AD-A078 241 PRINCETON UNIV NJ DEPT OF STATISTICS F/6 14/1
PREDICTING UNOBSERVABLE VALUES AND ESTIMATING MISSING ONES. A C--ETC(U)
JUL 79 A M HOUTMAN N00014-79-C-0322
UNCLASSIFIED TR-156-SER-2 NL

| OF |
ADA
078241

END
DATE
FILMED
1-80
DDC

LEVEL II

PREDICTING UNOBSERVABLE VALUES AND ESTIMATING MISSING ONES. A COORDINATE-FREE APPROACH. PART 2.

by

Anne Houtman
Princeton University

Technical Report No. 156, Series 2
Department of Statistics
Princeton University
July 1979

Research supported in part by a contract with the Office of Naval Research, No. N00014-79-C-0322, awarded to the Department of Statistics, Princeton University, Princeton, New Jersey.

406873

Summary

The solution to the missing value equation in designed experiments with general covariance structure is shown to be identical to the "best" predictor of the missing data based on the observed data.

1. Introduction

Between 1952 and 1957 a number of papers appeared (Snedecor and Williams (1952,1953), Nelder (1954), Tukey (1954), Norton (1955) and Fairfield Smith (1957)) discussing the role and meaning of estimates of missing yields in designed experiments. These followed the answer by Snedecor to a query concerning a negative value obtained as replacement for a missing observation on the number of flies caught with different types of baits. It was stated that although the negative value obtained was the solution of the missing value equation, and that the usual analysis performed using the completed set of data led to the right estimates of the effects, the value was not meant to "estimate" the missing yield. The appearance of an impossible value for the replacement was to be considered more as evidence that the data did not conform to the model assumed than as a defect in the missing value procedure. In Snedecor's example, a transformation to the logarithm of the data turned out to be more appropriate.

Further light was thrown on this problem by Fairfield Smith who noted that one can indeed regard the replacement value as an estimate - either of the actual missing yield or of its expected value under the model. His main point was that the variance one ascribes to the estimate depends on what it is regarded as estimating, being larger when the yield itself is being estimated as the lost value was clearly not identical to its expectation but deviated from it by a random error whose variance must be added to that of the estimator of the mean.

The preceding discussion took place entirely in the context of uncorrelated observations, problems concerning incomplete data under models with more complex error structure (e.g. split-plots, BIBD's) having generally received little attention. Early references in this area include Anderson (1946) who gave estimates based on minimizing the subplot error sum of squares

in a split-plot experiment, and a series of papers by Cornish (1943, 1944, 1956) dealing with the recovery of interblock information with incomplete data for a variety of block designs. Contributions framed within analysis of covariance can also be found in Coons (1957) and Truitt and Fairfield Smith (1956).

In this paper we discuss missing value estimates in models with general error structure and show what is almost obvious with uncorrelated observations, that the "best" predictor of the missing yield coincides with the correct replacement value. Our analysis combines results from a recent paper (Houtman and Speed, 1979) examining missing value problems in models with general error structure with ones from Houtman (1979) concerned with best linear unbiased prediction, and the notation and terminology of this second paper (referred to as [H]) will be used in what follows.

2. Best linear unbiased prediction

The problem of prediction of one random variable based on others has received a lot of attention in the time series literature and also, but to a lesser extent, within the standard linear model. This work goes back at least to Henderson (1963), more recent references being G.S. Watson (1972), Searle (1974), Harville (1976) and [H] .

The n-dimensional space $\mathcal{D}$ of *full data* arrays may be decomposed into a direct sum of the space $\mathcal{D}_1$ of *observed* and $\mathcal{D}_2$ of *unobserved data*, and we denote by D_1 and D_2 the projections onto $\mathcal{D}_1$ and $\mathcal{D}_2$ orthogonal with respect to the inner product $\langle x,y \rangle = x^*y$. As in [H] write

$$y = D_1 y + D_2 y \equiv y_1 + y_2 \ , \ y \in \mathcal{D} \ .$$

If we suppose our full data satisfies

$$\begin{aligned} &\mathbf{E}y = \tau \in \mathcal{T} \text{ , } \mathcal{T} \text{ a subspace of } \mathcal{D} \text{ ,} \\ &\mathbf{Var}\ y = V \text{ , } V \text{ known, positive-definite,} \end{aligned} \tag{1}$$

then the observed data y_1 has

$$\begin{aligned} &\mathbf{E}y_1 = D_1\tau \equiv \tau_1 \in D_1\mathcal{T} \equiv \mathcal{T}_1 \text{ ,} \\ &\mathbf{Var}\ y_1 = D_1VD_1 \text{ ;} \end{aligned} \tag{2}$$

the unobserved data y_2 satisfies

$$\begin{aligned} &\mathbf{E}y_2 = D_2\tau \equiv \tau_2 \in D_2\mathcal{T} \equiv \mathcal{T}_2 \text{ ,} \\ &\mathbf{Var}\ y_2 = D_2VD_2 \text{ ;} \end{aligned} \tag{3}$$

and

$$\mathbf{cov}\,(y_1,y_2) = D_1VD_2 \text{ .}$$

A best linear unbiased predictor (BLUP) of y_2 based on y_1 is an array $\tilde{y}_2 = \tilde{A}y_1$ where $\tilde{A}$ is a linear transformation on $\mathcal{D}$ such that

$$\tilde{A}\tau_1 = \tau_2 \quad \forall \tau \in \mathcal{T}$$

and

$$\min_{A\tau_1=\tau_2} \mathbf{E}\|Ay_1 - y_2\|^2$$

is attained at $A = \tilde{A}$. The solution is unique whenever $\dim \mathcal{T}_1 = \dim\mathcal{T}$ - this will be assumed to be the case in the sequel - and can be written

$$\tilde{y}_2 = \tilde{\tau}_2 + B(y_1 - \tilde{\tau}_1) \qquad (4)$$

where $\tilde{\tau}_2$ is the best linear unbiased estimator (BLUE) of τ_2 based on y_1, B is the product of the covariance D_2VD_1 with an effective inverse of D_1VD_1, and $y_1 - \tilde{\tau}_1$ is the residual after fitting of the observed model. At this stage we can observe that if y_1 and y_2 are uncorrelated, then the BLUP of y_2 is identical to the BLUE of the expected value of y_2.

3. The missing value equations

It was suggested by R.A. Fisher (see Yates, 1933) that replacements for missing yields in a designed experiment can be obtained by minimizing the residual sum of squares when unknowns are substituted for them and the validity of this process was shown in Yates (1933).

Using the notations introduced for the prediction problem, let $y_1 \in \mathcal{D}_1$ denote the observed yields and $y_2 \in \mathcal{D}_2$ the missing ones, expectations and covariances continuing to be given by (1), (2) and (3). Still following the idea of Fisher, let us fit the model $\mathcal{T}$ to $y_1 + y_2^*$, where y_2^* denotes a set of parameters replacing the lost yields. The fitted value $\hat{\tau}$ is then the weighted projection of $y_1 + y_2^*$ onto $\mathcal{T}$:

$$\hat{\tau} = P_{\mathcal{T}}^{V}(y_1 + y_2^*) , \qquad (5)$$

where $P_A^{\ddagger}$ is used to denote the weighted projection onto a subspace A of $\mathcal{D}$, orthogonal with respect to the inner product $\langle x,y \rangle_{\ddagger} = x^*\ddagger^{-1}y$. The missing values estimates are then obtained by minimizing

$$\|y_1 + y_2^* - \hat{\tau}\|_V^2$$

over $\mathcal{D}_2$, where $\|\cdot\|_V^2$ is the norm associated with the inner product $\langle \cdot,\cdot \rangle_V$ described above. By least squares theory, the solution is given by

$$y_2^* = P_{\mathcal{D}_2}^{V}(\hat{\tau} - y_1) \qquad (6)$$

where $\hat{\tau}$ is given by (5). On substituting (5) into (6) we conclude that the solution y_2^* must satisfy the equation

$$y_2^* = P_{\mathcal{D}_2}^V \left[P_{\mathcal{T}}^V (y_1 + y_2^*) - y_1 \right] . \tag{7}$$

The solution to (7) is unique whenever $\mathcal{T} \cap \mathcal{D}_2 = \{0\}$, i.e. whenever $\dim \mathcal{T}_1 = \dim \mathcal{T}$ (see [H]). Equivalently $\|y_1 + y_2 - \tau\|_V^2$ can be minimized over $\mathcal{D}_2$ first and over $\mathcal{T}$ next, leading to equations (6) and (5). Now by substituting (6) into (5) we obtain an equation for the fitted values:

$$\hat{\tau} = P_{\mathcal{T}}^V \left[y_1 + P_{\mathcal{D}_2}^V (\hat{\tau} - y_1) \right] . \tag{8}$$

The following result, whose proof can be found in Houtman and Speed, shows that the fitted values obtained from the completed set of data give the correct fit for the observed model:

Theorem: The BLUE $P_{\mathcal{T}_1}^{D_1 V D_1} y_1$ of τ_1 based on y_1 coincides with $D_1 \hat{\tau}$ where $\hat{\tau}$ is a solution of (8). Equivalently, using (6)

$$P_{\mathcal{T}_1}^{D_1 V D_1} y_1 = D_1 P_{\mathcal{T}}^V (y_1 + y_2^*)$$

where y_2^* satisfies equation (7).

4. Missing value estimate and BLUP are identical

We now organize the formulae from the preceding two sections to provide proof for our main assertion, namely that with V known up to a scalar, the BLUP and the missing value estimate coincide.

Let M denote the linear operator on $\mathcal{D}_1$ such that

$$D_1(Mz) = z, \; z \in \mathcal{T}_1 \; ; \; Mu = 0 \; , \; u \in \mathcal{D}_1 \ominus \mathcal{T}_1 \; .$$

Then if $\tilde{\tau}_1 = P_{\mathcal{J}_1}^{D_1 DV_1} y_1$ is the BLUE of τ_1 based on y_1, $M\tilde{\tau}_1 = \tilde{\tau}$ is the BLUE of τ based on y_1 and $D_2 M\tilde{\tau}_1 = \tilde{\tau}_2$ is the BLUE of τ_2 based on y_1. If $(D_1VD_1)^-$ denotes an effective inverse of D_1VD_1, then $\tilde{y}_2$ has representation

$$\tilde{y}_2 = D_2 M P_{\mathcal{J}_1}^{D_1 VD_1} y_1 + (D_2 VD_1)(D_1 VD_1)^- (I - P_{\mathcal{J}_1}^{D_1 VD_1}) y_1$$

$$= \left[D_2 - (D_2 VD_1)(D_1 VD_1)^- \right] \left[MP_{\mathcal{J}_1}^{D_1 VD_1} y_1 - y_1 \right],$$

$$= P_{\mathcal{J}_2}^{V} [\tilde{\tau} - y_1] \qquad (11)$$

where $\tilde{\tau}$ is such that

$$D_1 \tilde{\tau} = \tilde{\tau}_1 .$$

Using the theorem of section 3 it follows that

$$\tilde{\tau} = \hat{\tau}$$

where $\hat{\tau}$ satisfies (8) and hence, by comparing (11) with (6), we conclude that the solution to the missing value equation is exactly the best linear unbiased predictor of the missing observations obtained from the existing ones.

This conclusion can be re-expressed as follows:
the problem of finding $\hat{y}_2 = Ay_1$ such that

$$E\|Ay_1 - y_2\|^2$$

is minimum subject to $A\tau_1 = \tau_2$, $\forall \tau \in \mathcal{J}$, is equivalent to that of finding y_2 such that

$$\|y_1 + y_2 - \tau\|_V^2$$

is minimum over all $\tau \in \mathcal{T}$ and over all $y_2 \in \mathcal{D}_2$.

We close with two remarks. Firstly it is clear that whenever the procedures just discussed are applied in practice, an estimate of V must be used. Ways of doing this are explained in Houtman and Speed (1979). And finally we point out that the interpretation of solutions of missing value equations as predictors of those values provides a strong argument for the unsuitability of the underlying model whenever unreasonable replacements arise.

Acknowledgments: Grateful acknowledgment is made to Dr G.S. Watson and Dr T.P. Speed for their help and encouragement on the subject of these papers and to the Universities of Princeton and Western Australia where this work was done.

References

Anderson, R.L. (1946). Missing-plot techniques. *Biometrics*, 2, 41-47.

Coons, I. (1957). The analysis of covariance as a missing-plot technique. *Biometrics*, 13, 387-405.

Cornish, E.A. (1943). The recovery of inter-block information in quasi-factorial designs with incomplete data. 1. Square, triple and cubic lattices. *C.S.I.R. Bull.* No. 158.

(1944). The recovery of inter-block information in quasi-factorial designs with incomplete data. 2. Lattice squares. *C.S.I.R. Bull.* No. 175.

(1956). The recovery of inter-block information in quasi-factorial designs with incomplete data. 3. Balanced incomplete blocks. *C.S.I.R.O. Divn. Math. Stats. Tech. paper*, No. 4.

Harville, D. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Ann. Stat.* 4, No. 2, 384-395.

Henderson, C.R. (1963). Selection index and expected genetic advance. *Statistical genetics and Plant Breeding*, NAS-NRC Publication No. 982, 141-163.

Houtman, A.M. (1979). Predicting unobservable values and estimating missing ones. A coordinate-free approach. Part I. Manuscript submitted for publication.

Houtman, A.M. and Speed, T.P. (1979). Missing values in multistrata designed experiments. Paper in preparation.

Nelder, J.A. (1954). A note on missing plot values. *Biometrics 10*, 400-401.

Norton, H.W. (1955). A further note on missing data. *Biometrics 11*, 100.

Searle, S.R. (1974). Prediction, mixed models, and variance components. *Reliability and Biometry* (F. Proschan and R.J. Serfling, eds.) 229-266. SIAM, Philadelphia.

Smith, H.F. (1957). Missing plot estimates. Note 125. *Biometrics 13*, 115-118.

Snedecor, G.W. and Williams, C.B. (1952 - 1953). Queries 96 and 103. *Biometrics 8*, 384 and 9, 425-427.

Truitt, J.T. and Smith, H.F. (1956). Adjustment by covariance and consequent tests of significance in split-plot experiments. *Biometrics 12*, 23-39.

Tukey, J.W. (1954). Query 111, comment on queries 96 and 103. *Biometrics 10*, 412-413.

Watson, G.S. (1972). Prediction and the efficiency of least squares. *Biometrika 59*, 91-98.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal Exp. Agriculture*, 1 No. 2, 129-142.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Tech. Report No. 156, Ser.2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* PREDICTING UNOBSERVABLE VALUES AND ESTIMATING MISSING ONES. A COORDINATE-FREE APPROACH. PART 2.		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Anne M. Houtman		8. CONTRACT OR GRANT NUMBER(s) N00014-79-C-0322
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Princeton University Princeton, New Jersey 08540		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research (Code 436) Arlington, Virginia 22217		12. REPORT DATE July 1979 13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

79 12 14 094

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The solution to the missing value equation in designed experiments with general covariance structure is shown to be identical to the "best" predictor of the missing data based on the observed data.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*