Serdica
Bulgariacae mathematicae
publicationes

Сердика
Българско математическо
списание

ESTIMATION OF MULTIPLE MISSING VALUES IN FIXED EFFECTS EXPERIMENTAL DESIGNS

SHAWNM ABDUL-KADIR

This paper deals with the estimation of multiple missing values for some unifactor experimental designs. The method of minimizing the residual sum of squares is used and an example is given.

Introduction. The statistical analysis of unifactor experimental designs is complicated very often by the lacking of some observations. When some values are missing, the usual method of computing the various sums of squares cannot be used unless the missing values are estimated through the existing data.

Nowadays a number of computer algorithms are known for estimating missing values in experimental designs. Some of them use iterative methods (Pearce & Jeffers [3], Preece [4]), others are non-iterative (see Draper [2], Rubin [5, 6]). Usually such algorithms are based primarily on subroutines used for complete data set analysis. In most of the methods the vector of estimated residuals is obtained as a part of the computer analysis.

This paper describes a procedure for estimating multiple missing values in Cross-Over design and Graeco-Latin square design by minimizing the residual sum of squares and shows that for a (3×3) Graeco-Latin square design the estimates of the missing values are always zero, i. e. $\{\theta_h=0\}_{h=1}^{T}$. Also an example is given.

1. Cross-Over design. The model is as follows:

$$Y_{ijk}=\mu+T_i+R_j+C_k+\varepsilon_{ijk}, \tag{1.1}$$

$$1\leq i\leq t,\quad 1\leq j\leq r,\quad 1\leq k\leq c,\quad r=t,\quad r\leq c.$$

Here (Y_{ijk}) denotes the value observed in the i-th treatment, j-th row and k-th column; $Y_{ijk}=0$ stands for the missing value in the cell (i, j, k). Now let us estimate the missing values for a Cross-Over design by minimizing the residual sum of squares (MSSR).

The following notations shall be used.

Let $\{\theta_h\}_{h=1}^{m}$ be the unkown missing values, and let A_1, A_2 and A_3 be a treatment, a row and a column factor respectively. The (i, j, k) cell in the design matrix is the only one comprizing 1 and the remaining values are zero.

Let us define the dummy variable corresponding to θ_h, as follows:

$a_{hi}=1$ if θ_h is in treatment i
$\quad 0$ otherwise,
$b_{hj}=1$ if θ_h is in row j
$\quad 0$ otherwise,
$c_{hk}=1$ if θ_h is in column k
$\quad 0$ otherwise.

The following three dummy variables will be defined by the equation (1. 2) if θ_g and θ_h, $h \neq g$ are missing.

$$\varnothing_{gh}(A_1)=1 \text{ if } \theta_g \text{ and } \theta_h \text{ are from one and the same treatment}$$
$$0 \text{ otherwise,}$$

(1.2)
$$\varnothing_{gh}(A_2)=1 \text{ if } \theta_g \text{ ang } \theta_h \text{ appear in the same row}$$
$$0 \text{ otherwise,}$$
$$\varnothing_{gh}(A_3)=1 \text{ if } \theta_g \text{ and } \theta_h \text{ are in the same column}$$
$$0 \text{ otherwise,}$$

Also let

(1.3)
$T_h(A_1)$ be the total sum of all observations for the treatment where θ_h appears,
$T_h(A_2)$ be the total sum of all observations for the row for which θ_h appears,
$T_h(A_3)$ be the total sum of all observations for the column for which θ_h appears.

Theorem 1.1. *For the Cross-Over design given by (1.1), the missing values* $\{\theta_h\}_{h=1}^{m}$ *are obtained by solving the following system of equations:*

$$(1.4) \qquad (c-1)(t-2)\theta_h+\sum_{g\neq h}\theta_g\{2-t[\Sigma\varnothing_{gh}(A_1)+\Sigma\varnothing_{gh}(A_3)]-c\Sigma\varnothing_{gh}(A_2)\}$$
$$=c\Sigma T_h(A_2)+t[\Sigma T_h(A_1)-\Sigma T_h(A_3)]-2G$$

for $1\leq h\leq m$, when $(t-1)(r-2)>m$.

Proof. The marginal totals are, as follows:

$$T_i=\sum_i Y_{i\,..}+\sum_{h=1}^{T}\theta_h a_{hi}, \quad i=1,\ldots,t,$$

(1.5)
$$R_j=\sum_j Y_{.\,j\,.}+\sum_{h=1}^{m}\theta_h b_{hj}, \quad i=1,\ldots,r,$$

$$C_k=\sum_k Y_{..\,k}+\sum_{h=1}^{m}\theta_h c_{hk}, \quad k=1,\ldots,c,$$

$$G=\sum_i\sum_j\sum_k Y_{ijk}+\sum_h\theta_h.$$

Here $Y_{i\,..}$, for example, means averaging over the indices replaced by points while the rest index remains fixed.

The usual sum of squares for residual is

$$\text{SSE}=\text{SSG}-\text{SST}-\text{SSR}-\text{SSC},$$

where

$$\text{total sum of squares}=\text{SSG}=\sum_i\sum_j\sum_k Y_{ijk}^2-\frac{Y^2_{...}}{tc}$$

$$\text{sum of squares for treatment}=\text{SST}=\sum_i\frac{Y^2_{i\,..}}{r}-\frac{Y^2_{...}}{tc}$$

$$\text{sum of squares for row}=\text{SSR}=\sum_j\frac{Y^2_{.\,j\,.}}{c}-\frac{Y^2_{...}}{tc}$$

$$\text{sum of squares for column}=\text{SSC}=\sum_k\frac{Y^2_{..\,k}}{t}-\frac{Y^2_{...}}{tc}$$

The values of θ_h which minimize SSE are to be found from the equations:

$$\frac{\partial \, SSE}{\partial \, \theta_h} = 2\theta_h - \frac{2}{r} \sum_i T_i a_{hi} - \frac{2}{c} \sum_j R_j b_{hj} - \frac{2}{t} \sum_k C_k c_{hk} + \frac{4Y\ldots}{tc}$$

for any h, $1 \leq h \leq m$.

Replacing T_i, R_j, C_k and G by (1.5), we obtain the following equations:

$$(1.6) \qquad (c-1)(t-2)\,\theta_h + \sum_{g \neq h} \theta_g [2 - t(\Sigma a_{hi} a_{gi} + \Sigma c_{hk} c_{gk}) - c b_{hj} b_{gj}]$$

$$= t \Sigma b_{hj} \Sigma Y_{..k} + c(\Sigma a_{hi} \Sigma Y_{i..} + \Sigma c_{hk} \Sigma Y_{.j.}) - 2Y_{...}$$

for fixed h, $1 \leq h \leq m$.

The desired equation (1.4) is obtained by using relations (1.2) and (1.3) in (1.6), i. e.

$$(c-1)(t-2)\,\theta_h + \sum_{g \neq h} \theta_g \{2 - t\,[\Sigma \varnothing_{gh}(A_1) + \Sigma \varnothing_{gh}(A_3)] - c\,\Sigma \varnothing_{gh}(A_2)\}$$

$$= c\,\Sigma\, T_h(A_2) + t\,[\Sigma\, T_h(A_1) + \Sigma\, T_h(A_3)] - 2G$$

for $(c-1)(t-2) > m$.

2. Graeco-Latin square design. The mathematical model for Graeco-Latin square design is given, as follows:

$$(2.1) \qquad Y_{ijkl} = \mu + T_i + R_j + C_k + G_l + \varepsilon_{ijkl},$$

for $1 \leq i \leq t$, $1 \leq j \leq r$, $1 \leq k \leq c$, $1 \leq l \leq s$, and $t = r = c = s$, $r > 3$.

Usually $Y_{ijkl} = 0$ for the missing value in cell (i, j, k, l).

We define dummy variables for treatments, rows, columns and Greek-Letters, as follows:

$a_{hi} = 1$ if θ_h is in treatment i
$\quad 0$ otherwise,
$b_{hj} = 1$ if θ_h is in row j
$\quad 0$ otherwise,
$c_{hk} = 1$ if θ_h is in column k
$\quad 0$ otherwise,
$d_{hl} = 1$ if θ_h is in Greek-Letter l
$\quad 0$ otherwise.

The following four dummy variables will be defined by using equation (2.2) if θ_g and θ_h, $h = g$ are missing.

(2.2)
$\varnothing_{gh}(A_1) = 1$ if θ_g and θ_h are from one and the same treatment
$\quad 0$ otherwise,
$\varnothing_{gh}(A_2) = 1$ if θ_g and θ_h appear in the same row
$\quad 0$ otherwise,
$\varnothing_{gh}(A_3) = 1$ if θ_g and θ_h are in the same column
$\quad 0$ otherwise,
$\varnothing_{gh}(A_4) = 1$ if θ_g and θ_h are in the same Greek-Letter
$\quad 0$ otherwise,

and A_1, A_2, A_3 and A_4 are treatment, row, column and Greek-Letter factors respectively.

Also let

$T_h(A_1)$ be the total sum of all observations for the treatment where θ_h appears,

(2.3) $T_h(A_2)$ be the total sum of all observations for the row for which θ_h appears,
$T_h(A_3)$ be the total sum of all observations for the column for which θ_h appears,
$T_h(A_4)$ be the total sum of all observations for the Greek-Letter for which θ_h appears.

Theorem 2.1. *For the Graeco-Latin square design the missing values* $\{\theta_h\}_{h=1}^{m}$ *are obtained by solving the following system of equations:*

(2.4)
$$(t-1)(r-3)\theta_h + \sum_{g \neq h} \theta_g [3 - r \sum_i \varnothing_{gh}(A_i)] = r \Sigma_i T_h(A_i) - 3G$$

for $1 \leq h \leq m$ *when* $(t-1)(r-3) > m$.

Proof. The marginal totals are found, as follows:

(2.5)
$$T_i = \sum_i Y_{i\ldots} + \sum_{h=1}^{m} \theta_h a_{hi}, \quad i = 1, \ldots, t$$
$$R_j = \sum_j Y_{.j..} + \sum_{h=1}^{m} \theta_h b_{hj}, \quad j = 1, \ldots, r$$
$$C_k = \sum_k Y_{..k.} + \sum_{h=1}^{m} \theta_h c_{hk}, \quad k = 1, \ldots, c$$
$$G_l = \sum_l Y_{\ldots l} + \sum_{h=1}^{m} \theta_h d_{hl}, \quad l = 1, \ldots, s$$
$$G = \sum_i \sum_j \sum_k \sum_l Y_{ijkl} + \sum_{h=1}^{m} \theta_h .$$

The usual sum of squares for residual is

$$SSE = SSG - SST - SSR - SSC - SSGL,$$

where

$$\text{total sum of squares} = SSG = \sum_i \sum_j \sum_k \sum_l Y_{ijkl}^2 - \frac{Y_{\ldots\ldots}^2}{tr}$$

$$\text{sum of squares for treatment} = SST = \sum_i \frac{{}^2Y_{i\ldots}}{r} - \frac{Y_{\ldots\ldots}^2}{tr}$$

$$\text{sum of squares for row} = SSR = \sum_i \frac{Y_{.j..}^2}{c} - \frac{Y_{\ldots\ldots}^2}{tr}$$

$$\text{sum of squares for column} = SSC = \sum_k \frac{Y_{..k.}^2}{t} - \frac{Y_{\ldots\ldots}^2}{tr}$$

$$\text{sum of squares for Greek-Letter} = SSCL = \sum_l \frac{Y_{\ldots l}^2}{s} - \frac{Y_{\ldots\ldots}^2}{tr} .$$

The values of θ_h which minimize SSE are to be found from the equations:

$$\frac{\partial\, SSE}{\partial\, \theta_h} = 0, \quad h = 1, \ldots, m,$$

i. e.

$$\frac{\partial\,\mathrm{SSE}}{\partial\,\theta_h}=2\theta_h-\frac{2}{t}\sum_i T_i\,a_{hi}-\frac{2}{r}\sum_j R_j\,b_{hj}-\frac{2}{c}\sum_k C_k\,c_{hk}-\frac{2}{s}\sum_l C_k\,c_{hl}+\frac{6G}{r^2}=\mathbf{0}.$$

Replacing T_i, R_j, C_k, G_l and G by (2.5), we obtain the following equations:

$$(2.6)\qquad (t-1)(r-3)\,\theta_h+\sum_{g\neq h}\theta_g\,[3-r\,(\Sigma a_{hi}\,a_{gi}+\Sigma b_{hj}\,b_{gj}+\Sigma c_{hk}\,c_{gk}+\Sigma d_{hl}\,d_{gl})]$$

$$=r\,[\Sigma a_{hi}\,\Sigma Y_{i\,\ldots}+\Sigma b_{hj}\,\Sigma Y_{.\,j\,..}+\Sigma c_{hkj}\,\Sigma Y_{..\,k\,.}+\Sigma d_{hl}\,\Sigma Y_{...\,l}]-3\Sigma Y_{\,\ldots}$$

for any h, $1\leqq h\leqq m$.

The desired system of equations (2.4) is obtained by using relations (2.2) and (2.3) in (2.6), i. e.

$$(t-1)(r-3)\,\theta_h+\sum_{g\neq h}\theta_g\,[3-r\sum_i \varnothing_{gh}(A_i)]=r\sum_i T_h\,(A_i)-3G$$

for $1\leqq h\leqq m$, when $(t-1)(r-3)>m$.

Corollary. *For a* (3×3) *Graeco-Latin Square design the missing values are always zero, i. e.* $\{\theta_h=0\}_{h=1}^{m}$.

Proof. The general total is found by the equation

$$\sum_h\sum_i T_h(A_i)=(T_i'+R_j'+C_k'+G_l')=G.$$

Also, any two missing values, θ_h and θ_g, $h=g$, are in the same treatment or in the same row or in the same column or in the same Greek-Letter.

For $r=3$ by using (2.4), we obtain the identity

$$0+0=0,$$

and so any value of θ_h is acceptable, which leads to an uncertain situation.

Thus any missing value θ_h has to be set equal to zero.

3. Example. 1. Let us illustrate the MSSS method for estimating the missing values by giving the following example (see [1, p. 315]):

Let us have a (5×5) Graeco-Latin square for estimating the response of Hylotrupes larvae to four nutrients, each at five equally spaced log concentrations; y = average log weight after 103 ± 5 days on the experimental diet; rows are yeast extract, columns are poptone levels, treatments are cholesterol levels and Greek-Letters are riboflavin levels.

Let the treatment observation be denoted by D, and let the two observations $D\varepsilon$ and $D\beta$ in columns 2 and 5 and rows 3 and 5, respectively, be missing, where ε and β are Greek letters.

The missing observations are denoted by θ_1 and θ_2 with $m=2$. Then the dummy variable $\varnothing_{12}$ is determined, as follows:

$$\varnothing_{12}(A_1)=1,\quad \varnothing_{12}(A_2)=0,\quad \varnothing_{12}(A_3)=0\ \text{and}\ \varnothing_{12}(A_4)=0.$$

Since $r=t=5$, the corresponding totals are

$$T_1(A_1)=1.45,\quad T_1(A_2)=2.63,\quad T_1(A_3)=1.95\ \text{and}\ T_1(A_4)=2.25,$$
$$T_2(A_1)=1.45,\quad T_2(A_2)=2.67,\quad T_2(A_3)=3.14\ \text{and}\ T_2(A_4)=2.11,$$
$$G=13.21.$$

The system of linear equations for calculating the missing values is obtained immediately from (2.4):

$$8\theta_1-2\theta_2=1.77,$$
$$-2\theta_1+8\theta_2=7.22$$

which yields $\theta_1=0.477$, $\theta_2=1.022$.

Discussion. The method of minimizing the residual sum of squares has been applied to estimating multiple missing values in Cross-Over design and Graeco-Latin square design.

The solution is obtained directly. The technique may be applied to a number of designs. It is usefui for the Randomized Block design where blocks are one of the factors. The formula holds also for a $(r\times r)$ Latin-Square design. For this model the system of m linear equations is the following:

$$(r-1)(r-2)\,\theta_h+\sum_{g=h}\theta_g\,[2-r\sum_i\varnothing_{gh}(A_i)]=r\sum_i T_h(A_i)-2G$$

for $h=1,\ldots,m$ and $(r-1)(r-2)>m$.

Acknowledgements. I wish to thank my research supervisor Professor D. Vandev for our discussions. I am much obliged to Professor L. Boneva and Dr I. Tzankova for their recommendations. The referees' comments are most gratefully acknowledged.

REFERENCES

1. G. I. Bliss. Statistics in Biology, New York, 1968.
2. N. R. Draper. Missing values in response surface design. *Technometrics*, 3. 1961, 389-398.
3. S. C. Pearce, J. N. R. Jeffers. Block designs and missing data. *J. R. Statist.*, 33, 1971, 131-136.
4. D. A. Preece. Iterative procedures for missing values in experiments. *Technometrics*, 13, 1971, 743-753.
5. D. B. Rubin. A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Appl. Statist.*. 21, 1972, 136-141.
6. D. B. Rubin. Noniterative least squares estimates, standard error and F-test for analyses of variance with missing data. *J. R. Statist.*, 33, 1976, 270-274.

Institute of Mathematics
P. O. Box 373
Sofia, Bulgaria

Received 29.01. 1991
Revised 15. 06. 1991