

RT-MAE 9628

**BEST LINEAR UNBIASED PREDICTOR
IN THE MIXED MODEL
WITH INCOMPLETE DATA**

by

**Lúcia P. Barroso, Wilton O. Bussab
and
Martin Knott**

Palavras-Chave: BLUP (Best Linear Unbiased Predictor), missing observation, mixed
(Key words) linear model

Classificação AMS: 62K99.
(AMS Classification)

Best Linear Unbiased Predictor in the Mixed Model with Incomplete Data

Lúcia P. Barroso

Departamento de Estatística, Universidade de São Paulo

Caixa Postal 66281, São Paulo, CEP: 05315-970, Brazil

Wilton O. Bussab

Fundação Getúlio Vargas

Av. Nove de Julho, 2029, São Paulo, CEP: 01313-902, Brazil

Martin Knott

Statistics Department, London School of Economics

Houghton Street, WC2A 2AE, London, UK

Summary

The problem of predicting individual measurement is considered. This paper develops the Best Linear Unbiased Predictor (BLUP) of the fixed and random effects and the missing observations, under a mixed linear model. The mean square errors are also obtained.

Some key words: BLUP (Best Linear Unbiased Predictor), missing observation, mixed linear model

1. Introduction

According to Little and Rubin (1987) there are basically three ways to analyse incomplete data: elimination of the units partially observed, reweighting of units and imputation. This last approach predicts the missing values and then analyses the completed data. In "hot

deck" imputation, the prediction of values is made several times without a specific purpose of predicting any parameter. The idea is to complete the set of data to use in future studies, with different objectives.

The mixed model has been intensively used. Searle (1988) has a survey of the random effects model or variance components model. Among the authors that use this model with complete data are Henderson (1975), Wang (1983) and Jeske and Harville (1988). Others, such as Chi and Reinsel (1989) and Harville and Carriquiry (1992) consider the model when data are not balanced which suits the case of non-response. The mixed model was also used by Box and Tiao (1968) with a Bayesian approach and by Lindstrom and Bates (1988) who implemented the Newton-Raphson and EM algorithms.

In this paper, we consider the mixed model with incomplete data and derive the Best Linear Unbiased Predictor (BLUP) for fixed and random effects and for the missing values.

2. The model

We consider the mixed linear model

$$y = X\alpha + Zb + e \quad (2.1)$$

where y is a vector of d observed random variables, α is a vector of p unknown fixed parameters (fixed effects), b is a vector of q unobservable random variables (random effects), X and Z are known matrices and e is a vector of d unobservable random variables (measurement errors) such that $E(b) = 0$, $E(e) = 0$ and

$$Var \begin{pmatrix} b \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \sigma^2,$$

where G and R are known positive definite matrices, with dimension $(q \times q)$ and $(d \times d)$, respectively, and σ^2 is a positive constant. We suppose that G and R are matrices of full rank.

Generally, it is assumed that, except for the parameter σ^2 , the variance-covariance structure is known. The variance-covariance matrix of y is

$$Var(y) = (R + ZGZ')\sigma^2 = W\sigma^2, \quad (2.2)$$

where

$$W = R + ZGZ'.$$

The BLUP (Best Linear Unbiased Predictor), used in this context is the usual one with M-optimality, defined as in Subramani (1991), by:

Definition 2.1 - Unbiased Predictor: the predictor $\hat{\theta}$ of θ is said to be unbiased for θ if

$$E(\hat{\theta} - \theta) = 0.$$

Definition 2.2 - Mean Square Error (MSE): the mean square error of $\hat{\theta}$, as a predictor of θ is defined as

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'].$$

Definition 2.3 - BLUP: let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two linear unbiased predictors of θ and A_1 and A_2 the respective matrices of mean square error. $\hat{\theta}_1$ is said to be better than $\hat{\theta}_2$ if $A_2 - A_1$ is a non negative definite matrix.

3. Derivation of BLUP

The BLUP can be derived in several different ways. Henderson(1950) derived the BLUP by maximizing the joint density of b and y with respect to α and b and suggested the name *joint maximum likelihood estimate*. Goldberger (1962) and Henderson(1963) applied Lagrange multipliers to obtain the BLUP of linear combinations of the effects and Harville (1977) showed that the BLUP of a linear combination of fixed and random effects is the same linear combination applied to the BLUP predictors of these effects.

Mathematically, the BLUP of α and b are the solutions of the following equations, called equations of the mixed model, given by Henderson (1950)

$$X'R^{-1}X\hat{\alpha} + X'R^{-1}Z\hat{b} = X'R^{-1}y \quad (3.1)$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} \hat{\alpha} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\hat{\mathbf{b}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} . \quad (3.2)$$

The solutions, known as solutions of the mixed model, are:

$$\hat{\alpha} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} \quad (3.3)$$

and

$$\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}[\mathbf{Z}'\mathbf{R}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}]\mathbf{y} . \quad (3.4)$$

3.1. Estimation and prediction with known variances

In this section we derive the BLUP of fixed and random effects and of the missing values, when the variance components are known. We consider the case in which m observations are missing at random and the corresponding errors of observed and missing measurements have no correlation. In the derivation, we combine the ideas of Henderson (1950), Singh and Pratap (1989) and Harville (1990).

We suppose that m of the d observations are missing and, without loss of generality, suppose that these are the last observations of the vector \mathbf{y} .

Thus, the model (2.1) can be written as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_o \\ \mathbf{y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}_o \\ \mathbf{X}_m \end{pmatrix} \alpha + \begin{pmatrix} \mathbf{Z}_o \\ \mathbf{Z}_m \end{pmatrix} \mathbf{b} + \begin{pmatrix} \mathbf{e}_o \\ \mathbf{e}_m \end{pmatrix} . \quad (3.5)$$

The BLUP of α , \mathbf{b} and \mathbf{y}_m can be obtained by minimizing the expression \mathbf{H} , where

$$\mathbf{H} = \begin{pmatrix} \mathbf{b} \\ \mathbf{y}_o - \mathbf{X}_o\alpha - \mathbf{Z}_o\mathbf{b} \\ \mathbf{y}_m - \mathbf{X}_m\alpha - \mathbf{Z}_m\mathbf{b} \end{pmatrix}' \begin{pmatrix} \mathbf{G}^{-1} & 0 & 0 \\ 0 & \mathbf{R}_o^{-1} & 0 \\ 0 & 0 & \mathbf{R}_m^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ \mathbf{y}_o - \mathbf{X}_o\alpha - \mathbf{Z}_o\mathbf{b} \\ \mathbf{y}_m - \mathbf{X}_m\alpha - \mathbf{Z}_m\mathbf{b} \end{pmatrix} . \quad (3.6)$$

These estimators are the following:

$$\begin{aligned}\hat{y}_m = & \{I_m - X_m(X'W^{-1}X)^{-1}[X'_m - X'R^{-1}ZAZ'_m]R_m^{-1} - Z_mA[Z'_m - Z'R^{-1}X(X'W^{-1}X)^{-1} \\ & \cdot (X'_m - X'R^{-1}ZAZ'_m)]R_m^{-1}\}^{-1}\{X_m(X'W^{-1}X)^{-1}[X'_o - X'R^{-1}ZAZ'_o] \\ & + Z_mA[Z'_o - Z'R^{-1}X(X'W^{-1}X)^{-1}(X'_o - X'R^{-1}ZAZ'_o)]\}R_o^{-1}y_o.\end{aligned}\quad (3.7)$$

$$\hat{\alpha} = (X'W^{-1}X)^{-1}X'W^{-1}\hat{y}.\quad (3.8)$$

and

$$\hat{b} = (Z'R^{-1}Z + G^{-1})^{-1}(Z'R^{-1} - Z'R^{-1}X(X'W^{-1}X)^{-1}X'W^{-1})\hat{y}\quad (3.9)$$

where

$$A = (Z'R^{-1}Z + G^{-1})^{-1}$$

and

$$\hat{y} = \begin{pmatrix} y_o \\ \hat{y}_m \end{pmatrix}.$$

Note that (3.8) and (3.9) are the same expressions of $\hat{\alpha}$ and \hat{b} for complete data ((3.3) and (3.4)), but applied at completed data.

3.2. Simplification of the predictors

The predictors given in previous section can be simplified using matrices of indicators of observed and missing values. Let E be the matrix of indicators of missing values and F the matrix of indicators of observed values. The dimension of E is $(d \times m)$, where each column corresponds at one of the missing data. Each column of E has $(d - 1)$ zero's and one 1, located in the row corresponding to the missing value. The matrix F has dimension $(d \times (d - m))$ and is analogous to E , but indicates the observed data. Using this notation,

$$y_o = F'y \quad y_m = E'y \quad (3.10)$$

$$X_o = F'X \quad X_m = E'X \quad (3.11)$$

$$Z_o = F'Z \quad Z_m = E'Z \quad (3.12)$$

$$R_o^{-1} = F'R^{-1}F \quad R_m^{-1} = E'R^{-1}E \quad (3.13)$$

and

$$E'E = I_m \quad F'F = I_{d-m} \quad (3.14)$$

$$E'F = 0 \quad EE' + FF' = I_d \quad (3.15)$$

$$FF'R^{-1}FF' + EE'R^{-1}EE' = R^{-1} \quad (3.16)$$

Thus,

$$\hat{y}_m = (I_m - E'QEE'R^{-1}E)^{-1}E'QFF'R^{-1}FF'y \quad (3.17)$$

where

$$Q = R[W^{-1}X(X'W^{-1}X)^{-1}X'W^{-1} + R^{-1} - W^{-1}]R \quad (3.18)$$

If the errors are independent, $R = I_d$, the expressions in (3.17) and (3.18) become

$$\hat{y}_m = (I_m - E'QE)^{-1}E'QFF'y \quad (3.19)$$

where

$$Q = W^{-1}X(X'W^{-1}X)^{-1}X'W^{-1} + I_d - W^{-1} \quad (3.20)$$

3.3. The Mean Square Error

The mean square errors of the predictors are:

$$MSE(\hat{y}_m) = [(I_m - E'QEE'R^{-1}E)^{-1}E'QFF'R^{-1}FF' - E']W \\ [FF'R^{-1}FF'QE(I_m - E'R^{-1}EE'QE)^{-1} - E]\sigma^2 \quad (3.21)$$

$$MSE(\hat{\alpha}) = BWB'\sigma^2 \quad (3.22)$$

$$MSE(\hat{b}) = (CWC' - CZG - GZ'C' + G)\sigma^2 \quad (3.23)$$

where

$$B = (X'W^{-1}X)^{-1}X'W^{-1}R[I_d + EE'R^{-1}E(I_m - E'QEE'R^{-1}E)^{-1}E'Q]FF'R^{-1}FF'$$

and

$$C = AZ'[I_d - R^{-1}X(X'W^{-1}X)^{-1}X'W^{-1}R][I_d + EE'R^{-1}E(I_m - E'QEE'R^{-1}E)^{-1}E'Q]FF'R^{-1}FF'$$

If the errors are independent,

$$MSE(\hat{y}_m) = [(I_m - E'QE)^{-1}E'QFF' - E'W[FF'QE(I_m - E'QE)^{-1} - E]\sigma^2,$$

$$B = (X'W^{-1}X)^{-1}X'W^{-1}[I_d + E(I_m - E'QE)^{-1}E'Q]FF',$$

$$C = AZ'[I_d - X(X'W^{-1}X)^{-1}X'W^{-1}][I_d + E(I_m - E'QE)^{-1}E'Q]FF'.$$

4. Unknown variances

When the variance components are unknown, the traditional method consists of estimating these parameters and using the estimates as the true values. This approach is known as empirical BLUP and it can present problems when the estimate of the ratio of variance is near to zero. According to Harville and Carriquiry (1992), "empirical BLUP is satisfactory - or can be made satisfactory by introducing appropriate modifications - unless the estimate of the variance ratio is imprecise and is close to zero, in which case more sensible point and interval predictions can be obtained by adopting a Bayesian approach".

The main methods of estimation considered in the literature are Analysis of Variance (ANOVA), Estimation by Maximum Likelihood (EML), Restricted Estimation by Maximum Likelihood (REML) and Minimum Norm Quadratic Unbiased Estimation (MINQUE). References about these methods are Winer (1971), Searle (1971), Patterson and Thompson (1971), Patterson and Thompson (1975), Rao (1970, 1971a, 1971b, 1972, 1979) and Rao and Kleffe (1988).

The estimation of variance components based on analysis of variance is one of the methods commonly suggested when the set of data is balanced. It is best used on the complete part of the data. The problem is that negative estimates may occur, in which case they are taken to be zero.

For balanced data in some cases of the model (2.1), the estimator provided by MINQUE theory, under the Euclidean norm, is the same estimator obtained from analysis of variance

and this, truncated at zero, is the same as obtained from REML, under the assumption of a normal distribution for random effects and errors. This all suggests using the estimator from analysis of variance, truncated at zero.

References

- Box, G.E.P. and Tiao, G.C. (1968). Bayesian estimation of means for the random effects model. *Journal of the American Statistical Association*, **63**, 174-181.
- Chi, E.M. and Reinsel, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, **84**, 452-459.
- Goldberger, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, **57**, 369-375.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-340.
- Harville, D.A. (1990). BLUP (Best Linear Unbiased Prediction) and beyond. *Advances in Statistical Methods for Genetic Improvement of Livestock*. D. Gianola and K. Hammond, eds. Springer, NY, 239-276.
- Harville, D.A. and Carriquiry, A.L. (1992). Classical and Bayesian prediction as applied to an unbalanced mixed linear model. *Biometrics*, **48**, 987-1003.
- Henderson, C.R. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics*, **21**, 309-310.
- Henderson, C.R. (1963). Selection index and expected genetic advance. *Statistical Genetics and Plant Breeding*. 141-163. Nat. Acad. Sci., Nat. Res. Council, Publication 982, Washington, D.C.

- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.
- Jeske, D.R. and Harville, D.A. (1988). Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models. *Communications in Statistics A*, 17, 1053-1087.
- Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014-1022.
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, NY, 278 p.
- Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554.
- Patterson, H.D. and Thompson, R. (1975). Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference*, 197-207.
- Rao, C.R. (1970). Estimation of heteroscedastic variances in a linear model. *Journal of the American Statistical Association* 65, 161-172.
- Rao, C.R. (1971a). Estimation of variance and covariance components - MINQUE Theory. *Journal of Multivariate Analysis* 1, 257-275.
- Rao, C.R. (1971b). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis*, 1, 445-456.
- Rao, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.
- Rao, C.R. (1979). MINQUE-theory and it relation to ML and MML estimation of variance components. *Sankhyā B* 41, 138-153.

- Rao, C.R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications*, North-Holland, Amsterdam, 370 p.
- Searle, S.R. (1971). *Linear Models*, John Wiley & Sons, NY, 532 p.
- Searle, S.R. (1988). Mixed models and unbalanced data: wherefrom, whereat and whereto? *Communications in Statistics A*, 17, 935-968.
- Singh, S.P. and Pratap, M. (1989). A non-iterative method for least squares estimation of missing values in any design. *Journal of the Indian Statistical Association* 27, 63-71 (*).
- Subramani, J. (1991). On invariant quadratic unbiased estimation of variance components. *Communications in Statistics A*, 20, 1705-1730.
- Wang, C.M. (1983). On the analysis of multivariate repeated measures designs. *Communications in Statistics A*, 12, 1647-1659.
- Winer, B.J. (1971). *Statistical Principles in Experimental Design*, 2ª edição, McGraw-Hill, NY, 907 p.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

- 9601 - MENTZ, R.P.; MORETTIN, P.A. and TOLOI, C.M.C. Bias correction for estimators of the residual variance in the ARMA (1,1) Model. São Paulo, IME-USP, 1996. 21p. (RT-MAE-9601)
- 9602 - SINGER, J.M.; PERES, C.A.; HARLE, C.E. Performance of Wald's test for the Hardy-Weiberg equilibrium with fixed sample sizes. São Paulo, IME-USP, 1996.15p. (RT-MAE-9602).
- 9603 - SINGER, J.M. and E. SUYAMA, E. Dispersion structure, Hierarchical models, Random effects models, Repeated measures. São Paulo, IME-USP, 1996. 21p. (RT-MAE-9603).
- 9604 - LIMA, A.C.P. and SEN, P.K. A Matrix-Valued Counting Process with First-Order Interactive Intensities. São Paulo, IME-USP, 1996, 25p. (RT-MAE-9604)
- 9605 - BOTTER, D.A. and SINGER, J.M. Experimentos com Intercâmbio de Dois Tratamentos e Dois Períodos: Estratégias para Análise e Aspectos Computacionais, São Paulo, IME-USP, 1996. 18p. (RT-MAE-9605)
- 9606- MORETTIN, P.A. From Fourier to Wavelet Analysis of Time Series. São Paulo, IME-USP, 1996. 11p. (RT-MAE-9606)
- 9607 - SINGER, J.M.; HO, L.L. Regression Models for Bivariate Counts. São Paulo, IME-USP, 1996. 20p. (RT-MAE-9607)
- 9608 - CORDEIRO, G.M. and FERRARI, S.L.P. A method of mements for finding Bartlett-type corrections. São Paulo, IME-USP, 1996. 11p. (RT-MAE-9608)
- 9609 - VILCA-LABRA, F.; ARELLANO-VALLE, R.B.; BOLFARINE, H. Elliptical functional models. São Paulo, IME-USP, 1996. 20p. (RT-MAE-9609)
- 9610 - BELITSKY, V.; FERRARI, P.A.; KONNO, M. A Refinement of Harris-FKG Inequality for Oriented Percolation. São Paulo, IME-USP, 1996. 13p. (RT-MAE-9610)

- 9611 - GIMENEZ, P.; BOLFARINE, H.** Unbiased Score Functions in Error-In-Variables Models. São Paulo, IME-USP, 1996. 23p. (RT-MAE-9611)
- 9612 - BUENO, V.C.** Comparing component redundancy allocation in K-out-of-n system. IME-USP, 1996. 10p. (RT-MAE-9612)
- 9613 - GALEA, M.; PAULA, G.A.; BOLFARINE, H.** Local influence in elliptical linear regression models. IME-USP, 1996. 14p. (RT-MAE-9613)
- 9614 - LIMA, C.R.; BOLFARINE, H.; SANDOVAL, M.C.** Linear calibration in multiplicative measurement error models. IME-USP, 1996. 12p. (RT-MAE-9614)
- 9615 - AVERBACH, M.; CUTAIT, R.; WECHSLER, S.; CORRÊA, P.; BORGES, J.L.A.** Aplicação da inferência bayesiana no estudo das probabilidades de diagnóstico, por colonoscopia, das afecções colorretais em portadores da síndrome da imunodeficiência adquirida com diarreia. IME-USP, 1996. 10p. (RT-MAE-9615)
- 9616 - YOSHIDA, O.S.; LEITE, J.G.; BOLFARINE, H.** Inferência bayesiana do número de espécies de uma população. IME-USP, 1996. 33p. (RT-MAE-9616)
- 9617 - CARMONA, S.C.; TANAKA, N.I.** Exponential estimates for "Not Very Large Deviations" and ware front propagation for a class of reaction-diffusion equations. IME-USP, 1996. 30p. (RT-MAE-9617)
- 9618 - IRONY, T.Z.; PEREIRA, C.A.B.; TIWARI, R.C.** On The Comparison Between Two Correlated Proportions in 2×2 Tables. IME-USP, 1996. 15p. (RT-MAE-9618)
- 9619 - ANDRÉ, C.D.S.; ELIAN, S.N.; NARULA, S.C.; AUBIN, E.C.** Stepwise procedures for selecting variables in the minimum sum of absolute errors regression. IME-USP, 1996. 9p. (RT-MAE-9619)
- 9620 - ANDRÉ-C.D.S.; NARULA, S.C.; PERES, C.A.P.; VENTURA.G.A.** Asymptotic properties of the MSAE estimators in the dose-response model. IME-USP, 1996. 9p. (RT-MAE-9620)
- 9621 - FERRARI, S.L.P.; CRIBARI-NETO, F.** On bootstrap and analytical bias corrections. IME-USP, 1996. 9p. (RT-MAE-9621)

9622 - NARULA, S.C.; SALDIVA, P.H.N.; ANDRÉ, C.D.S.; ELIAN, S.N.; AUBIN, E.C.Q. MSAE regression: a more robust alternative to the least squares regression. IME-USP, 1996. 10p. (RT-MAE-9622)

9623 - BRANCO, M.; BOLFARINE, H.; IGLESIAS, P. Bayesian calibration under a student-t model. IME-USP, 1996. 14p. (RT-MAE-9623)

9624 - GASCO, L.; BOLFARINE, H.; SANDOVAL, M.C. Regression estimators under multiplicative measurement error superpopulation models. IME-USP, 1996. 11p. (RT-MAE-9624)

9625 - BOTTER, D.A.; CORDEIRO, G.M. Improved estimators for generalized linear models with dispersion covariates. IME-USP, 1996. 12p. (RT-MAE-9625)

9626 - SUYAMA, E.; SINGER, J.M. Identification of random effects models for longitudinal data. IME-USP, 1996. 22p. (RT-MAE-9626)

9627 - ELIAN, S.N. Simple forms of the best linear unbiased predictor in the linear regression model. IME-USP, 1996. 6p. (RT-MAE-9627)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP

Caixa Postal 66.281
05315-970 - São Paulo, Brasil