

MISSING PLOT TECHNIQUES

262
by

349 5839

CHING-LAN WU

B. A., National Taiwan University, 1969

A MASTER'S REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE


Department of Statistics

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1973

Approved by:


Major Professor

**THIS BOOK
CONTAINS
NUMEROUS
PAGES WITH
THE ORIGINAL
PRINTING ON
THE PAGE BEING
CROOKED.**

**THIS IS THE
BEST IMAGE
AVAILABLE.**

LD
2668
R4
1973
W78
C.2

ii

TABLE OF CONTENTS

Document	page
I. INTRODUCTION.....	1
II. ANALYSIS OF COVARIANCE TECHNIQUE.....	6
III. LEAST SQUARES ANALYSIS TECHNIQUE FOR LINEAR MODEL.....	13
IV. ANALYSIS OF COVARIANCE TECHNIQUE FOR LINEAR MODEL.....	17
V. NUMERICAL ILLUSTRATIONS.....	20
VI. DISCUSSION.....	29
VII. REFERENCE.....	31
VIII. ACKNOWLEDGMENT.....	33

I. INTRODUCTION

The occurrence of missing data in a statistically designed experiment requires some modification of the usual statistical techniques because the orthogonality, or balance, of the design is destroyed.

One of the first papers on the subject of estimating the yield of a missing unit in the field experimental work is stated by Anderson (1) to have been published by Allan and Wishart. They derived formulae and illustrated their uses for a single missing plot in a randomized block and in a latin square experiment. These estimation methods were expanded by Yates (16) to cover several missing units in a given experiment. The procedure he used is to minimize the error variance obtained when unknowns are substituted for the missing yields. The formula given by Yates for estimating the yield of a single missing unit in a randomized complete block experiment is

$$y = \frac{rB + tT - G}{(r - 1)(t - 1)}$$

where r = the number of blocks, t = the number of treatments in the experiment, B = the total yield of the remaining units in the block where the missing unit appears, T = the total of the yields of the treatment with missing unit, and G = the grand total. Similarly for a single missing unit in a latin squares,

$$y = \frac{r(R + C + T) - 2G}{(r - 1)(r - 2)}$$

where r = the number of rows, columns, or treatments; and R , C , and T represent the total yields of the remaining units in the row, column and treatment containing the missing unit. G is the grand total. He used these formulae for several missing units by means of iterative methods.

Yates also showed that in a complete analysis of variance of the augmented data, the treatment sum of squares is overestimated but may be corrected by subtracting a correction for bias. The formulas for correction for bias in two designs with one missing unit are given. For a randomized block experiment, the correction for bias is

$$\frac{(B - (t - 1)y)^2}{t(t - 1)},$$

which is subtracted from the treatment sum of squares.

The correction for bias in a latin square experiment is to subtract

$$\frac{(G - R - C - (r - 1)T)^2}{(r - 1)^2(r - 2)^2}$$

from the treatment sum of squares.

Anderson (1) derived some formulas for missing plots in split-plot experiments by minimizing the error variance. His covariance methods are used in the derivations which follow in order to furnish an easy means for correcting the bias in the treatment sum of squares, and of estimating the missing data. We assume that we have a split-plot experiment with r replications, a whole-plot, and b sub-plot treatments so that the total number of units is $N = rab$. Let the single missing sub-unit be that for the whole-plot treatment a_1 , sub-plot treatment b_1 and replication r_1 . Also let A_1 be the total yield of all existing units with treatment a_1 , B_1 the total yield of all existing units with treatment b_1 , R_1 the total yield in replication r_1 , (A_1B_1) the total yield of all existing units with both a_1 and b_1 , (R_1A_1) the total yield of all existing units with both r_1 and a_1 and G the grand total. Set $x=0$ and $y=$ the actual yield for the existing units and $x=-1$ and $y=0$ for the

missing unit. In the analysis of covariance table, any sum of squares, $S(x^2)$, equals its degrees of freedom divided by N , in all cases. The best estimate of the yield of the missing unit in order to minimize the sub-plot error mean square is simply the error b) regression coefficient,

$$y = \frac{r(R_1 A_1) + b(A_1 B_1) - A_1}{(r - 1)(b - 1)} .$$

As this yield is used for the missing unit, all sums of squares except that for error b) are over-estimated. The unbiased estimate of any sum of squares is found by computing a new $S(x_1^2)$ and $S(x_1 y)$. Where $S(x_1^2)$ and $S(x_1 y)$ are the $S(x^2)$ and $S(xy)$ plus error b) respectively. Then the new regression coefficient is

$$y_1 = \frac{S(x_1 y)}{S(x_1^2)} .$$

The bias in estimating the sum of squares under consideration is:

$(y - y_1)^2 S(x_1^2)$. The bias is always positive; that is, the sum of squares is always over-estimated in the analysis of variance.

Thus, for the treatment B,

$$y_1 = \frac{ra(R_1 A_1) + ab(A_1 B_1) - aA_1 - bB_1 + G}{(b - 1)(ra - a + 1)} ,$$

and $S(x_1^2) = (b - 1)(ra - a - 1)/rab$.

For the interaction AB,

$$y_1 = \frac{ra(R_1 A_1) + bB_1 - G}{(ra - 1)(b - 1)} ,$$

and $S(x_1^2) = (ra - 1)(b - 1)/rab$.

One possibility of obtaining more exact estimates of the sums of squares for treatment A and for error a) will be to minimize the sum of squares for error a) and calculate the true sum of squares for treatment A on this basis. The estimate of the missing value is

$$y' = - \frac{ra(R_1A_1) - rR_1 - aA_1 + G}{(r - 1)(a - 1)} .$$

However, the adjusted sum of squares is no longer independent of error sum of squares; hence the F-test can not be used to test the significance of the A differences.

Coons (3) pointed out that the usual work required to estimate the missing values and correct the resulting bias may be very tedious. Also, situations may arise for which no general formula is available. A general method of handling missing observations in any situation is presented. The technique employs the computational procedures of a covariance analysis using a dummy X covariate. This method was originally given in a paper by Bartlett (2) and also was described by Anderson, but had not been exploited fully by them. The advantages of the covariance method are its generality of application and the ease with which 'exact' tests of significance may be obtained.

Wilkinson (15) presented methods for setting up and solving equations for missing values of several experimental designs. Tables for determining the coefficients for missing value equations are presented for some standard designs, such as randomized block, latin square and lattice square designs. A second paper (13) gives correction formulae for the treatment sum of squares when estimated data are computed for the designs with a two-way restriction. Also a basic procedure of analysis of covariance with incomplete data was described in another paper (14). He assumed some observations on y are

missing and the corresponding measurements of p concomitant variates are x_1, x_2, \dots, x_p . The analysis procedures are stated as follows:

- 1). discard all measurements of x_1, x_2, \dots, x_p that correspond to the missing observations on y ,
- 2). fit a set of missing values for y as though for an ordinary analysis (ignoring covariance),
- 3). with exactly parallel calculations fit sets for missing values for x_1, x_2, \dots, x_p to replace those discarded,
- 4). carry out the covariance analysis on the completed data for y and x_1, x_2, \dots, x_p .

The steps provided the exact analysis, and formal justification of the procedure is given in (13).

This report contains a presentation of an analysis of covariance technique for analyzing univariate experiments with missing data. Also a least squares analysis technique is described, and some general results for linear model are given. Numerical examples of latin square experiments with missing data will be analyzed by three techniques for illustrations.

II. ANALYSIS OF COVARIANCE TECHNIQUE

1). General Use of Covariance to Deal with Missing Data

Some properties will be listed here which give the justification for the computational procedures to be described. Property 1 is due to Fisher, property 2 is due to Bartlett (2), property 3 has been implicitly assumed by many authors, properties 4, 5, 6 have been obtained by Kempthorne (see (3)), but probably are known to a number of workers. Those properties are as follows:

1. If an analysis of variance is made with symbols $\beta_1, \beta_2, \dots, \beta_q$ in place of missing observations, then the best linear unbiased estimates (BLUE's) of the missing observations are the quantities $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_q$ which minimize the error sum of squares.

2. Let the data be the observed data where obtained, and be zero where missing. Introduce a concomitant variable X_m ($m = 1 \dots q$) corresponding to the m th missing observations. Let X_m take the value $-v$ for the m th missing observation and zero for all others, missing or not. If the error partial regression coefficients obtained from an analysis of covariance are denoted by $\hat{\beta}_{1E}, \hat{\beta}_{2E}, \dots, \hat{\beta}_{qE}$, then $v\hat{\beta}_{1E}, v\hat{\beta}_{2E}, \dots, v\hat{\beta}_{qE}$ are the BLUE's of the missing observations.

3. Given that, with full data (y_1, y_2, \dots, y_n) , the BLUE of some linear function of the parameters is $v_1 y_1 + v_2 y_2 + \dots + v_n y_n$, then the best estimate of that function with missing data is obtained by replacing the missing y 's with the missing value estimates.

4. Estimates of functions of data with missing observations, and of variances and covariances for these estimates may be obtained by the routine application of the formulae for adjusted means in the analysis

of covariance; i.e., by regarding the zero yields supplied in the analysis of covariance procedure as having variances σ^2 . The above statement applies to functions of the augmented data; the variance of a missing observation per se is given by statement 5 following.

5. Denote the error sum of squares of X_i by E_{ii} and the error sum of products of X_i and X_j by E_{ij} . Then the variance of the i th missing value estimate is $(v^2\mu_{ii} - 1)\sigma^2$ and the covariance of the i th and j th missing value estimates is $v^2\mu_{ij}\sigma^2$, where

$$\begin{pmatrix} E_{11} & E_{12} & \cdots & E_{1q} \\ E_{21} & & & \\ \vdots & & & \\ E_{q1} & & & E_{qq} \end{pmatrix} \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1q} \\ \mu_{21} & & & \\ \vdots & & & \\ \mu_{q1} & & & \mu_{qq} \end{pmatrix} = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ 0 & & & 1 \end{pmatrix}$$

6. The sum of squares for treatments obtained by analyzing the data augmented by the missing value estimates is always greater than or equal to the exact sum of squares for treatment, so that a correction for this bias is needed.

2). Application of Covariance Technique to One Missing Observation

Let n equal the total number of observations in the experiment including the missing one. Consider the original data as the dependent variable Y of the covariance analysis and insert the value of zero in the cell which has the missing observation. Set up a concomitant variable X which takes the value of $-n$ in the cell corresponding to the substituted zero value and the value of zero elsewhere.

The usual computational procedures of the covariance analysis automatically provide unbiased tests of significance. However estimates of functions of the Y data, such as treatment means, must be adjusted to the value of zero for the concomitant variable rather than to the observed average value (grand average) of X as is usually done. For example, an adjusted treatment mean is not estimated by $\bar{Y}_{.j} - \hat{\beta}_E(\bar{X}_{.j} - \bar{X}_{..})$, the covariance formula generally used, but rather by the formula

$$\text{adj. } \bar{Y}_{.j} = \bar{Y}_{.j} - \hat{\beta}_E \bar{X}_{.j} , \quad (1)$$

where

$$\hat{\beta}_E = \frac{E_{xy}}{E_{xx}}$$

and E_{xy} and E_{xx} are the error sum of cross products and sum of squares, respectively, in the analysis of covariance.

From statements 2 and 5 above, the missing observation estimate and its variance are,

$$\text{missing observation} = n\hat{\beta}_E = n \frac{E_{xy}}{E_{xx}}$$

$$\text{var.}(\hat{n\beta_E}) = \sigma^2 \left(\frac{n^2}{E_{xx}} - 1 \right),$$

where σ^2 is estimated by s^2 which is the residual error mean square resulting from routine application of the analysis of covariance. That is,

$$s^2 = \frac{1}{\text{d.f.}} \left(E_{yy} - \frac{E_{xy}^2}{E_{xx}} \right)$$

where E_{yy} is the error sum of squares of the y's, and the number of degrees of freedom is equal to the number of the error degrees of freedom with full data, less one.

The variance of (1) is given by:

$$\text{var.}(\text{adj.}\bar{Y}.j) = \text{var.}(\bar{Y}.j) + (\bar{X}.j)^2 \text{var.}(\hat{\beta_E}),$$

where $\text{var.}(\hat{\beta_E})$ is taken to be σ^2/E_{xx} .

As comparison among the adjusted treatment means, such as:

$$\text{adj.Tm} - \text{adj.Tn} = (\bar{Y}.m - \bar{Y}.n) - \hat{\beta_E}(\bar{X}.m - \bar{X}.n)$$

may in general be designated by the notation

$$D = D_y - \hat{\beta_E} D_x$$

Since D_y and $\hat{\beta_E}$ are independent, $\text{var.}(D) = \text{var.}(D_y) + D_x^2 \text{var.}(\hat{\beta_E})$, where $\text{var.}(D_y)$ is calculated assuming no missing observations and $\text{var.}(\hat{\beta_E})$ is taken to be σ^2/E_{xx} .

The computation required by the use of an X covariate is relatively slight, due to the simple nature of the X data. In most situations, the correct sum of squares of X attributable to any given source of variation ($\sum x_i^2$) will simply be $n \times (\text{d.f. for the given source of variation})$.

The procedure, which gives unbiased estimates of both treatment and error sum of squares, can be simplified as follows:

1. Set $Y = 0$ for the missing observation.
2. Define a covariate as $X = 0$ for an observed Y , and $X = -n$ for $Y = 0$.
3. Carry out the analysis of covariance.
4. Compute $\hat{\beta}_E = E_{xy}/E_{xx}$ and multiply by n to estimate the missing value.

The estimate of the missing value, $n\hat{\beta}_E$, is essentially an adjustment to the so-called observation $Y = 0$, to give an estimate of the Y that would have been obtained if X had been zero instead of $-n$.

3). Application of Covariance Technique to More Than One Missing Observation

With more than one missing observation a multiple covariance analysis is required. Again let n equal the number of Y observations in the experiment including the missing ones, and assign a value of zero to the Y observations which are missing. Set up concomitant variables X_m for each missing observation. Each of these X_m will be zero in all cells except in that cell corresponding to the missing observation with which the given X_m is associated; in that one cell it will have a value of $-n$.

With q missing observations, a multiple covariance analysis must be performed on Y and the q covariates X_m .

A complete analysis of covariance will provide unbiased tests for treatment effects. Each missing observation Y_m may be estimated by $\hat{n\beta_{mE}} = n \times (\text{error estimate of the } \beta \text{ associated with that missing observation})$.

In order to obtain estimates of the β_{mE} , the solution of a set of simultaneous equations of size $q \times q$ is required. Each unbiased test of adjusted treatment effects requires the solution of an additional set of simultaneous equations of size $q \times q$. In general the solution of these equations will be easy, even by an iterative technique, because usually the off-diagonal terms of the matrix of coefficient will be small.

The computations required to obtain the sum of products — $\sum x_m x_n$ and $\sum x_m y$ — are simple, since each x_m is associated with a single missing value and therefore has only one non-zero cell. In computing $\sum x_m x_n$, two situations may be encountered.

1. The two missing values associated with x_m and x_n occur in the same level of the given source of variation. The results are exactly the same as those obtained for $\sum x^2$; i.e., $\sum x_m x_n = n \times (\text{d.f. for the given source of variation})$.

2. The two missing values occur in different levels of the given source of variation. Then, for most cases, $\sum x_m x_n = -nr$, where r is dependent upon the hierarchical (nested) classification which is used. When no hierarchical classification is involved, $r = 1$. When the given source of variation is an interaction effect, then the corresponding main effects and lower order interactions must be subtracted from the above $\sum x_m x_n = -nr$.

III. LEAST SQUARES ANALYSIS TECHNIQUE FOR LINEAR MODEL

Consider an experiment involving n experimental units where the observations on q of the experimental units are missing. Without loss of generality, it is assumed that the first q experimental units correspond to the missing observations. The resulting missing data model is

$$\begin{pmatrix} \underline{y}_0 \\ \underline{y}_1 \end{pmatrix} = \begin{pmatrix} \underline{X}_0 \\ \underline{X}_1 \end{pmatrix} \underline{\beta} + \underline{e} \quad , \quad (3.1)$$

where \underline{y}_0 is the $q \times 1$ vector corresponding to the missing observations, \underline{y}_1 is the $\overline{n-q} \times 1$ vector of observations, \underline{X}_0 is the $q \times p$ matrix of constants forming the design matrix corresponding to the missing data, \underline{X}_1 is the $\overline{n-q} \times p$ matrix of constants forming the design matrix corresponding to the observed data, $\underline{\beta}$ is a $p \times 1$ vector of unknown parameters, and \underline{e} is a $n \times 1$ unobserved random vector assumed to have mean zero and covariance matrix $\sigma^2 \underline{I}_n$ where σ^2 is unknown. The matrices \underline{X}_0 and \underline{X}_1 may or may not be of full rank.

Consider the partition of model (3.1) corresponding to the observed data, i.e.,

$$\underline{y}_1 = \underline{X}_1 \underline{\beta} + \underline{e}_1 \quad . \quad (3.2)$$

This is a linear model and the analysis is well known (for example, see Graybill (4) or Scheffe (10)). The partitioning of the sums of squares for this model is given in the following analysis of variance table,

SOURCE	DF	SUM OF SQUARES
Total	$n-q$	$\underline{y}_1' \underline{y}_1$
Estimation	r	$\underline{y}_1' \underline{X}_1 \underline{X}_1^{-1} \underline{y}_1$
Error	$n-q-r$	$\underline{y}_1' (\underline{I} - \underline{X}_1 \underline{X}_1^{-1}) \underline{y}_1$

There are various ways to compute the sum of squares due to a testable hypothesis $\underline{A} \underline{\beta} = \underline{0}$. The principle of conditional error will be used. It is equivalent to the likelihood ratio procedure. The procedure is: First obtain the sum of squares due to error for model (3.2). Next obtain a restricted model by imposing the hypothesis on model (3.2) and compute the sum of squares due to error for the restricted model. The sum of squares due to the hypothesis is the sum of squares due to error for the restricted model minus the sum of squares due to error for model (3.2). The restricted model is

$$\underline{y}_1 = \underline{X}_1 (\underline{I} - \underline{A}^{-1} \underline{A}) \underline{\beta} + \underline{e} ,$$

and the sum of squares due to error is

$$SSE_R = \underline{y}_1' (\underline{I} - \underline{X}_1 (\underline{I} - \underline{A}^{-1} \underline{A}) (\underline{X}_1 (\underline{I} - \underline{A}^{-1} \underline{A}))^{-1}) \underline{y}_1 .$$

The sum of squares due to the testable hypothesis $\underline{A} \underline{\beta} = \underline{0}$ is given by

$$SSH_0 = \underline{y}_1' (\underline{X}_1 \underline{X}_1^{-1} - \underline{X}_1 (\underline{I} - \underline{A}^{-1} \underline{A}) (\underline{X}_1 (\underline{I} - \underline{A}^{-1} \underline{A}))^{-1}) \underline{y}_1$$

The above analyses provide what we call the "exact analysis" of the observed data model (3.2). A missing data technique will be considered adequate only if it provides the same results as the exact analysis.

Now consider the problem of estimating the missing values, \underline{y}_0 , of

model (3.1). At this point replace y_0 by $\underline{\eta}$ in model (3.1), where $\underline{\eta}$ is the vector of unknown parameters as $\underline{\eta} = E(y_0) = \underline{X}_0\beta$. A technique for estimating $\underline{\eta}$ is to select values for the elements of $\underline{\eta}$ that minimize the error sum of squares under model (3.1). The general instructions put forth in methods books are: (1) replace the missing data by the symbols $\underline{\eta}$; (2) write the usual sum of squares due to error $\underline{y}'(\underline{I} - \underline{X}\underline{X}^{-})\underline{y}$, which is a function of $\underline{\eta}$ since $\underline{y}' = (\underline{\eta}', \underline{y}_1')$ and (3) minimize the sum of squares due to error by choice of $\underline{\eta}$. Let $\hat{\underline{\eta}}$ denote that value of $\underline{\eta}$ which minimizes the sum of squares due to error. For the missing data model (3.1), if the linear combinations $\underline{\eta} = \underline{X}_0\beta$ are estimable under model (3.2), the error sum of squares is minimized by $\hat{\underline{\eta}} = \underline{X}_0\hat{\beta} = \underline{X}_0\underline{X}_1^{-}\underline{y}_1$ which is the unique BLUE of $\underline{\eta}$. Thus the augmented data model is

$$\begin{pmatrix} \hat{\underline{\eta}} \\ \underline{y}_1 \end{pmatrix} = \begin{pmatrix} \underline{X}_0 \\ \underline{X}_1 \end{pmatrix} \underline{\beta} + \underline{e}, \quad (3.3)$$

where \underline{y}_1 , \underline{X}_0 , \underline{X}_1 , $\underline{\beta}$ and \underline{e} are defined as in model (3.1) and $\hat{\underline{\eta}} = \underline{X}_0\hat{\beta}$.

The error sum of squares for this model is equal to the error sum of squares of the observed data model $\underline{y}_1 = \underline{X}_1\beta + \underline{e}$, i.e., $SSE_a = SSE$.

When you compute the sum of squares due to the testable hypothesis $\underline{A}\underline{\beta} = \underline{0}$ by standard computational techniques for the model (3.3) that sum of squares will be biased. The principle of conditional error may be used to circumvent the problem when the restricted model is easy to work with computationally. The restricted model is

$$\begin{pmatrix} \underline{\eta} \\ \underline{y}_1 \end{pmatrix} = \underline{X}(\underline{I} - \underline{A}^{-}\underline{A})\underline{\beta} + \underline{e} , \quad (3.4)$$

i.e., the model $\begin{pmatrix} \underline{\eta} \\ \underline{y}_1 \end{pmatrix} = \underline{X} \underline{\beta} + \underline{e}$ is restricted by the hypothesis $\underline{A} \underline{\beta} = \underline{0}$ where the $\underline{A} \underline{\beta}$ are estimable with respect to observed model (3.2). Estimate $\underline{\eta}$ under the restricted model (3.4), say by $\underline{\hat{\eta}}$, to obtain the augmented restricted model

$$\begin{pmatrix} \underline{\hat{\eta}} \\ \underline{y}_1 \end{pmatrix} = \underline{X}(\underline{I} - \underline{A}^{-}\underline{A})\underline{\beta} + \underline{e} . \quad (3.5)$$

Compute the error sum of squares as usual under model (3.5) as

$$SSE_{ra} = \underline{y}_1' (\underline{I} - \underline{X}_1 (\underline{I} - \underline{A}^{-}\underline{A}) (\underline{X}_1 (\underline{I} - \underline{A}^{-}\underline{A}))^{-}) \underline{y}_1$$

which is not biased.

From SSE_{ra} subtract the error sum of squares computed from the augmented data model (3.3). i.e., $SSE_{ra} - SSE_a$; that yields the exact value for the sum of squares due to the hypothesis $\underline{A} \underline{\beta} = \underline{0}$.

For example, suppose in a randomized complete block design, the missing observations can be estimated and the error sum of squares may be obtained from the augmented data. Under the hypothesis there is no difference between the treatment effects, the restricted model is a one-way classification design and the corresponding error sum of squares is easily computed. The difference between the two error sums of squares is the sum of squares due to the hypothesis that there is no difference between the treatment effects.

IV. ANALYSIS OF COVARIANCE TECHNIQUE FOR LINEAR MODEL

Milliken and McDonald (7) state and prove the results for the linear model to analyze experiments with incomplete data. Some general results will be presented here without proof.

The covariance model is

$$\begin{aligned} \underline{y} &= \underline{X} \underline{\beta} + \underline{Z} \underline{\gamma} + \underline{e} = \begin{pmatrix} \underline{X} & \underline{Z} \end{pmatrix} \begin{pmatrix} \underline{\beta} \\ \underline{\gamma} \end{pmatrix} + \underline{e} \\ &= \underline{W} \underline{\xi} + \underline{e} , \end{aligned} \quad (4.1)$$

where \underline{y} is the $n \times 1$ vector of observations, \underline{X} is the $n \times p$ design matrix of rank $r \leq p$, \underline{Z} is the $n \times s$ matrix of covariates of rank $v \leq s$, $\underline{\beta}$ is a $p \times 1$ vector of unknown design parameters, $\underline{\gamma}$ is a $s \times 1$ vector of unknown parameters associated with covariates, and \underline{e} is a $n \times 1$ unobserved random normal vector with mean zero and covariance matrix $\sigma^2 \underline{I}$, where σ^2 is unknown. Assume that the rank of \underline{W} is $t \leq r+v < n$.

We use the least square procedure by choosing $\hat{\underline{\beta}}$ and $\hat{\underline{\gamma}}$ to minimize the quantity $(\underline{y} - \underline{X} \underline{\beta} - \underline{Z} \underline{\gamma})'(\underline{y} - \underline{X} \underline{\beta} - \underline{Z} \underline{\gamma})$. Differentiating this quantity with respect to $\underline{\beta}$ and $\underline{\gamma}$, we obtain, respectively, the two equations

$$\underline{X}'\underline{X} \hat{\underline{\beta}} - \underline{X}' \underline{y} + \underline{X}'\underline{Z} \hat{\underline{\gamma}} = \underline{0} \quad (4.2)$$

$$\underline{Z}'\underline{Z} \hat{\underline{\gamma}} - \underline{Z}'\underline{y} + \underline{Z}'\underline{X} \hat{\underline{\beta}} = \underline{0}. \quad (4.3)$$

The sum of squares due to estimation in the analysis of variance table will be partitioned into two parts, one due to $\underline{\gamma}$ and one due to $\underline{\beta}$ after adjusting for $\underline{\gamma}$.

To analyze the missing data model (3.1) using the covariance model (4.1), construct a model as

$$\begin{pmatrix} \underline{n} \\ \underline{y}_1 \end{pmatrix} = \begin{pmatrix} \underline{x}_0 \\ \underline{x}_1 \end{pmatrix} \underline{\beta} + \begin{pmatrix} \underline{D} \\ \underline{0} \end{pmatrix} \underline{\gamma} + \underline{e}, \quad (4.4)$$

where \underline{n} is $q \times 1$ vector of arbitrary constants, \underline{D} is a $q \times q$ diagonal matrix with arbitrary nonzero diagonal elements, $\underline{\gamma}$ is the $q \times 1$ vector defined in model (4.1), $\underline{0}$ is a $\overline{n-q} \times q$ null matrix, and $\underline{y}_1, \underline{x}_0, \underline{x}_1, \underline{\beta}$, and \underline{e} are defined as for model (3.1). The analysis of the model (4.4) provides the exact analysis, regardless of the values selected for \underline{n} and the diagonal elements of \underline{D} , that is

1. the least squares estimate of $\underline{\beta}$ is $\hat{\underline{\beta}} = (\underline{x}_1' \underline{x}_1)^c \underline{x}_1' \underline{y}_1$
2. the sum of squares due to $\underline{\beta}$ adjusted for $\underline{\gamma}$ is $SS(\underline{\beta}|\underline{\gamma}) = \underline{y}_1' \underline{x}_1 \underline{x}_1' \underline{y}_1$

and

3. the error sum of square is $SSE = \underline{y}_1' (\underline{I} - \underline{x}_1 \underline{x}_1') \underline{y}_1$,

The analysis of model (4.4) provides the exact sum of squares due to a testable hypothesis $\underline{A} \underline{\beta} = \underline{0}$, regardless of the choice of \underline{n} and \underline{D} . The sum of squares due to this hypothesis is

$$SSH_0 = \underline{y}_1' (\underline{x}_1 \underline{x}_1' - \underline{x}_1 (\underline{I} - \underline{A} \underline{A}) (\underline{x}_1 (\underline{I} - \underline{A} \underline{A}))^{-1}) \underline{y}_1$$

The above results show that one can insert an arbitrary values for \underline{n} and for the diagonal elements of \underline{D} and do a conventional analysis of covariance using model (4.4) to obtain the exact analysis.

For general application of the covariance method, one needs a general analysis of covariance computer program that uses the above formulae involving conditional and/or generalized inverses.

After the analysis of variance has been completed, it may be desirable to

estimate the means of the missing observations. If the $\underline{X}_0 \underline{\beta}$ are estimable linear combinations for model (3.2), then the first q elements of the residual vector $\underline{y}_r = \underline{y} - \underline{Z} \hat{\underline{y}}$ are the BLUE's of $\underline{X}_0 \underline{\beta}$. The residual vector can be computed after obtaining the least squares estimate of \underline{y} from equations (4.3) using the solution $\hat{\underline{\beta}} = (\underline{X}'_1 \underline{X}_1)^c \underline{X}'_1 \underline{y}_1$.

The residual vector \underline{y}_r is

$$\underline{y}_r = \begin{pmatrix} \underline{X}_0 (\underline{X}'_1 \underline{X}_1)^c \underline{X}'_1 \underline{y} \\ \underline{y}_1 \end{pmatrix}$$

The elements of the vector $\underline{X}_0 (\underline{X}'_1 \underline{X}_1)^c \underline{X}'_1 \underline{y}$ are the unique BLUE's of the linear combinations $\underline{X}_0 \underline{\beta}$ if and only if the $\underline{X}_0 \underline{\beta}$ are estimable for model (3.2).

V. NUMERICAL ILLUSTRATIONS

Table 1 shows the field layout and yields of a 5×5 latin square experiment on the effects of spacing (treatment) on yields of millet plants (12).

Table 1.

Yields (Grams) of Plots of Millet Arranged in A Latin Square

Row	Column					Sum
	1	2	3	4	5	
1	B : 257	E : 230	A : 279	C : 287	D : 202	1255
2	D : 245	A : 283	E : 245	B : 280	C : 260	1313
3	E : 182	B : 252	C : 280	D : 246	A : 250	1210
4	A : 203	C : 204	D : 227	E : 193	B : 259	1086
5	C : 231	D : 271	B : 266	A : 334	E : 338	1440
Sum	1118	1240	1297	1340	1309	6304
Summary by Spacing						
	A : 2"	B : 4"	C : 6"	D : 8"	E : 10"	
Sum	1349	1314	1262	1191	1188	6304
Mean	269.8	262.8	252.4	238.2	237.6	252.2

Case 1. One Missing Observation

To illustrate, suppose that in Table 1, the yield, 338 grams in row 5, column 5 and treatment E, were missing. Let y_{555} denotes this missing value.

1). Use of formulae to analyze data:

In a latin square experiment, the formula for calculating this missing value is

$$\hat{y}_{555} = \frac{r(R + c + T) - 2G}{(r - 1)(r - 2)}$$

where $R = 1440 - 338 = 1102$

$C = 1309 - 338 = 971$

$T = 1188 - 338 = 850$

$G = 6304 - 338 = 5966,$

Then $\hat{y}_{555} = 223.6$

The correction for bias is

$$\frac{(G - R - C - (r-1)T)^2}{(r - 1)^2 (r - 2)^2} = 1687.8$$

After putting the estimated missing yield, 223.6 grams, in the table, the analysis of variance gives

Treatment Sum of Squares	9581.8
Less Bias	1687.8
Unbiased treatment S.S.	7894.0

The final analysis is

	<u>D.F.</u>	<u>S.S.</u>	<u>M.S.</u>	<u>F ratio</u>
Rows(unadj.)	4	7495.1	-	-
Columns(unadj.)	4	6034.5	-	-
Treatments(adj.)	4	7894.0	1973.5	3.39
Error	11	6385.5	580.5	-
Total	23	-	-	-

2). Analysis of covariance technique:

The procedures are as following:

- (i). Set $y_{555} = 0$ in the missing cell.
- (ii). Define a covariate as $X = 0$ for an observed Y , and
 $X_{555} = -25$ for $y_{555} = 0$.
- (iii). Carry out the analysis of covariance which is given in table 2, with all the computations relative to a given source of variation listed on a single line. Column (1), (2) and (3) are the sums of squares and products of x and y . Column (4) shows the correct sum of squares for error (S_E) and sum of squares for treatment + error (S_{T+E}). S_E is subtracted from S_{T+E} to obtain adjusted treatment sum of squares S_T . Thus the completed analysis is obtained and exact tests of significance have been made. Furthermore adjusted means and their variances can be calculated.
- (iv). Compute $\hat{\beta}_E = \frac{E_{xy}}{E_{xx}} = \frac{2683}{25(12)} = 8.94$, and multiply by 25 to obtain the missing value estimate $\hat{y}_{555} = 223.6$.

Table 2.

Analysis of Latin Square with One Missing Observation

Source of Variation	D.F.	(1) $\sum x^2$	(2) $\sum xy$	(3) $\sum y^2$	(4) $\sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$	D.F.	Mean Square	F Ratio
Row	4	25(4)	456	7652.6	-	-	-	-
Column	4	25(4)	1111	17908.6	-	-	-	-
Treatment	4	25(4)	1716	32278.2	-	-	-	-
Error	12	25(12)	2683	30378.4	$S_E = 6383.77$	11	580.34	-
Treatment+Error	16	25(16)	4399	62656.6	$S_{T+E} = 14278.60$	15	-	-
Adjusted Treatment	4	-	-	-	$S_T = 7894.83$	4	1973.70	3.40

3). Least square analysis technique

The model for a latin square experiment is

$$y_{ijk} = \mu + \alpha_i + \delta_j + \gamma_k + e_{ijk}; \quad i, j \text{ and } k=1\dots 5; \quad e_{ijk} \sim N(0, \sigma^2).$$

where α, δ, γ indicate treatment, row, and column effects.

Refer to missing data model (3.1), y_0 is the scalar corresponding to this one missing observation, y_1 is the 24×1 vector of observations, X_0 is the 1×16 vector of constants forming the design matrix corresponding to the missing data, X_1 is the 24×16 matrix of constants forming the design matrix corresponding to the observed data, β is a 16×1 vector of unknown parameters, and e is a 25×1 random vector assumed to have mean zero and covariance matrix $\sigma^2 I_{25}$.

By imposing the hypothesis $\underline{A} \underline{\beta} = \underline{0}$ on the model (3.3), and using the results we already described, the exact value for the sum of squares due to hypothesis will be obtained.

In this problem, we have used least squares analysis computer program (5) to analyze the data. The results are as follows:

<u>Source</u>	<u>D.F.</u>	<u>Sum of Squares</u>	<u>Mean Squares</u>	<u>F Ratio</u>
Rows	4	7014.175	-	-
Columns	4	5917.428	-	-
Treatments	4	7895.148	1973.787	3.39
Residual	11	6387.605	580.691	-
Total	23	28895.832		

Case 2. Two Missing Observations

Suppose that in Table (1), the first item, 257 grams and the last item, 338 grams were missing. Let y_{112} and y_{555} denote the two missing observations.

1). Use of formulae to analyze data:

Snedecor (12.2) describes a method for making an unbiased test for latin square design, with two missing values. The steps are as follows,

- (i). Supply estimates of the missing values. The recursion method is applied to the missing plot formula for the latin square. Results: $\hat{y}_{112} = 225.36$, $\hat{y}_{555} = 228.85$.
- (ii). Analyze the variance of the augmented square (that is, including the two estimated values \hat{y}_{112} and \hat{y}_{555}). The only part needed is the sum of squares for error, $SSE_S = 5915.7$.

- (iii). Treating the original data as randomized blocks in rows and columns, ignoring the treatments, supply a new pair of estimates by use of the formula for missing plots in randomized blocks: $\hat{y}_{112} = 206.7$, $\hat{y}_{555} = 278.1$.
- (iv). Analyze the variance of the augmented randomized blocks.
What is needed is the sum of squares for error $SSE_B = 12668.2$.
- (v). Analyze the variance of the latin square, using $SSE_B - SSE_S = 12668.2 - 5915.7 = 6752.5$ as the sum of squares for treatment.

<u>Source of Variation</u>	<u>D.F.</u>	<u>S.S.</u>	<u>M.S.</u>	<u>F Ratio</u>
Unbiased Treatment	4	6752.5	1688.00	2.8
Error	10	5915.7	591.57	-

2). Analysis of covariance technique:

A multiple covariance analysis is used to handle the problem of two missing observations.

- (i). Assign the value of zero to the two missing observations,
 $y_{112} = 0$ and $y_{555} = 0$.
- (ii). Set up two concomitant variables X_1 and X_2 for each missing observation. Each of $X_1 = 0$ in all cells except in that cell corresponding to y_{112} , in that one cell $X_1 = -25$. Similarly, each of $X_2 = 0$ in all cells except in that cell corresponding to y_{555} , in that one cell $X_2 = -25$.
- (iii). The computation of a multiple covariance analysis is given in table 3. Column (6) shows how to obtain adjusted sum of squares for treatment. The methods of obtaining the correct error term is just using the usual covariance formula

Table 3

Analysis of Latin Square Design with Two Missing Observations

Source of variation	D.F.	(1) $\sum x_m^2$	(2) $\sum x_1 x_2$	(3) $\sum x_1 y$	(4) $\sum x_2 y$	(5) $\sum y^2$	(6) $\sum y^2 - \hat{\beta}_1 \sum x_1 y - \hat{\beta}_2 \sum x_2 y$	D.F.	M.S.	F ratio
Row	4	25(4)	-25	719	199	11867.4	-	-	-	-
Column	4	25(4)	-25	1404	854	36207.0	-	-	-	-
Treatment	4	25(4)	-25	424	1459	30427.8	-	-	-	-
Error	12	25(12)	50	3162	3197	63685.6	$S_E = 5915.886$	10	591.589	2.8
Treatment + Error	16	25(16)	25	3586	4656	94113.4	$S_{T+E} = 12673.400$	14	-	-
Adjusted Treatment	4	-	-	-	-	-	$S_T = 6757.500$	4	1689.375	-

 $\hat{\beta}_m$ estimates:

The $\hat{\beta}_{mE}$ satisfy: $\begin{cases} 25(12)\beta_{1E} + 50\beta_{2E} = 3162 \\ 50\beta_{1E} + 25(12)\beta_{2E} = 3197 \end{cases}$ Therefore $\begin{cases} \hat{\beta}_{1E} = 9.0143, \\ \hat{\beta}_{2E} = 9.1543. \end{cases}$

The $\hat{\beta}_{m(T+E)}$ satisfy: $\begin{cases} 25(16)\beta_{1(T+E)} + 25\beta_{2(T+E)} = 3586 \\ 25\beta_{1(T+E)} + 25(16)\beta_{2(T+E)} = 4656 \end{cases}$ Therefore $\begin{cases} \hat{\beta}_{1(T+E)} = 8.2698, \\ \hat{\beta}_{2(T+E)} = 11.1221. \end{cases}$

$E_{yy} - \sum_{m=1}^2 \hat{\beta}_{mE} (E_{x_m y})$. In order to obtain the exact sum of squares for treatment, the following adjustment must be applied to treatment+error:

$$\sum_{m=1}^2 \hat{\beta}_{m(T+E)}^{(T+E)} x_m y$$

where the $\hat{\beta}_{m(T+E)}$ must satisfy the equation (see table 3):

$$25(16)\hat{\beta}_{1(T+E)} + 25\hat{\beta}_{2(T+E)} = 3586,$$

$$25\hat{\beta}_{1(T+E)} + 25(16)\hat{\beta}_{2(T+E)} = 4656.$$

Therefore

$$\hat{\beta}_{1(T+E)} = 8.2698,$$

$$\hat{\beta}_{2(T+E)} = 11.1221.$$

Hence the adjusted (treatment+error) sum of squares is

$$\sum y^2 - \sum_{m=1}^2 \hat{\beta}_{m(T+E)}^{(T+E)} x_m y = 12673.4.$$

Subtraction of the correct error term, $S_E = 5915.886$, from this gives the treatment sum of squares, 6757.5, which will provide an exact test for treatment effect.

(iv). Two missing observations are estimated by

$$\hat{y}_{112} = n\hat{\beta}_{1E} = 225.36$$

$$\hat{y}_{555} = n\hat{\beta}_{2E} = 228.85.$$

3). Least square analysis technique:

In this two-missing-observations case ($q=2$), we apply the same linear model as one missing observation example, except that $q=1$ is replaced by $q=2$. The same least squares computer program is used to analyze the data.

<u>Source</u>	<u>D.F.</u>	<u>Sum of squares</u>	<u>Mean squares</u>	<u>F ratio</u>
Rows	4	7023.415	-	-
Columns	4	6266.660	-	-
Treatments	4	6752.386	1688.096	2.85
<u>Residuals</u>	<u>10</u>	<u>5920.738</u>	<u>592.074</u>	<u>-</u>
Total	22			

VI. DISCUSSION

When one missing observation exists in an experimental design, one may do analysis by using formulae which are available in some designs. If more than one missing observation, the iterative methods can be used to estimate the missing values, the unbiased test will be performed by a special method which depends on different experimental design, e.g., the method we described in the previous example for latin square design. Since the procedures required to estimate the several missing values and correcting for bias will become complicated, the other two techniques — least squares analysis and covariance analysis seem preferable as general missing plot procedures.

Both of these techniques provide an exact analysis. The covariance analysis technique is generally computationally easier than least squares, and also provides the missing value estimates. If one is analyzing the data by a high speed computing device, there would be no point in working out the partition of the sum of squares of x or products of x and y algebraically. Even when the expected values of the missing data are not linearly estimable, the covariance analysis still provides the exact analysis, but the standard computational formulae break down and thus new formulae, involving pseudo-inverses of certain matrices, are given (7).

The advantage of the least squares analysis is that when the computer program of least squares analysis is available, one just needs to use the observed data to do an analysis which provides an exact test. Because covariance analysis uses several covariates to analyze the data, it may take longer computing time than least squares procedure.

Techniques for the estimation of missing data in the multivariate linear model are suggested and the subsequent analysis of the "complete" data is considered by McDonald (6). Those techniques are generalizations of the procedures for the analysis of univariate experiments in which some of the observations are missing. The techniques require only computational procedures which are already available in the literature for univariate experiments.

VII. REFERENCE

1. Anderson, R. L. (1946) Missing Plot Techniques. *Biometrics*, 2: 41-47.
2. Bartlett, M. S. (1937) Some Examples of Statistical Methods of Research in Agriculture and Applied Biology. *J. Roy. Stat. Soc. Suppl.*, 4: 137-170.
3. Coons, I. (1957) The Analysis of Covariance as A Missing Plot Technique. *Biometrics*, 13: 387-405.
4. Graybill, F. A. (1969) An Introduction to Linear Statistical Model, Vol. I, McGraw-Hill, New York.
5. Kemp, K. E. (1972) Least Squares Analysis of Variance, A Procedure, A Program, and Examples of Their Use. Research Paper 7, Dept. of Statistics, Kansas State Univ., June 1972.
6. McDonald, L. (1971) On the Estimation of Missing Data in The Multivariate Linear Model, *Biometrics*, 27: 535-543.
7. Milliken, G. A. and McDonald L. (1971) Techniques for Analysis Experiments with Missing or Incomplete Data. Technical Report No. 14, Dept. of Statistics, Kansas State Univ.
8. Nelder, J. A. (1954) A Note on Missing Plot Values. *Biometrics*, 10: 400-401.
9. Norton, H. W. (1955) A Further Note on Missing Data. *Biometrics*, 11: 110.
10. Scheffe, H. (1959) The Analysis of Variance, Wiley, New York.
11. Smith H. F. (1957) Missing Plot Estimates. *Biometrics*, 13: 115-118.
12. Snedecor, G. W. and Cochran, W. G. (1957/56) Statistical Methods, 6th/5th editions, Iowa State Univ. Press, Ames, Iowa.

13. Wilkinson, G. N. (1958) The Analysis of Variance and The Derivation of Standard Errors for Incomplete Data. *Biometrics*, 14: 360-384.
14. Wilkinson, G. N. (1957) The Analysis of Covariance with Incomplete Data. *Biometrics*, 14: 363-372.
15. Wilkinson, G. N. (1958) Estimation of The Missing Values for The Analysis of Incomplete Data. *Biometrics*, 14: 257-286.
16. Yates, F. (1933) The Analysis of Replicated Experiments When Field Results Are Incomplete. *The Empire J. Exp. Agri.*, 1: 129-142.

VIII. ACKNOWLEDGMENT

The writer wishes to express her sincere appreciation to her major professor, Dr. Holly C. Fryer of the Department of Statistics, for reviewing this manuscripts and offering advice.

MISSING PLOT TECHNIQUES

by

CHING-LAN WU

B. A., National Taiwan University, 1969

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1973

The occurrence of missing data in a statistically designed experiment requires some modification of the usual statistical techniques because the orthogonality or balance of the design is destroyed. One of the first papers on the subject of estimating the missing yield was published by Allan and Wishart (1930), and had been expanded by Yates (1933) to cover several missing units in given experiment.

Bartlett (1937), Anderson (1946) and Coons (1957) had used the analysis of covariance model to analyze the experiments with missing data. The technique employs the computational procedures of a covariance analysis using a dummied X covariate. With $q(>1)$ missing observations, a multiple covariance analysis is required. Set up concomitant variables X_m for each missing observation. Each of these X_m will be zero in all cells except in that cell corresponding to the missing observation with which the given X_m is associated; in that one cell it will have a value of $-n$. A multiple covariance analysis will be performed on y and the q covariates X_m . A complete analysis of covariance will provide unbiased tests for treatment effects.

Another missing plot technique is the least squares analysis which produces what we call the exact analysis. The least squares estimators obtained are those that minimize the residual sum of squares. Usually, the least squares computations are more complicated than analysis of covariance.

Either of these two techniques is suggested to be used as a missing plot technique, because of its generality of application and the ease with which 'exact' tests of significance will be obtained.