LEVEL III

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.

This document has been approved for public release and sale; its distribution is unlimited.
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

\[
y = D_1 y + D_2 y = y_1 + y_2 , \quad y \in \mathcal{D}.
\]
If we suppose our full data satisfies

\[ \mathbb{E} y = \tau \in \mathcal{J}, \mathcal{J} \text{ a subspace of } \mathcal{D}, \]

(1)

\[ \operatorname{Var} y = V, V \text{ known, positive-definite}, \]

then the observed data \( y_1 \) has

\[ \mathbb{E} y_1 = D_1 \tau = \tau_1 \in D_1 \mathcal{J} = \mathcal{J}_1, \]

(2)

\[ \operatorname{Var} y_1 = D_1 V D_1; \]

the unobserved data \( y_2 \) satisfies

\[ \mathbb{E} y_2 = D_2 \tau = \tau_2 \in D_2 \mathcal{J} = \mathcal{J}_2, \]

(3)

\[ \operatorname{Var} y_2 = D_2 V D_2; \]

and

\[ \operatorname{cov}(y_1, y_2) = D_1 V D_2. \]

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

\[ \tilde{A} \tau_1 = \tau_2, \quad \forall \tau \in \mathcal{J}, \]

and

\[ \min_{\tilde{A}} \mathbb{E} \| \tilde{A} y_1 - y_2 \|^2 \]

is attained at \( A = \tilde{A} \). The solution is unique whenever \( \dim \mathcal{J}_1 = \dim \mathcal{J} \) - this will be assumed to be the case in the sequel - and can be written
\[ \tilde{y}_2 = \tilde{\tau}_2 + B(y_1 - \tilde{\tau}_1) \]  

where \( \tilde{\tau}_2 \) is the best linear unbiased estimator (BLUE) of \( \tau_2 \) based on \( y_1 \), \( B \) is the product of the covariance \( D_2 V D_1 \) with an effective inverse of \( D_1 V D_1 \), 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 \( \mathcal{J} \) 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 \( \mathcal{J} \):

\[ \hat{\tau} = \mathbb{P}_\mathcal{J}(y_1 + y_2^*) \]  

where \( \mathbb{P}_\mathcal{J} \) is used to denote the weighted projection onto a subspace \( \mathcal{A} \) of \( \mathcal{D} \), orthogonal with respect to the inner product \( \langle x, y \rangle_\mathcal{A} = x^* y^\dagger \).

The missing values estimates are then obtained by minimizing

\[ \| y_1 + y_2^* - \hat{\tau} \|^2_\mathcal{V} \]

over \( \mathcal{D}_2 \), where \( \| \cdot \|^2_\mathcal{V} \) is the norm associated with the inner product \( \langle \cdot, \cdot \rangle_\mathcal{V} \) described above. By least squares theory, the solution is given by

\[ y_2^* = \mathbb{P}_{\mathcal{D}_2}^{\mathcal{V}} (\hat{\tau} - y_1) \]  

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

\[
y_2^* = P_2^V \left[ P_y^V (y_1 + y_2^*) - y_1 \right].
\]  

(7)

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

\[
\hat{\tau} = P_y^V \left[ y_1 + P_2^V (\hat{\tau} - y_1) \right].
\]  

(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 \( P_1^V y_1 \) of \( \tau_1 \) based on \( y_1 \) coincides with \( D_1 \hat{\tau} \) where \( \hat{\tau} \) is a solution of (8). Equivalently, using (6)

\[
P_1^V y_1 = D_1 P_1^V (y_1 + y_2^*)
\]

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

\[
D_1(Mz) = z, \ z \in \mathcal{I}_1 \ ; \ Mu = 0 \ , \ u \in \mathcal{D}_1 \in \mathcal{I}_1.
\]
Then if \( \tilde{\tau}_1 = P_{\mathcal{J}_1} y_1 \) is the BLUE of \( \tau_1 \) based on \( y_1 \), \( M_{\mathcal{J}_1} = \tilde{\tau}_1 \) is the BLUE of \( \tau \) based on \( y_1 \) and \( D_2 M_{\mathcal{J}_1} = \tilde{\tau}_2 \) is the BLUE of \( \tau_2 \) based on \( y_1 \). If \( (D_1 V D_1)^{-} \) denotes an effective inverse of \( D_1 V D_1 \), then \( \tilde{\tau}_2 \) has representation

\[
\tilde{\tau}_2 = D_2 M_{\mathcal{J}_1} y_1 + (D_2 V D_1) (D_1 V D_1)^{-} (I - P_{\mathcal{J}_1}) y_1
\]

\[
= \left[ D_2 - (D_2 V D_1)(D_1 V D_1)^{-} \right] \left[ M_{\mathcal{J}_1} y_1 - y_1 \right],
\]

\[
= P_{\mathcal{J}_2} [\tilde{\tau}_2 - y_1]
\]

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 \( \hat{y}_2 = A y_1 \) such that

\[
\mathbb{E}(\|A y_1 - y_2\|^2)
\]

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 \( \tau \in J \) 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


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