Robustness of some designs against missing data

R. SRIVASTAVA[1], **V. K. GUPTA**[1] **& A. DEY**[2], [1] *Indian Agricultural Statistics Research Institute, New Delhi, and* [2] *Indian Statistical Institute, New Delhi*

SUMMARY *This article studies the robustness of several types of designs against missing data. The robustness of orthogonal resolution III fractional factorial designs and second-order rotatable designs is studied when a single observation is missing. We also study the robustness of balanced incomplete block designs when a block is missing and of Youden square designs when a column is missing.*

1 Introduction

The robustness of statistical designs against missing observations has been studied by several researchers, e.g. Hedayat & John (1974), John (1976), Ghosh (1978, 1980, 1981, 1982a, b) and Dey & Dhall (1988). Ghosh (1978) introduced a criterion of robustness of designs in the following manner: a design is robust against missing observations if all the parameters are still estimatable under an assumed model when a given number of observations are missing. Furthermore, it was observed by Ghosh (1982b) that even in robust designs (in the above sense), some observations are more informative than others and, consequently, if a more informative observation is lost accidentally, the overall loss in efficiency is larger than that in the case when a less informative observation is lost.

The purpose of this paper is to extend the study of Ghosh to several types of design. We study the robustness of orthogonal resolution III designs and second-order rotatable designs when a single observation is missing. We also study the robustness of balanced incomplete block (BIB) designs when all the observations in a block are missing, and of Youden square designs when all the observations in a column are lost.

2 Some preliminary results

Consider the usual Gauss–Markov linear model

$$E(\mathbf{y}) = \mathbf{X}\theta \qquad D(\mathbf{y}) = \sigma^2 \mathbf{I}_n \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ observations vector, $\mathbf{X}$ the $n \times p$ design matrix, θ the p-component vector of parameters, σ^2 the per observation variance and $\mathbf{I}_n$ the nth-order identity matrix. When the order of the matrix is clear from the context, we shall simply write $\mathbf{I}$ instead of $\mathbf{I}_n$. For the present, assume that $\mathbf{X}$ has full column rank p. Suppose d is a robust design as per the criterion given in the Introduction against one missing observation. Let d_1 be the residual design when an observation (collected through d) is lost. Obviously, there are n possible designs $\{d_1\}$. Let $\mathbf{b}'$ represent the row in $\mathbf{X}$ corresponding to the missing observation. Then writing

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{b}' \end{bmatrix}$$

we have

$$\mathbf{X}_1'\mathbf{X}_1 = \mathbf{X}'\mathbf{X} - \mathbf{b}\mathbf{b}' \tag{2}$$

and

$$\det(\mathbf{X}_1'\mathbf{X}_1) = [1 - \mathbf{b}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{b}] \det(\mathbf{X}'\mathbf{X}) \tag{3}$$

If $\hat{\theta}_d (\hat{\theta}_{d_1})$ denotes the best linear unbiased estimator of θ using d (d_1), then

$$D(\hat{\theta}_d) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad D(\hat{\theta}_{d_1}) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} \tag{4}$$

and

$$\det[D(\hat{\theta}_{d_1})] = [1 - \mathbf{b}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{b}]^{-1} \det[D(\hat{\theta}_d)] \tag{5}$$

In equation (4), $D(\ )$ stands for the dispersion matrix.

The amount of information contained in the unavailable observation in d is therefore

$$I(\mathbf{b}) = \mathbf{b}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{b} \tag{6}$$

It is easy to see that

$$\sum_{\mathbf{b}} I(\mathbf{b}) = \operatorname{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = p$$

where tr ( ) stands for the trace of a square matrix. It should be noted that for a robust design d, $0 < I(\mathbf{b}) < 1$ for all $\mathbf{b}$. This measure of information $I(\ )$ is a reasonable one, as argued by Ghosh (1982b). Clearly, a good design should have a small $I(\mathbf{b})$ for all $\mathbf{b}$ so that the loss of an observation does not result in a large loss in efficiency of the residual design.

3 Robustness of orthogonal main effect plans

In this section, we consider the robustness of orthogonal main effect fractional factorial plans (or orthogonal resolution III plans). Recall that a fractional factorial plan is said to be an orthogonal main effect plan if it permits the estimation of the mean and all main effects with zero correlation under the assumption that all interactions are absent. For details on these fractions, a reference may be made to Dey (1985).

Let d be an n-run orthogonal main effect plan for an $s_1 \times s_2 \times \ldots \times s_k$ factorial. For notational convenience, we write $r_i = s_i - 1$, for $i = 1, 2, \ldots, k$ and $p = 1 + \Sigma r_i$. For the analysis of an orthogonal main effect plan, we postulate the linear model

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\theta} = \mu\mathbf{1} + \mathbf{A}\boldsymbol{\xi} \qquad D(\mathbf{y}) = \sigma^2\mathbf{I} \tag{7}$$

where μ is the general mean, $\boldsymbol{\xi}$ the vector of main effects, $\mathbf{1}$ a column of all unities and $\mathbf{I}$ an identity matrix; the orders of $\mathbf{1}$ and $\mathbf{I}$ should be clear from the context. In equation (7) $\mathbf{A}$ is the design matrix. The design d will be robust against a missing observation under equation (7) if the $(n-1) \times p$ matrix, obtained from $\mathbf{X}$ by deleting a row, has rank p. A main effect plan is called *saturated* if $n = p$. Obviously then, no saturated plan can be robust against missing observations under model equation (7). Alternatively, for saturated plans, we consider the robustness under the model

$$E(\mathbf{y}) = \mathbf{A}\boldsymbol{\xi} \qquad D(\mathbf{y}) = \sigma^2\mathbf{I} \qquad \text{rank}(\mathbf{A}) = p - 1 \tag{8}$$

Remembering that the columns of $\mathbf{X}$ in the case of orthogonal main effect plans (other than the first column of all unities) are orthogonal polynomials, one can show the following: (i) that the saturated plans under equation (8) are robust against one missing observation; (ii) that the amount of information contained in any observation is $(n-1)/n$ and is thus a constant for given n.

If an orthogonal main effect plan is not saturated, one can study the robustness under the full model (equation (7)). However, the robustness of a plan depends on the design and no general conclusions can be drawn. To illustrate this, we consider two orthogonal main effect plans for 3^7 experiments, one in 18 runs and the other in 16 runs. The details of these designs can be found, for example, in Dey (1985, pp. 31, 36). It is seen that while the 18-run plan is robust against the loss of one observation, the 16-run plan is not.

4 Robustness of balanced incomplete block designs

In this section, the robustness of BIB designs is studied when all the observations pertaining to a block are missing. The criterion of robustness is the same as that of Ghosh (1978), according to which a design is robust if the residual design permits the estimation of all treatment contrasts.

Let d be a BIB design with parameters v, b, r, k and λ, and incidence matrix $\mathbf{N}$. Without loss of generality, assume that the labels of the treatments appearing in the missing block are $1, 2, \ldots, k$. Thus the incidence matrix of the residual design, $\mathbf{N}^*$, can be written as

$$\begin{aligned} \mathbf{N}^* = [\boldsymbol{\phi} \quad \mathbf{N}_1] &= \mathbf{N} - \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \boldsymbol{\phi} & \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{1} & \mathbf{N}_1 \\ \boldsymbol{\phi} & \end{bmatrix} - \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \boldsymbol{\phi} & \end{bmatrix} \end{aligned} \tag{9}$$

where $\mathbf{N}_1$ is the $v \times (b-1)$ incidence matrix of the unaffected blocks in d, $\boldsymbol{\phi}$ is a $(v-k)$-component null vector and $\mathbf{0}$ is a $v \times (b-1)$ null matrix. Let $\mathbf{N}_1' = [\mathbf{N}_{11}' : \mathbf{N}_{12}']$, where $\mathbf{N}_{11}$ is the $k \times (b-1)$ incidence matrix of the 'affected' treatments versus the $(b-1)$ 'unaffected' blocks of d, and $\mathbf{N}_{12}$ is a similar matrix of the $(v-k)$ 'unaffected'

treatments. From the properties of a BIB design, the following results are easily verified:

$$\begin{aligned}\mathbf{N}_{11}\mathbf{N}_{11}' &= (r-\lambda)\mathbf{I}+(\lambda-1)\mathbf{J}_k \\ \mathbf{N}_{11}\mathbf{N}_{12}' &= \lambda\mathbf{J}_{k,v-k} \\ \mathbf{N}_{12}\mathbf{N}_{12}' &= (r-\lambda)\mathbf{I}+\lambda\mathbf{J}_{v-k}\end{aligned} \tag{10}$$

where $\mathbf{J}_{mn}$ is an $m \times n$ matrix in which all the entries are ones; if $m=n$, we write $\mathbf{J}_m$. If d_1 is the residual design when all observations of a block in d are missing, then d_1 has $v^*=v$ treatments, $b^*=b-1$ blocks and the replication- and block-size-vectors are, respectively, given by

$$\begin{aligned}\mathbf{r}^* &= [(r-1)\mathbf{1} : r\mathbf{1}]' \\ \mathbf{k}^* &= k\mathbf{1}\end{aligned} \tag{11}$$

The C matrix of the reduced intrablock normal equations for d_1 is given by

$$\mathbf{C}^* = \begin{bmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D}\end{bmatrix} \tag{12}$$

where

$$\begin{aligned}\mathbf{A} &= k^{-1}[(\lambda v-k)\mathbf{I}-(\lambda-1)\mathbf{J}_k] \\ \mathbf{B} &= -\lambda k^{-1}\mathbf{J}_{k,v-k} \\ \mathbf{D} &= \lambda v k^{-1}(\mathbf{I}-v^{-1}\mathbf{J}_{v-k})\end{aligned} \tag{13}$$

A generalized inverse of $\mathbf{C}^*$ is

$$\mathbf{C}^{*-} = \begin{bmatrix}\mathbf{U} & \mathbf{V} \\ \mathbf{V}' & \mathbf{W}\end{bmatrix} \tag{14}$$

where

$$\begin{aligned}\mathbf{U} &= k(\lambda v-k)^{-1}[\mathbf{I}+(\lambda v)^{-1}(v-k)^{-1}(2\lambda v-v-k)\mathbf{J}_k] \\ \mathbf{V} &= k(\lambda v)^{-1}(v-k)^{-1}\mathbf{J}_{k,v-k} \\ \mathbf{W} &= k(\lambda v)^{-1}\mathbf{I}\end{aligned} \tag{15}$$

Let the treatment effect vector be partitioned as $\mathbf{t}'=[\mathbf{t}_1' : \mathbf{t}_2']$ where $\mathbf{t}_1$ represents the effects of the k affected treatments and $\mathbf{t}_2$ that of the remaining treatments. If $\mathbf{p}'\mathbf{t}$ is a linear function of treatment effects, then $\mathbf{p}'\mathbf{t}$ is estimatable using the design d_1 if and only if

$$\mathbf{p}'\mathbf{C}^{*-}\mathbf{C}^* = \mathbf{p}' \tag{16}$$

Partitioning $\mathbf{p}'$ in conformity with that of $\mathbf{t}$ as $\mathbf{p}'=[\mathbf{p}_1' : \mathbf{p}_2']$, we have from equations (14) and (15)

$$\mathbf{p}'\mathbf{C}^{*-}\mathbf{C}^* = [\mathbf{p}_1' : -(v-k)^{-1}\mathbf{p}_1'\mathbf{J}_{k,v-k}+\mathbf{p}_2'(\mathbf{I}-(v-k)^{-1}\mathbf{J}_{v-k})] \tag{17}$$

Thus for the estimatability of $\mathbf{p}'\mathbf{t}$ we must have

$$-(v-k)^{-1}\mathbf{p}_1'\mathbf{J}_{k,v-k}+\mathbf{p}_2'(\mathbf{I}-(v-k)^{-1}\mathbf{J}_{v-k})=\mathbf{p}_2'$$

or

$$-\mathbf{p}_1'\mathbf{J}_{k,v-k}=\mathbf{p}_2'\mathbf{J}_{v-k}$$

or

$$-\mathbf{p}_1'\mathbf{1}=\mathbf{p}_2'\mathbf{1} \tag{18}$$

It is easy to see that equation (18) is always satisfied if $\mathbf{p}'\mathbf{t}$ is an elementary contrast. Thus all elementary contrasts are estimatable using d_1, which in turn implies that d is robust against the non-availability of all the observations in a block.

5 Robustness of Youden square designs

Consider a Youden square design in v treatments, arranged in k rows and v columns. Suppose all the observations pertaining to a column, collected through this design, are lost. The reduced normal equations for estimating linear functions of treatment effects for the residual design are $\mathbf{F}\mathbf{t}=\mathbf{Q}$ where

$$\mathbf{F}=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}' & \mathbf{D}\end{bmatrix} \tag{19}$$

$$\mathbf{A}=v(k-2)\mathbf{I}/(v-1)-(k^2-v-k)\mathbf{J}_k/(vk-k)$$

$$\mathbf{B}=-(k-1)\mathbf{J}_{k,v-k}/(v-1) \tag{20}$$

$$\mathbf{D}=v(k-1)[\mathbf{I}-v^{-1}\mathbf{J}_{v-k}]/(v-1)$$

$\mathbf{t}$ is the vector of treatment effects and $\mathbf{Q}$ the vector of adjusted treatment totals. Routine calculations show that a generalized inverse of $\mathbf{F}$ is

$$\mathbf{F}^-=\begin{bmatrix}\mathbf{U} & \mathbf{V}\\ \mathbf{V}' & \mathbf{W}\end{bmatrix} \tag{21}$$

where

$$\mathbf{U}=(v-1)\{\mathbf{I}/(vk-2v)+(2k^2-3k-v)\mathbf{J}_k/[vk(k-1)(k-2)(v-k)]\}$$

$$\mathbf{V}=(v-1)\mathbf{J}_{k,v-k}/[v(k-1)(v-k)] \tag{22}$$

and

$$\mathbf{W}=(v-1)\mathbf{I}/(vk-v)$$

Let $\mathbf{p}'\mathbf{t}=(\mathbf{p}_1':\mathbf{p}_2')\mathbf{t}$ be a contrast of treatment effects where $\mathbf{p}_1$ has k components and $\mathbf{p}_2$, $v-k$ components. Since

$$\mathbf{F}^-\mathbf{F}=\begin{bmatrix}\mathbf{I} & -\mathbf{J}_{k,v-k}\\ \mathbf{0} & \mathbf{I}-\mathbf{J}_{v-k}/(v-k)\end{bmatrix}$$

it follows that for the estimatability of $\mathbf{p}'\mathbf{t}$, we must have

$$\mathbf{p}_2'\mathbf{1}=-\mathbf{p}_1'\mathbf{1} \tag{24}$$

which holds if $\mathbf{p}'\mathbf{t}$ is an elementary contrast. Consequently, a Youden square design is robust against the loss of all the observations in a column.

TABLE 1. Values of $I(\mathbf{b})$ for central composite designs

No. of factors	No. of cube points	No. of star points	No. of centre points	Missing point	$I(\mathbf{b})$
4	16	8	2	Centre	0.50
				Cube	0.58
				Star	0.58
5	16	10	—	Cube	0.98
				Star	0.67
5	16	10	1	Centre	0.78
				Cube	0.88
				Star	0.61
6	64	12	—	Cube	0.34
				Star	0.67
6	64	12	1	Centre	0.57
				Cube	0.32
				Star	0.57
7	64	14	—	Cube	0.43
				Star	0.60
8	64	16	2	Centre	0.50
				Cube	0.55
				Star	0.55
9	256	18	—	Cube	0.18
				Star	0.55
10	256	20	—	Cube	0.18
				Star	0.51

6 Rotatable designs

For a second-order rotatable design, we compute in this section the loss of information resulting from the loss of one observation. It is clear that for a rotatable design the information measure $I(\mathbf{b})$ given by equation (6) is simply proportional to the variance of the estimated response at $\mathbf{b}$. To get an idea about the information contained at different points, we present in Table 1 results for central composite rotatable designs. From Table 1 it is clear that the information contained in all types of points is quite appreciable, especially for designs with small number of factors. As such, although the rotatable designs appear to be robust, when one observation is lost, the loss of information is considerable.

Correspondence: A. Dey, Indian Statistical Institute, Delhi Centre, 7 SJS, Sansanwal Marg, New Delhi 110016, India.

REFERENCES

DEY, A. (1985) *Orthogonal Fractional Factorial Designs* (New Delhi, Wiley Eastern).

DEY, A. & DHALL, S. P. (1988) Robustness of augmented BIB designs, *Sankhya B*, 50, pp. 376–381.

GHOSH, S. (1978) On robustness of designs against incomplete data, *Sankhya B*, 40, pp. 204–208.

GHOSH, S. (1980) On robustness of optimal balanced resolution V designs, *Journal of Statistical Planning and Inference*, 4, pp. 313–319.

GHOSH, S. (1981) Robustness of three dimensional designs, I, *Sankhya B*, 43, pp. 222–229.

GHOSH, S. (1982a) Robustness of BIB designs against non availability of data, *Journal of Statistical Planning and Inference*, 6, pp. 29–32.

GHOSH, S. (1982b) Information in an observation in robust designs, *Communications in Statistics A*, 11, pp. 1173–1184.

HEDAYAT, A. & JOHN, P. W. M. (1974) Resistant and susceptible BIB designs, *Annals of Statistics*, 2, pp. 148–158.

JOHN, P. W. M. (1976) Robustness of incomplete block designs, *Annals of Statistics*, 4, pp. 960–962.