that there is a difference of means among these  $k$  fixed treatments. On the other hand, if we consider this as a "random-effects model", we interpret a difference

among sample means as not indicating a difference among  $k$  fixed treatments only, but rather indicative of a difference among all possible treatments which could have been examined (e.g., all possible routes which led to work in the first example of comparing three particular routes). That is, we take this difference as an inference of fixed differences among individual treatments of a particular generic type.

Since the questions to be answered by the data are different in each of these two cases, the models are interpreted differently. Although the algebra involved is the same, the assumptions underlying each differ somewhat.

Assumptions

The algebraic procedure employed to construct the analysis of variance results is true, no matter what the numbers involved represent. Hence, it holds for both fixed- and random-effects models.

Consider  $k \cdot n$  numbers arranged in a matrix of  $k$  rows and  $n$  columns, and  $X_{ij}$  denotes the number occurring in the  $i$ th row and the  $j$ th column of this array. If we border the array with row means we obtain a configuration like the one below.

Table 2.

	1	2	...	j	...	n	row means
1	$X_{11}$	$X_{12}$	...	$X_{1j}$	...	$X_{1n}$	$\bar{X}_{1.}$
2	$X_{21}$	$X_{22}$	...	$X_{2j}$	...	$X_{2n}$	$\bar{X}_{2.}$
...	...	...	...	...	...	...	...
i	$X_{i1}$	$X_{i2}$	...	$X_{ij}$	...	$X_{in}$	$\bar{X}_{i.}$
...	...	...	...	...	...	...	...
k	$X_{k1}$	$X_{k2}$	...	$X_{kj}$	...	$X_{kn}$	$\bar{X}_{k.}$
							$\bar{X}_{..}$

### Fixed-effects Model

When the formulas and procedures of analysis of variance are used to summarize certain properties of the data on hand and nothing more, no assumptions are needed since it is just an employment of algebra calculating and comparing means. However, if from the data on hand in the samples inferences about properties of the "populations" from which the data was drawn are to be made, then certain assumptions about the populations, and about how the samples were obtained, must be made if the inferences are to be valid.

No statistical inferences can be made from the numbers  $X_{ij}$  unless they are assumed to be observations of random variables of some sort. So that must be the first assumption, that (1) the numbers  $X_{ij}$  are (observed values of) random variables that are distributed about mean values  $m_{ij}$ , ( $i=1,2,\dots,k; j=1,2,\dots,n$ ), that are fixed constants.

It is possible to arrange the parameters  $m_{ij}$  into a table form like Table 2 for the  $X_{ij}$ , bordered by the row means,  $m_i$ . It is apparent now that the value  $X_{12} - X_{52}$  gives an unbiased estimate of  $m_{12} - m_{52}$ , so now we can draw inferences from the data concerning the means of the populations from which the data was drawn. In fact, assumption 1 allows that the unbiased estimate of any linear combination of the  $m_{ij}$  is provided by the same combination of the  $X_{ij}$ .

If the true mean values  $m_{ij}$  in such a table are additive functions of the row means and grand mean, that is, if

$$\begin{aligned}
 m_{ij} &= (m_{ij} - m_{i.}) + m_{..} \\
 &= m_{..} + (m_{i.} - m_{..}) + (m_{ij} - m_{i.})
 \end{aligned}
 \tag{1}$$

then the inferences that can be drawn from the data are much more general. If equation (1) is satisfied by the model, that is, if equation (1) represents the relationship among the data, then the difference between any two row means,  $m_{1.} - m_{2.}$  for example, which is estimated by  $\bar{X}_{1.} - \bar{X}_{2.}$ , is a comprehensive estimate of the difference between the two row means. However, if equation (1) did not hold, then  $\bar{X}_{1.} - \bar{X}_{2.}$  would estimate the difference between the two row means for this configuration only, with the column-wise data ( $j=1,2,\dots,n$ ) in the particular order they are in and the rows in their present order. This is because if the additivity (equation (1)) does not hold, "interaction" effects between certain rows and columns are present, adding a hidden effect not represented in observation/row-mean deviation or in row-mean/grand-mean deviation.

Therefore, in order to be able to draw general inferences concerning sample, or row, population means, regardless of the particular order they happen to be in for the experiment, we assume that (2) the parameters  $m_{ij}$  (and hence estimates from the  $X_{ij}$ ) are related to the means  $m_{i.}$  and  $m_{..}$  as in equation (1), namely,

$$m_{ij} = m_{..} + (m_{i.} - m_{..}) + (m_{ij} - m_{i.}),$$

for  $i=1,2,\dots,k$  and  $j=1,2,\dots,n$ .

With assumptions 1 and 2 satisfied, the estimate of the difference between any two row means from the observations is an unbiased estimate of the general average difference between the two row populations concerned, regardless of column order or row order because we assume there are no non-additive effects, and the addi-

tive effects can be summed in any order.

If we want to be able to say something about the variance of the  $X_{ij}$ , and from that the variance of the population means and linear combinations of them, to get some idea of the precision of our estimates, we must go further with the assumptions.

In general, it is not possible to derive unbiased estimates of the variances of the  $X_{ij}$ , nor linear combinations of them either, using regular analysis of variance techniques unless assumptions 1, 2, and 3 given below, are satisfied.

We assume that (3) the random variables  $X_{ij}$  all have a common variance,  $\sigma^2$ , and that they are mutually uncorrelated.

Usually uncorrelatedness is a very reasonable assumption. This means that the amount of error in one observation does not affect the amount of error in another observation. A special approach, called randomization, is used to help ensure this uncorrelatedness. The experimenter selects experimental units at random and measures them separately. Hence, the error for any one sample is independent of that for any other sample.

Cochran and Cox [6] (pg. 8) make the following remark concerning randomization: "Randomization is somewhat analogous to insurance, in that it is a precaution against disturbances that may or may not occur and that may or may not be serious if they do occur. It is generally advisable to take the trouble to randomize even when it is not expected that there will be any serious bias from failure to randomize.

In order to have a simple analysis of variance table compu-

tationally it is desirable that the errors have the same variance from one population to another, and aren't effected by different "treatments" (row differences). This is probably one of the more critical assumptions, and one of the hardest to be sure of. Several tests for homogeneity of variance exist. One of the more common is Bartlett's test. This test is given in some detail in Ostle [34] .

When these three assumptions all are satisfied, an unbiased estimate of the difference between two row means, and the variance, can be calculated. If assumption 3 does not hold, then the covariances between the  $X_{ij}$  are not zero, and the estimates of the variances of combinations of the data become complex weighted averages of variances and covariances.

We now have a means by which the variances of row means, and other combinations of data may be identified and estimated, and so we have a method for judging whether real differences exist between population (row) means, which is the objective of the analysis of variance. However, we now have a means by which we can tell the accuracy, or significance, of our judging, and exactly how "sure" we are that differences among true row populations do exist. To be able to do this, i.e., assign some kind of quantitative probability level reflecting the "sureness" or significance of the variance estimates, we must know something of the joint distribution of the  $X_{ij}$ . Fortunately, "normality", in addition to assumptions one through three, allows us to make exact tests of significance. Therefore, we assume (4) that the  $X_{ij}$  are

jointly distributed in a multivariate normal distribution.

This assumption allows us to use such tests of significance as the t test and the F test, as will be shown later. This assumption is probably least likely to be completely true; however, much of analysis of variance can be used without using this assumption. The parts requiring normality have been shown to be fairly robust, that is, a fair amount of departure from normality can be tolerated without the accuracy being greatly affected.

Note that with assumption 4 made, assumption 1 is nearly covered, serving mainly now to define the means,  $m_{ij}$ . Also, the fact that the  $X_{ij}$  are assumed uncorrelated in assumption 3 taken together with the assumption that they are normally distributed in assumption 4 implies that the  $X_{ij}$  are mutually independent. Further results of these four assumptions will be shown as we continue with the analysis of variance model.

For the fixed-effects model, the basic assumptions, in summary, are:

- (1) The observations  $X_{ij}$  are (observed values of) random variables distributed about true means (expected values)  $m_{ij}$ , ( $i=1,2,\dots,k$ ;  $j=1,2,\dots,n$ ), which are fixed constants.
- (2) Additivity. That is, if we define the true grand mean as

$$\mu = \sum_i \sum_j m_{ij} / nk$$

and define a "row effect" as

$$\alpha_i = m_{i.} - \mu$$

then the parameters  $m_{ij}$  can be expressed as

$$m_{ij} = \mu + \alpha_i$$

(3) The random variables  $X_{ij}$  have a common variance  $\sigma^2$  (usually unknown).

(4) The  $X_{ij}$  are independently distributed in a multivariate normal distribution.

These assumptions are shown in construction of the model in the following way: If we assume that an observation may be represented as

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad ; \quad i=1,2,\dots,k, \quad j=1,2,\dots,n,$$

where

$$\sum_1 \alpha_i = 0 \quad (\mu \text{ is the grand mean}),$$

and the error terms  $\epsilon_{ij}$  are normally distributed with mean zero and a common variance  $\sigma^2$ , then we can validly apply analysis of variance techniques.

### Random-effects Model

In considering the random-effects model, although algebraically it is identical to the fixed-effects model, since different inferences are made from the results slightly different assumptions are made.

We assume (1) that the  $X_{ij}$  are, again, (observed values of) random variables distributed about a common mean value  $\mu$ , where  $\mu$  (defined as before) is a fixed constant.

(2) Additivity, that the random variables  $X_{ij}$  are sums of

component random variables, such that

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where  $\alpha_i$  and  $\epsilon_{ij}$  are both random variables.

(3) The random variables  $\alpha_i$  and  $\epsilon_{ij}$  are distributed with zero means, and variances  $\sigma_\alpha^2$  and  $\sigma^2$ , respectively.

(4) That the random variables  $\alpha_i$  and  $\epsilon_{ij}$  are independently and normally distributed.

### Results of Unsatisfied Assumptions

For a brief discussion of the consequences when certain of the assumptions are not satisfied, refer to Appendix A. For a more thorough discussion, refer to an article by Cochran [5], and to Cochran and Cox [6].

### Some Terminology

It will be useful to learn some conventional terminology which may help in describing the set-up of an experiment more precisely. The basic concept is that of a factor, which categorizes some property of the data according to which it -- the data -- will be classified. For example, in an agricultural experiment the factor may be fertilizer, to determine the effects it has on crop yield. In measuring the yield data, it would be classified according to the fertilizer used on that plot of ground. If the plots were in different parts of the country, then there would be (at least) two factors, the fertilizer and the climate, that we should consider. The term level refers to particular properties defining subgroups of the factor groups. Thus, the levels of the factor fertilizer would be the different fertili-

zers used. The levels for the climate factor might be set up as the regions in which the plots were located, e.g., Pacific Northwest, Southwest, etc., or they might be set up in some other manner, e.g., dry, moderately dry, humid, etc.

We can describe the structure of an experiment, or the experimental design, by describing the factors and the way in which the levels of the different factors are combined. In the example in which the driver compared driving times taking three different routes to work, there was only one factor we were considering, the route taken. The levels were the three different routes, route 1, route 2, route 3. Hence, we had a one-factor experiment with three levels.

### Preparation

#### Setting Up the Experiment

First, a realistic model must be set up so that the observations,  $X_{ij}$ , whatever they represent, can be obtained and are in a form that can be used in the analysis of variance techniques. (For example, in an experiment where results are obtained as shades of a certain color, these results would have to be quantified before they could be used.) The detail of how the experiment is then going to be run, once the model is decided on, should be outlined.

(1) The objectives of the experiment should be clearly defined. For instance, if the experiment is a preliminary one to determine what future experiments should be like, or if it is to get answers to immediate questions. Is it mainly to get "ball-park" estimates or is the experimenter mainly interested in tests

of significance (accuracy)? Over what range of conditions are the results to be extended?

(2) The experiment should be described in detail. The different factors to be considered, the levels the experimenter wants to test for each factor, the size of the experiment, and the material necessary to complete the experiment.

(3) An outline of the analysis to be done would be helpful to have as a guide before the experiment is started so that the data necessary to the analysis is obtained.

### Running the Experiment

The experimental techniques should be refined as much as possible.

(1) There should be a uniform method of applying different treatments to the experimental units.

(2) Control over external influences should be exercised as much as possible so that every treatment operates under as nearly the same conditions as possible.

(3) Unbiased methods of measuring results should be devised so that the results are as objective as possible and can be compared with results from other experiments. As an example, it is difficult to measure objectively educational progress, social standing, or socio-economic levels since these are by nature subjective judgements.

(4) If possible, checks should be set up to avoid making, and admitting to analysis, gross errors in experimental measurement, since one or two such errors could bias the entire results.

Further information along this line is included in many texts, including Cochran and Cox [6], Johnson and Leone [29], Snedecor [36], Ostle [34], or other standard statistics texts.

### Analysis of Variance

We will consider our models, and examples, as being fixed-effects models. Some authors vary notation to indicate which model they are working with. Notation here is fairly consistent with that used in Probability and Statistics for Engineers by Miller and Freund [32].

#### One-way Classification

The objective, once again, is to determine if the means of the populations from which the different samples have been taken are equal. We approach this by analyzing the variance among the sample, or treatment, means and attempting to judge whether the amount of variance is due only to chance variation, in selecting a sample from the population, or indicative of actual differences among means of the populations. Recall that our null hypothesis,  $H_0$ , was that the true (row, or treatment) population means were equal,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k = \mu \quad (\text{grand mean}).$$

To test the hypothesis of equal means, we first need to find the total variability of the combined data. This quantity is usually referred to as the "total sum of squares", and is denoted by SST.

$$SST = \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2$$

where  $\bar{X}_{..}$  is the overall mean of all observations,

$$\bar{X}_{..} = \frac{1}{(n \cdot k)} \sum_i \sum_j X_{ij} \quad .$$

If  $H_0$  is true, the SST is totally due to chance variation. If  $H_0$  is not true, then there is a contribution to variation due to differences among the means of the  $k$  populations.

To test this hypothesis that the  $k$  population means are equal, we compare two estimates of the variance  $\sigma^2$ , the variance of the observations  $X_{ij}$ . One estimate is based on the variation between the sample (or treatment) means, and one is based on the variation within the samples.

If the hypothesis is true that the population means are equal, then these are estimates of the same value,  $\sigma^2$ , and should be approximately the same. However, if the hypothesis is not true and in fact there is a difference between population means, this difference will "inflate" the estimate based on variation among sample means, making it larger than the estimate based only on variation within the samples. Hence, if the null hypothesis is false, that is, if all population means are not equal, we would expect the "between sample" estimate,  $s_B^2$  say, to exceed the "within sample" estimate,  $s_W^2$ . Forming a ratio of these two values, we reject the hypothesis, or say there is evidence to indicate the means are not all equal, if the ratio  $s_B^2/s_W^2$  is "too large", i.e., if the between sample estimate is larger than the within sample estimate by enough to indicate a difference of means among the populations.

The problem now is to determine what is "too large." Before

approaching this problem, let us look at the two variance estimates,  $s^2_B$  and  $s^2_W$ .

Since, by assumption, each sample comes from a population having variance  $\sigma^2$ , this variance can be estimated by any one of the sample variances ( $i=1,2,\dots,k$ )

$$s^2_{W_i} = \sum_{j=1}^n (X_{ij} - \bar{X}_{i.})^2 / (n-1)$$

and, hence, this variance is also estimated by the mean of all these  $k$  estimations,

$$s^2_W = \sum_i s^2_{W_i} / k = \sum_i \sum_j (X_{ij} - \bar{X}_{i.})^2 / k(n-1) .$$

The variance of the sample means is given by

$$s^2_{\bar{X}} = \sum_i (\bar{X}_{i.} - \bar{X}_{..})^2 / (k-1) ,$$

and if the null hypothesis is true and there is no difference among population means and this variation among sample means is due to chance variation, it estimates  $\sigma^2/n$ . (See Appendix B). Thus, an estimate of  $\sigma^2$  based on variation among the sample means is given by

$$s^2_B = n \cdot s^2_{\bar{X}} = n \cdot \sum_i (\bar{X}_{i.} - \bar{X}_{..})^2 / (k-1) .$$

It can be shown that we now have two independent estimates of the variance  $\sigma^2$ .

The next fact is what allows analysis of variance to work. The two estimates of the variance shown above can be obtained (except for the divisors  $(k-1)$  and  $k(n-1)$ ) by "breaking up" the total variance of the combined data into two parts. The total variance for the combined data is estimated by

$$s^2_T = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 / (nk-1) .$$

The decomposition of this total sum of squares (without denominators)

is shown in the theorem below.

$$SST = \sum_1 \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_1 \sum_j (X_{ij} - \bar{X}_{i.})^2 + n \sum_1 (\bar{X}_{i.} - \bar{X}_{..})^2$$

where  $\bar{X}_{i.}$  is the mean of the sample from the  $i$ th population, as defined before. (Proof of theorem in Appendix C).

What is remarkable is that this decomposition means that the total sum of squares is broken into component sums of squares which are, themselves, squares of, or sums of squares of, linear combinations of the observations,  $X_{ij}$ , that sum mutually distinct properties of the data, so that one sum is independent of the other. (See Appendix D).

Referring to the decomposition above, we see by inspection the first term on the right gives the variation due to differences of observations from their sample means, i.e., within sample variation. This is usually referred to as "within" variation, or error sum of squares, SSE. The term "error" expresses the idea that the quantity estimates random, or chance, error variation. The second term on the right gives the variation among sample, or treatment, means. This is referred to as the "between" variation, or the treatment sum of squares, SS(Tr).

Thus, we have partitioned the total variability of the combined data into two components: The first, SSE, measures chance variation within samples; regardless of whether the null hypothesis is true or not, it only measures variation due to chance, and is an unbiased estimator of  $\sigma^2$ . The second, SS(Tr), also measures chance variation when the hypothesis is true, but it is affected by the added variation among the population means when the null hypothesis is false.

In the variance estimators  $s_B^2$  and  $s_W^2$ , the divisors,  $(k-1)$  and  $k(n-1)$ , are called degrees of freedom. They represent here the number of independent observations of the data,  $X_{ij}$ . It can be shown (see Hoog and Craig [27] or Anderson and Bancroft [1] or Larson [31]) that the ratios (SSE/deg. of freedom) and (SS(Tr)/deg. of freedom) are independent, and both have  $\chi^2$  (Chi-square) distributions. It is this fact that allows us to make tests of significance. Therefore,

$$\frac{\left[ \frac{\text{SS(Tr)}}{\text{trt. deg. of free.}} \right]}{\left[ \frac{\text{SSE}}{\text{error deg. of free.}} \right]} = \frac{\left[ \frac{\text{SS(Tr)}}{k-1} \right]}{\left[ \frac{\text{SSE}}{k(n-1)} \right]} \sim F(k-1, k(n-1)) ,$$

where  $F$  is the Snedecor  $F$  distribution, with  $k-1$  and  $k(n-1)$  degrees of freedom as the two parameters, and where " $\sim$ " means "is distributed as." This is true because the ratio of two independent  $\chi^2$  distributions has an  $F$  distribution (see references given above). We can now test this ratio of variance estimates to see if it is "too large" by comparing it to a tabled value of the  $F$  distribution, to see if it is large enough to indicate a difference of population means.

An analysis of variance table is shown below for a one-factor experiment.

Source of variation	degrees of freedom	sums of squares	mean squares	F values
treatments	$k-1$	SS(Tr)	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$\frac{MS(Tr)}{MSE}$
error	$k(n-1)$	SSE	$MSE = \frac{SSE}{k(n-1)} = \hat{\sigma}^2$	
Total	$nk-1$	SST		

where  $\hat{\sigma}^2$  is the estimate of the population variance  $\sigma^2$ .

This analysis is referred to as one-way analysis of variance, or one-way classification, or one-factor analysis. This term (or terms) expresses the idea that we are studying the effect of only one source of variation (other than chance), namely, the possible added variation due to differences of treatment means.

### Two-way Classification

It is possible still, however, that SSE, the variability which we ascribe to chance (or experimental error) may actually be "inflated" by other identifiable sources of variation, other than just chance. This suggests an extension of the analysis applied to the one-factor experiment to experiments with more than one factor. As an example of this extension, we will consider a two-way analysis of variance, or two factor analysis, in which the total variability of the data is partitioned into one component which we ascribe to possible differences due to one factor (the treatments), and a second component which is ascribed to possible differences in a second factor (usually referred to as blocks, again due to the origin of this method in agriculture), and the remainder of the variability is ascribed to chance. We have broadened the model now to consider two factors. The treatments are levels of one factor, and the blocks are levels of the second factor. This lay-out could take the form illustrated on the following page.

It is possible to have only one value in each "cell", or multiple values in each cell, called "replicates." We do it here for the slightly less general case of one replicate per cell.

	Blocks				treatment means
	1	2	...	n	
treatment 1	$X_{11}$	$X_{12}$	...	$X_{1n}$	$\bar{X}_{1.}$
treatment 2	$X_{21}$	$X_{22}$	...	$X_{2n}$	$\bar{X}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
treatment k	$X_{k1}$	$X_{k2}$	...	$X_{kn}$	$\bar{X}_{k.}$
Block means	$\bar{X}_{.1}$	$\bar{X}_{.2}$	...	$\bar{X}_{.n}$	$\bar{X}_{..}$ grand mean

If  $X_{ij}$ , ( $i=1,2,\dots,k$ ;  $j=1,2,\dots,n$ ), are values of independent random variables having normal distributions with respective true means  $m_{ij}$ , and a common variance  $\sigma^2$ , we can write the model for a two-way analysis of variance (assuming there is no interaction between block effects and treatment effects) as

$$X_{ij} = \mu + \alpha_i + \delta_j + \epsilon_{ij}$$

where  $\mu$  is the overall grand mean

$\alpha_i$  is the  $i$ th treatment effect

$\delta_j$  is the  $j$ th block effect

$\epsilon_{ij}$  is the error component,

so that  $\mu_{ij} = \mu + \alpha_i + \delta_j$  ;  $\mu_{ij} \equiv m_{ij}$  .

Again, we assume additivity, so that

$$\sum_i \alpha_i = \sum_j \delta_j = 0 \quad .$$

Two ratios must now be tested. We wish to compare "between treatment" variation to "within" variation again to determine if there is a possible difference among treatment means, and we also wish to compare "between block" variation to "within" variation to determine if there is a possible contribution to variability

due to differences among block means.

For two-factor analysis, the total sum of squares is decomposed into

$$\begin{aligned} SST = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 &= \sum_i \sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2 \\ &+ n \sum_i (\bar{X}_{i.} - \bar{X}_{..})^2 + k \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2 \end{aligned}$$

(Proof in Appendix E).

The terms on the right hand side of the equation are, first, the new SSE with possible block effects removed, secondly, the treatment sum of squares,  $SS(\text{Tr})$ , and the third term is the block sum of squares,  $SSB$ , which gives a measure of the variation of the  $\bar{X}_{.j}$ , the block means.

We now have

$$SST = SS(\text{Tr}) + SSB + SSE ,$$

and it can be shown, as before, that the ratios ( $SS(\text{Tr})/\text{trt d.f.}$ ), ( $SSB/\text{block d.f.}$ ), and ( $SSE/\text{error d.f.}$ ), are mutually independent and distributed as  $\chi^2$  with the respective degrees of freedom. (See Hoog and Craig [27]). We now can test the null hypotheses that  $\alpha_i = 0$  for every  $i$  &  $\gamma_j = 0$  for every  $j$ ; that is, all treatment means are equal and not a source of added variation, and all block means are equal and not an added source of variation. To test this, the null hypotheses are rejected if

$$F_{\text{tr}} > F(\alpha/2, k-1, (n-1)(k-1)) \text{ and}$$

$$F_b > F(\alpha/2, n-1, (n-1)(k-1)),$$

where

$$F_{\text{tr}} = \frac{\left[ \frac{SS(\text{Tr})}{k-1} \right]}{\left[ \frac{SSE}{(n-1)(k-1)} \right]} \quad F_b = \frac{\left[ \frac{SSB}{n-1} \right]}{\left[ \frac{SSE}{(n-1)(k-1)} \right]}$$

and  $F(\frac{\alpha}{2}, k-1, (n-1)(k-1))$  and  $F(\frac{\alpha}{2}, n-1, (n-1)(k-1))$  are tabled values of the F distribution with the listed degrees of freedom and  $\alpha$  denoting the significance level of the test.

### Example

An example of two-way analysis of variance with a single replicate in each cell would perhaps be helpful, and is presented here.

Suppose we wish to compare several drill bit designs for speed of drilling core samples. Since the geology of the area in which the drilling is taking place could effect the speed of the drill, perhaps even more than the design of the drill bit, we should consider two factors: drill bit design and the geology of the different test areas. Each of four bit designs were tested in five different types of deposits including clay, shale, sandstone, limestone, and oolite. The times in minutes to drill a core of specified depth are recorded in the table below.

		Deposit					
		A	B	C	D	E	Totals
Bit design	1	22	26	25	25	31	129
	2	25	27	28	26	29	135
	3	26	29	33	30	33	152
	4	26	28	27	30	30	141
Totals		99	110	113	111	123	556

Drilling time in minutes for four different bit designs in five different deposits.

Using the short cut formulas for calculations, we get the results listed on the following page.

$$\begin{aligned} SST &= \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = \sum_i \sum_j x_{ij}^2 - [(x_{..})^2 / n \cdot k] \\ &= 22^2 + 26^2 + \dots + 30^2 - \frac{556^2}{5 \cdot 4} \\ &= 153.2 \end{aligned}$$

$$\begin{aligned} SS(\text{Tr}) &= k \cdot \sum_i (\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_i x_{i.}^2 / n - [(x_{..})^2 / n \cdot k] \\ &= \frac{99^2 + \dots + 123^2}{5} - \frac{556^2}{20} \\ &= 52.8 \end{aligned}$$

$$\begin{aligned} SSB &= n \cdot \sum_j (\bar{x}_{.j} - \bar{x}_{..})^2 = \sum_j x_{.j}^2 / k - [(x_{..})^2 / n \cdot k] \\ &= \frac{129^2 + \dots + 141^2}{4} - \frac{556^2}{20} \\ &= 73.2 \end{aligned}$$

$$\begin{aligned} SSE &= SST - SS(\text{Tr}) - SSB = 153.2 - 52.8 - 73.2 \\ &= 27.2 \end{aligned}$$

The analysis of variance table is constructed below.

Source of variation	degrees of freedom	sums of squares	mean square	F values
treatments (bits)	3	52.8	17.6	$\frac{17.6}{2.27} = 7.75$
blocks (deposits)	4	73.2	18.3	$\frac{18.3}{2.27} = 8.06$
error	12	27.2	2.27	
Total	19	153.2		

Since  $F_{tr} = 7.75$  exceeds  $F(.05, 3, 12) = 3.49$  and  $F_b = 8.06$  exceeds  $F(.05, 4, 12) = 3.26$ , we find that both hypotheses must be rejected. The differences between the means for the four bits are significant enough to indicate it is more than a result of

chance variation, and so are the differences between mean times for the different deposits. Note, however, that in each case all we have discovered is that an apparent difference among the means exists. To determine which mean or means differ from the rest, and if they are higher or lower than average, we would need to use multiple comparison techniques. (See David [11]).

This completes a brief introduction to some of the basic ideas of analysis of variance. These techniques, along with regression analysis are used in analysis of covariance.

ANALYSIS OF COVARIANCEIntroduction

As Sir R. A. Fisher, who first published this technique, expressed it, the analysis of covariance "combines the advantages and reconciles the requirements of the two very widely applicable procedures known as regression and analysis of variance."

This combination results in a more discriminating analysis than would be afforded by the straight application of analysis of variance.

Before discussing actual computational methods, an example of a situation where this technique would be applicable may be helpful. Suppose we are conducting an experiment to compare several methods of teaching by employing these different methods to teach the same material to different classes and the criterion for judgement of the methods is to be the final score,  $Y$ , obtained by the students, all of whom take the same examination at the end of the course. Before we judge the various methods of teaching, however, we realize that final scores could also be influenced by the intelligence of the students, and we might end up thinking one method superior when actually the results were better only because all the smart students were under that method. To eliminate this, we might want to adjust the  $Y$  values (scores) according to associated values,  $X$ , I.Q. ratings say, so that we could examine all the adjusted scores as being from stu-

dents with a standard (equal) intelligence (I.Q). We could then examine the data, the adjusted Y values, and ascribe any variation between groups to differences in teaching methods, having standardized the variation of intelligence. The method by which this adjustment and subsequent analysis is carried out is called analysis of covariance.

More generally, it sometimes happens that we wish to perform an analysis of variance on observations of the random variable Y; however, some additional factor, X, (or several additional factors) varies during the period in which observations were made on Y and, hence, any dependence of Y on X will tend to obscure, and possibly even make useless, any analysis of variance performed only on the original dependent variable Y, since its variation may be inflated by variation in the independent variable X, due to the dependence of Y on X. In such a case, then, we would assume a relationship between the variables and any analysis of the Y values would be preceded by an adjustment of the Y values to some standard value of the additional variable X, thus eliminating variation among the X values. Frequently, however, such a relationship is not known beforehand and must be estimated from the data present.

Conceptually, we wish to determine the regression relationship, correct the observations by means of this relationship to some standard condition, and examine the corrected values by the analysis of variance.

## Simple Covariance

### Introduction

Since analysis of covariance, even by its name, represents a combination of correlation analysis and analysis of variance, let us consider a problem where each technique is applied separately and then they are combined to perform analysis of covariance.

The term "simple covariance" means that for each value of  $Y$ , the variable of primary interest, we are examining only one additional variable, called the covariate or concomitant variable,  $X$ , corresponding to each  $Y$  value.

### Example

Suppose a company makes steel brackets which it sends to three different firms to have chrome plated. The characteristic we are interested in checking is the thickness of the chrome plating, the concern being to judge whether the three firms are plating the brackets equally.

To test this hypothesis that all the three firms plate the brackets with the same average thickness (the average of the thickness applied to all brackets), four brackets are sent to each of the firms. Data for the thicknesses of plating for these brackets from each of the three firms is recorded below.

### ANOVA Model

We can apply one-way (one-factor) analysis of variance to test the hypothesis that the plating thickness means from each of the three firms are equal.

		Brackets			
Firm	A	40	38	30	47
	B	25	32	13	35
	C	27	24	20	13

Chrome plating thicknesses in thousandths of an inch, from the 3 plating firms.

The analysis of variance model would be

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}^a \quad (1)$$

where  $Y_{ij}$  represents the thickness of plating on the (ij)th bracket,  $\mu$  is the overall mean,  $\alpha_i$  is the treatment, or firm, effect, and  $\epsilon_{ij}^a$  is the random error term for the analysis of variance model.

If we make the assumptions necessary ( $Y_{ij}$  random variables with true means, additivity, and the  $\epsilon_{ij}$  are independently and normally distributed with mean zero and common variance  $\sigma^2$ ) an analysis of variance can be performed validly and the table is constructed below.

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F values
Between (trtmnt)	2	665.2	332.6	5.51
Within (error)	9	543.5	60.4	
Total	11	1208.7		

The F value of 5.51 is greater than the tabled F value for  $F(.05, 2, 9)$ , showing the variance among the plating firm means is sufficiently larger than the "error" variance to indicate a real difference among the average plating thicknesses of the three firms.

However, in studying the results it was noticed that there might be some correlation between the thickness of the bracket and the thickness of the plating put on it. When the data is plotted, it appears as below.

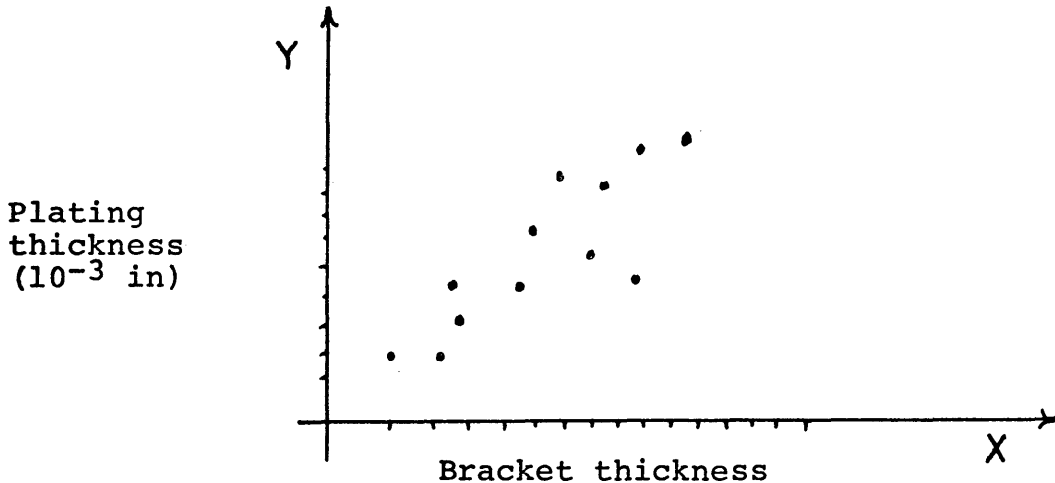


Figure 4.

The diagram indicates a positive correlation between the bracket thickness and the plating thickness.

The Problem: Hence, in analyzing variation among plating thicknesses, there is a contribution to the variance from variability among bracket thicknesses. That is, the variance in the ANOVA model we thought was attributable to differences among the plating techniques of the three firms may really be due to differences in the thicknesses of the brackets sent to them. This

is the problem mentioned earlier in which the variable of interest,  $Y$ , is dependent on another variable,  $X$ , which varies during the experiment. This dependence of  $Y$  on  $X$  causes  $Y$  to vary during the experiment also, "inflating" the variance of  $Y$  and obscuring any results from the analysis of that variance.

Approach to the Problem: This problem is susceptible to analysis because this (possible) source of added variation from the covariate  $X$  is identifiable and separable, using analysis of covariance.

#### Regression Model

Since for every plating thickness that is measured it is possible to measure the thickness of the bracket before it is plated, we can graph a set of paired values as shown in Figure 4. It is now possible, using regression analysis, to determine how the plating thickness,  $Y$ , varies with the bracket thickness,  $X$ .

A (linear) regression model would be

$$Y_{ij} = \mu + \beta(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}^r \quad (2)$$

where  $Y_{ij}$  is the thickness of the plating on the (ij)th bracket,  $\mu$  is the overall plating thickness mean,  $X_{ij}$  is the thickness of the (ij)th bracket,  $\epsilon_{ij}^r$  is the random error term for the regression model, and  $\beta$  is the true slope or linear regression coefficient between  $Y$  and  $X$ .  $\bar{X}_{..}$  is the mean bracket thickness.

#### Analysis of Covariance Model

It is possible to combine these two models, equations (1) and (2), to form the analysis of covariance model to determine variance due to treatment differences and variance due to var-

iation in bracket thicknesses, the covariate.

So model (1)

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}^a \quad (\text{ANOVA})$$

and model (2)

$$Y_{ij} = \mu + \beta(x_{ij} - \bar{X}_{..}) + \epsilon_{ij}^r \quad (\text{Reg.})$$

are combined to give the covariance model

$$Y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{X}_{..}) + \epsilon_{ij} \quad (3)$$

The error term for this model,  $\epsilon_{ij}$ , should be smaller than in the original analysis of variance model because some of the variation included there in  $\epsilon_{ij}^a$  has now been identified in the new regression term  $\beta(x_{ij} - \bar{X}_{..})$ . The superscripts were just to mark the fact that the error terms were not the same in the three different models.

Equation (3) can be written in a different way to aid in understanding its function. From the viewpoint of analysis of variance, we could write

$$[Y_{ij} - \beta(x_{ij} - \bar{X}_{..})] = \mu + \alpha_i + \epsilon_{ij} \quad (4)$$

In this form, (4) is the typical equation for analysis of variance of the quantities

$$[Y_{ij} - \beta(x_{ij} - \bar{X}_{..})] .$$

These quantities represent the deviations of the  $Y_{ij}$  from their linear regression on  $X_{ij}$ , that is, the values of  $Y_{ij}$  after adjustment for this linear regression. From this form, it is apparent that  $\alpha_i$  is the true effect of the  $i$ th treatment on  $Y_{ij}$ , after adjustment for variation in the  $X_{ij}$ . Thus, the analysis of covariance technique enables us to remove the part of an ob-

served treatment effect which can be attributed to a linear dependence on  $X_{ij}$ .

### Application of the Model

To determine just exactly how to remove from estimates of variance the variation due to the covariate  $X$ , the model (3) must be rearranged some. Recall that the regression model can be written in "deviation" form,

$$y_{ij} = bx_{ij} + e_{ij}$$

where  $y_{ij} = Y_{ij} - \bar{Y}..$ ,  $x_{ij} = X_{ij} - \bar{X}..$ ,  $e_{ij}$  is the residual, and  $b$  is the sample slope (estimate of  $\beta$ ). In the chapter on regression it was shown that

$$b = \frac{\sum \sum x_{ij} y_{ij}}{\sum \sum x_{ij}^2} = \frac{\sum xy}{\sum x^2}$$

and that

$$\sum e_{ij}^2 = \sum y^2 - [(\sum xy)^2 / \sum x^2] \quad (5)$$

where  $\sum$  represents summation over all points.

In this expression, the term  $[(\sum xy)^2 / \sum x^2]$  is the amount of reduction in the sum of squares of the  $Y$  variable due to its linear dependence (regression) on  $X$ . Therefore, this term may be removed from the variance estimate (for the  $Y$  population) because it is due to variation in the covariate. So if a term of this type is subtracted from the sum of squares of the dependent variable,  $Y$ , which estimates the variance in the  $Y$  population, then the result will be a corrected or adjusted estimate of the variance of  $Y$  which cannot be explained by, nor is dependent on, variation in the concomitant variable  $X$ .

Example Application

We will use this approach on the bracket problem. Below is the data on both variables for the previous problem.

		Brackets								Totals	
		1		2		3		4			
		X	Y	X	Y	X	Y	X	Y	X	Y
Firm (trtmnt)	A	110	40	75	38	93	30	97	47	375	155
	B	60	25	75	32	38	13	140	35	313	105
	C	62	27	90	24	45	20	59	13	256	84

Bracket thickness (X) and plating thickness (Y) for three plating firms.

The calculations of sums of products follow the general pattern of the analysis of variance. Recall that in the ANOVA model

$$SST = \sum y^2$$

$$SS(Tr) = n \cdot \sum_1 (Y_{i.} - \bar{Y}_{..})^2$$

$$SSE = SST - SS(Tr) \quad .$$

We perform each of these calculations for X, Y, and their product, introducing some new notation in the process.

The total sums of squares are:

$$\begin{aligned} \sum y^2 &= \sum_{ij} Y_{ij}^2 - [(Y_{..})^2 / n \cdot k] \\ &= 40^2 + 38^2 + \dots + 13^2 - 344^2 / 12 = 1208.7 \end{aligned}$$

$$\begin{aligned} \sum x^2 &= \sum_{ij} X_{ij}^2 - [(X_{..})^2 / n \cdot k] \\ &= 110^2 + 75^2 + \dots + 59^2 - 944^2 / 12 = 9240.7 \end{aligned}$$

$$\begin{aligned} \sum xy &= \sum_{ij} X_{ij} Y_{ij} - [(X_{..})(Y_{..}) / n \cdot k] \\ &= 110 \cdot 40 + 75 \cdot 38 + \dots + 59 \cdot 13 - (944)(344) / 12 = 2332.7 \end{aligned}$$

The treatment sums of squares are:

$$\begin{aligned} T_{YY} &= \sum_i Y_{i.}^2/k - (Y_{..})^2/n \cdot k \\ &= \frac{155^2 + 105^2 + 84^2}{4} - \frac{344^2}{12} = 665.2 \end{aligned}$$

$$\begin{aligned} T_{XX} &= \sum_i X_{i.}^2/k - (X_{..})^2/n \cdot k \\ &= \frac{375^2 + 313^2 + 256^2}{4} - \frac{944^2}{12} = 1771.2 \end{aligned}$$

$$\begin{aligned} T_{XY} &= \sum_i (X_{i.})(Y_{i.})/k - (X_{..})(Y_{..})/n \cdot k \\ &= \frac{375 \cdot 155 + 313 \cdot 105 + 256 \cdot 84}{4} - \frac{344 \cdot 944}{12} = 1062.2 \end{aligned}$$

The error sums of squares are:

$$\begin{aligned} E_{YY} &= \text{Total S.S.} - \text{Treatment S.S.} = \sum y^2 - T_{YY} \\ &= 543.5 \end{aligned}$$

$$E_{XX} = \sum x^2 - T_{XX} = 7469.5$$

$$E_{XY} = \sum xy - T_{XY} = 1270.5$$

Using this information, the sum of squares for the dependent variable, Y, can be adjusted for dependence on X. For the total sum of squares, the adjusted sums are (using equation (5))

$$SST^* = \left[ \text{adj. } \sum Y^2 \right] = \sum Y^2 - (\sum xy)^2 / \sum x^2 .$$

The adjusted sum of squares within treatments is

$$SSE^* = \left[ \text{adj. } E_{YY} \right] = E_{YY} - (E_{XY})^2 / E_{XX} ,$$

and by subtraction,

$$SS(\text{Tr})^* = \left[ \text{adj. } T_{YY} \right] = SST^* - SSE^* .$$

So, for the bracket example,

$$\begin{aligned} SST^* &= \text{adj. } \sum Y^2 = 1208.7 - [(2332.7)^2 / 9240.7] \\ &= 619.8 \end{aligned}$$

$$\begin{aligned} \text{SSE}^* &= \text{adj. } E_{yy} = 543.5 - [(1270.5)^2/7469.5] \\ &= 327.4 \end{aligned}$$

$$\text{SS(Tr)}^* = \text{SST}^* - \text{SSE}^* = 619.8 - 327.4 = 292.4$$

Arranging this in an analysis of covariance table, we get

Source of variation	d.f.	sums of products of			d.f.	adj. sums	M.S.
		x,x	x,y	Y,Y			
Total	11	9240.7	2332.7	1208.7			
Trtmnts	2	1771.2	1062.2	665.2			
Error	9	7469.5	1270.5	543.5	8	327.4	40.9
Trtmnt + Error					10	619.8	
Adj. trtmnt -- SS(Tr)*					2	292.4	146.2

The new adjusted F value is

$$F = \frac{\text{SS(Tr)}^*/\text{d.f.}}{\text{SSE}^*/\text{d.f.}} = \frac{\text{MS(Tr)}^*}{\text{MSE}^*} = \frac{146.2}{40.9} = 3.57 \quad F(2,8)$$

which is not significant. Hence, for this example we see that variation originally attributed to differences in plating techniques, leading us to believe a difference among treatment means existed, was actually due to variation in the covariate, X. When this variation is removed, the amount of variation left among the Y values indicates that no treatment mean differences exist after all.

#### Regression Slopes Used in Adjustment

Looking at the application of the covariance analysis, three different regressions of Y on X can be seen:

- 1 - The overall regression of all Y's on all X's.
- 2 - The within treatment regressions (assumed to be the same regression in each treatment, by setting each equal to the average).
- 3 - The regression of the Y group means on the X group means.

The "average within-treatment" regression slope is found by

$$b = \frac{E_{xy}}{E_{xx}} .$$

In adjusting the error sum of squares by this slope, and the total (treatment + error) sum of squares by the "overall" regression slope, given by

$$b = \frac{\sum xy}{\sum x^2} ,$$

we assume the two are approximately equal, that is, that the "pooled" or "average" within treatment regression slope is the same as the overall regression slope. The third regression, means on means, is not used.

The results of the covariance analysis are shown graphically below, using the chrome plating example to illustrate the adjustments. To examine the variance among treatment means adjusted for variation in X, the three treatment means are "slid" along lines parallel to the slope of the regression line. With this adjustment it is seen that by using analysis of covariance we are viewing the different treatment means as they would have been had all observations of Y been made at a standard value of

X, namely  $\bar{X}..$ . A comparison shows the adjusted means for this example are closer together than the unadjusted means.

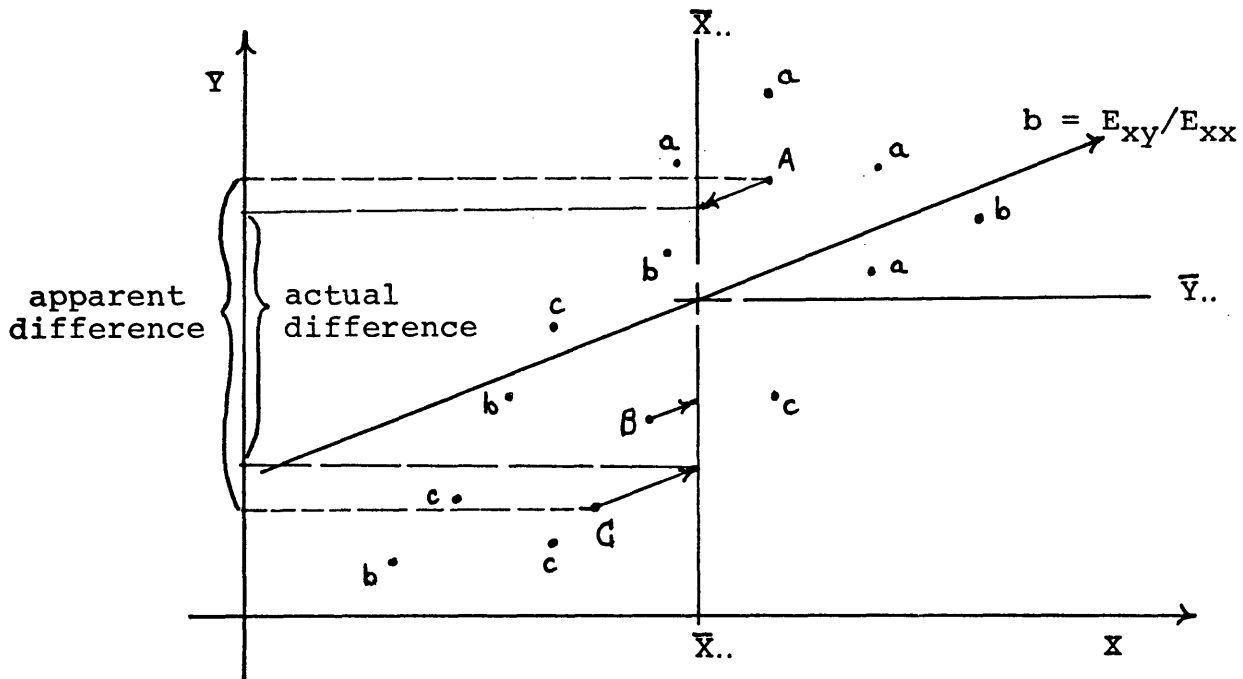


Figure 5. Plating thickness (Y) vs. bracket thickness (X). Lower case letters are values from that firm. Capital letters represent mean values by plating firms.

General Computational Procedure

Given here is a general computational approach as applied to a two-factor analysis design. The approach is general enough to illustrate the method employed for the completely general case.

If the data is put in table form it appears as

Block

		1	2	...	n	row sums	
Trtmnt	1	X <sub>11</sub> Y <sub>11</sub>	X <sub>12</sub> Y <sub>12</sub>	...	X <sub>1n</sub> Y <sub>1n</sub>	X <sub>1.</sub> Y <sub>1.</sub>	
	2	X <sub>21</sub> Y <sub>21</sub>	X <sub>22</sub> Y <sub>22</sub>	...	X <sub>2n</sub> Y <sub>2n</sub>	X <sub>2.</sub> Y <sub>2.</sub>	
	...	...	...	...	...	...	
	k	X <sub>k1</sub> Y <sub>k1</sub>	X <sub>k2</sub> Y <sub>k2</sub>	...	X <sub>kn</sub> Y <sub>kn</sub>	X <sub>k.</sub> Y <sub>k.</sub>	
Col. sums		X <sub>.1</sub> Y <sub>.1</sub>	X <sub>.2</sub> Y <sub>.2</sub>	...	X <sub>.n</sub> Y <sub>.n</sub>	X <sub>..</sub> Y <sub>..</sub>	grand sum

$n$  is the number of blocks (levels of factor 2) and  $k$  is the number of treatments (levels of factor 1).

An analysis of covariance table is set up by first constructing the regular analysis of variance table and deriving, in addition to this, the sums of squares and cross products of  $X$  and  $Y$  for each line in the analysis of variance table. The analysis of covariance table for testing for adjusted treatment means is constructed below for the two-factor design. Instead of blocks and treatments, the factors could be referred to as factor A and factor B as a more general heading, making extensions of this table more obvious. The notation is given below for the general case shown in the covariance table, Figure 6.

The total sums of squares are:

$$\begin{aligned}\sum x^2 &= \sum_i \sum_j x_{ij}^2 = \sum \sum (x_{ij} - \bar{x}_{..})^2 = \sum_i \sum_j x_{ij}^2 - [(x_{..})^2 / n \cdot k] \\ \sum y^2 &= \sum_i \sum_j y_{ij}^2 = \sum \sum (y_{ij} - \bar{y}_{..})^2 = \sum_i \sum_j y_{ij}^2 - [(y_{..})^2 / n \cdot k] \\ \sum xy &= \sum_i \sum_j x_{ij} y_{ij} = \sum \sum (x_{ij} - \bar{x}_{..}) (y_{ij} - \bar{y}_{..}) = \sum_i \sum_j y_{ij} x_{ij} - [x_{..} y_{..} / nk]\end{aligned}$$

The block sums of squares are:

$$\begin{aligned}B_{xx} &= n \sum_j (x_{.j} - \bar{x}_{..})^2 = \sum_j x_{.j}^2 / k - [(x_{..})^2 / nk] \\ B_{yy} &= n \sum_j (y_{.j} - \bar{y}_{..})^2 = \sum_j y_{.j}^2 / k - [(y_{..})^2 / nk] \\ B_{xy} &= n \sum_j (x_{.j} - \bar{x}_{..}) (y_{.j} - \bar{y}_{..}) = \sum_j x_{.j} y_{.j} / k - [(x_{..}) (y_{..}) / nk]\end{aligned}$$

The treatment sums of squares are:

$$\begin{aligned}T_{xx} &= k \sum_i (x_{i.} - \bar{x}_{..})^2 = \sum_i x_{i.}^2 / n - [(x_{..})^2 / nk] \\ T_{yy} &= k \sum_i (y_{i.} - \bar{y}_{..})^2 = \sum_i y_{i.}^2 / n - [(y_{..})^2 / nk] \\ T_{xy} &= k \sum_i (x_{i.} - \bar{x}_{..}) (y_{i.} - \bar{y}_{..}) = \sum_i x_{i.} y_{i.} / n - [(x_{..}) (y_{..}) / nk]\end{aligned}$$

The error sums of squares are:

Source of variation	d.f.	sum of product of		d.f.	Adj. S.S.	M.S.	F
		x, x	x, y				
TOTAL	nk-1	$x^2$	xy	$y^2$			
Blocks	n-1	$B_{xx}$	$B_{xy}$	$B_{yy}$			
Treatments	k-1	$T_{xx}$	$T_{xy}$	$T_{yy}$			
Error	$(n-1)(k-1)$	$E_{xx}$	$E_{xy}$	$E_{yy}$	$E_{xy}^2 - \frac{E_{xx} E_{yy}}{E_{xx}}$ = SSE*	MSE*	
Blocks + error	k(n-1)	$S_{xx}^b$	$S_{xy}^b$	$S_{yy}^b$	$S_{xy}^2 - \frac{S_{xx} S_{yy}}{S_{xx}}$ = $s_b^{*2}$		
Trtmnt + error	n(k-1)	$s_{xx}^t$	$s_{xy}^t$	$s_{yy}^t$	$s_{xy}^2 - \frac{s_{xx} s_{yy}}{s_{xx}}$ = $s_t^{*2}$		
Adj. Blocks					$s_b^{*2} - SSE^*$	MS (Bl) *	$F_b = \frac{MS(Bl)}{MSE^*}$
Adj. Trtmnts					$s_t^{*2} - SSE^*$	MS (Tr) *	$F_t = \frac{MS(Tr)}{MSE^*}$

Figure 6.

$$E_{xx} = \sum x^2 - B_{xx} - T_{xx}$$

$$E_{yy} = \sum y^2 - B_{yy} - T_{yy}$$

$$E_{xy} = \sum xy - B_{xy} - T_{xy}$$

These last represent the residual error for each of the variables. Although perhaps not readily apparent due to the different notation, we have performed a regular analysis of variance on the two variables, X and Y, separately and on their product. For example, writing  $E_{yy}$  in summation notation,

$$\begin{aligned} E_{yy} &= \sum y^2 - B_{yy} - T_{yy} \\ &= \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 - n \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 - k \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2, \end{aligned}$$

which is the form in the analysis of variance for the residual sum of squares, so that  $E_{yy}$  is the amount of variance in Y not ascribable to block effects or treatment effects. If we wished to test the null hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ , where  $\mu_i$  are unadjusted treatment means, it would be possible by forming the F statistic as before, using

$$F = \frac{\left[ \frac{T_{yy}}{k-1} \right]}{\left[ \frac{E_{yy}}{(n-1)(k-1)} \right]}$$

To test the hypothesis

$$H_0: \mu_{x,1} = \mu_{x,2} = \dots = \mu_{x,k},$$

the hypothesis that there is no real difference from treatment to treatment in the average value of the covariate, we form the ratio

$$F = \frac{\left[ \frac{T_{xx}}{k-1} \right]}{\left[ \frac{E_{xx}}{(k-1)(n-1)} \right]}$$

If the test is not significant and there is no reason to believe the average value of the covariate varied from treatment to treatment, then we can say that the dependent variable,  $Y$ , was observed at the same average value of  $X$  in each treatment. This means variation in the covariate,  $X$ , was not significant and the regression slope,  $b$ , is zero (refer back to Figure 5); and that variation in the  $Y$ 's was not affected by variation in the  $X$ 's. In this case, analysis of covariance would not have been necessary.

In comparing adjusted treatment means, as shown before, we adjust the error, or residual, sum of squares,  $E_{yy}$ , to account for variation attributable to the dependence of  $Y$  on  $X$ . The adjusted error sum of squares, once more, is

$$E_{yy} - [(E_{xy})^2/E_{xx}] = SSE^* ,$$

where  $b_{yx}$  is the regression coefficient for error,  $Y$  regressed on  $X$ ,

$$b_{yx} = E_{xy}/E_{xx}$$

following the form found in equation (5).

$SSE^*$  is the adjusted error sum of squares, and it represents the amount of error sum of squares unexplained by block effects, treatment effects, or by variation in the covariate  $X$ . Hence,

$$MSE^* = SSE^*/(\text{error d.f.}) = s_{y \cdot x}^2$$

is an estimate of the variance within each (and every, since homogeneity of variance is assumed) population of  $Y$  values. Since it is an unbiased estimate of the variance in the data which is

unexplained by any known factors (blocks, treatments, or covariate variation), we use this as the estimate of the actual "within treatment" variance,  $\sigma^2$ , of the Y.

Again using the regression reduction equation (5) for sum of squares, we can solve for the adjusted sum of squares for treatment + error, yielding

$$S_{yy}^t - [(S_{xy}^t)^2 / S_{xx}] = SS(T+E)^* .$$

This represents the amount of sum of squares that would be contributed by both variation between treatments and variation within treatments, if all observations of Y had been made at a standard value of X.

To find the sum of squares for adjusted treatment means, the adjusted sum of squares due to error alone is subtracted from the adjusted sum of squares due to error plus treatment, so that

$$SS(Tr)^* = SS(T+E)^* - SSE^* .$$

Remember that for the given adjustments to be valid, it requires that the linear regression of Y on X be homogeneous within all treatments. This can be seen by referring back to Figure 5 and noticing that the three treatment means were adjusted by "sliding" them along lines parallel to the regression slope determined from all the data grouped together.

The hypothesis of no difference among treatment means (or block means) for Y adjusted for the regression of Y on X can be tested. The F statistic is given by

$$F = \frac{MS(Tr)^*}{MSE^*}$$

and compared to  $F(\alpha/2, k-1, (n-1)(k-1)-1)$  where  $\alpha$  is the level of

significance desired in the test.

Note that the degrees of freedom for error decreases by one. This is because another parameter must be estimated from the data, the regression slope  $[(E_{xy}) / E_{xx}]$ .

### General Computational Procedure

Before giving examples of analysis of covariance applied to other experimental designs, it is worthwhile to note that the general computational approach will remain essentially the same as used here.

1 - Set up the usual analysis of variance table as given in the second chapter, but include the sum of squares for X and the sum of cross products (XY) as well as the sum of squares for Y for each line in the analysis.

2 - Compute the error line by subtraction for  $E_{yy}$  (SSE),  $E_{xy}$ , and  $E_{xx}$ .

3 - Compute the adjusted SSE, denoted now as

$$SSE^* = E_{yy} - [(E_{xy})^2 / E_{xx}],$$

and the adjusted mean square

$$MSE^* = s_{y \cdot x}^2 = SSE^* / (\text{error d.f.} - 1).$$

4 - Compute the (treatment + error) line ( $S_{yy}^t$ ) by adding the treatment line and the error line.

5 - Compute the adjusted (treatment + error) sum of squares by

$$S_{yy}^* = S_{yy} - [(S_{xy})^2 / S_{xx}] = SS(T+E)^*$$

6 - Compute the adjusted sum of squares for treatments,

$$SS(Tr)^* = S_{yy}^* - E_{yy}^* = SS(T+E)^* - SSE^*$$

and the adjusted mean square

$$MS(Tr)^* = SS(Tr)^*/(k-1) .$$

$$7 - F = MS(Tr)^*/MSE^*$$

$$= \frac{SS(Tr)^*/(k-1)}{SSE^*/(n-1)(k-1)-1} .$$

### Specific Results

#### Reduction in Variance

There is also an additional result which was not mentioned in the analysis, that being the analysis of covariance is often effective in reducing the error variance by adjusting the dependent Y values for chance or random changes in X. Even though the means of the X's may not vary from treatment to treatment, as before noted, there still exists chance variation which could effect variation in the Y values, inflating the variance slightly. D. J. Finney (in Anderson and Bancroft [1]) shows that the average variance of the difference between two adjusted means is

$$\overline{S_d^2} = \frac{2 \cdot MSE^*}{n} \left[ 1 + \frac{T_{XX}}{(k-1)E_{XX}} \right] .$$

So we can, if we wish, measure the efficiency of the analysis of covariance in reducing the error variance by comparing this adjusted variance with the unadjusted estimate.

#### Adjustment of Means

Having estimated the overall regression slope by b, and assuming this value remains essentially unchanged from treatment to treatment, it is possible to calculate the adjusted treatment means. These are the values the treatment means would have been had all values of Y been taken at a common value for X, namely  $\bar{X}$ .

$$\begin{aligned} \text{adjusted } \bar{Y}_{i.} &= \bar{Y}_{i.}^* = \bar{Y}_{i.} - b(\bar{X}_{i.} - \bar{X}_{..}) \\ &= \bar{Y}_{i.} - [E_{xy}/E_{xx}](X_{i.} - X_{..}) \end{aligned}$$

If it should be determined, and it will be shown later how, that the regression slope is not fairly homogeneous for all treatment groups, the same adjustment relation holds as is shown above with the change of  $b_i$  for  $b$ ,  $b_i$  being the regression slope for the  $i$ th treatment group.

### Assumptions

With the general procedure outlined, it is important that several assumptions must be satisfied to ensure that inferences drawn from the analysis of covariance are valid. These include:

1 -  $Y_{ij}$  are (observed values of) random variables and  $X_{ij}$  are (observed values of) random variables, all distributed about true means.

2 - A linearly-additive model, i.e.,

$$[\text{Total S.S.}] = \left[ \sum_i \text{Factor}_i \text{ S.S.} \right] + [\text{Residual S.S.}]$$

3 - The error terms are independently and normally distributed with mean zero and a common variance  $\sigma^2$ .

In addition to these assumptions from the analysis of variance model, it must further be assumed that:

4 - The regression of  $Y$  on  $X$  is linear.

5 - The slope of the regression line is not zero (i.e., analysis of covariance was necessary).

6 - The covariate  $X$ 's are measured without error, so that the only error in measurement is among the  $Y$ 's and is included

in the error term  $\epsilon$ .

7 - The regression coefficients within each group do not vary from factor to factor, that is, the regression coefficients are homogeneous so that the "average" or "pooled data" within groups regression estimate can be used for all groups.

8 - The independent variable(s),  $\{X_i\}$ , is (are) not affected by the treatments given to the groups of data (nor are the treatments affected by the covariate).

### Tests of Assumptions

The linearity assumption may be tested if the experiment is planned so that there is more than one Y observation for each X. If this is not done, less precise estimates that the regression is approximately linear must suffice.

To test the hypothesis that the true slope,  $\beta$ , for the regression, is zero, consider whether or not the reduction in the error sum of squares is significant when compared to the adjusted error sum of squares itself. Both of these estimates of variance are distributed as  $\chi^2$ , and taking their ratio, an F test can be applied (see [3]):

$$F = \frac{(E_{xy})^2/E_{xx}}{SSE^*/[(n-1)(k-1)-1]}$$

and is compared to  $F(\frac{\alpha}{2}, 1, (n-1)(k-1)-1)$ . If the F value is significantly large, we reject the hypothesis that  $\beta = 0$ .

This hypothesis can also be tested, as shown before, by testing the hypothesis  $H_0: \mu_{x,1} = \mu_{x,2} = \dots = \mu_{x,k}$ .

To test the hypothesis that all regression coefficients are

equal, compute each sample coefficient and adjust the sum of squares within each group by its own regression coefficient. For example, consider the problem at the beginning of this chapter with the three plating companies. We would compute:

$$b_A = \frac{\sum_A xy}{\sum_A x^2}$$

for X and Y values from company A. Since this was the first treatment, we would have

$$b_A = \sum_j x_{1j} y_{1j} / \sum_j x_{1j}^2 .$$

Then calculate the adjusted error sum of squares for the first treatment alone:

$$SSE_A^* = \sum_j y_{1j}^2 - (\sum_j x_{1j} y_{1j})^2 / (\sum_j x_{1j}^2) .$$

In the same manner, calculate

$$b_B, SSE_B^* ;$$

$$b_C, SSE_C^* .$$

Remember that estimating an adjusted error sum of squares within each group decreases the degrees of freedom by one additional unit. That is, each group has n (the number of blocks for each treatment) degrees of freedom, less one for estimating b, and less one for estimating the adjusted sum of squares for error.

Add the new adjusted sums of squares, giving a total "within sample" estimate of the adjusted error sum of squares based on  $k(n-2)$  degrees of freedom. We compare this to the "within sample" adjusted error sum of squares adjusted by a "pooled" within sample regression, based on  $[(n-1)(k-1)-1]$  degrees of freedom. If there are significant differences in regression among the

treatments from the overall pooled regression estimate, this would show up as a difference between these two estimates. We can again use an F test, as shown below

$$F = \frac{\text{SSE}^* \left[ \begin{array}{l} \text{based on} \\ \text{pooled-overall} \\ \text{estimate} \end{array} \right] - \text{SSE}^* \left[ \begin{array}{l} \text{based on} \\ \text{separate} \\ \text{reg. estimates} \end{array} \right]}{(k-1)} \\ \frac{\text{SSE}^* \left[ \begin{array}{l} \text{based on} \\ \text{separate} \\ \text{reg. estimates} \end{array} \right]}{k(n-2)}$$

and compare to  $F(\frac{\nu}{2}, k-1, k(n-2))$ . If the test is not significant, that is, the difference is not significantly large, we conclude the regression coefficients are homogeneous.

The final assumption that the treatments do not effect the covariate,  $X$ , is usually reasonable. If the values of  $X$  are actually influenced by the treatments but can be measured without error, an analysis of covariance can still be run but interpretations are often quite difficult. See the section on pitfalls in application at the end of the next section for a more thorough coverage of this.

It can be seen that satisfying all the assumptions can be a problem, although some laxity in some cases does not often greatly affect results (see [6]).

### Multiple Covariance

In the same way that simple covariance adjusts Y values for dependence on a single covariate, X, by using linear regression of Y on X, multiple covariance adjusts Y values for dependence on more than one covariate,  $X_1, X_2, \dots, X_m$ , by using multiple linear regression of Y on the m fixed covariates. Before covering the procedure followed in multiple covariance, a brief review of multiple linear regression may be helpful.

### Multiple Linear Regression

In simple covariance, we determined the amount of the error sum of squares which could be attributed to variation in the covariate X, by finding the regression of Y on X, which yielded

$$\text{Adj. } \sum Y^2 = \sum Y^2 - [(\sum xy)^2 / (\sum x^2)] ,$$

so the term  $[(\sum xy)^2 / (\sum x^2)]$  was the sum of squares attributable to regression in the single covariate case. The problem now is to determine what is the amount of the error sum of squares attributable to variation among the covariates when Y is dependent on not one, but m fixed covariates. To find this term (the sum of squares attributable to regression) it is necessary to work with the multiple regression model.

Suppose that the conditional expected value of Y can be approximated by the linear relation

$$E(Y|X_1, X_2, \dots, X_m) \doteq \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

so that Y can be approximated by

$$Y = \alpha + \sum_1 \beta_i X_i + \epsilon$$

where the regression coefficients  $\{\beta_i\}$  are determined from  $n$  observations of  $Y$  and the  $X_i$ , so the data consists of  $n$   $(m+1)$ -tuples:  $\{(Y_j, X_{1j}, X_{2j}, \dots, X_{mj}), j=1,2,\dots,n\}$ . This model can be rewritten in the form

$$Y_j = \bar{Y} + \sum_1 b_i x_{ij} + e_j \quad j=1,2,\dots,n$$

where  $\{b_i\}$  are the estimates of the  $\{\beta_i\}$ ,  $\bar{Y}$  is the mean of the  $Y$  values,  $x_i = X_i - \bar{X}$ , and  $e_i$  is the residual.

Rewriting this in deviation form, or residual form,

$$e_j = Y_j - \bar{Y} - \sum_1 b_i x_{ij} = y_j - \sum_1 b_i x_{ij}$$

Squaring both sides and summing over all  $n$  data points, we obtain

$$\sum e_j^2 = \sum (y_j - \sum_1 b_i x_{ij})^2$$

The objective is to solve for the  $\{b_i\}$  so as to minimize the sum of the squares of the  $\{e_j\}$ , which gives the least squares estimate of the regression coefficients  $\{\beta_i\}$ . This is done, as before, by taking partial derivatives of the sum of squares equation with respect to the  $\{b_i\}$ , ( $i=1,2,\dots,m$ ), yielding the following  $m$  equations:

$$\begin{aligned} b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \dots + b_m \sum x_1 x_m &= \sum x_1 y \\ b_1 \sum x_2 x_1 + b_2 \sum x_2^2 + \dots + b_m \sum x_2 x_m &= \sum x_2 y \\ \vdots & \\ b_1 \sum x_m x_1 + b_2 \sum x_m x_2 + \dots + b_m \sum x_m^2 &= \sum x_m y \end{aligned}$$

where  $\sum$  represents summation over all the data points of the indicated product (e.g.,  $\sum x_1 x_2$  represents  $\sum_j^n x_{1j} x_{2j}$ ).

These equations, analogous to the single regression model, are referred to as the "normal" equations.

These equations are solved simultaneously (in any one of several different ways -- see Appendix F for a solution) to ob-

tain values for the  $\{b_i\}$ . The solution yields

$$b_k = \sum_{i=1}^m \left[ c_{ki} \sum_{j=1}^n x_{ij} y_j \right]$$

where  $c_{kj}$  are elements of the inverse of the product matrix of the normal equations (see Appendix F). Having a method to solve for the  $\{b_i\}$ , we can turn our attention to the error sum of squares. It can be shown that

$$\begin{aligned} \text{SSE} &= \sum e^2 = \sum [y - \sum_i b_i x_i]^2 \\ &= \sum Y^2 - \left[ \sum [b_i \sum x_i y] \right] \end{aligned}$$

(refer to Appendix G).

Hence, the amount of the sum of squares attributable to the dependence of  $Y$  on the  $m$  fixed covariates,  $X_1, X_2, \dots, X_m$ , is given by the term

$$\sum_{i=1}^m [b_i (\sum x_i y)] ,$$

where  $\sum$  represents summation over all data.

A new term, multiple correlation coefficient,  $R$ , is defined by

$$\sum_1 [b_i (\sum x_i y)] \triangleq R^2 \cdot \sum Y^2 ,$$

so that

$$R^2 = \frac{\sum_1 [b_i \sum x_i y]}{\sum Y^2} = \left[ \begin{array}{l} \text{percent of } \sum Y^2 \\ \text{attributable to} \\ \text{regression} \end{array} \right]$$

Using this notation

$$\sum e^2 = \sum Y^2 - [R^2 \sum Y^2] = \sum Y^2 (1 - R^2) .$$

The multiple correlation coefficient,  $R$ , measures the closeness with which the regression plane (hyper-plane) fits the observed points. It is the correlation between the observed  $Y$ 's and the linear regression function predicted  $Y$ 's,  $\hat{Y}$ . So  $R$  meas-

ures the combined effect of the  $m$  independent variables,  $X_1, X_2, \dots, X_m$ , on the dependent variable,  $Y$ . (See Appendix H for a more intuitive derivation for  $R^2$ ).

### Adjustment for Regression

We can write the total sum of squares as a sum of two parts,

$$SST = \sum y_j^2 = \sum_1 [b_i \sum x_i y] + [\text{residual sum of squares}].$$

The total measure of variation,  $SST$ , is composed of variation due to dependence of  $Y$  on the  $m$  covariates and of variation not due to this dependence, i.e., chance variation, called the residual sum of squares. It is the residual sum of squares which gives an unbiased estimate of the "error", or true,  $Y$  population variance,  $\sigma^2$ . Rewriting the above for the residual sum of squares,

$$\begin{aligned} \text{residual S.S.} &= \sum y^2 - \sum_1 [b_i \sum x_i y] \\ &= \sum y^2 - R^2 \sum y^2 \\ &= \sum y^2 (1 - R^2) \end{aligned} \tag{6}$$

Equation (6) shows  $R^2$ , ( $0 \leq R^2 \leq 1$ ), as the fraction of the total sum of squares due to dependence of  $Y$  on the covariates.  $R^2$  can be viewed as the "percent reduction" in the variance when this effect is removed.

### Multiple Covariance Model

Now that it is known how to adjust total variance to remove that amount which is "inflation" caused by variation in the concomitant variables upon which the variable of interest is dependent, multiple covariance can be used in the analysis of covariance. Although the computations become more involved, the objective remains identical to that in simple covariance. That is, to

determine the regression relationship between  $Y$  and the independent  $X_i$ , adjust the observations of  $Y$  by means of this relationship to some standard values of the  $X_i$ , and then examine the adjusted values of  $Y$  by the analysis of variance. In the multiple covariance case the "adjustment" is a little more involved, but the objectives are the same.

General Example; 2 Factors, 3 Covariates

Consider an example in which the dependent variable,  $Y$ , is affected by three covariates,  $X_1, X_2, X_3$ . Such an example could be an analysis of drilling speeds for different drilling bits in which we wish to standardize measurements of speed as if they were for standard conditions of bit pressure, water (cooling) pressure, and rock hardness. After standardizing the measurements of drill speed,  $Y$ , by regressing  $Y$  on the three covariates, the characteristics of interest in  $Y$  are one, do the different types of bits have different speeds, and two, do average speeds vary from operator to operator.

In simple covariance, a two-factor experiment (assuming no interaction effect) would have the model

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \beta(x_{ij} - \bar{X}) + \epsilon_{ij}$$

where  $\mu$  is the grand true mean of the  $Y_{ij}$

$\alpha_i$  is the treatment effect (first factor)

$\gamma_j$  is the block effect (second factor)

$\beta(x_{ij} - \bar{X})$  is the regression relation which adjusts  $Y$  for variation in  $X$ , and  $\epsilon_{ij}$  is the error associated with  $Y_{ij}$ . For the multiple covariance case in which there are  $m$  covariates, this

model becomes

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \sum_{k=1}^m \beta_k X_{kij} + \epsilon_{ij}$$

where  $m=3$  for this example.  $Y$  is adjusted for  $m$  concomitant variables rather than a single, as can be seen by rewriting this as

$$\left[ Y_{ij} - \sum_k \beta_k (X_{kij} - \bar{X}) \right] = \mu + \alpha_i + \gamma_j + \epsilon_{ij}$$

which is the analysis of variance relation for the adjusted  $Y$  observations. This approach is not computationally feasible, time-wise, however.

General Computational Approach

After the data is collected and put in table form, it would appear as below, with one replicate, or observation, per cell.

Blocks

		Block 1				Block 2				Block n			
		Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>
t r t m n t s	1	Y <sub>11</sub>	X <sub>111</sub>	X <sub>211</sub>	X <sub>311</sub>	Y <sub>12</sub>	X <sub>112</sub>	X <sub>212</sub>	X <sub>312</sub>	Y <sub>1n</sub>	X <sub>11n</sub>	X <sub>21n</sub>	X <sub>31n</sub>
	2	Y <sub>21</sub>	X <sub>121</sub>	X <sub>221</sub>	X <sub>321</sub>	Y <sub>22</sub>	X <sub>122</sub>	X <sub>222</sub>	X <sub>322</sub>	Y <sub>2n</sub>	X <sub>12n</sub>	X <sub>22n</sub>	X <sub>32n</sub>
	.	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	k	Y <sub>k1</sub>	X <sub>1k1</sub>	X <sub>2k1</sub>	X <sub>3k1</sub>	Y <sub>k2</sub>	X <sub>1k2</sub>	X <sub>2k2</sub>	X <sub>3k2</sub>	Y <sub>kn</sub>	X <sub>1kn</sub>	X <sub>2kn</sub>	X <sub>3kn</sub>

We make exactly the same assumptions concerning the data in this case as were made for the simple analysis of covariance. The computational procedure is also begun in the same way.

Construct a regular analysis of variance table and in addition, for each line of the table, calculate the sums of squares and cross-products (for total, block, treatment, and any other factors). Then compute the error line for each sum of squares

and cross-products, by subtracting each of the factors in a column from the total sum of squares for that column.

Finally the (treatment plus error), or in general the factor plus error, line is calculated by addition of the factor's sums of squares and cross products to the corresponding sums for the error line. These values are then used in adjusting the sums of squares. An analysis of covariance table has been constructed on the next page. Some of the notation is explained below.

$$\sum x_k x_1 = \sum_1^k \sum_j^n x_{kij} x_{1ij} = \sum_1^k \sum_j^n x_{kij} x_{1ij} - [(x_{k..}) (x_{1..}) / nk]$$

$$B_{x_k x_1} = \sum_j x_{k..j} x_{1..j} / k - [(x_{k..}) (x_{1..}) / nk] \quad [l \neq k]$$

$$T_{x_k x_1} = \sum_i x_{ki} \cdot x_{1i} / n - [(x_{k..}) (x_{1..}) / nk]$$

$$E_{x_k x_1} = \sum x_k x_1 - B_{x_k x_1} - T_{x_k x_1}$$

$$S_{x_k x_1} = T_{x_k x_1} + E_{x_k x_1} \cdot$$

The  $\{b_i\}$  are calculated from the error sum of squares values. The  $\{b'_i\}$  are calculated using the [treatment + error] sum of squares values. Also, recalling from multiple regression the calculation of the multiple correlation coefficient as being given as

$$\sum_j^m [b_j \sum x_j y] = R_e^2 \cdot \sum_1^n y_i^2,$$

we see we can rewrite the adjusted error sum of squares as

$$E_{yy} - \sum_1^m b_i E_{x_i y} = E_{yy} - R_e^2 \cdot E_{yy} = E_{yy} (1 - R_e^2)$$

where  $R_e^2$  is the multiple correlation coefficient for the error line. This also holds true for the [error + treatment] line, and we have

		Sum of product of										df	adj. sums	M.S.
Source	d.f.	$X_1, X_1$	$X_1, X_2$	$X_1, X_3$	$X_2, X_2$	$X_2, X_3$	$X_3, X_3$	$X_1, Y$	$X_2, Y$	$X_3, Y$	$YY$			
TOTAL	$nk-1$	$\sum X_1^2$	$\sum X_1 X_2$	$\sum X_1 X_3$	$\sum X_2^2$	$\sum X_2 X_3$	$\sum X_3^2$	$\sum X_1 Y$	$\sum X_2 Y$	$\sum X_3 Y$	$\sum Y^2$			
Blocks	$n-1$	$B_{X_1, X_1}$	$B_{X_1, X_2}$	$B_{X_1, X_3}$	$B_{X_2, X_2}$	$B_{X_2, X_3}$	$B_{X_3, X_3}$	$B_{X_1, Y}$	$B_{X_2, Y}$	$B_{X_3, Y}$	$B_{YY}$			
Treatmt	$k-1$	$T_{X_1, X_1}$	$T_{X_1, X_2}$	$T_{X_1, X_3}$	$T_{X_2, X_2}$	$T_{X_2, X_3}$	$T_{X_3, X_3}$	$T_{X_1, Y}$	$T_{X_2, Y}$	$T_{X_3, Y}$	$T_{YY}$			
Error	$(n-1)(k-1)$	$E_{X_1, X_1}$	$E_{X_1, X_2}$	$E_{X_1, X_3}$	$E_{X_2, X_2}$	$E_{X_2, X_3}$	$E_{X_3, X_3}$	$E_{X_1, Y}$	$E_{X_2, Y}$	$E_{X_3, Y}$	$E_{YY}$		$E_{YY} - \sum_{i=1}^k b_i E_{X_i Y}$ (SSE*)	$MSE^* = S^2_y$
Int + Error	$n(k-1)$	$S_{X_1, X_1}^T$	$S_{X_1, X_2}^T$	$S_{X_1, X_3}^T$	$S_{X_2, X_2}^T$	$S_{X_2, X_3}^T$	$S_{X_3, X_3}^T$	$S_{X_1, Y}^T$	$S_{X_2, Y}^T$	$S_{X_3, Y}^T$	$S_{YY}^T$		$S_{YY}^T - \sum_{i=1}^k b_i' S_{X_i Y}$ (SS(T+E)*)	
Adj. Trt.													$[S_{YY}^T - \sum b_i S_{X_i Y}] - [E_{YY} - \sum b_i E_{X_i Y}]$ ( $SS(T+E)^* - SSE^*$ )	$MS(T)^*$

$$F_T = \frac{MS(T)^*}{MSE^*} \sim F_{k-1, (n-1)(k-1)-1}$$

Figure 7.

$$S_{YY} - \sum_1^m b'_i \cdot S_{x_i Y} = S_{YY} - R_{e+t}^2 \cdot S_{YY} = S_{YY}(1 - R_{e+t}^2)$$

where  $R_{e+t}^2$  is the multiple correlation coefficient for the [error + treatment] line.

We can now find the sum of squares, and mean square, for the adjusted treatment means. This is given by

$$\text{Adjusted SS(Tr)} = \text{SS(Tr)}^* = \text{SS(T+E)}^* - \text{SSE}^*$$

where  $\text{SS(T+E)}^*$  is the adjusted sum of squares for the [error + treatment] line, and  $\text{SSE}^*$  is the adjusted sum of squares for the error line.

The adjusted mean square error is given by

$$\text{MSE}^* = s_{Y \cdot x_1, x_2, \dots, x_m}^2 = \text{SSE}^* / (\text{error d.f.} - \text{no. of fixed var.}),$$

and the adjusted mean square for treatments is given by

$$\text{MS(Tr)}^* = \text{SS(Tr)}^* / (k-1) = (\text{SS(T+E)}^* - \text{SSE}^*) / (k-1) .$$

The F test ratio for the adjusted treatment means is then

$$F = \text{MS(Tr)}^* / \text{MSE}^*$$

and is compared to  $F(\frac{\alpha}{2}, k-1, (n-1)(k-1) - \text{no. of fixed variates})$ , completing our analysis.

It may be worthwhile here to give the steps again which were taken computationally. This method remains essentially unchanged regardless of the number of factors or the experimental design.

1 - Set up the usual analysis of variance table but derive a table of sums of squares and cross-products for total, blocks, and treatments (and any other component in the analysis).

2 - Compute the error line for each sum of squares and

cross products by subtracting factor sums of squares from the total sum of squares. Use these error values to compute the  $\{b_i\}$  and  $R_e^2$ , the squared multiple correlation coefficient for the error line.

3 - Compute an [error + treatment] line and from this line compute  $R_{e+t}^2$ , the squared multiple correlation coefficient for the [error + treatment] line.

4 - The adjusted error sum of squares is

$$SSE^* = SSE(1 - R_e^2)$$

The adjusted [error + treatment] S.S. is

$$SS(T+E)^* = SS(T+E)(1 - R_{e+t}^2) .$$

5 - The estimate of the variance in the Y's which is not attributable to block effect, treatment effect, or variation in the X's is

$$s_{yx}^2 = MSE^* = SSE^*/(\text{error d.f.} - \text{no. of fixed variates}).$$

6 - The mean square for the adjusted treatment means is

$$MS(Tr)^* = [SS(T+E)^* - SSE^*]/(k-1) .$$

The F test ratio is

$$F = MS(Tr)^*/MSE^* .$$

An F test can be performed on any factor the experimenter wants to test by, first, forming the [factor + error] sum of squares line. Second, forming the adjusted [factor + error] sum of squares by computing  $R_{e+f}^2$ , the multiple correlation factor for the [factor + error] line. Third, subtracting the adjusted error sum of squares from the adjusted [factor + error] sum of squares, yielding the adjusted factor sum of squares. Fourth,

forming the F ratio from the adjusted factor mean square and the adjusted mean square error.

### Uses

#### Direct Applications

Two broad areas for applications of analysis of covariance may be distinguished.

(1) To increase the precision in randomized experiments. This is accomplished by eliminating any bias in the data because of an uneven distribution of the fixed covariate,  $X$ , among the various treatments.

(2) The adjustment of treatment means for differences in the covariate among intact groups, that is, groups which are naturally occurring and which the experimenter cannot randomize, when the covariate is unrelated to the "treatments."

For application under category (1), the experimenter must be sure the data is randomly assigned to the structure of the experiment to ensure independence of the error terms. In addition, it must be certain that the treatment does not effect the covariate value, either because

- (a) the covariate is measured before the treatment is applied, or
- (b) the nature of the treatment or of the covariate precludes effect.

Also, the covariate does not effect the treatment. This is the primary, least dangerous, and most straight forward application.

Increase in precision. It is obvious that covariance increases precision of the data (the Y values) even if the dependence of Y on X (the covariate) is slight. The variance in a sample group is given by  $\sigma^2/n$ . There are two ways to increase the precision and lower the variance. The first is to make the denominator larger, usually decreasing the variance. Analysis of covariance (ANCOVA) works with the numerator, eliminating all sources of variation other than within the data itself. After application of ANCOVA, a more accurate estimate of the variance among treatment means can be made, as shown at the end of the Simple Covariance section.

Aid in the interpretation of treatment effects. There are many instances where the concomitant variable might be part of the means by which the treatments produce their effects on the data. A covariance analysis offers the possibility of exploring to see if this is true. This can be done by performing ANCOVA to determine what proportion of the treatment sum of squares is attributable to the covariate. Also, by examining to determine variation of the covariate from treatment to treatment. By measuring this variation before and after the experiment, an estimate of the interaction of the treatment and the covariate (which are supposed to be independent) can be obtained.

Another difficulty with this application, besides the fact that the treatment may effect the covariate, is that sometimes the adjusted values of the Y's have no physical meaning, or at least no meaning of interest, which makes the results very hard to interpret.

To adjust treatment means of the dependent variable (Y) for differences in the independent variable (X) is another use under category (1). This is probably the primary use in which ANCOVA is applied. The examples in the main body are of this nature. The comparison of adjusted treatment means is accomplished by adjusting the data to what it would have reflected had it been measured at a common standard value of the covariate. Actual adjustment of the treatment means is accomplished by performing the calculations shown at the end of the section on simple covariance.

$$\bar{Y}_{i.}^* = \bar{Y}_{i.} - b_{yx}(\bar{X}_{i.} - \bar{X}_{..})$$

where  $b_{yx}$  is the regression coefficient for Y on X, determined by  $E_{xy}/E_{xx}$ .

Category (2) uses. Analysis of covariance was introduced by R. A. Fisher in this context of usage. It was to be applied when each treatment had to be applied to an intact group which was not formed by random sampling. In such a circumstance there are likely to be prior differences between groups, and such differences are completely confounded with treatment effects. There are certain experiments where the "treatments" are defined as naturally occurring groups. Such a case might be a comparison of different strata of society on some characteristic, say amount of money in U.S. Savings Bonds. The "treatments" might be Democrats, Republicans, Socialists and Independents. Since these are naturally occurring, there might be other differences which influence the data,  $\bar{Y}$ , of amount in U.S. Bonds. For example,

Republicans, on the average, might tend to be richer than Socialists, and so would have more bonds of every type, not just the one of interest. In this case ANCOVA could be used to remove that portion of the treatment sum of squares which is attributable to prior differences. We would want to standardize groups with respect to wealth and view them as if they all had the same amount of money and see what each did with it. The covariate must be unrelated to the treatments (e.g., they are not Republican because they are richer).

The first illustration of the covariance analysis in literature [18] involved an example of the type of problem that would fall in this second category.  $Y$  was the tea yield from certain plots after application of different treatments. The covariate,  $X$ , was the yield of tea per plot in a period preceding the experiment. Adjustment of the responses  $Y$  for their regression on  $X$  removes the effects of variations in base yields of the different plots from the experimental errors. In this example these effects might be due either to inherent differences in the tea bushes in different plots or to differences in soil fertility that were permanent enough to remain through the experiment. Although the plots initially may have been randomly planted (not all the good bushes in one plot, etc.), after some time of growing they can in one sense be viewed as intact groups not susceptible to further randomization, and there may be inherent differences which ANCOVA can be used to standardize by taking measurements prior to an experiment.

A second use in this category is to remove effects of disturbing variables in strictly observational studies. In some research work, experiments are performed where any randomization is impossible (or at least infeasible). However, one may observe two or more groups differing in some characteristic in hopes of discovering if there is some association between this characteristic and some (experimentally defined) response  $Y$ . A good example of this is the comparison of heights (the response) of rural and urban (the characteristic) school children.

In a comparison of heights of children from two different schools, Greenberg noticed that the two groups differed slightly in mean age. Analysis of covariance, using age as the covariate, was used to make an age adjustment on height which resulted in a more sensitive comparison.

#### Indirect Applications

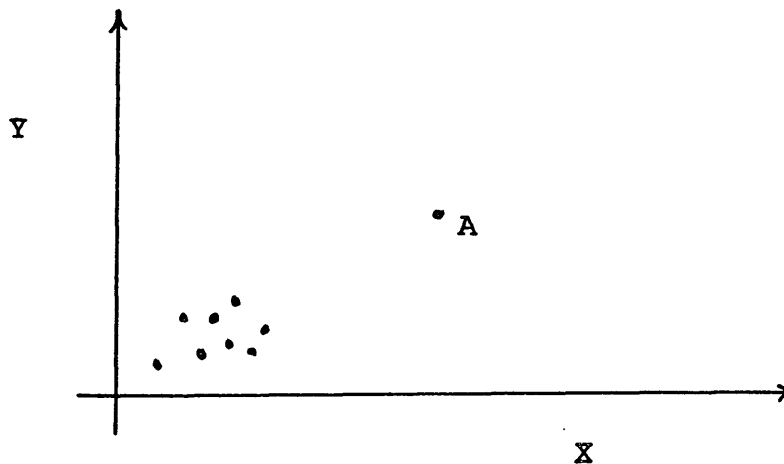
(1) Estimation of missing data, or to analyze data when some observations are missing.

If some data is missing, the first approach may be to use other data present to estimate what it might have been. Most formulas given for estimating missing data result in a minimum residual sum of squares, but it results in a biased treatment sum of squares. By using analysis of covariance an estimate of the missing data can be obtained which results in an unbiased estimate of both treatment and error sums of squares. So ANCOVA can be used to give an exact analysis of variance when some observations are missing. See Bartlett [2], pg. 137-183, or Coons [8].

As a sub-result, analysis of covariance can also be performed with incomplete data. See G. N. Wilkenson's article [41].

(2) To fit regressions in multiple classifications. ANCOVA can be used to compare the regression of  $Y$  on  $X$  (or  $\{X_i\}$ ) in different groups (e.g., treatments or blocks or cells, etc.). By standard techniques, as mentioned before in showing how to test for homogeneity of regression among groups, we can (i) fit a separate regression of  $Y$  on  $X$  within each class, (ii) test whether the slopes or positions of the lines differ from class to class, and (iii) if advisable, make a combined estimate of a common slope. Bennett and Franklin [3] go into some detail on this application, although their notation is very difficult to follow. The basic idea is in the section on simple covariance, and it is not difficult to apply.

(2a) To handle "outlier" points.

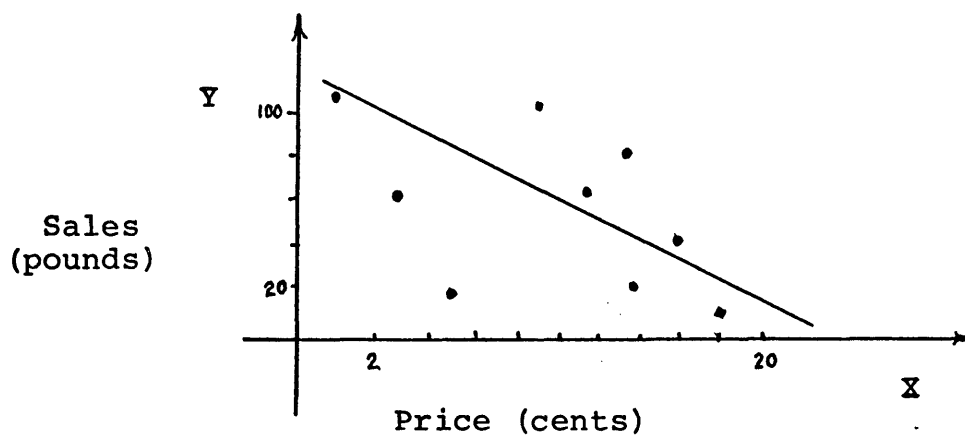


The graph above illustrates (hypothetically) a possible "outlier" problem. If the point A is in the same class as the other data then there is a significant regression of  $Y$  on  $X$ . However, if the point A is not in the same class, but from ano-

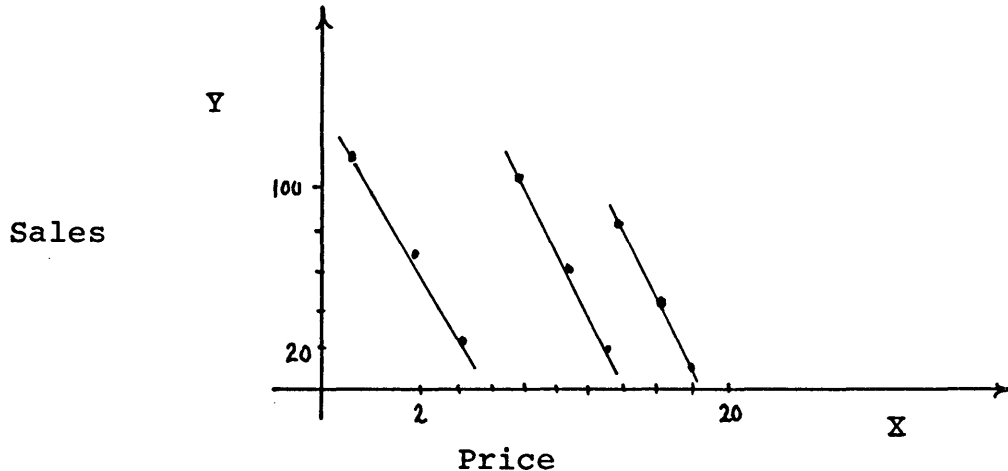
ther class, then there is no significant regression of  $Y$  on  $X$ .

The difference here is between "within" class regression and "between" class regression. To check on this the experimenter should return, if possible, to the source of the data to see if there is any reason to think point A may be from a different class. If so, other points from this class should be collected and a comparison of the "within" class regression and "between" class regression for homogeneity made.

(2b) A problem closely related to this can arise when the data collected may actually be from several different classes, and if the experimenter is unaware of this and attempts a straight application of regression analysis, he may get spurious results. An interesting example of this is regression analysis applied to data showing sales as related to product price, when the data come from several different years. This occurred in applying analysis to sales of frozen orange juice as related to price. The initial regression (not allowing for separate classes among the data) appears below. (See Henderson, et. al. [25]).

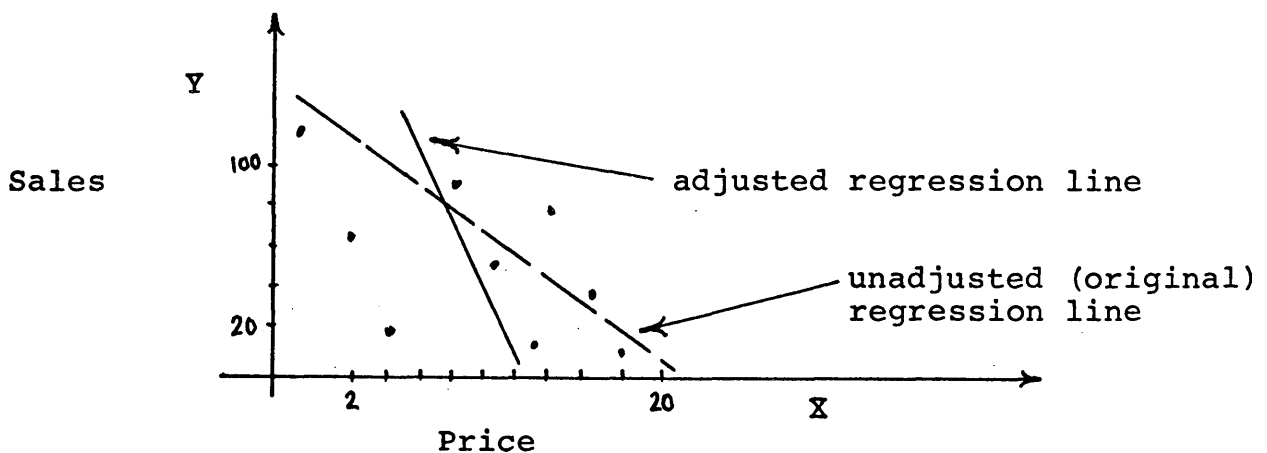


The data, however, is from three different years, and if the regression is performed on the data year by year, the results are quite different.



It can be seen that the regression within year groups is homogeneous, and different from the overall regression.

By combining different classes of data and finding an overall regression, the result has been a regression which partly reflects "within" class dependence, and partly reflects "between" class dependence, where the two are not the same. If analysis of covariance is used to adjust Y for X, the following is obtained.

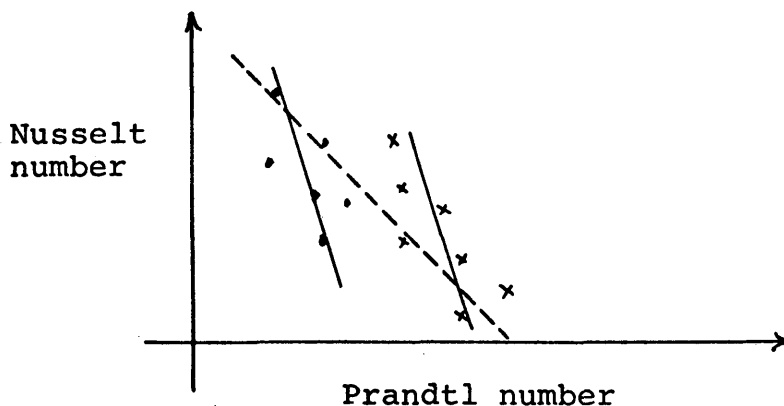


The application of ANCOVA removes the effect of differences in the average value of the covariate  $X$  from class to class.

(2c) The problem of pooling data.

Most research projects are preceded by an examination of results obtained previously in the same area of research by other workers in the field. In looking for results, or relationships between certain variables in the research, it would be helpful to be able to combine the data gathered by several different workers at different times under different conditions. This has at least two advantages. The range of the variables of interest are increased, and relationships are more easily detectable with the increased amount of data.

The problem is that when the data is combined and the independent variable(s) vary greatly from worker to worker, it makes it possible, as illustrated in (2b) above, to accept a regression found by an overall regression of  $Y$  on  $X$  as the regression of "within" class  $Y$  on  $X$  when it is influenced by a "between" class relationship. A hypothetical example is constructed below. Suppose we wish to see if there is a relationship between the Nusselt and Prandtl numbers in liquids.



The data has been differentiated between x's, from oil, and •'s, from water. Although a simplistic example, we see that if the data were combined indiscriminately, a relationship results which does not reflect the true relationship, which is the "within" class relationship.

A two part approach can be taken (i) to determine regression coefficients within each class then (ii) use the analysis of covariance techniques to test for homogeneity of within class regression slopes, and if they are homogeneous (as they appear to be in the example above), compare the within class slope to the overall regression to see if the two are equal.

(3) Analysis of covariance can also be used to estimate individual yields of a group of plots whose productions have inadvertently been combined so that only a total result from the entire group is known. Refer to the article by K. R. Nair [33].

#### Applications in Special Experimental Design Situations

For assistance in the application of the analysis of covariance in some regularly occurring, but not common, experimental designs, the articles below may be of some help.

Split-plot Design.

Kempthorne, [30].

Incomplete block with recovery of interblock information.

Cox, Eckhardt, and Cochran [7].

Two-way analysis with unequal numbers in the cells.

Hazel, [24]; Das, [10].

With missing data.

Wilkinson, [41]; Coons, [8].

For some other special references refer to:

Federer, [17], (Appendix, pp. 46-47 cites references to special experimental designs).

### Pitfalls in Application

Violations in satisfying certain of the assumptions result in the same consequences in errant results as are outlined at the end of the chapter on analysis of variance, since most of the assumptions necessary to ANCOVA are from, or are related to, the ANOVA and linear regression models. There is one assumption underlying the analysis of covariance which is unique to the model, and which, when violated, leads to spurious results.

We assume that the treatment does not effect the covariate. There is a class of problems to which analysis of covariance is applied wrongly. This is because the treatments have an effect, or are correlated with, the covariate, resulting in a confounding of the covariate variation with variation of the independent variable so that they are not separable. These difficulties arise because when the treatments effect the covariate, the assumption of homogeneity of within- and between-group regression also falls, resulting in improper adjustment of the variate sum of squares.

The class of problems in which the analysis cannot be carried out properly is in attempting to remove some of the treatment sum of squares as being attributable to variation in  $X$ , when the treatments have an effect on the covariate  $X$ . This

problem arises in two ways usually:

- (a) The correlation may occur accidentally under direct application category (1) (pg. 79).
- (b) Intentionally, through misapplication under direct application category (2).

An example of error (b) would be to use analysis of covariance to remove part of a treatment effect, or to adjust naturally occurring groups (e.g., social classes) for differences in one characteristic (e.g., intelligence) while comparing them on another attribute (e.g., professional aspirations). The covariate (intelligence) can't be randomly paired with "treatments" (naturally occurring social classes), which indicates that the two (treatment and covariate) are correlated.

This situation frequently arises when the experimenter adjusts data in a latter phase of an experiment for differences brought about as the result of an earlier phase of the experiment among the same data.

The results of this correlation of treatment and covariate effects several assumptions in the model. We have assumed a linearly additive model in which the additive components (e.g., treatment effect, covariate component, and error component) are statistically independent, and sum to the whole.

If the treatment is correlated with the covariate there is a confounding of variation due to the error variance in  $Y$ , and variation due to the covariate, so that the components are no longer independent. This also precludes homogeneity of within-group regression coefficients, since the treatments will cause

covariates under one treatment to vary more than under a different treatment, and the treatment means may be overadjusted or underadjusted. In either case, the results are wrong.

The robustness of ANCOVA with respect to violations of the independence of the components is small. Small correlations that might occur under usage (1) by accident may not be too harmful, but direct errors of misapplication can, and do, ruin results.

In testing the assumptions, a test for homogeneity of within-group regressions should yield one of the following results:

(a) The regressions are not homogeneous, and ANCOVA is not directly applicable.

(b) The regressions are homogeneous, and the overall regression is zero. That means the treatment means have no extra component which is predictable from the covariate, so no adjustment is required to remove this component, and ANCOVA is not necessary.

Either result would rule out usage of ANCOVA in such a situation.

USE OF THE COMPUTER PROGRAMPurpose

The program performs an analysis of covariance on up to a 5-way, or 5-factor, experimental design without interaction. More than 5 factors may be included if the number of levels in the factors is small enough to keep storage requirements within computer capabilities. The program will handle up to 5 covariates in multiple covariance problems. If an increase in capacities is desired in either case above (factors or covariates), the dimension statements must be changed accordingly.

Documentation

Variable names and functions in the program are listed below.

Non-dimensioned Variables.

Cov. F4 (Main)

MI -- Input device code number for reading data.

MO -- Output device code number.

N -- Number of observations.

M -- Number of variables. (Variate plus covariates).

NS -- Number of problem selections on data file.

XMEAN -- Mean square values for covariates

XEMEAN -- Mean square error for covariates.

FX -- F values for the covariate variation analysis.

KF -- Number of factors.

KC -- Number of covariates.

KC2 -- Number of covariates plus one , quantity squared

KC1 -- Number of covariates plus one (Number of variables).

ASSE -- Adjusted error sum of squares.

EDF -- Floating point storage for error degrees of freedom.

NDEP -- Position of variate in observation data rows.

DET -- Value of the determinate of inverted correlation matrix.

RE2 -- Multiple correlation coefficient for factor and error sums.

NEF -- Unadjusted error degrees of freedom.

#### Corre

FN -- Floating point storage for number of observations.

#### Avdat

M -- Product of (Level (I) + 1). Total storage for spaced data.

#### Avcal

SUM -- Summing variable for variate Y.

#### Mean

KC2 --  $KC * KC$  = number of data in covariate matrix.

L -- Designated factor for sums of squares.

GMEAN -- Grand overall mean.

FN -- Floating point storage for N, number of observations.

FN1 -- Correction divisor for sums of squares.

FN2 -- Number of degrees of freedom.

ND2 -- Fixed-point storage for number of degrees of freedom.

MSUM -- Check sum for number of factors with sums of squares calculated.

#### Subt

YSSE -- Error sum of squares for variate.

Covar

EDF -- Floating point storage for NEDG, error degrees of freedom.

FKC -- Floating point storage for KC, number of covariates.

VAR -- Adjusted mean square error.

ASSE -- Adjusted error sum of squares.

Multr

RM -- Multiple correlation coefficient, squared.

Ronorm

RONORM -- Row norm value.

SUM -- Summing variable for absolute values of row elements.

Invert

TOL -- .001, Check value for singularity of matrix.

TEST -- Storage location for possible pivot elements.

DET -- Determinant of matrix A.

INT -- Counting variable for interchanges to unscramble the matrix.

## Dimensioned Variables.

AFSSE(I) -- Adjusted (factor plus error) sums of squares.

B(I) -- Work vector (corre). I=number of variables.

D(I) -- Work array to store data observations. I=number of variables. (Data).

DF(I) -- Floating point variable of D.F. used in division for F values.

F(I) -- F values for (adj factor MS/adj. error MS).

HEAD(I) -- Name or label for factors. May be alphanumeric.

ISTEP(I) -- Counting index. Spacing for factors.

KOUNT(I) -- Counting index. Spacing for levels.

- LASTS(I) -- Counting index. Indicating last data position per factor.
- LEVEL(I) -- Number of levels for factor I. I = number of factor.
- NDF2(I) -- Unadjusted factor and error degrees of freedom.
- NDF(I) -- Degrees of freedom for factor sums of squares.
- PR1(I) -- Problem name of label. Alphanumeric demensioned for 3A2 format.
- R(I) -- Work matrix used in calculation fo correlation coefficients (in main program).
- RX(I) -- Matrix for containing intercorrelations of covariates. For matrix of size (KC+1) x (KC+1).
- RY(I) -- Vector for intercorrelations of covariates with variate Y. (I is position of covariate).
- R2(I) -- Multiple correlation coefficient for (factor + error) lines. I is the factor number.
- SMEAN(I) -- Mean square for factors.
- STD(I,J) -- Work array used to store standard deviations (in main).
- SUMSQX(I,J) -- Factor sums of squares for covariate products. I gives position in product vector.  $J=2^k-1$ .
- SUMXY(I,J) -- Factor sums of squares for cross products. I gives position in product vector.  $J=2^k-1$ .
- SUMX(I) -- Summing variable for covariates (AVCAL).
- SX(I) -- Storage array for sums of squares of covariates, used in analysis of variation among covariates.
- S(I) -- Internal work array, (INVERT).
- T(I) -- Work vector (CORRE). I=number of variables.
- X(I,J) -- Covariates. I=number of covariate; J=number of observations.
- XSSE(I) -- Error sums of squares matrix for covariates (X's). Matrix size KC x KC.
- XYSSE(I) -- Error sums of squares vector for cross products. Vector length KC + 1.

XFPE(I,J) -- Factor plus error sums of squares for covariates.  
 I=position in product matrix. J=factor number.

XYFPE(I,J) -- Factor plus error sums of squares for cross product  
 terms. I=position in product matrix. J=factor number.

Y(I) -- The variate dimensioned for number of observations.

YFPE(I) -- Factor plus error sums of squares for variates.  
 I=factor number.

The arranging of the data arrays, and formation of sums of squares is primarily data manipulation, multiplication, and addition loops performed in subroutines CORRE, AVDAT, AVCAL, and MEAN.

After subroutine MEAN the sums of squares for the factors have been computed, and the sums of squares for these lines for the covariate products and cross-products (xx), and the covariate-variate products ( $x_i y$ ). These sums are combined in subroutine SUBT to yield the error sum of squares line and (factor + error) sums of squares lines for all the above inter-products.

The next part of the main program arranges these values in the following matrices:

XSSE(I) - Error sum of squares values.

XFPE(I,J) - Factor plus sums of squares, where J is the factor indicator.

To find the multiple correlation coefficients, the Gauss-Jordan method is used, as explained in detail in Statistics in Research by Bernard Ostle [30], chapter 8.

The correlation coefficients are calculated for the two matrices (XSSE and XFPE), these coefficients are rearranged into proper matrix form in subroutine ORDER, the correlation

matrix is inverted in subroutine INVERT, and the multiple correlation coefficient (squared), is calculated in subroutine MULTR, using the procedure outlined in Ostle, first for the error sums, and then for the KF factor plus error lines. The multiple correlation coefficients are used in subroutine COVAR to adjust the sums of squares and calculate the F values.

The results are written on data file FOR16.DAT in the last part of the main (COV.F4) program.

### Input

A sample input file has been constructed for the problem given below from Statistical Methods by George Snedecor [32], p. 427.

Analysis of covariance applied to a problem where

Y - Yield of wheat in Great Britain

X<sub>1</sub> - Height of shoots at ear emergence

X<sub>2</sub> - Number of plants at tillering

Year	Variate	Place
1933	X <sub>1</sub>	25.6 25.4 30.8 33.0 28.5 28.0
	X <sub>2</sub>	14.9 13.3 4.6 14.7 12.8 7.5
	Y	19.0 22.2 35.3 32.8 25.3 35.8
1934	X <sub>1</sub>	25.4 28.3 35.3 32.4 25.9 24.2
	X <sub>2</sub>	7.2 9.5 6.8 9.7 9.2 7.5
	Y	32.4 32.2 43.7 35.7 28.3 35.3
1935	X <sub>1</sub>	27.9 34.3 32.5 27.5 23.7 32.9
	X <sub>2</sub>	18.6 22.2 10.0 17.6 14.4 7.9
	Y	26.2 32.7 40.0 29.6 20.6 47.2

The program is set up to be run from the teletype.

The input data must be put on a data file prior to running the program. The program then reads the data from the storage file during execution. A file can be created either by using cards and inputting them through BATCH or by using the teletype.

The easiest way to store the data is by creating a data file, using the name FORXX.F4, where XX is a two-digit number between 13 and 20. The data can then be read by the program by telling it to read at location number XX, e.g., READ (XX,5), which instructs it to read data file XX according to format statement number 5.

The input file appears below, and is explained line by line.

SN427	18	3	1	2	2
SEA	3				
PLA	6				
25.6	14.9	19.0			
25.4	7.2	32.4			
27.9	18.6	26.2			
25.4	13.3	22.2			
28.3	9.5	32.2			
34.4	22.2	34.7			
30.8	4.6	35.3			
35.3	6.8	43.7			
32.5	10.0	40.0			
33.0	14.7	32.8			
32.4	9.7	35.7			
27.5	17.6	29.6			
28.5	12.8	25.3			
25.9	9.2	28.3			
23.7	14.4	20.6			
28.0	7.5	35.8			
24.2	7.5	35.2			
32.9	7.9	47.2			

## Explanation.

SN427      18      3      1      2      2

SN427 - Designation of problem name.

18 - Number of observations made on variate and covariates.

3 - Number of variables (covariates plus variates).

1 - Number of selections, i.e., complete sets of data for problem solution. If the entire experiment were performed three times, then the data would be put on a data file and NS would be set equal to 3 to run the three sets of data through the program.

2 - Number of factors.

2 - Number of covariates.

This data is read in the main program (COV.F4) according to format number 3, (3A2, 5G). The problem name is read from the first six spaces on the line. No other values may appear in the first six spaces, and the entire name must be in the first six spaces, otherwise, letters from the name will be read into the program in one of the other five locations and run errors will occur. The next five values are read according to a free-field G format, and all that is required is that they be separated by a space or spaces or some other non-numeric symbol which would not ordinarily be in a data format (e.g., a comma, a slash, etc.)

SEA 3

PLA 6

SEA - Label for first factor.

3 - Number of levels in first factor.

PLA - Label for second factor.

6 - Number of levels in second factor.

The heading names and number of levels are read according to format number 5, where the format (A3,I) is given. The heading name is read in the first three spaces of the data line. The number of levels is read according to free field integer format, so the value may appear anywhere on the line after the name, so long as it is separated from the name by a space or some other non-numeric symbol which would not ordinarily be in a data field.

25.6	14.9	19.0
.	.	.
:	:	:
.	.	.
32.9	7.9	47.2

This is the experimental data. The order of the data in each row should be

$$X_1 \ X_2 \ \cdots \ X_m \ Y \ .$$

The covariate observations first, and then the variate observations.

This data is read one line at a time according to a free-field G format, so all that is required is that the values be separated by spaces or a comma or some other non-numeric symbol as in the other free-field reads.

Since the data is entered into a one-dimensional array, the order of the data must be determined so as to designate the factor and level location of the observations. The program reads the data and stores it by column. The data should be read in so that the first subscript varies most rapidly.

For example, for a two-factor analysis with the data as given below:

	block					
trtmnt	Y <sub>11</sub>	X <sub>111</sub>	X <sub>211</sub>	Y <sub>12</sub>	X <sub>112</sub>	X <sub>212</sub>
	Y <sub>21</sub>	X <sub>121</sub>	X <sub>221</sub>	Y <sub>22</sub>	X <sub>122</sub>	X <sub>222</sub>

the data would be put on the file as

```

X111   X211   Y11
X121   X221   Y21
X112   X212   Y12
X122   X222   Y22

```

In this case, the first subscript on the X's refers to the covariate, whether it is the first, second or third, which is why it doesn't vary. The second two subscripts identify the location of the covariate value in the experimental design. The data is arranged according to these subscripts.

### Execution

The main program and all subprogram names have been placed in a command file. The program is executed by giving the command

```
·EX @COV
```

where COV is the name of the command file, COV.CMD, and · is the teletype response indicating it is ready for a command.

This set up allows the user to list any of the several programs separately, if desired, yet not have to execute them all by name individually.

Also, if the user desires to have information printed out which is not being printed out currently, print statements can

be inserted where desired without requiring the entire set of programs to be recompiled, only the subprogram changed.

After the program is compiled and execution is begun, the following is typed out on the teletype.

```
ENTER I/O DEVICE NUMBERS:
```

The first number typed should give the location of the data file which was created prior to running the program. If the data is stored in a data file, e.g., FOR17.DAT, then the first number input would be 17. This tells the program where to read the input data. The output number directs the program where to print out in-progress messages, should there be any. This should be number 4, designating the teletype unit, if the program is being executed from there. These two numbers are read according to a free-field format. In this case, the program users response would appear as

```
ENTER I/O DEVICE NUMBERS: 17,4 ↵
```

where " ↵ " is the carriage return, instructing the program to continue.

### Output

All output data is written on the data file FOR16.DAT. The user has the option of printing this information out on the teletype, achieved by the command TYPE FOR16.DAT, or having the file printed out by the line printer, by giving the command PRINT FOR16.DAT.

The output from the sample problem from G. W. Snedecor's book is shown below.

+++PROBLEM: SN427

SELECTION: 1

FACTOR: SEA LEVELS: 3

FACTOR: PLA LEVELS: 6

MULTIPLE CORRELATION COEFFICIENT  
FOR ERROR SUMS OF SQUARES

0.8772542

MULTIPLE CORRELATION COEFFICIENTS FOR  
FACTOR PLUS ERROR SUMS OF SQUARES

SEA + ERROR 0.6601446

PLA + ERROR 0.8495038

.....

UNADJUSTED VARIATION ANALYSIS

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F VALUES
SEA	124.41444	2	62.20722	2.720
PLA	629.21777	5	125.84355	5.503
ERROR	228.66556	10	22.86656	

.....

ADJUSTED VARIATION ANALYSIS

ADJUSTED +FACTOR PLUS ERROR+ VARIATION

SOURCE OF VARIATION	ADJUSTED SUMS OF SQUARES	DEGREES OF FREEDOM
SEA + ERROR	119.99615	10
PLA + ERROR	129.10816	13

## ADJUSTED FACTOR VARIATION

SOURCE OF VARIATION	ADJUSTED SUMS OF SQUARES	DEGREES OF FREEDOM	F VALUES
SEA	91.92842	2	13.10094
PLA	101.04043	5	5.75931
ERROR	28.06773	8	

## SIGNIFICANCE OF VARIATION OF COVARIATES

## COVARIATE 1

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F VALUES
SEA	6.25333	2	3.12667	0.265
PLA	106.33167	5	21.26633	1.803
ERROR	117.94000	10	11.79400	

## COVARIATE 2

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F VALUES
SEA	139.41444	2	69.70722	9.395
PLA	171.46444	5	34.29289	4.622
ERROR	74.19222	10	7.41922	

The explanation for the output is now given.

+++PROBLEM: SN427

This is the problem name given by the user in the input.

SELECTION: 1

This indicates the experiment selection being analyzed. If the experiment had been performed completely three times, then all three sets of data would have been entered in sequence and there would have been three sets of analysis, each with the same form as appears for one analysis here, but headed by SELECTION: 2 and SELECTION: 3, as each was completed and printed out.

FACTOR: SEA                   LEVELS: 3

FACTOR: PLA                   LEVELS: 6

This tells the number of factors, the name label which has been given to each one, and the number of levels for that factor.

MUPTIPLE CORRELATION COEFFICIENT FOR ERROR SUMS OF SQUARES

This is the multiple correlation coefficient computed from the  $E_{**}$  line in the ANCOVA table. It indicates the "within" group regression of Y on the covariates. It is of no special computational value to the user except in the case in which "within" group regressions are to be compared from group to group and data is read in for only one experimental unit at a time.

In the single covariate case, the output is changed to read "adjustment for error sum of squares" and "adjustment for factor

plus error sums of squares" rather than "correlation coefficients" as shown below. This is because multiple correlation coefficients are not used in the single covariate case. Except for this heading change, the rest of the output is the same.

MULTIPLE CORRELATION COEFFICIENTS FOR FACTOR PLUS ERROR SUMS OF SQUARES

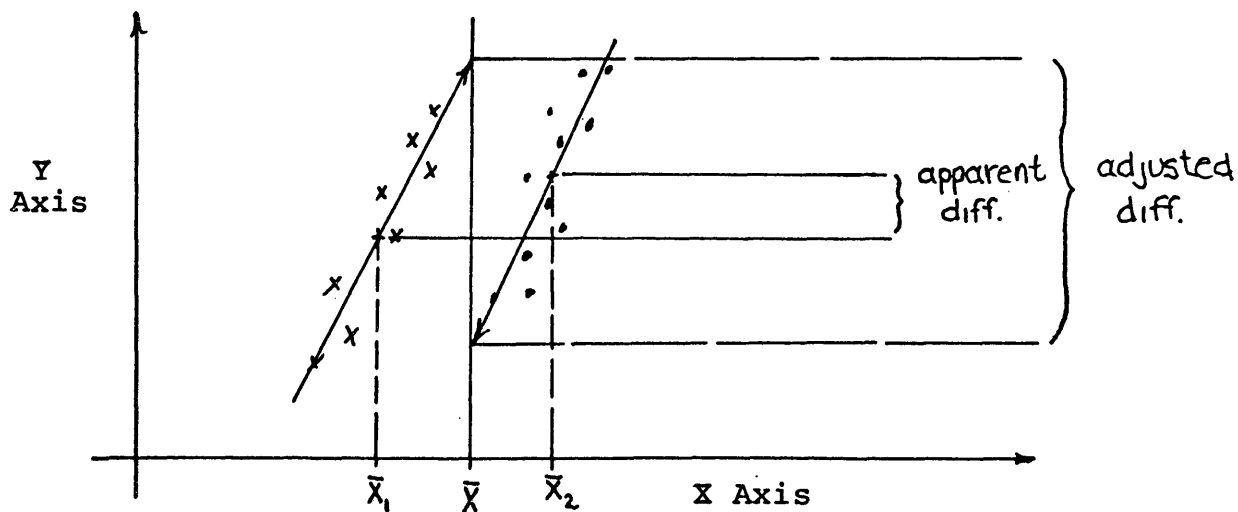
These are the correlation coefficients computed for the  $S_{**}^f$  lines in the ANCOVA and represent the percent adjustment in the factor plus error sums of squares.

UNADJUSTED VARIATION ANALYSIS

An analysis of variance table is given for the variate unadjusted for any dependence on the covariates.

This can be used in judging the effect of dependence of Y on varying covariates, by comparing unadjusted variation to adjusted variation.

It is sometimes the case, as in this example, that differences in the covariate values tend to dampen out differences among variate means. Such an example is illustrated below for comparison of treatment means.



## ADJUSTED VARIATION ANALYSIS

The adjusted sums of squares are given below this heading. First the adjusted values are given for the factor plus error sums of squares, and finally, the adjusted sums of squares for the factors, which is what the experimenter is interested in. The degrees of freedom are also the adjusted values.

The adjusted factor sums of squares represent the variation among factor means which is not due to variation among covariate values. The new F values represent a comparison of adjusted factor sums of squares to an adjusted error sum of squares.

## SIGNIFICANCE OF VARIATION OF COVARIATES

This indicates the differences which exist between average covariate values from factor level to factor level.

This information can be used in two ways. First, it indicates which covariate may be the significant contributor of added variation to the Y values. For this example, variation among covariate 1 values is very small for both factors, whereas covariate 2 varies quite a bit. Secondly, it can be used as a test to make sure the covariates are independent of the treatments which are applied if the experiment is of such a nature that allows this testing. This can be done by comparing variation among the covariates prior to application of the treatments, and then following the application. If there is a marked difference, the experimenter should investigate a possible dependency.

Variation among covariate 1 is computed as a straight analysis of variance on the variable  $X_1$ . This is performed using the

factor sums of squares and error sum of squares from the  $X_1X_1$  column of the ANCOVA table. The same pattern is used for the others.

If there is additional information the user wants which is not included in the output, it can be obtained by locating that information in the program using the explanation of the program variables given before, and inserting print statements. These should be taken out after the user is finished using the program.

APPENDICES

Appendix A

Consequences of assumptions for analysis of variance not being satisfied.

One of the more basic underlying assumptions of the analysis of variance model not mentioned in the section involves the experiment itself, and is that the different types of effects can be identified and controlled enough to draw realistic inferences from the analysis.

Three types of effects are generally recognized. These are:

- (a) Treatment effects: The effects of procedures deliberately introduced by the experimenter.
- (b) "Environmental" effects: The certain features of the environment which the analysis model enables us to measure. E.g., by checking block effect in a one-factor analysis, or the effects of replication in a two-factor analysis.
- (c) Experimental Error: All elements of variation that are not, or cannot be, taken account of in (a) and (b).

Added to these are the four assumptions listed in the section.

These were:

- (1) The observations are observed values of random variables, distributed about true means.
- (2) Additivity. The treatment and environmental effects must be additive. E.g., in a randomized block (2 factor)

design with no interaction, the observations can be written

$$X_{ij} = \mu + \alpha_i + \delta_j + e_{ij}.$$

and in general

$$SST = \left[ \sum_i \text{factor}_i \text{ S.S.} \right] + \left[ \text{error S.S.} \right].$$

- (3,4) The experimental error terms are all distributed independently and normally with mean zero and common variance  $\sigma^2$ .

Some of the consequences follow.

- (1) The Effects of Non-normality -- Extensive experimentation.

As mentioned in the body, a substantial departure from normality can be tolerated without great effect on the practical results. No serious error is introduced by non-normality in significance levels of the F-test, or the two-tailed t-test if this is being used. Non-normality may cause a loss of efficiency in the estimate of the treatment effects, and affect the F- and t-tests this way. However, this effect is not often very great.

If there is to be extensive experimentation, certain precautions could be taken, such as,

- (i) From prior information estimate the distribution of the e's.
- (ii) By trial and error determine a transformation to be used on the data so the e's are approximately normally distributed.

If it is only one experiment, it is usually pretty safe to assume that the error terms are close enough to normally distributed so as not to greatly bias results.

(2) Non-additivity.

This is covered in the last chapter on pitfalls in using analysis of covariance.

(3) Heterogeneity of Errors.

The immediate effects are a biasing of treatment estimates and a loss of sensitivity of the significance tests. These can be obviated by replacing the usual ANOVA table variance by a weighted average analysis in which each observation is weighted in proportion to the inverse of its error variance, although this becomes very involved.

This difficulty can arise through mishaps or damage to some part of the experiment, through less carefully controlled conditions, or through the use of less homogeneous materials as the experiment progresses. These are things to watch for in conducting the experiment.

(4) Correlation among the errors.

This usually makes the treatment means less accurate while at the same time, by decreasing overall combined data variance, makes the estimates appear more accurate than usual. Hence, it biases error estimates and impairs the t-test.

This is largely corrected by randomization, allowing us to treat the errors as if they were independent. For more detailed methods, see "The Formation of the Latin Squares for use in the Field", by F. Yates in the Empire Journal of Experimental Agriculture, vol. 1 (1933), pp. 235-244.

(5) Gross Errors.

The results of this error are pretty obvious:

- (i) The between treatment variance estimates are biased.
- (ii) The standard error of the estimates will be effected.

There are several tests which can be used to determine if a data point which appears to be an outlier is actually a gross error or not. See the chapter on uses of covariance, or see the article by E. S. Pearson and C. C. Sekar in Biometrika, vol. 28, pp. 308-320.

According to W. G. Cochran [5], "The factors most liable to cause the most severe disturbances are extreme skewness, the presence of gross errors, anomalous behavior of certain treatments or parts of the experiment, marked departures from the additivity relationship and changes in the error variance, either related to the mean or to certain treatments or to parts of the experiment. The principal methods for an improved analysis are omission of certain observations, treatments, or replicates, subdivision of the error variance, and transformation to another scale before analysis."

For a more complete coverage of difficulties and solutions, refer to the above reference by Cochran, or to Cochran and Cox [6].

Appendix B

Show that the estimate of variance based on variations among treatment means estimates  $\sigma^2/n$ .

$$s_{\bar{X}}^2 = \frac{\sum_{i=1}^k (\bar{X}_{i.} - \bar{X}_{..})^2}{(k-1)}$$

estimates the variance of the variable  $\bar{X}_i$ , where  $k$  is the number of treatments. Since we assume the  $X_{ij}$  are normally and independently distributed with a common variance  $\sigma^2$ , then there is a theorem we may apply which states that if  $X_1, X_2, \dots, X_n$  are mutually independent random variables, normally distributed with variances  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ , then the random variable

$$Y = K_1 X_1 + K_2 X_2 + \dots + K_n X_n$$

is normally distributed with variance  $\sum_1 [K_i^2 \cdot \sigma_i^2]$ .

In this case, we set

$$\begin{aligned} Y &= \bar{X} \\ &= \sum_i \bar{X}_{i.} / n \\ &= \bar{X}_{i1}/n + \bar{X}_{i2}/n + \dots + \bar{X}_{in}/n \end{aligned}$$

So  $K_i = 1/n$  for all  $i$ , and  $\sigma_i^2 = \sigma^2$  by assumption, so that the variance of  $\bar{X}$  is given by

$$\sum_{j=1}^n (1/n)^2 \cdot \sigma^2 = (1/n^2) \cdot n \cdot \sigma^2 = \sigma^2/n ,$$

hence,  $s_{\bar{X}}^2$  estimates the variance  $\sigma^2/n$ .

Appendix C

Prove the decomposition for the sum of squares for a one-factor design.

We use the algebraic identity

$$x_{ij} - \bar{x}_{..} = (x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}_{..}) ,$$

obtained by adding and subtracting  $\bar{x}_{i.}$  to the right side.

Squaring both sides, and then summing over  $i$  and  $j$ , we obtain

$$\begin{aligned} \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 &= \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 + \sum_i \sum_j (\bar{x}_{i.} - \bar{x}_{..})^2 \\ &\quad + 2 \sum_i \sum_j (x_{ij} - \bar{x}_{i.}) (\bar{x}_{i.} - \bar{x}_{..}) . \end{aligned}$$

Note that

$$\begin{aligned} \sum_i \sum_j (x_{ij} - \bar{x}_{i.}) (\bar{x}_{i.} - \bar{x}_{..}) &= \sum_i (\bar{x}_{i.} - \bar{x}_{..}) \cdot \sum_j (x_{ij} - \bar{x}_{i.}) \\ &= 0 \cdot 0 \\ &= 0 \end{aligned}$$

since  $\bar{x}_{i.}$  is the mean of the  $i$ th sample, and  $\bar{x}_{..}$  is the mean of the  $i$  sample means. Lastly, note that the second term in the right hand side of the expanded squared value does not include the subscript  $j$ , hence,

$$\sum_i \sum_j (\bar{x}_{i.} - \bar{x}_{..})^2 = n \cdot \sum_i (\bar{x}_{i.} - \bar{x}_{..})^2$$

and it then follows that the theorem is true.

Appendix D

Mutually distinct properties of the data.

To illustrate what is meant by "mutually distinct", or in geometric terms, "orthogonal", properties, note that in the case of numbers arranged as they are in Table 1, if we add an arbitrary fixed constant to each number in column one, a different arbitrary fixed constant to each number in column two, and so forth until we have added a different constant to each of the  $n$  columns, then it is obvious the column means will all be altered by different amounts, according to the constant added to the column, but the row means will all be altered by the same amount, so the difference between any two arbitrary row means is unchanged by the manipulations on the columns. Also, the values such as  $(\bar{X}_{i.} - \bar{X}_{..})$  and the error sum of square differences  $(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})$  in the two-factor case, remain unchanged. This "unchangingness" is because the differences in row means (or the differences among the "error" or "residual" terms) are orthogonal, or are mutually distinct from, differences among column means or differences of column means from the general mean. They summarize mutually distinct properties of the numbers involved in the analysis.

Appendix E

Prove the decomposition for the sum of squares for a two-factor design.

We use the identity

$$x_{ij} - \bar{x}_{..} = (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}) + (\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..})$$

obtained by adding and subtracting  $\bar{x}_{i.}$ ,  $\bar{x}_{.j}$ , and  $\bar{x}_{..}$  to the above left hand side to obtain the form on the right. The approach is then essentially the same as in Appendix C. Square both sides and sum over  $i$  and  $j$ .

Appendix F

Solution of normal equations for multiple regression.

The normal equations (no connection with the normal distribution) are given once again, as

$$\begin{array}{r} b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \cdots + b_m \sum x_1 x_m = \sum x_1 y \\ b_1 \sum x_2 x_1 + b_2 \sum x_2^2 + \cdots + b_m \sum x_2 x_m = \sum x_2 y \\ \vdots \qquad \qquad \qquad \vdots \qquad \qquad \qquad \vdots \qquad \qquad \qquad \vdots \\ b_1 \sum x_m x_1 + b_2 \sum x_m x_2 + \cdots + b_m \sum x_m^2 = \sum x_m y \end{array}$$

where " $\Sigma$ " represents summation over all data points.

If we make two substitutions, letting

$$\begin{aligned} a_{ij} &= \sum x_i x_j \quad ( = a_{ji} ) \\ g_i &= \sum x_i y \end{aligned}$$

we can rewrite the above system of equations as

$$\begin{array}{r} a_{11}b_1 + a_{12}b_2 + \cdots + a_{1m}b_m = g_1 \\ a_{21}b_1 + a_{22}b_2 + \cdots + a_{2m}b_m = g_2 \\ \vdots \qquad \qquad \qquad \vdots \qquad \qquad \qquad \vdots \qquad \qquad \qquad \vdots \\ a_{m1}b_1 + a_{m2}b_2 + \cdots + a_{mm}b_m = g_m \end{array}$$

It now remains to solve for the  $\{b_i\}$ . This can be done directly by solving the system of equations simultaneously, but as the number of independent variables increases, this way becomes very time consuming. An easier approach involves the use of a matrix inversion technique.

Define the following matrices as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$$

$$B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix}$$

Using matrix notation we can rewrite the system of normal equations in the following form.

$$\begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 & \cdots & \sum x_1 x_m \\ \sum x_2 x_1 & \sum x_2^2 & \cdots & \sum x_2 x_m \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_m x_1 & \sum x_m x_2 & \cdots & \sum x_m^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum x_1 y \\ \sum x_2 y \\ \vdots \\ \sum x_m y \end{bmatrix}$$

as

$$A \cdot B = G .$$

Since we want to solve for the  $\{b_i\}$ , to get them by themselves we multiply both sides of the above equation by  $A^{-1}$ , the inverse of the matrix A.

Then

$$A^{-1} \cdot A \cdot B = A^{-1} \cdot G ,$$

and since  $A^{-1} \cdot A = I$ , the identity matrix,

$$A^{-1} \cdot A \cdot B = I \cdot B = B ,$$

and therefore

$$B = A^{-1} \cdot G$$

Let

$$A^{-1} = C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix}$$

Then

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = B = C \cdot G = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix}$$

Since two matrices are equal only if corresponding elements are equal, if we multiply the two matrices on the right and set the product equal to the B matrix, we obtain

$$b_k = c_{k1} \sum_{j=1}^n x_{1j} y_j + c_{k2} \sum_j x_{2j} y_j + \cdots + c_{km} \sum_j x_{mj} y_j$$

or

$$b_k = \sum_{i=1}^m \left[ c_{ki} \sum_j x_{ij} y_j \right] \quad , \quad k=1,2,\dots,m .$$

Appendix G

Residual sum of squares derivation.

We have that

$$SSE = \sum e^2 = \sum \left[ y - \sum_{i=1}^m b_i x_i \right]^2 .$$

Expanding the right-hand side,

$$SSE = \sum y^2 - \sum_i \left[ b_i \cdot \sum x_i y \right] + \sum_{i=1}^m b_i \left[ \sum_{k=1}^m b_k \cdot \left[ \sum x_i x_k - \sum x_i y \right] \right]$$

But the values in parenthesis in the third term are the normal equations, where

$$\sum_{k=1}^m \left[ b_k \cdot \sum x_i x_k \right] = \sum x_i y$$

which reduces the third term to zero, leaving

$$\sum e^2 = \sum y^2 - \sum_{i=1}^m \left[ b_i \cdot \sum x_i y \right] ,$$

where " $\sum$ " represents summation over all  $n$  data points of the indicated product.

Appendix H

Derivation for  $R^2$ .

In the regression chapter,  $\hat{y}$  was defined as  $(\hat{Y} - \bar{Y})$ , the difference between the estimated Y value and the mean.

It can be shown that

$$\begin{aligned}\sum_j \hat{y}_j^2 &= \sum_j (\hat{Y}_j - \bar{Y})^2 \\ &= b_1 \sum x_1 Y + b_2 \sum x_2 Y + \cdots + b_k \sum x_k Y \\ &= R^2 \sum (Y - \bar{Y})^2 \\ &= R^2 \sum Y^2\end{aligned}$$

so that

$$R^2 = (\sum \hat{y}^2) / (\sum y^2) .$$

$R^2$  is then the ratio of deviations of the estimated Y (by regression) from its' mean value. It is these deviations from the mean, accounted for by regression for Y's dependence on the  $X_i$ , which can be identified as coming from variation in the  $X_i$ , and can be removed. Hence,  $R^2$  represents the variation in Y accounted for by regression, over the total amount of variation, yielding the proportion of variation in Y due to variation in the  $X_i$ , upon which it is dependent.

Appendix I

The computer program listing follows.

C.....  
C  
C  
C  
C  
C  
C  
C  
C  
C  
C.....

COMMAND FILE COV.CMD, CONTAINING THE NAMES OF THE  
MAIN PROGRAM AND ALL SUBPROGRAMS NECESSARY FOR EX-  
ECUTION OF THE ANALYSIS OF COVARIANCE PROGRAM  
PACKAGE.

- COV.F4
- CORRE.F4
- DATA.F4
- AVDAT.F4
- AVCAL.F4
- MEAN.F4
- SUBT.F4
- ORDER.F4
- INVERT.F4
- PIVOT.F4
- RONORM.F4
- MULTR.F4
- COVAR.F4

C\*\*\*\*\*  
 C THE MAIN PROGRAM FOR THE ANALYSIS OF COVARIANCE  
 C

C SUBPROGRAMS REQUIRED!  
 C CORRE SUBT AVCAL AVDAT  
 C PIVOT MEAN DATA RON  
 C COVAR ORDER MULTR INFERT  
 C\*\*\*\*\*

C THE FOLLOING DIMENSIONS MUST BE GREATER THAN OR EQUAL  
 C TO THE NUMBER OF FACTORS

C Y IS THE VARIATE, DIMENSIONED FOR THE CUMULATIVE PRO-  
 C DUCT OF EACH FACTOR PLUS ONE, (LEVEL(I)+1), X(I,J) IS  
 C THE COVARIATE, DIMENSIONED WITH I=NUMBER OF COVARIATES  
 C AND J=SAME AS DIMENSION FOR Y

C THE FOLLOWING DIMENSIONS ARE EITHER FOR K, THE NUMBER  
 C FACTORS, OR 2 TO THE K-TH POWER MINUS 1, ((2\*\*K)-1).

DIMENSION Y(100),X(5,100),HEAD(5),LEVEL(5),ISTEP(5),KOUNT(5),  
 1LASTS(5),SMEAN(5),XBAR(6),NDF2(5),UF(5),SX(36),  
 2STD(6,7),D(6),RY(6),B(6),T(6),RX(40),R(40),XSSE(30),XYSSE(30),  
 3SUMSQX(30,32),SUMXY(30,32),SUMSQ(32),NDF(6),  
 4XFPE(30,5),XYFPE(30,5),YFPE(6),R2(5),DF(5),AFSSE(5),F(5),  
 5PR1(3)  
 WRITE(4,1)

C IN AND OUT DEVICES, IN THAT ORDER, ENTERED BY TELETYPE.  
 C THE OUT DEVICE RECEIVES IN-PROGRESS MESSAGES  
 C THE IN DEVICE TELLS PROGRAM WHERE TO READ DATA  
 C MI - IN DEVICE NUMBER  
 C MO - OUT DEVICE

1 FORMAT(26H ENTER I/O DEVICE NUMBERS: \$)  
 READ(4,2)MI,MO  
 2 FORMAT(2G)

C .....  
 C READ(MI,3) PR1,N,M,NS,KF,KC

C PROBLEM PARAMETER CARDS  
 C PR1.....PROBLEM NAME OR NUMBER  
 C N.....NUMBER OF PIECES OF DATA (Y)  
 C M.....NUMBER OF VARIABLES (VARIATE PLUS COVARIATES)  
 C NS.....NUMBER OF PROBLEMS ENTERED ON DATA FILE  
 C KF.....NUMBER OF FACTORS  
 C KC.....NUMBER OF COVARIATES

```

      IF(NS)108,108,109
108  WRITE(M0,4)
      4  FORMAT(51H NUMBER OF SELECTIONS NOT SPECIFIED, JOB TERMINATED)
      GO TO 500
109  DO 200 I=1,NS
      READ(MI,5)((HEAD(J),LEVEL(J)),J=1,KF)
C
C      HEAD....FACTOR NAMES OR LABELS
C      LEVEL...NUMBER OF LEVELS FOR FACTOR J
C.....
C
      3  FORMAT(3A2,5G)
C
C      READ ALL DATA AND CALCULATE TOTAL SUMS OF SQUARES
C
      CALL CORRE(N,M,RX,R,B,D,T,MI)
      REWIND 15
      DO 110 J=1,M
      L=J+(M*M-M)
110  RY(J)=RX(L)
      KC2=(KC+1)*(KC+1)
      5  FORMAT(A3,I)
C
C      BEGIN CREATION OF OUTPUT FILE FOR16.DAT
C
      WRITE(16,6)PR1
      6  FORMAT(/12H+++PROBLEM: ,3A2)
      WRITE(16,17)NS
      17  FORMAT(/2X,12H SELECTION: ,I3)
      WRITE(16,7)((HEAD(I),LEVEL(I)),I=1,KF)
      7  FORMAT(/2X,9H FACTOR: ,A3,5X,7HLEVELS:,I3)
      N=LEVEL(1)
      DO 102 J=2,KF
102  N=N*LEVEL(J)
C
C      REREAD DATA IN VARIATE,COVARIATE ARRAYS Y(N),X(KC,N)
      DO 400 K=1,N
C
      400 READ(15,8)(X(J,K),J=1,KC),Y(K)
      8  FORMAT(6F8.2)
      CALL AVDAT(KF,KC,LEVEL,N,Y,X,L,ISTEP,KOUNT)
      CALL AVCAL(KF,KC,LEVEL,Y,X,L,ISTEP,LASTS)
      CALL MEAN(KF,KC,LEVEL,Y,X,GMEAN,SUMSQ,NDF,SMEAN,MSTEP,KOUNT,
1  LASTS,SUMSQX,SUMXY)
      CALL SUBT(N,RX,RY,KF,KC,XSSE,XYSSE,YSSE,SUMSQX,SUMXY,SUMSQ,
1  NEDF,NDF,XFPE,XYFPE,YFPE)
      KC1=KC+1
      J=0
      IF(KC-1)70,71,70
C

```

```

C      ANALYSIS FOR SINGLE COVARIATE PROBLEMS
C
71  DO 72 J=1,KC
    RE=(XYSSE(J)*XYSSE(J))/XSSE(J)
    WRITE(16,155)RE
155  FORMAT(/36H ADJUSTMENT FOR ERROR SUM OF SQUARES//G)
72  ASSE=YSSE-RE
    WRITE(16,156)
156  FORMAT(/30H ADJUSTMENT(S) FOR FACTOR PLUS/22H ERROR SUMS
1 OF SQUARES//)
    DO 73 J=1,KF
    RM=(XYFPE(1,J)*XYFPE(1,J))/XFPE(1,J)
    WRITE(16,57)HEAD(J),RM
    AFSSE(J)=YFPE(J)-RM
    SX(J)=XSSE(J)
    EDF=NEDF
    DF(J)=NDF(J)
73  F(J)=((AFSSE(J)-ASSE)/DF(J))/(ASSE/EDF)
    GO TO 19

```

```

C
C      ANALYSIS FOR MULTIPLE COVARIANCE PROBLEMS
C
C . . . . .
C

```

```

C      REARRANGE ERROR SUMS OF SQUARES ARRAYS
70  DO 201 K=1,KC
    JK=KC*KC+KC+K
    XSSE(JK)=XYSSE(K)
    J=J+(KC+1)
    XSSE(J)=XYSSE(K)
201  CONTINUE
    K=(KC+1)*(KC+1)
    XSSE(K)=YSSE
    DO 202 J=1,K
202  SX(J)=XSSE(J)

```

```

C      CALCULATE CORRELATION COEFFICIENTS FOR
C      ERROR SUMS OF SQUARES
C

```

```

    KK=0
    JK=0
    DO 220 J=1,KC1
    JK=JK+KK*KC1+1
    STD(J,7)=SQRT(ABS(XSSE(JK)))
    KK=1
220  CONTINUE
    KK=0
    DO 230 J=1,KC1
    DO 230 K=J,KC1
    JK=J+(K*K=K)/2
    L=KC1*(K=1)+J

```

```

XSSE(JK)=XSSE(L)/(STD(J,7)*STD(K,7))
230 CONTINUE
C
C   REARRANGE FACTOR PLUS ERROR SUM OF SQUARES ARRAYS
C
DO 280 L=1,KF
J=0
DO 300 K=1,KC
JK=KC*KC+KC+K
XFPE(JK,L)=XFPE(K,L)
J=J+KC1
XFPE(J,L)=XFPE(K,L)
300 CONTINUE
K=KC1+KC1
XFPE(K,L)=XFPE(L)
C
C   CALCULATE CORRELATION COEFFICIENTS FOR
C   FACTOR PLUS ERROR SUMS OF SQUARES
C
KK=0
JK=0
DO 320 J=1,KC1
JK=JK+KK*KC1+1
STD(J,L)=SQRT(ABS(XFPE(JK,L)))
KK=1
320 CONTINUE
DO 330 J=1,KC1
DO 330 K=J,KC1
JK=J+(K-K)/2
L1=KC1*(K-1)+J
XFPE(JK,L)=XFPE(L1,L)/(STD(J,L)*STD(K,L))
330 CONTINUE
280 CONTINUE
335 M=KC1
NDEP=KC1
K=KC
L=(M*(M+1))/2
C
C   CALCULATE MULTIPLE CORRELATION COEFFICIENT FOR ERROR
C
DO 340 J=1,L
340 R(J)=XSSE(J)
CALL ORDER(M,R,NDEP,K,RX,RY)
CALL INVERT(RX,KC,KC,DET)
IF(DET) 112,111,112
111 WRITE(MO,9)
9   FORMAT(40H MATRIX SINGULAR. THIS SELECTION SKIPPED)
GO TO 200
112 CALL MULTR(N,K,RX,RY,RM)
WRITE(16,55)RM
55  FORMAT(/33H MULTIPLE CORRELATION COEFFICIENT /26H

```

```

1FOR ERROR SUMS OF SQUARES// 10X,G)
  RE2=RM
  WRITE(16,56)
56  FORMAT(//38H MULTIPLE CORRELATION COEFFICIENTS FOR/
134H FACTOR PLUS ERROR SUMS OF SQUARES//)
  DO 350 J=1,KF
  M=KC1
  NDEP=KC1
  L=(M*(M+1))/2
C
C      CALCULATE MULTIPLE CORRELATION COEFFICIENT
C      BQR FACTOR PLUS ERROR SUMS
C
  DO 355 K=1,L
355  R(K)=XFPE(K,J)
  CALL ORDER(M,R,NDEP,KC,RX,RY)
  CALL INVERT(RX,KC,KC,DET)
  IF(DET) 113,111,113
113  CALL MULTR(N,KC,RX,RY,RM)
  WRITE(16,57)HEAD(J),RM
  57  FORMAT(2X,A3,8H + ERROR,G/)
  R2(J)=RM
350  CONTINUE
C
C      CALCULATE ADJUSTED SUMS OF SQUARES
C
  CALL COVAR(ASSE,YSSE,AFSSE,YFPE,NEDF,NDF,VAR,F,RE2,R2,KF)
C
C      PRINT OUTPUT ON OUTPUT FILE FOR16,DAT
C
19  WRITE(16,66)
66  FORMAT(//'.....
1.....'//20X,30H UNADJUSTED VARIATION
2ANALYSIS)
  WRITE(16,222)
20  FORMAT(//10H SOURCE OF,18X,7HSUMS OF,10X,10HDEGREES OF,9X,4HMEAN/
110H VARIATION,18X,7HSQUARES,11X,7HFREEDOM,10X,7HSQUARES/)
222  FORMAT(//10H SOURCE OF,8X,7HSUMS OF,8X,10HDEGREES OF,8X,
14HMEAN,9X,1HF/10H VARIATION,8X,7HSQUARES,9X,7HFREEDOM,
29X,7HSQUARES,5X,6HVALUES/)
  NEF=NEDF+KC
  EMEAN=YSSE/FLOAT(NEF)
  DO 600 J=1,KF
  UF(J)=SMEAN(J)/EMEAN
600  WRITE(16,21)HEAD(J),SUMSQ(J),NDF(J),SMEAN(J),UF(J)
21  FORMAT(2X,A3,1X,F20.5,5X,16,F20.5,F11.3)
  WRITE(16,29)YSSE,NEF,EMEAN
29  FORMAT(2X,5HERROR,F19.5,5X,16,F20.5)
  WRITE(16,64)
28  FORMAT(//10H SOURCE OF,10X,13HADJUSTED SUMS,11X,10HDEGREES
1 OF/10H VARIATION,11X,11HOF SQUARES,13X,7HFREEDOM/)

```

```

WRITE(16,67)
WRITE(16,28)
64  FORMAT(//'.....
1.....'//21X,28H ADJUSTED VARIATION ANALYSIS)
67  FORMAT(//39H ADJUSTED (FACTOR PLUS ERROR) VARIATION)
22  FORMAT(//10H SOURCE OF,10X,13HADJUSTED SUMS,11X,10HDEGREES
1 OF,10X,1HF/10H VARIATION,11X,11HOF  SQUARES,13X,7HFREEDOM
2,10X,6HVALUES/)
DO 610 J=1,KF
NDF2(J)=NDF(J)+NEDF
610 WRITE(16,23)HEAD(J),AFSSE(J),NDF2(J)
23  FORMAT(1X,A3,8H + ERROR,F19.5,12X,I6,F20.5)
WRITE(16,68)
68  FORMAT(////26H ADJUSTED FACTOR VARIATION)
WRITE(16,22)
DO 601 J=1,KF
AFSSE(J)=AFSSE(J)-ASSE
601 WRITE(16,32)HEAD(J),AFSSE(J),NDF(J),F(J)
32  FORMAT(2X,A3,6X,F20.5,12X,I6,F21.5)
WRITE(16,31)ASSE,NEDF
31  FORMAT(2X,5HERROR,4X,F20.5,12X,I6)
WRITE(16,61)
61  FORMAT(////'-----
1-----'//15X,39HSIGNIFICANCE OF VARIATI
2ON OF COVARIATES)
JK=0
KK=0
JT=0
DO 69 J=1,KC
WRITE(16,62)J
62  FORMAT(////10H COVARIATE,I3/)
WRITE(16,222)
JK=JK+KK*KC1+1
JT=JT+KK*KC+1
DO 63 K=1,KF
XMEAN=SUMSQX(JT,K)/NDF(K)
FX=XMEAN/(SX(JK)/FLOAT(NEF))
63  WRITE(16,21)HEAD(K),SUMSQX(JT,K),NDF(K),XMEAN,FX
XEMEAN=SX(JK)/FLOAT(NEF)
WRITE(16,29)SX(JK),NEF,XEMEAN
KK=1
WRITE(16,61)
81  FORMAT(////'-----
1-----')
69  CONTINUE
200 CONTINUE
500 STOP
END

```

SUBROUTINE CORRE(N,M,RX,R,B,D,T,MI)

PURPOSE: COMPUTES SUMS OF SQUARES AND CROSS PRODUCTS  
FOR VARIATE AND COVARIATES.

NEW PARAMETERS:

N            -NUMBER OF OBSERVATIONS  
M            -NUMBER OF VARIABLES  
RX           -OUTPUT MATRIX (M X M) CONTAINING SUMS OF  
              SQUARES AND CROSS PRODUCTS  
R            -INTERNAL WORKING ARRAY  
B,D,T        -WORKING VECTORS, LENGTH M  
MI           -INPUT DEVICE NUMBER (FOR SUBROUTINE DATA)

SUBPROGRAMS REQUIRED:  
SUBROUTINE DATA(M,D,MI)

.....  
DIMENSION RX(40),R(40),D(6),T(6),B(6)

INITIALIZATION

DO 100 J=1,M  
B(J)=0.0  
100 T(J)=0.0  
K=((M\*M)+M)/2  
DO 102 I=1,K  
102 R(I)=0.0

FN=N  
L=0

READ M OBSERVATIONS AND CALCULATE TEMPORARY  
MEANS IN T(J)

IF(N=M) 130,130,135  
130 KK=N  
GO TO 137  
135 KK=M  
137 DO 140 I=1,KK  
CALL DATA(M,D,MI)  
DO 140 J=1,M  
T(J)=T(J)+D(J)  
L=L+1  
140 RX(L)=D(J)  
FKK=KK  
DO 150 J=1,M

```

150 T(J)=T(J)/FKK
C
C      CALCULATE SUMS OF CROSS PRODUCTS OF DEVIATIONS FROM
C      TEMPORARY MEANS FOR M OBSERVATIONS
C
      L=0
      DO 180 I=1,M
      JK=0
      DO 170 J=1,M
      L=L+1
170 D(J)=RX(L)-T(J)
      DO 180 J=1,M
      B(J)=B(J)+D(J)
      DO 180 K=1,J
      JK=JK+1
180 R(JK)=R(JK)+D(J)*D(K)
C
      IF(N-KK) 205,205,185
C
C      READ REMAINING OBSERVATIONS AND CALCULATE SUMS
C      OF CROSS PRODUCTS OF DEVIATIONS
C
185 KK=N-KK
      DO 200 I=1, KK
      JK=0
      CALL DATA(M,D,MI)
      DO 190 J=1,M
      D(J)=D(J)-T(J)
190 B(J)=B(J)+D(J)
      DO 200 J=1,M
      DO 200 K=1,J
      JK=JK+1
200 R(JK)=R(JK)+D(J)*D(K)
205 JK=0
C
C      SUBTRACT CORRECTION FACTOR SUMS
C
      DO 210 J=1,M
      DO 210 K=1,J
      JK=JK+1
210 R(JK)=R(JK)-B(J)*B(K)/FN
C
C      STORE CROSS PRODUCTS (AND PRODUCTS) IN RX ARRAY(M X M)
C
      DO 230 J=1,M
      DO 230 K=J,M
      JK=J+(K-K=K)/2
      L=M*(J-1)+K
      RX(L)=R(JK)
      L=M*(K-1)+J
230 RX(L)=R(JK)

```

RETURN  
END

SUBROUTINE DATA(M,D,MI)

PURPOSE: READ AN OBSERVATION (VARIATE AND CORRESPOND-  
ING COVARIATES - M VALUES) FROM THE INPUT DEVICE,  
CALLED BY SUBROUTINE CORRE.

SUBPROGRAMS REQUIRED: NONE

PARAMETERS:

M -NUMBER OF VARIABLES IN AN OBSERVATION  
D -OUTPUT VECTOR OF LENGTH M CONTAINING THE  
OBSERVATION VALUES  
MI -INPUT DEVICE NUMBER

.....  
DIMENSION D(6)

1 FORMAT(6G)

READ(MI,1) (D(I),I=1,M)

OBSERVATION WRITTEN ON DISK FILE 15 TO BE REREAD LATER

WRITE(15,2)(D(I),I=1,M)

2 FORMAT(6F8.2)

RETURN

END

```

SUBROUTINE AVDAT(KF,KC,LEVEL,N,Y,X,L,ISTEP,KOUNT)
C
C   PURPOSE:  TO SPACE DATA IN THE ARRAYS ACCORDING TO
C             FACTOR AND LEVEL CLASSIFICATION
C
C   NEW PARAMETERS
C       L       -COUNTING VARIABLE(DATA)
C       ISTEP   -COUNTING VARIABLE(FACTORS)
C       KOUNT   -COUNTING VARIABLE(LEVELS)
C
C.....
C
C   DIMENSION LEVEL(5),Y(100),X(5,100),ISTEP(5),KOUNT(5)
C   M=LEVEL(1)+1
C   DO 105 I=2,KF
105  M=M*(LEVEL(I)+1)
C
C       CALCULATE THE TOTAL AREA REQUIRED FOR SPACED DATA
C
C   N1=M+1
C   N2=N+1
C   DO 107 I=1,N
C   N1=N1-1
C   N2=N2-1
C   Y(N1)=Y(N2)
C   DO 106 J=1,KC
106  X(J,N1)=X(J,N2)
107  CONTINUE
C
C       COMPUTE COUNTING VARIABLES
C
C   ISTEP(1)=1
C   DO 110 I=2,KF
110  ISTEP(I)=ISTEP(I-1)*(LEVEL(I-1)+1)
C   DO 115 I=1,KF
115  KOUNT(I)=1
C
C       SPACE DATA
C
C   N1=N1-1
C   DO 135 I=1,N
C   L=KOUNT(1)
C   DO 120 J=2,KF
120  L=L+ISTEP(J)*(KOUNT(J)-1)
C   N1=N1+1
C   Y(L)=Y(N1)
C   DO 121 J=1,KC
121  X(J,L)=X(J,N1)
C

```

```
C      INCREMENT COUNTERS
C
DO 130 J=1,KF
IF(KOUNT(J)-LEVEL(J)) 124,125,124
124 KOUNT(J)=KOUNT(J)+1
GO TO 135
125 KOUNT(J)=1
130 CONTINUE
135 CONTINUE
RETURN
END
```

```

SUBROUTINE AVCAL(KF,KC,LEVEL,Y,X,L,ISTEP,LASTS)
C
C   PURPOSE:  CALCULATE GROUP SUMS OF SQUARES FOR VARIATE
C             AND COVARIATES FOR COMBINATION INTO FACTOR SUMS OF
C             SQUARES IN SUBROUTINE MEAN.
C
C   NEW PARAMETERS:
C
C       LASTS  -IDENTIFIES THE LAST DATA POSITION FOR
C              EACH FACTOR
C
C   .....
C
C   DIMENSION LEVEL(5),Y(100),X(5,100),ISTEP(5),LASTS(5),SUMX(5)
C   KT=1
C   LASTS(1)=L+1
C
C   CALCULATE THE LAST DATA POSITION FOR EACH FACTOR
C
C   DO 145 I=2,KF
145  LASTS(I)=LASTS(I-1)+ISTEP(I)
150  DO 175 I=1,KF
    L=1
    LL=1
    SUM=0.0
    DO 149 J=1,KC
149  SUMX(J)=0.0
    NN=LEVEL(I)
    FN=NN
    INCRE=ISTEP(I)
    LAST=LASTS(I)
C
C   COMPUTE SUMMATIONS OF DATA
C
C   DO 160 J=1,NN
155  SUM=SUM+Y(L)
    DO 156 MN=1,KC
156  SUMX(MN)=SUMX(MN)+X(MN,L)
    L=L+INCRE
160  CONTINUE
    Y(L)=SUM
    DO 157 MN=1,KC
157  X(MN,L)=SUMX(MN)
C
C   COMPUTE CORRECTION FACTOR SUMMATIONS
C
C   DO 165 J=1,NN
    Y(LL)=FN*Y(LL)-SUM
    DO 161 MN=1,KC

```

```
161 X(MN,LL)≡FN*X(MN,LL)-SUMX(MN)
C
C      RESET SUM VARIABLES AND INCREMENT COUNTERS
C
165 LL≡LL+INCRE
    SUM≡0.0
    DO 166 MN=1,KC
166 SUMX(MN)≡0.0
    IF(L=LAST) 167,175,175
167 IF(L=LAST+INCRE) 168,168,170
168 L=L+INCRE
    LL≡LL+INCRE
    GO TO 155
170 L=L+INCRE+1-LAST
    LL≡LL+INCRE+1-LAST
    GO TO 155
175 CONTINUE
    RETURN
    END
```



```

IF(KOUNT(I)-LASTS(I)) 210,250,210
210 IF(L) 220,220,240
220 KOUNT(I)=KOUNT(I)+1
IF(KOUNT(I)-LEVEL(I)) 230,230,250
230 L=L+MSTEP(I)
GO TO 260
240 IF(KOUNT(I)-LEVEL(I)) 230,260,230
250 KOUNT(I)=0
260 CONTINUE
IF(L) 285,285,270
270 DO 271 I=1,KF
IF(L.EQ.MSTEP(I))KL=KL+1
271 CONTINUE
IF(KL-1) 273,272,273
272 JJ=1
SUMSQ(L)=SUMSQ(L)+Y(NN)*Y(NN)
JK=1
JL=0
274 JL=JL+1
SUMSQX(JL,L)=SUMSQX(JL,L)+X(JJ,NN)*X(JK,NN)
JJ=JJ+1
IF(JJ-KC)274,274,276
276 JJ=1
IF(JK-KC)278,277,277
278 JK=JK+1
GO TO 274
277 DO 280 J=1,KC
280 SUMXY(J,L)=SUMXY(J,L)+X(J,NN)*Y(NN)
273 NN=NN+1
GO TO 200
C CALCULATE THE GRAND OVERALL MEAN
285 FN=N
GMEAN=Y(NN)/FN
C
C CALCULATE DEGREES OF FREEDOM (LEVELS - 1) FOR
C EACH FACTOR
C
DO 310 I=2,KF
310 MSTEP(I)=0
NN=0
MN=1
MSTEP(1)=1
320 ND1=1
ND2=1
DO 340 I=1,KF
IF(MSTEP(I)) 330,340,330
330 ND1=ND1*LEVEL(I)
ND2=ND2*(LEVEL(I)-1)
340 CONTINUE
FN1=ND1
FN1=FN*FN1

```

```
      FN2=ND2
      NN=NN+1
      SUMSQ(NN)=SUMSQ(MN)/FN1
      NDF(NN)=ND2
      SMEAN(NN)=SUMSQ(NN)/FN2
342  DO 342 J=1,KC2
      SUMSQX(J,NN)=SUMSQX(J,MN)/FN1
      DO 343 J=1,KC
343  SUMXY(J,NN)=SUMXY(J,MN)/FN1
      MN=MN*2
      IF(NN-KF) 345,370,370
C
C      INCREMENT COUNTERS FOR NEXT FACTOR
C
345  DO 360 I=1,KF
      IF(MSTEP(I)) 347,350,347
347  MSTEP(I)=0
      GO TO 360
350  MSTEP(I)=1
      GO TO 365
360  CONTINUE
365  MSUM=0
      DO 366 I=1,KF
366  MSUM=MSUM+MSTEP(I)
      IF(MSUM-1) 345,320,345
370  RETURN
      END
```

```
SUBROUTINE SUBT(N,RX,RY,KF,KC,XSSE,XYSSE,YSSE,SUMSQX,  
1SUMXY,SUMSQ,NDF,NDF,XFPE,XYFPE,YFPE)
```

```
PURPOSE:  TO CALCULATE THE ERROR SUMS OF SQUARES LINE  
AND THE [FACTOR PLUS ERROR] SUMS OF SQUARES LINE FOR  
THE ANCOVA TABLE
```

```
NEW PARAMETERS:  
XSSE,XYSSE,YSSE  -ERROR SUMS OF SQUARES FOR COVARIATES,  
CROSS PRODUCTS, AND VARIATE, RESPECTIVELY  
XFPE,XYFPE,YFPE  -[FACTOR PLUS ERROR] SUMS OF SQUARES  
NDF              -ERROR DEGREES OF FREEDOM
```

```
.....  
DIMENSION RX(40),RY(6),XSSE(30),XYSSE(30),SUMSQX(30,32),  
1SUMXY(30,32),SUMSQ(32),NDF(6),XFPE(30,5),XYFPE(30,5),YFPE(6)  
L=0
```

```
INITIALIZE ERROR S,S. AS TOTAL S,S. BEFORE SUBTRACTION
```

```
DO 3 J=1,KC  
DO 2 I=1,KC  
L=L+1  
2 XSSE(L)=RX(L)  
L=L+1  
3 CONTINUE  
M=0  
L=0
```

```
COMPUTE ERROR SUMS OF SQUARES
```

```
DO 7 I=1,KC  
DO 6 J=1,KC  
M=M+1  
L=L+1  
DO 5 K=1,KF  
5 XSSE(L)=XSSE(L)-SUMSQX(M,K)  
6 CONTINUE  
L=L+1  
7 CONTINUE  
DO 10 I=1,KC  
10 YYSSE(I)=RY(I)  
DO 15 I=1,KC  
DO 15 J=1,KF  
15 YYSSE(I)=YYSSE(I)-SUMXY(I,J)  
KC1=KC+1  
YSSE=RY(KC1)
```

```

DO 40 I=1,KF
40  YSSE=YSSE+SUMSQ(I)
C
C      CALCULATE ERROR DEGREES OF FREEDOM
C
NEDF=N-1
DO 50 I=1,KF
50  NEDF=NEDF-NDF(I)
NEDF=NEDF-KC
M=0
L=0
C
C      CALCULATE FACTOR PLUS ERROR SUMS OF SQUARES
C
DO 60 I=1,KF
DO 56 J=1,KC
DO 55 K=1,KC
M=M+1
L=L+1
55  XFPE(M,I)=XSSE(M)+SUMSQX(L,I)
M=M+1
56  CONTINUE
M=0
L=0
60  CONTINUE
DO 70 J=1,KF
DO 65 K=1,KC
65  XYFPE(K,J)=XYSSE(K)+SUMXY(K,J)
70  CONTINUE
DO 85 I=1,KF
85  YFPE(I)=YSSE+SUMSQ(I)
RETURN
END

```

SUBROUTINE ORDER(M,R,NDEP,K,RX,RY)

PURPOSE: REORDER CORRELATION COEFFICIENTS TO FORM  
MATRIX OF INTERCORRELATIONS AMONG COVARIATES AND  
A VECTOR OF INTERCORRELATIONS BETWEEN COVARIATES  
AND THE VARIATE, ( RX AND RY RESPECTIVELY),

NEW PARAMETERS:

R -INPUT MATRIX OF CORRELATION COEFFICIENTS,  
STORED AS A ONE-DIMENSIONAL ARRAY  
RX -OUTPUT MATRIX (KC X KC) OF INTERCORRELATIONS  
OF COVARIATES  
RY -OUTPUT VECTOR (LENGTH KC) OF INTERCORRELATIONS  
OF VARIATE AND COVARIATES  
NDEP -POSITION OF VARIATE IN OBSERVATION DATA ROWS

DIMENSION R(40),RX(40),RY(6)

MM=0

DO 130 J=1,K

L=J+(NDEP\*NDEP-NDEP)/2

RELOCATE COVARIATE/VARIATE INTERCORRELATION

125 RY(J)=R(L)

DO 130 I=1,K

IF(I=J) 127,127,128

127 L=I+(J+J-J)/2

GO TO 129

128 L=J+(I+I-I)/2

129 MM=MM+1

RELOCATE COVARIATE INTERCORRELATIONS

130 RX(MM)=R(L)

RETURN

END

SUBROUTINE INVERT(A,N,M,DET)

PURPOSE: INVERT A SQUARE (OR RECTANGULAR, IF THE  
NUMBER OF COLUMNS NOT LESS THAN THE NUMBER OF ROWS)  
MATRIX.

PARAMETERS:

A	-INPUT	MATRIX TO BE INVERTED
	OUTPUT	RESULTING INVERTED MATRIX
N	-INPUT	NUMBER OF ROWS IN MATRIX
M	-INPUT	NUMBER OF COLUMNS IN MATRIX
DET	-OUTPUT	DETERMINANT OF FIRST N ROWS

SUBPROGRAMS REQUIRED:

SUBROUTINE PIVOT  
FUNCTION RONORM

.....

DIMENSION A(N,M),S(50)  
INTEGER R(50),C(50)  
DATA TOL/0.001/  
DET=1.

START INVERSION PROCESS FOR N CYCLES

DO 40 K=1,N  
TEST=0.  
DO 30 I=1,N  
DO 20 J=1,N

BEGIN SEARCH FOR PIVOT ELEMENT

IF(K.EQ.1)GO TO 15  
DO 10 L=1,K-1  
IF(I.EQ.R(L))GO TO 30  
10 IF(J.EQ.C(L))GO TO 20  
15 IF(ABS(A(I,J)).LE.TEST)GO TO 20  
R(K)=I  
C(K)=J  
TEST=ABS(A(I,J))  
20 CONTINUE  
30 CONTINUE  
IF(TEST/RONORM(A,N,M).GT.TOL)GO TO 35  
DET=0.  
RETURN  
35 DET=DET\*A(R(K),C(K))  
40 CALL PIVOT(A,N,M,R(K),C(K))

```

C
C      AFTER INVERSION IS COMPLETE, UNSCRAMBLE THE ARRAY
C
      DO 60 J=1,M
      DO 50 I=1,N
50     S(C(I))=A(R(I),J)
      DO 60 I=1,N
60     A(I,J)=S(I)
      DO 80 I=1,N
      DO 70 J=1,N
70     S(R(J))=A(I,C(J))
      DO 80 J=1,N
80     A(I,J)=S(J)
      DO 90 K=1,N
90     S(R(K))=C(K)
      INT=0
      DO 100 I=1,N-1
      DO 100 J=I+1,N
      IF(S(J).GE.S(I))GO TO 100
      TEST=S(I)
      S(I)=S(J)
      S(J)=TEST
      INT=INT+1
100    CONTINUE
C
C      SET SIGN ON DETERMINANT
C
      IF(INT/2*2.NE.INT)DET=-DET
      RETURN
      END

```

SUBROUTINE PIVOT(A,M,N,R,C)

PURPOSE: PERFORMS ONE STANDARD PIVOT OPERATION IN  
PLACE FOR EACH ENTRY INTO THE ROUTINE(MATH PROGRAM  
LIBRARY MA6006)

PARAMETERS:

A -INPUT MATRIX TO BE PIVOTED  
OUTPUT PIVOTED MATRIX  
M -INPUT NUMBER OF ROWS IN MATRIX  
N -INPUT NUMBER OF COLUMNS IN MATRIX  
R -INPUT ROW INDEX OF PIVOT ELEMENT  
C -INPUT COLUMN INDEX OF PIVOT ELEMENT

```

.....
DIMENSION A(M,N)
INTEGER R,C
A(R,C)=1./A(R,C)
DO 10 J=1,N
10 IF(J.NE.C)A(R,J)=A(R,J)*A(R,C)
DO 30 I=1,M
IF(I.EQ.R)GO TO 30
DO 20 J=1,N
20 IF(J.NE.C)A(I,J)=A(I,J)-A(I,C)*A(R,J)
30 CONTINUE
DO 40 I=1,M
40 IF(I.NE.R)A(I,C)=-A(I,C)*A(R,C)
RETURN
END

```

```
FUNCTION RONORM(A,M,N)
C
C      PURPOSE:  SELECTS THE ROW NORM BY ADDING ABSOLUTE
C      VALUES OF THE ELEMENTS IN EACH ROW AND SELECTING
C      THE LARGEST OF THESE VALUES FOR THE ROW NORM VALUE.
C      EMPLOYED IN SUBROUTINE INVERT.
C
C      DIMENSION A(M,N)
C      RONORM=0,
C
C      COMPUTE M SUMS BY ROW OF ROW ELEMENTS
C
C      DO 20 I=1,M
C      SUM=0
C      DO 10 J=1,N
10    SUM=SUM+ABS(A(I,J))
C
C      STORE LARGEST ABSOLUTE SUM TO THIS POINT IN RONORM
C
20    IF(SUM.GT.RONORM)RONORM=SUM
C      RETURN
C      END
```

SUBROUTINE MULTR(N,K,RX,RY,RM)

PURPOSE: CALCULATE THE MULTIPLE CORRELATION COEFFICIENT SQUARED USING INPUT MATRIX RX AND VECTOR RY.

NEW PARAMETERS:

RM -MULTIPLE CORRELATION COEFFICIENT, SQUARED  
B -INTERNAL WORK VECTOR

.....  
DIMENSION RX(40),RY(6),B(6)

CALCULATE WEIGHTS

DO 100 J=1,K  
100 B(J)=0.0  
DO 110 J=1,K  
L1=K\*(J-1)  
DO 110 I=1,K  
L=L1+I  
110 B(J)=B(J)+RY(I)\*RX(L)  
RM=0.0

CALCULATE MULTIPLE CORRELATION COEFFICIENT

DO 120 I=1,K  
RM=RM+B(I)\*RY(I)  
120 CONTINUE  
RETURN  
END

```

SUBROUTINE COVAR(ASSE,YSSE,AFSSE,YFPE,NEDF,NDF,VAR,F,RE2,R2,KF)
C
C   PURPOSE:  CALCULATE ADJUSTED SUMS OF SQUARES AND
C             F STATISTIC VALUES
C
C   NEW PARAMETERS:
C     ASSE      -ADJUSTED ERROR SUM OF SQUARES
C     AFSSE     -ADJUSTED FACTOR PLUS ERROR SUMS OF SQUARES
C     RE2,R2    -CORRELATION COEF. SQUARED FOR SSE, SS(F+E)
C     F         -F VALUES
C
C.....
C   DIMENSION R2(5),NDF(6),AFSSE(5),YFPE(6),F(5),DF(5)
C
C   ADJUST ERROR AND [FACTOR PLUS ERROR] SUMS OF SQUARES
C
C   ASSE=YSSE*(1.-RE2)
C   DO 360 J=1,KF
360  AFSSE(J)=YFPE(J)*(1.-R2(J))
C   EDF=NEDF
C   FKC=KC
C   VAR=ASSE/EDF
C   DO 363 I=1,KF
363  DF(I)=NDF(I)
C
C   CALCULATE F VALUES FOR ADJUSTED SUMS OF SQUARES
C
C   DO 365 J=1,KF
365  F(J)=((AFSSE(J)-ASSE)/DF(J))/VAR
C   RETURN
C   END

```

LITERATURE CITED

1. Anderson, R.L., and Bancroft, T.A., 1952, Statistical Theory in Research: New York, McGraw-Hill.
2. Bartlett, M.S., 1937, "Some Examples of Statistical Methods of Research in Agriculture": Jour. Roy. Stat. Soc. Supple., vol. 4, pp. 581-588.
3. Bennett, C.A., and Franklin, N.L., 1954, Statistical Analysis in Chemistry and the Chemical Industry: New York, John Wiley.
4. Cochran, W.G., 1946, "The Analysis of Covariance": Inst. Stat. Mim. Ser., vol. 6.
5. Cochran, W.G., 1947, "Consequences of Failing to Satisfy the Assumptions in the Analysis of Variance": Biometrics, vol. 3, no. 1.
6. Cochran, W.G., and Cox, G.M., 1950, Experimental Design: New York, John Wiley and Sons.
7. Cochran, W.G., Cox, G.T., and Eckhardt, "The Analysis of Lattice and Triple Lattice Experiments in Combined Varietal Tests": Iowa Agr. Expt. Station and Res. Bull., p. 281.
8. Coons, I., 1957, "The Analysis of Covariance as a Missing Plot Technique": Biometrics, vol. 13, no. 3, pp. 387-404.
9. Crampton, E.W., and Hopkins, J.W., 1934, "The Use of the Method of Partial Regression in Analysis of Comparative Feeding Trial Data, Part II": Journal of Nutrition, vol. 8, p. 329.
10. Das, 1953, "Analysis of Covariance in Two-Way Classification with Disproportionate Cell Frequencies": Jour. Ind. Soc. Agric. Stat., vol. 5, pp. 161-178.
11. David, H.A., 1963, The Method of Paired Comparisons: New York, Hafner.

12. DeLury, B., 1948, "The Analysis of Covariance":  
Biometrics, vol. 4, pp. 153-170.
13. Dixon, W.J., and Massey, F.J., 1957, Introduction to  
Statistical Analysis: New York, McGraw-Hill.
14. Dwyer, P.S., 1951, Linear Computations: New York, John  
Wiley.
15. Eisenhart, C., 1947, "The Assumptions Underlying the  
Analysis of Variance": Biometrics, vol. 3, no. 1.
16. Evans, S.H., and Anastasio, E.J., 1970, "Misuse of the  
Analysis of Covariance When Treatment Effect and  
Covariate are Compounded: Readings in Statistics  
for the Behavioral Sciences; Heerman, E., and  
Braskamp, L., ed., New Jersey, Prentice-Hall.
17. Federer, W.T., 1955, Experimental Design: New York,  
MacMillan Co.
18. Fisher, R.A., 1963, Statistical Methods for Research  
Workers, 13th Edition: London, Oliver and Boyd.
19. Freund, J.E., 1971, Mathematical Statistics: New Jersey,  
Prentice-Hall.
20. Goldberger, A., 1968, Topics in Regression Analysis:  
New York, MacMillan Co.
21. Graybill, F.A., 1961, An Introduction to Linear Stat-  
istical Models, vol. 1: New York, McGraw-Hill.
22. Hadley, G., 1968, Introduction to Business Statistics:  
San Francisco, Holden-Day.
23. Hahn, G.J., and Shapiro, S.S., 1967, Statistical Models  
in Engineering: New York, John Wiley.
24. Hazel, L.N., 1953, "The Covariance Analysis of Multiple  
Classification Tables with Unequal Subclass  
Numbers": Biometrics, vol. 2, pp. 21-25.
25. Henderson, P.L., Brown, S.E., and Hind, J.F., 1965,  
"Nonquantified Adjustment of Seasonality in  
Time Series Data": Readings in Applied Statistics;  
Peters, W.S., ed., New Jersey, Prentice-Hall.
26. Hicks, C.R., 1965, "The Analysis of Covariance": Readings  
in Applied Statistics; Peters, W.S., ed., New Jersey,  
Prentice-Hall.

27. Hoog, R.V., and Craig, A.T., 1970, Introduction to Mathematical Statistics: New York, MacMillan and Co.
28. Johnson, N.L., and Leone, F.C., 1964, Statistics and Experimental Design, vol. 1: New York, John Wiley and Sons.
29. Johnson, N.L., and Leone F.C., 1964, Statistics and Experimental Design, vol. 2: New York, John Wiley and Sons.
30. Kempthorne, 1952, The Design and Analysis of Experiments: New York, John Wiley and Sons.
31. Larson, H.J., 1973, Introduction to the Theory of Statistics: New York, John Wiley and Sons.
32. Miller, I, and Freund, J., 1965, Probability and Statistics for Engineers: New Jersey, Prentice-Hall.
33. Nair, K.R., 1939, "The Application of Covariance Technique to Field Experiments with Missing or Mixed-Up Yields": Sankhya, vol. 4, pp. 581-588.
34. Ostle, B., 1954, Statistics in Research: Iowa, Iowa State College Press.
35. Outhwaite, A.D., and Rutherford, 1955, "Covariance as an Alternative to Stratification in the Control of Gradients": Biometrics, vol. 11, pp. 431-440.
36. Snedecor, G.W., 1956, Statistical Methods: Iowa, Iowa State University Press.
37. Steel, R.G., 1954, "Which Dependent Variate? Y, or Y-X?": Mimeo Series BU-54-M, Biometrics Unit, New York, Cornell University.
38. Steel, R.G., and Torrie, J.H., 1960, Principles and Procedures of Statistics: New York, McGraw-Hill.
39. Villars, D.S., 1951, Statistical Design and Analysis of Experiments for Development Research: Iowa, W. C. Brown Co.
40. Weinburg, G.H., and Schumaker, J.A., 1962, Statistics, An Intuitive Approach: California, Brooks/Cole Pub. Co.
41. Wilkensen, G.N., 1957, "Analysis of Covariance with Incomplete Data": Biometrics, vol. 13, pp. 363-372.

42. Wishart, J., 1936, "Tests of Significance in Analysis of Covariance": Suppl. Jour. Roy. Stat. Soc., col. 3, pp. 79-82.