Pertanika 9(1), 1 – 5 (1986)

# Analysis of Covariance For Repeated Measures Design with Missing Observations

AHMAD BIN ALWI and C.J. MONLEZUN[1]
*Guthrie Research Chemara,*
*Seremban, Negeri Sembilan.*

**Key words: Repeated measures design; subspaces; non-centrality; orthonormal basis; analysis of covariance.**

## ABSTRAK

*Analisis kovarians dianggap sebagai salah satu kaedah yang paling kurang difahami dan diajar di antara semua kaedah statistik gunaan. Kebanyakan buku kaedah tidak menyentuh sama sekali analisis kovarians (Guttman* et. al., *1982), atau dengan sepintas lalu (Brownlee, 1965), dan ada yang memberi perhatian dengan banyaknya seperti Federer (1955), Snedecor dan Cochran (1980), Steel dan Torrie (1980), dan Winer (1971). Kebanyakan daripada buku-buku tersebut menumpukan kepada data yang seimbang, khususnya data yang mempunyai bilangan cerapan yang sama dalam subkelas. Apa jadi jika data yang diperolehi tidak seimbang dan malahan pula ada di antara cerapan yang hilang? Sudah tentu analisis kovarians menjadi lebih rumit. Penggunaan geometri dalam analisis kovarians diharap dapat memberi kefahaman dan juga meluaskan lagi jenis-jenis kaedah yang boleh dipertimbangkan bagi menyelesaikan masalah seumpama ini.*

## ABSTRACT

*Analysis of covariance might be one of the most misunderstood and inadequately taught of all applied statistical methods. Many methods books do not deal with it at all (Guttman* et. al.*, 1982), or sparingly (Brownlee, 1965), and most of those that treat it substantially, such as Federer (1955), Snedecor and Cochran (1980), Steel and Torrie (1980), and Winer (1971), concentrate on balanced data, namely those which have equal numbers of observations in the subclasses. What happens if the data are not balanced and moreover if some of the observations are missing? The missing observations complicate computations and affect what is estimable. The analysis of covariance would become more complex. The application of geometry in the analysis of covariance may offer an understanding of the analysis as well as broaden the variety of methods that can be considered. When there are no missing observations on the repeated measures factor(s), computational algorithms can be used (see Henderson and Henderson, 1979).*

## INTRODUCTION

The design considered here is a two-factor Repeated Measures Design. Let $Y_{ijk}$ be the measurement made on subject $i$ $(1 \leqslant i \leqslant n_j)$ at level $j$ $(1 \leqslant j \leqslant a)$ of factor A and level $k$ $(1 \leqslant k \leqslant b)$ of factor B. For every $Y_{ijk}$, there is a concomitant measurement $X_{ijk}$. If we let $a = 3$, $b = 4$, $n_1 = 3$, $n_2 = 2$, $n_3 = 4$; we can tabulate a data table as in Table 1.

We arbitrarily set the observations $(Y_{113}, X_{113})$, $(Y_{312}, X_{312})$, $(Y_{123}, X_{123})$, $(Y_{232}, X_{232})$, $(Y_{233}, X_{233})$ and $(Y_{334}, X_{334})$ be missing.

[1]Ass. Prof., Dept. of Exp. Statistics, LSU, U.S.A.

TABLE 1
Data table for observations

| | $B_1$ | $B_2$ | $B_3$ | $B_4$ |
|---|---|---|---|---|
| | $Y_{111}, X_{111}$ | $Y_{112}, X_{112}$ | $Y_{113}, X_{113}^*$ | $Y_{114}, X_{114}$ |
| $A_1$ | $Y_{211}, X_{211}$ | $Y_{212}, X_{212}$ | $Y_{213}, X_{213}$ | $Y_{214}, X_{214}$ |
| | $Y_{311}, X_{311}$ | $Y_{312}, X_{312}^*$ | $Y_{313}, X_{313}$ | $Y_{314}, X_{314}$ |
| $A_2$ | $Y_{121}, X_{121}$ | $Y_{122}, X_{122}$ | $Y_{123}, X_{123}^*$ | $Y_{124}, X_{124}$ |
| | $Y_{221}, X_{221}$ | $Y_{222}, X_{222}$ | $Y_{223}, X_{223}$ | $Y_{224}, X_{224}$ |
| | $Y_{131}, X_{131}$ | $Y_{132}, X_{132}$ | $Y_{133}, X_{133}$ | $Y_{134}, X_{134}$ |
| | $Y_{231}, X_{231}$ | $Y_{232}, X_{232}^*$ | $Y_{233}, X_{233}^*$ | $Y_{234}, X_{234}$ |
| $A_3$ | $Y_{331}, X_{331}$ | $Y_{332}, X_{332}$ | $Y_{333}, X_{333}$ | $Y_{334}, X_{334}^*$ |
| | $Y_{431}, X_{431}$ | $Y_{432}, X_{432}$ | $Y_{433}, X_{433}$ | $Y_{434}, X_{434}$ |

*The observations assumed missing.

## MODEL DESCRIPTION

In the model where there is no covariate, the model used by A. Ahmad and C.J. Monlezun (1984) is given by

$$Y_{ijk} = U_{jk} + S_{ij} + E_{ijk}$$

where $U_{jk}$ is the cell mean for the level $j$ of factor A and the level $k$ of factor B, $S_{ij}$ is the effect of subject $i$ in the level $j$ of factor A, and $E_{ijk}$ is the random error component. With the additional concomitant measurement $X_{ijk}$, a model can now be written as

$$Y_{ijk}(X) = U_{jk}(X) + S_{ij} + E_{ijk}$$

where $U_{jk}(X) = \beta_0^{(jk)} + \beta_1(X_{ijk} - \overline{X}_{....})$ and $\beta_0^{(jk)}$ is the intercept for group $(j,k)$ and $\beta_1$ is the common slope of all lines. Let $Y_x$ be the observational vector and it can be written as illustrated in Table 2.

An alternative way of writing the model is

$$Y_X \sim MVN\,(E(Y_X) \in C_X = C \boxplus X,\ \sigma_E^2 I + \sigma_S^2 J).$$

Note that the covariance structure remains the same as in the ANOVA case discussed by A. Ahmad and C.J. Monlezun (1984). All subspaces defined in A. Ahmad and C.J. Monlezun (1984) will be used throughout this paper.

## HYPOTHESES TESTING

In the ANOVA case, we have shown that there is no exact test for testing no main effect A. A similar situation prevails in the case of analysis of covariance. Therefore we are interested in testing the following hypotheses:

$$H_B : U_{.k}(X) = U_{.k'}(X) \Leftrightarrow \beta_0^{(.k)} = \beta_0^{(.k')}$$

$$H_{AB} : U_{jk}(X) - U_{j'k}(X) = U_{jk'}(X) - U_{j'k'}(X)$$

$$\Leftrightarrow \beta_0^{(jk)} - \beta_0^{(j'k)} = \beta_0^{(jk')} - \beta_0^{(j'k')}$$

In the analysis of variance model, all observations for the cell $(j, k)$ have the same mean response $U_{jk}$. This is not so with the covariance model, since the mean response here depends on the $(j, k)$ combinations and also on the value of the concomitant variable $X_{ijk}$ for the experimental unit. Thus the expected response for the $(j, k)$ cell with the covariance model is given by a regression line:

$$U_{jk}(X) = \beta_0^{(jk)} + \beta_1(X_{ijk} - \overline{X}_{...})$$

where

$$\beta_0^{(jk)} = U_{..} + A_j + B_k + (AB)_{jk}.$$

TABLE 2
Observational vector

$$Y_{ij}X_{ij} = \begin{bmatrix} Y_{ij1} & X_{ij1} \\ Y_{ij2} & X_{ij2} \\ Y_{ij3} & X_{ij3} \\ Y_{ij4} & X_{ij4} \end{bmatrix} \qquad \text{for } ij = 21, 22, 13, 43$$

$$Y_{11}X_{11} = \begin{bmatrix} Y_{111} & X_{111} \\ Y_{112} & X_{112} \\ Y_{114} & X_{114} \end{bmatrix} \qquad Y_{12}X_{12} = \begin{bmatrix} Y_{121} & X_{121} \\ Y_{122} & X_{122} \\ Y_{124} & X_{124} \end{bmatrix}$$

$$Y_{31}X_{31} = \begin{bmatrix} Y_{311} & X_{311} \\ Y_{313} & X_{313} \\ Y_{314} & X_{314} \end{bmatrix} \qquad Y_{33}X_{33} = \begin{bmatrix} Y_{331} & X_{331} \\ Y_{332} & X_{332} \\ Y_{333} & X_{333} \end{bmatrix}$$

$$Y_{23}X_{23} = \begin{bmatrix} Y_{231} & X_{231} \\ Y_{234} & X_{234} \end{bmatrix}$$

$$Y_X = [\, Y_{11}X_{11},\ Y_{21}X_{21},\ Y_{31}X_{31} \mid Y_{12}X_{12},\ Y_{22}X_{22} \mid Y_{13}X_{13},\ Y_{23}X_{23},\ Y_{33}X_{33},\ Y_{43}X_{43} \,]$$

Note that $\beta_0{}^{(.k)} = U.. + B_k$ for $\sum_j A_j = \sum_j (AB)_{jk} = 0$ If we want to measure the difference at any convenient point $X_{ijk}$, say $X_{ijk} = \bar{X} \ldots$, then $(U\ldots + B_1) - (U\ldots + B_2) = B_1 - B_2$.

Thus $B_1 - B_2$ measures how much higher the mean response is with $B_1$ than with $B_2$ for any value of $X_{ijk}$. Therefore for testing no main effect B in $U_{jk}(X)$, the regression lines must have equal slopes and the test for main effect B is $B_i - B_i = 0$. In other words all of the $B_k$'s have to be equal. *Fig. 1* illustrates an experiment with four levels of factor B, and how these regression lines might appear.

In constructing the statistics, we first need to define subspaces for the numerator space and the error space. For the error space, we need a space that is orthogonal to $C_X$ and $W_S$. Recall from ANOVA case that the smallest subspace containing both C and $W_S$ is $W_S \boxplus B \boxplus AB$, and thus the smallest subspace that contains only X, C, and $W_S$ is

$$P_E X \oplus (W_S \boxplus B \boxplus AB).$$

Therefore the error space that we desire must be

$$E_X = (P_E X \oplus (W_S \boxplus B \boxplus AB))^{\perp} = E \ominus P_E X$$

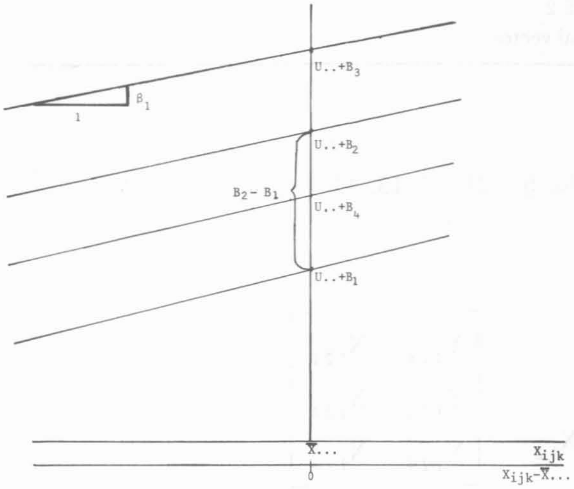


*Fig. 1: Regression lines of an experiment with four levels of factor B.*

From the ANOVA case, we have also defined the following subspaces:

$T = N_B \oplus (W_S \boxplus AB)$ for $H_B$

$T = N_{AB} \oplus (W_S \boxplus B)$ for $H_{AB}$

Let $M_B = (W_S \boxplus AB)$ and $M_{AB} = (W_S \boxplus B)$. The smallest subspace containing $W_S$ and $W_{D,X}$ is

$$(P_{ND} X + P_E X) \oplus M_D \quad \text{for } D = B, AB$$

The numerator space for testing $H_D$ has to be orthogonal to $E_X$, $W_S$ and $W_{D,X}$. Therefore we can define the numerator space by

$$N_{D,X} = (P_E X \oplus N_D \oplus M_D) \ominus ((P_{N_D} X + P_E) \oplus M_D) = \eta_{D,X} \oplus \tau_{D,X}$$

where $\eta_{D,X} = N_D \ominus P_{N_D} X$ and

$$\tau_{D,X} = \frac{P_{N_D} X}{X' P_{N_D} X} - \frac{P_E X}{X' P_E X}$$

## DISTRIBUTION OF SUM OF SQUARES/ TEST STATISTICS

Here we are interested to determine the distribution of the numerator and the error sum of squares and hence derive the test statistics for testing $H_B$ and $H_{AB}$. For simplicity, we let D = B, AB. The sum of squares for effect D is defined as

$$SSD = Y_X' P_{N_{D,X}} Y_X \quad \text{with } d \text{ degree of freedom}$$

$$\text{where } d = \begin{cases} (b-1) & \text{if } D = B \\ (a-1)(b-1) & \text{if } D = AB \end{cases}$$

Let $\{v_k\}$ be an orthonormal basis for $N_{D,X}$. Then the numerator sum of squares can be rewritten as $\sum_{k=1}^{d} (Y_X' v_k)^2$. We note that $Y_X' v_k$ is normally distributed with mean $E(Y_X)' v_k$ (nonzero since $E(Y_X) \in C_X$ and $v_k \in C_X$) and variance $\sigma^2_E$. Why is the variance equal to $\sigma^2_E$? Is any two $(Y_X' v_k, Y_X' v_m)$ independent? To answer both the questions we look at the following calculations of variance and covariance:

$$\begin{aligned} Var(Y_X' v_k) &= v_k' \, Var(Y_X) \, v_k \\ &= v_k' [\sigma^2_E I + \sigma^2_S J] v_k \\ &= \sigma^2_E v_k' v_k + \sigma^2_S v_k' J v_k \\ &= \sigma^2_E \quad (J v_k = 0 \text{ since } v_k \in W_S^{\perp}) \end{aligned}$$

$$\begin{aligned} Cov(Y_X' v_k, Y_X' v_m) &= v_k' [\sigma^2_E I + \sigma^2_S J] v_m \\ &= \sigma^2_E v_k' v_m + \sigma^2_S v_k' J v_m \\ &= 0 \end{aligned}$$

Since the covariance of any two $Y_X' v_k$ is zero, this would imply that they are independent random variables, and hence we can have the distribution of sum of squares for effect D as

$$\begin{aligned} SSD &= \sum_{k=1}^{d} [N(E(Y_X)' v_k, \sigma^2_E)]^2 \\ &= \sigma^2_E \chi^2 (d; \text{noncentral}) \end{aligned}$$

where the noncentrality parameter is given by

$$\frac{E(Y_X)' P_{N_{D,X}} E(Y_X)}{2\sigma^2_E}$$

Similarly, the distribution of the sum of squares error can be derived as

$$SSE = \sum_{1}^{t-ab-n.+a} [N(0, \sigma^2_E)]^2$$

$$= \sigma^2_E \chi^2 \text{ (t–ab–n.+a; central)}$$

After knowing the distribution of the sum of squares, we can now write a test statistic for testing no effect D in $U_{jk}(X)$ as

$$\frac{SSD \div d}{SSE \div t-ab-n.+a} = \frac{\chi^2 \text{ (d; noncentral)}}{\chi^2 \text{ (t–ab–n.+a; central)}}$$

The test statistic above is distributed as a noncentral F – distribution and when the null hypothesis is true, the test statistic becomes central F – distribution.

## REFERENCES

A. AHMAD and C.J. MONLEZUN, (1984): A Geometric Look at Repeated Measures Design with Missing Observations. *PERTANIKA* **7(2):** 71 – 81.

GRAYBILL, F.A., (1976): Theory and application of the Linear Models. New York: Holt, Rhinehart, and Winston, Inc.

HENDERSON, C.R. Jr., (1982): Analysis of Covariance in the mixed model: higher-level, nonhomogeneous, and random regressions.

HENDERSON, C.R. Jr. and HENDERSON, C.R., (1979): Analysis of Covariance in mixed models with unequal subclass numbers. *Comm. in Statistics,* series A, **8:** 751 – 787.

NETER, J. and WASSERMAN, W., (1974): Applied Linear Statistical Models. Homewood: Richard D. Irwin Inc.

SCHWERTMAN, N.C., (1978): A Note on the Geisser-Greenhouse correction for incomplete data split-plot analysis. *Jour. Amer. Stat. Ass.* **73:** 393 – 396.

SEARLE, S.R., (1971): Linear Models. New York: John Wiley and Sons, Inc.

SNEDECOR, G.W. and COCHRAN, W.G., (1980): Statistical Methods. The IOWA State Univ. Press, Ames, IOWA.

STEEL, R.G.D. and TORRIE, J.H., (1980): Principles and procedures of Statistics. New York: McGraw-Hill Book Company, Inc.

TIMM, N.H., (1975): Multivariate analysis with applications in Education and Psychology. Monterey: Books/Cole publishing company.

WINER, B.J., (1971): Statistical principles in experimental design. New York: McGraw-Hill Book Company, Inc.

*(Received 7 February 1985)*