BU-888-M

RECOVERY OF INTERBLOCK INFORMATION FOR INCOMPLETE BLOCK DESIGNS

O. B. Allen

Departments of Animal and Poultry Science
and Mathematics and Statistics
University of Guelph, Guelph, Ontario

Key Words and Phrases: partially balanced designs; missing data; SAS; BMDP.

ABSTRACT

This paper was motivated by the limited statistical software available for the analysis of incomplete block designs which recovers interblock information, particularly for partially balanced designs. The paper shows how one can achieve Yates (1940) estimates by making an appropriate transformation of the response vector, Y, and the design matrix, X. The method is general and will work for any balanced, partially balanced or unbalanced design, including lattices. The method is illustrated using SAS.

INTRODUCTION

A model for the recovery of interblock information from an incomplete block design with t treatments and b blocks of size k is

$$Y_{ij} = \mu + \beta_j + \tau_i + \varepsilon_{ij} \qquad \begin{array}{l} i = 1,\cdots,t \\ j = 1,\cdots,b \end{array} \qquad (1)$$

but where only certain combinations of i and j occur, depending on the design. We shall assume that the ε_{ij} are independently distributed, with mean 0 and variance σ^2. To recover interblock information a further assumption must be made – that the β_j are independently distributed, with mean 0 and variance σ_β^2. Further, the ε_{ij} and β_j are assumed independent of one another. Finally, μ and τ are regarded as fixed effects (Cochran and Cox, 1957, p. 382). We will set $\mu = 0$ in order that the remaining parameters are estimable. The τ's then represent the treatment means. If the variance components are to be estimated using maximum likelihood or restricted maximum likelihood (REML) (Patterson and Thompson, 1971), a further distributional assumption is required. Typically, the ε_{ij} and β_j are assumed to be normally distributed.

An equivalent formulation is to state that the observation vector, Y, follows a multivariate normal distribution with $E(Y_{ij}) = \mu + \tau_i$, $Var(Y_{ij}) = \sigma_\beta^2 + \sigma^2$. Further, for two observations in the same block, $Cov(Y_{ij}, Y_{kj}) = \sigma_\beta^2$. Observations in separate blocks are uncorrelated. In matrix notation, the covariance matrix of Y (if the observations are ordered by blocks) is block diagonal and can be written using direct product notation as $\Sigma = I_b * A_{k\times k}$ where $A = \sigma_\beta^2 J_k$ and I_k, J_k are the k×k identity matrix and the k×k matrix of 1's.

If σ^2 and σ_β^2 were known (or their ratio was known) the optimal estimate (minimum variance unbiased, maximum likelihood) of the treatment means would be the generalized least squares (GLS) estimator

$$\hat{\tau} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

and (2)

$$cov(\hat{\tau}) = (X'\Sigma^{-1}X)^{-1}$$

where X is the usual design matrix for the model

$$Y_{ij} = \tau_i + \varepsilon_{ij} \quad . \tag{3}$$

Any contrast among the treatments, $c'\tau$, is estimated by $c'\hat{\tau}$ with a variance of $c'(X'\Sigma^{-1}X)^{-1}c$.

Rao (1947), Cochran and Cox (1957), Kempthorne (1952), Federer (1955) and others, following Yates (1940), recommend estimating the variance components σ^2_β and σ^2 by equating the mean square for blocks adjusted for treatments and the intrablock error mean square with their expectations. Should the estimate of σ^2_β be negative, it is generally recommended that it be set to 0. This amounts to ignoring the incomplete blocks and simply using the unadjusted means to estimate treatment effects. The variance components are then treated as though they are known exactly and used in (2) to obtain estimates with recovery of interblock information. In this regard, Cox (1956) recommends recovering interblock information only if there are at least 10 blocks and if the efficiency factor is less than .85. The effect of inaccuracies in the estimated variance components has been reported by Yates (1940) and Kempthorne (1952, p. 468). In no case considered (with as few as 8 degrees of freedom to estimate the block mean square) did the percent loss of information exceed 4.6 percent. The observed and expected mean square (MS) for Blocks and Error are equated and estimates of σ^2 and σ^2_β obtained. Provided the experiment is reasonably large, these estimates can be taken to be the exact values without serious error (Cochran and Cox, 1957, p. 399).

Newer methods, such as maximum likelihood (ML) and restricted maximum likelihood (REML) (Patterson and Thompson, 1971), exist for estimating variance components. These methods generally have more satisfactory theoretical properties than the method of moments (Harville, 1977), although the price is greater computational complexity. BMDP P3V (Dixon, 1981) provides ML or REML estimates and computes the adjusted means without special programming. Nevertheless, the procedures given in this paper are valuable when BMDP is not available or when the problem is prohibitively large for P3V.

THE TRANSFORMATION

It is well known that one can perform a GLS analysis using a simple least squares procedure by first transforming the response vector, Y, and the design matrix, X (from model (3)). The argument is outlined briefly here. Because Σ^{-1} is positive definite, it can be written $\Sigma^{-1} = \Lambda D \Lambda'$ where D is a diagonal matrix whose diagonal elements are the eigenvalues of Σ^{-1} and the columns of Λ are the corresponding eigenvectors, normalized so that $\Lambda^{-1} = \Lambda'$ (Searle, 1982, p. 290). Writing $D = D^{\frac{1}{2}} \cdot D^{\frac{1}{2}}$, (2) can be written $\hat{\tau} = (W'W)^{-1}W'Z$ where $W = D^{\frac{1}{2}}\Lambda'X$ and $Z = D^{\frac{1}{2}}\Lambda'Y$.

Because $\Sigma^{-1} = I_b * A^{-1}$ with A^{-1} positive definite, we can write $\Sigma^{-1} = I_b * G\ V\ G'$ where V is a k×k diagonal matrix whose diagonal elements are the eigenvalues of A^{-1} and the columns of G are the corresponding eigenvectors. Consequently, W and Z can be written

$$W = \begin{pmatrix} V^{\frac{1}{2}}\ G'\ X_1 \\ \vdots \\ V^{\frac{1}{2}}\ G'\ X_b \end{pmatrix} \qquad Z = \begin{pmatrix} V^{\frac{1}{2}}\ G'\ Y_1 \\ \vdots \\ V^{\frac{1}{2}}\ G'\ Y_b \end{pmatrix}$$

where Y_j is the observation vector for block j and X_j is the k×t matrix composed of the rows of the design matrix corresponding to these observations. Note that this means that the transformation can be performed one block at a time.

The eigenvalues of A^{-1} can be shown to be $(\sigma^2 + k\sigma_\beta^2)^{-1}$ and σ^{-2} (with multiplicity k-1) (Searle, 1982, p. 292). The corresponding eigenvectors are $k^{-\frac{1}{2}}(1\cdots1)'$ and any set of (k-1) orthonormal contrasts.

MISSING DATA

The approach outlined above works equally well for any incomplete block design, whether balanced, partially balanced or unbalanced. However, it does assume that there are k observations in each block. If the numbers of observations per block varies due to missing data, for example, the approach can still be used but requires the following modification. The transformation on each block now depends on the number of observations in the block. So, for example, if block j has only k-1 observations, the eigenvalues and eigenvectors are obtained in the same way but with k-1 substituted in place of k.

ILLUSTRATION USING SAS

The above methodology will now be illustrated using the Statistical Analysis System (SAS Institute, 1981). For this purpose, data from a partially balanced incomplete block design (k=4, b=15, t=15, r=4) given by Cochran and Cox (1957, p. 456) will be used.

The first step is to perform an analysis of variance, obtaining the block MS adjusted for treatments, the error MS and their expectations. The proc step, assuming the data are arranged one observation per line, is given in Table I. Rao (1947) has pointed out that the expectation of the block MS adjusted for treatments will be $\sigma^2 + \sigma^2_\beta\,(bk-t)/(b-1)$ for any blocked design with b blocks and t treatments in which the block size, k, is constant. (This excludes designs with missing plots.) If the design is resolvable into replications, the expectation of the block within-replicate MS adjusted for treatments is $\sigma^2 + \sigma^2_\beta\,(t-k)(r-1)/(b-r)$.

TABLE I

Data and Proc Steps to Obtain Estimates of σ^2 and σ^2_β

```
data whole;
input block treat y;
cards;
     (data here)
proc glm;
classes treat block;
model y = treat block/ss1 ss2;
random block;
```

One may then solve for the variance components σ^2 and σ^2_β. With these in hand, the adjusted treatment means and their covariance matrix can be obtained. The proc step is given in

TABLE II. The vector EIGENVAL is the vector of eigenvalues of A (namely σ_β^2 and σ^2 with multiplicity k-1). The columns of the matrix EIGENVEC are the corresponding eigenvectors. These are generated using the orthogonal polynomial function, ORPOL. The procedure makes the appropriate transformation of both X and Y for each block of data, then concatenates these vertically. The resulting observation vector Z and design matrix W are then output as a SAS data set to which proc glm is applied.

TABLE II

Proc Step to Obtain Estimates of Treatment Means and Their Covariance Matrix Recovering Interblock Information

```
proc matrix;
fetch y data=whole(keep=y);
fetch t data=whole(keep=treat);
x=design(t);
eigenvec=orpol(1 2 3 4);
eigenval=.2765 .0866 .0866 .0866;
delta=eigenvec*diag(eigenval##-0.5);
w=delta'*x(1:4,);
z=delta'*y(1:4,);
do i=2 to 15;
z=z//delta'*y(4*i-3:4*i,);
w=w//delta'*x(4*i-3:4*i,);
end;
output w data=x;
output z data=y(rename=(col1=y));
data whole;
merge x y;
proc glm;
model y=col1-col15/ss1 i noint;
```

With the flexibility of proc matrix, there is no need to output the data set and use proc glm. Table III gives a shorter proc step using only proc matrix to obtain the same analysis. The variance components $\hat{\sigma}^2 = .0866$ and $\hat{\sigma}_\beta^2 = .04747$ are needed to compute A^{-1} (AINV). This procedure does not actually compute W and Z but computes

$$X'\Sigma^{-1}X = \sum_{j=1}^{b} X_j' A^{-1} X_j$$

and

$$X'\Sigma^{-1}Y = \sum_{j=1}^{b} X_j' A^{-1} Y_j$$

directly. The former procedure, however, may be more suited to the software packages which may allow transformation but not extensive matrix manipulation. An added advantage of using proc matrix is that it is easy to compute contrasts among the treatments and their variances. This is illustrated in Table III by a contrast (contl) to compare the average of the first seven treatments with the average of the last eight.

TABLE III

Alternative Proc Step to Obtain Estimates of Treatment Means and Their Covariance Matrix Recovering Interblock Information

```
proc matrix;
fetch y data=whole(keep=y);
fetch t data=whole(keep=treat);
x=design(t);
ainv=(.0866*i(4)+.04747*j(4))**-1;
ww=0;
wz=0;
do i=1 to 15;
ww=ww+x(4*i-3:4*i,)'*ainv*x(4*i-3:4*i,);
wz=wz+x(4*i-3:4*i,)'*ainv*y(4*i-3:4*i,);
end;
cov=ww**-1;
means=cov*wz;
cl=(8 8 8 8 8 8 8 -7 -7 -7 -7 -7 -7 -7 -7)#/56;
contl=cl*means;
varcontl=cl*cov*cl';
print cov means contl varcontl;
```

ACKNOWLEDGEMENTS

This research was supported in part by a Natural Sciences and Engineering Research Council of Canada grant. The author gratefully acknowledges a referee who pointed out BMDP program P3V. The final revision of the paper was prepared while the author was a visiting fellow of Cornell University.

BIBLIOGRAPHY

Cochran, W. G. and Cox, G. M. (1957). *Experimental Design,* 2nd edition. New York: Wiley.

Cox, D. R. (1956). *Planning of Experiments.* New York: Wiley.

Dixon, W. J. (Ed.) (1981). *BMDP Statistical Software.* Los Angeles: University of California Press.

Federer, W. T. (1955). *Experimental Design: Theory and Application.* New York: The Macmillan Company.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* 72:320-338.

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.

Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* 58:545-554.

Rao, C. R. (1947). General methods of analysis for incomplete block designs. *J. Amer. Statist. Assoc.* 42:541-561.

SAS Institute Inc. (1982). *SAS User's Guide: Statistics, 1982 Edition.* Cary, NC: SAS Institute Inc.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* New York: Wiley.

Yates, F. (1940). The recovery of interblock information in balanced incomplete block designs. *Ann. of Eugenics* 10:317-325.