MISSING AND MIXED UP OBSERVATIONS IN ANALYSIS OF COVARIANCE

BY

D. RAGHAVARAO AND HAR VISHAN LAL

Punjab Agricultural University, Ludhiana

(Received in October, 1969)

1. INTRODUCTION

Missing and mixed up observations often arise in experimental work due to various reasons and methods of handling such situations is available in literature only when they occur in analysis of variance problems [1, 2, 3, 5]. In the present paper, we discuss the methods of analysing data when missing and mixed up observations occur in covariance analysis.

2. MISSING OBSERVATIONS IN COVARIANCE ANALYSIS

In experiments designed to use covariance analysis, missing observations may arise on (*i*) random variable (or study variable), (*ii*) concomitant variable, or (*iii*) both random and concomitant variables.

If some observations are missing on the random variable alone then Yate's method or Bartlett's technique can be employed to handle the situation. Yates' method of estimating the missing observations cannot be applied when values on the concomitant variables are missing. However, Bartlett's technique of defining more concomitant variables can be suitably utilised to handle suitations at (*ii*) and (*iii*).

Let there be s concomitant variables $z_1, z_2, \ldots, z_s$ and let y be the random variable. Let $m+n$ be the total experimental units and data on the random variable and all concomitant variables be available on the first n units while either y value or some of $z_1, z_2 \ldots, z_s$

values be missing on the remaining m plots. The technique consists of defining m more pseudo concomitant variables which take zero values on the units with no missing observations and take values as shown in Table 1 on the m units with missing observations :

TABLE 1

Missing units	*Pseudo concomitant variables*				*Random variable y*
	z_{s+1}	z_{s+2}	z_{s+3}	z_{s+m}	
1	k_1	0	0 ...	0	0
2	0	k_2	0 ...	0	0
3	0	0	k_3 ...	0	0
.	.	.	. ...	.	.
.	.	.	. ...	.	.
.	.	.	. ...	.	.
m	0	0	0 ...	k_m	0

$k_1, k_2, \ldots, k_m$ are non zero real numbers

We now mathematically prove that the analysis based on pseudo concomitant variables gives the same analysis as if the existing observations are analysed.

Let $X=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be the $(n+m)\times p$ design matrix of known constants, where X_1 corresponds to the units with no missing observations, and let $Z=\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$ be the $(n+m)\times s$ matrix corresponding to concomitant observations where Z_1 corresponds to units with no missing observations. Let y be a $n\times 1$ column vector of the observations on the random variable. If the first n units are taken into account, we have the model

$$E(y)=X_1\beta+Z_1\gamma, \qquad \ldots(2{\cdot}1)$$

where E stands for mathematical expectation, β is a $p\times 1$ column vector of unknown parameters and γ is a $s\times 1$ column vector of regression coefficients with the usual assumptions. The estimated $\hat{\beta}$

and $\hat{\gamma}$ are solutions of the normal equations

$$\begin{bmatrix} X_1'X_1 & X_1'Z_1 \\ Z_1'X_1 & Z_1'Z_1 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X_1'y \\ Z_1'y \end{bmatrix}, \qquad \ldots(2\cdot2)$$

and the residual sum of squares R_0^2 will be given by

$$R_0^2 = y'y - \hat{\beta}'X_1'y - \hat{\gamma}'Z_1'y. \qquad \ldots(2\cdot3)$$

With the newly added pseudo concomitant variables, the model becomes

$$E\begin{bmatrix} y \\ O_{m,\,1} \end{bmatrix} = \begin{bmatrix} X_1 & Z_1 & O_{n,\,m} \\ X_2 & Z_2 & K \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \\ \delta \end{bmatrix}, \qquad \ldots(2\cdot4)$$

where $K = \text{diag}\,(k_1, k_2, \ldots, k_m)$, $O_{m,\,n}$ is an $m \times n$ null matrix and δ is $m \times 1$ column vector of the additional regression coefficients. The normal equations for the set up (2·4) are given by

$$\begin{bmatrix} X_1'X_1 + X_2'X_2 & X_1'Z_1 + X_2'Z_2 & X_2'K \\ Z_1'X_1 + Z_2'X_2 & Z_1'Z_1 + Z_2'Z_2 & Z_2'K \\ KX_2 & KZ_2 & K^2 \end{bmatrix} \begin{bmatrix} \beta^* \\ \gamma^* \\ \delta^* \end{bmatrix} = \begin{bmatrix} X_1'y \\ Z_1'y \\ O_{m,\,n} \end{bmatrix}, \qquad \ldots(2\cdot5)$$

where β^*, γ^* and δ^* are the estimates of the respective parameters with $m+s$ cocomitant variables.

Since K is non-singular, the third equation of (2·5) gives

$$K\delta^* = -X_2\beta^* - Z_2\gamma^*. \qquad \ldots(2\cdot6)$$

Substituting the value of δ^* of (2·6) in the first and second equations of (2·5), we observe that β^* and γ^* satisfy the same normal equations as $\hat{\beta}$ and $\hat{\gamma}$ satisfy. Thus the estimates of parameters and regression coefficients remain the same with the additional pseudo concomitant variables as if the analysis was performed with the existing observations. We can also see that similar results hold under any null hypothesis and the analysis will remain as if the affected plots are ignored in the analysis.

Since case (*ii*) is a particular case of (*iii*), similar results also hold for case (*ii*).

3. MIXED UP OBSERVATIONS IN COVARIANCE ANALYSIS

Nair (1940) gave a method of analysing the mixed up data in analysis of variance by using Bartlett's technique. Chakrabarti (1963) in reply to a query on mixed up observations established the equivalence of Nair's technique and the least squares technique.

In the field experimentation due to various reasons the boundaries between the plots may disappear and in such cases the data on individual plots may not be available but total yields or total concomitant variable data will be available for a group of plots. Sometimes, investigators may commit certain errors resulting in mixed up observations. In covariance analysis the mixed up data may arise with respect to (*i*) random variable or yield, (*ii*) concomitant variable, or (*iii*) both random and concomitant variables. Situation (*i*) can be handled as in analysis of variance. We shall give here the mathematics of case (*iii*) as it includes case (*ii*) also.

Let the observational set up be

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & Z_1 \\ X_2 & Z_2 \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix}. \qquad \ldots(3.1)$$

where y_1, y_2, X_1, X_2, Z_1, Z_2, β and γ are of orders $n \times 1$, $m \times 1$, $n \times p$, $m \times p$, $n \times s$, $m \times s$, $p \times 1$ and $s \times 1$ respectively. Let the m plots yield corresponding to the vector y_2 be mixed up giving the total U and let each of the s concomitant variables be mixed up on these m plots leaving totals $U_1, U_2, \ldots, U_s$ respectively. By defining $(m-1)$ more pseudo concomitant variables which take the value 0 for unaffected plots and take the values as indicated in Table 2 for the affected plots, we will show that we get the correct analysis of covariance.

TABLE 2

Affected plots	*Original concomitant variables* z_1	z_2	...	z_s	*Additional Pseudo variables* z_{s+1}	z_{s+2}	...	z_{s+m-1}	*Yield* y
1.	U_1/m	U_2/m	...	U_s/m	1	1	...	1	U/m
2.	U_1/m	U_2/m	...	U_s/m	$1-m$	1	...	1	U/m
3.	U_1/m	U_2/m	...	U_s/m	1	$1-m$	...	1	U/m
.	.	.	...	.	.	.	...	.	.
.	.	.	...	.	.	.	...	.	.
.	.	.	...	.	.	.	...	.	.
m	U_1/m	U_2/m	...	U_s/m	1	1	...	$1-m$	U/m

The least squares set up for meeting the present situation will be

$$E\begin{bmatrix} y_1 \\ \frac{1}{\sqrt{m}} E_{1,m}\, y_2 \end{bmatrix} = \begin{bmatrix} X_1 & Z_1 \\ \frac{1}{\sqrt{m}} E_{1,m}\, X_2 & \frac{1}{\sqrt{m}} E_{1,m}\, Z_2 \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix}, \quad \ldots(3\,2)$$

where $E_{m,n}$ denotes an $m \times n$ matrix with $+1$ everywhere, and normal equations estimating β and γ are given by

$$\begin{bmatrix} X_1'X_1 + \frac{1}{m} X_2' E_{m,m}\, X_2 & X_1'Z_1 + \frac{1}{m} X_2' E_{m,m}\, Z_2 \\ Z_1'X_1 + \frac{1}{m} Z_2' E_{m,m}\, X_2 & Z_1'Z_1 + \frac{1}{m} Z_2' E_{m,m}\, Z_2 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix}$$

$$= \begin{bmatrix} X_1'y_1 + \frac{1}{m} X_2' E_{m,m}\, y_2 \\ Z_1'y_1 + \frac{1}{m} Z_2' E_{m,m}\, y_2 \end{bmatrix}. \quad \ldots(3\cdot3)$$

With the added pseudo concomitant variables, the set up will be

$$E\begin{bmatrix} y_1 \\ \frac{U}{m} E_{m,1} \end{bmatrix} = \begin{bmatrix} X_1 & Z_1 & O \\ X_2 & \frac{1}{m} E_{m,1}\alpha' & M \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \\ \delta \end{bmatrix}, \quad \ldots(3\cdot4)$$

where $\alpha' = (U_1, U_2, \ldots, U_m)$, $M = \begin{bmatrix} E_{1,m-1} \\ E_{m-1,m-1} - mI_{m-1} \end{bmatrix}$,

and δ is a $m \times 1$ column vector of regression coefficients corresponding to the newly added concomitant variables.

The normal equations for model (3·4) are

$$\begin{bmatrix} X_1'X_1 + X_2'X_2 & X_1'Z_1 + \frac{1}{m} X_2' E_{m,1}\alpha' & X_2'M \\ Z_1'X_1 + \frac{1}{m}\alpha\, E_{1,m}\, X_2 & Z_1'Z_1 + \frac{1}{m}\alpha\,\alpha' & O_{s,m-1} \\ M'X_2 & O_{m-1,s} & M'M \end{bmatrix} \begin{bmatrix} \beta^* \\ \gamma^* \\ \delta^* \end{bmatrix}$$

$$= \begin{bmatrix} X_1'y_1 + \frac{U}{m} X_2'E_{m,1} \\ Z_1'y_1 + \frac{U}{m}\alpha \\ \frac{U}{m} M'E_{m,1} \end{bmatrix}. \qquad \ldots(3{\cdot}5)$$

From the third equation of (3·5) we have

$$\delta^* = -(M'M)^{-1} M' \left(\frac{U}{m} E_{m,1} - X_2\beta^* \right). \qquad \ldots(3{\cdot}6)$$

Substituting the value of δ^* in first and second equations of (3·5) and simplifying using

$$M(M'M)^{-1} M' = \left(I_m - \frac{1}{m} E_{m,m} \right), \qquad \ldots(3{\cdot}7)$$

we can easily verify that β^* and γ^* satisfy the same normal equations (3 3) as satisfied by $\hat{\beta}$ and $\hat{\gamma}$. Similar results can be verified under any null hypothesis and thus the least square analysis and the present analysis with pseudo concomitant variables remains the same.

REFERENCES

1. Anderson, R.L. (1946) : Missing Plot Technique, Biometrics, 2, 41—47.
2. Bartlett, M.S. (1937) : Some examples of statistical methods of research in agricultural and applied biology. Journal Royal Statistical Society, Suppl. 4, 137—183.
3. Chakrabarti, M.C. (1963) : Answer to the query. Journal Indian Statistical Association, 1, 50—52.
4. Har Bishan Lal (1969) : Missing plot technique in analysis of covariance. Unpublished M.Sc. thesis, Punjab Agricultural University.
5. Nair, K.R. (1940) : The application of the technique of analysis of covariance to field experiments with several missing or mixed up plots. Sankhya, 4, 581-88.