

# Simplifying General Least Squares

Kevin HAYES and John HASLETT

We present an approach to the problem of general least squares estimation of the general linear model in terms of constrained optimization, which is in turn solved via Lagrange multipliers. We demonstrate that one system of equations is sufficiently versatile to cover not only the estimation of new observations, of fixed parameters in regression and of fixed and random effects in mixed models, but also of the diagnostics associated with conditional and marginal residuals and of subset deletion.

KEY WORDS: Best linear unbiased estimation; Best linear unbiased prediction; Conditional residuals; Cook's distance; Cross-validation residuals; DFBETA; Kriging; Mixed models; Residuals; Updating formulas.

## 1. INTRODUCTION

Robinson's (1991) review of best linear unbiased prediction (BLUP), together with the subsequent discussion, has emphasized the very considerable range of models that may be addressed via the general least squares (GLS) solution to the general linear model $Y = X\beta + \varepsilon$, where $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = V$. These include linear mixed models, geostatistics, time series and multivariate regression. The texts by Christensen (1996, 1991) and the connections to modern topics of image analysis, quality analysis, Bayesian methods, and splines (all in Robinson and discussion) make it an eminently suitable topic for teaching in any course concerning statistical linear models. Nevertheless some of the matrix algebra that results from solving the normal equations for individual specifications of the general linear model will be daunting, and far from intuitive for many students, even those who are at home in linear space.

The conventional approach to prediction and estimation from data $Y$ associated with covariates $X$ via the general linear model $Y = X\beta + \varepsilon$ is essentially a two-stage process. The first stage is to determine the best—in the GLS sense—estimator $\hat{\beta}$ of $\beta$; and subsequently to determine everything else from this.

The estimator is said to be best if it minimizes the generalization of the sum of squares $\hat{e}^t V^{-1} \hat{e}$, where $\hat{e} = Y - X\hat{\beta}$. It is straightforward to show that $\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y = BY$ and at the minimum the sum of squares is $Y^t(V^{-1} - V^{-1}X(X^tV^{-1}X)^{-1}X^tV^{-1})Y = Y^tQY$.

The purpose of this note is to give emphasis to one derivation, based on Lagrange multipliers, which leads to a system of equations that is very intuitive and lends itself readily to specialization. This approach is in fact standard in the geostatistical treatment of kriging (see Matheron 1962; Journel and Huijbregts 1981; Ripley 1981; Cressie 1993). In the genetics literature it is associated with the name of Henderson (1983); or in the classical statistical literature Hocking (1996, p. 73) is a suitable reference. The approach based on Lagrange multipliers deemphasizes the explicit determination of $\hat{\beta}$ and leads to a clearer understanding of the complementary (but for some confusing) tasks known as best linear unbiased estimation (BLUE) and best linear unbiased prediction (BLUP). Regrettably, Robinson—despite offering four derivations, and having as his main concern the interplay of BLUP and BLUE—gives it little prominence. It has recently been discussed by Searle (1997, p. 278) who said that it makes another approach (Searle, Casella, and McCulloch 1992, p. 271) seem "obtuse and unnecessarily complicated." By contrast, our treatment emphasizes the fact that it leads to a single set of equations whose solution sheds simplifying light on very many issues in general least squares. *The American Statistician*'s Teacher's Corner (e.g., McLean, Sanders, and Stroup 1991; Puntanen and Styan 1989) has already played host to previous attempts to simplify the explanation of such topics.

Various authors (Christensen, Pearson, and Johnson 1992; Haslett and Hayes 1998; Martin 1992) have visited the more specialized area of diagnostics and have developed down-dating (leave-$k$-out) formulas. The conventional approach here is via tricky identities based on the inverses of partitioned matrices. Here again the Lagrange system of equations leads to a much simplified and—we claim—much more intuitive derivation of these more technical results.

The essence of the approach is to seek that linear combination of the available data $Y$ which is best for the estimation of $Z$ among those linear estimators which are constrained to be unbiased. We adopt therefore a constrained minimization approach, using Lagrange multipliers. By best we mean that combination $\hat{Z}(Y) = \lambda_z^t Y$ which has least mean square error $E(Z - \lambda_z^t Y)^2$, and by unbiased we mean $E(Z - \lambda_z^t Y) = 0$. Here $Z$ denotes that scalar which is to be the objective of the estimation. This estimator is written as $\hat{Z}(Y)$ to make its dependence on $Y$ explicit. Note that the term "best" is applied in the context of minimizing the prediction variance $\mathrm{var}(Z - \hat{Z}(Y))$.

We shall see that $Z$ may be used to denote either a random variable or an unknown parameter, and that it will be sufficient to specify $Z$ via $E[Z]$ and $\mathrm{cov}(Z, Y)$. If $Z$ is not a random variable then of course the latter is zero and $E[Z] = Z$. We establish—very simply, as below—a general solution in terms of $A$ and $\mathrm{cov}(Z, Y)$ and achieve particular tasks by identification of these. Our presentation is for a scalar $Z$, but the notation facilitates generalization to vector $Z$. We note that Robinson (1991) stated "A convention has

Kevin Hayes is Lecturer in Statistics, University of Limerick, Limerick, Ireland (Email: kevin.hayes@ul.ie). John Haslett is Professor, Trinity College, Dublin 2, Ireland.

 

somehow developed that estimators of random effects are called *predictors* while estimators of fixed effects are called *estimators*." We agree that this distinction is confusing and indeed unnecessary.

We seek $\hat{Z}(Y) = \lambda_z^t Y$, where $\lambda_z$ is an $n \times 1$ vector of estimation coefficients. It is convenient to specify $E[Z] = A\beta$ for known $A$. In this context $A$ denotes a row vector, but we generalize this in the following. The constraint requiring $\hat{Z}(Y)$ to be unbiased now reduces to $(A - \lambda_z^t X) = 0$. A solution is found by minimizing $\text{var}(Z - \lambda_z^t Y) + \gamma_z^t (X^t \lambda_z - A^t)$, where $\gamma_z$ is a $p \times 1$ vector of Lagrange multipliers, where $p$ is the length of the parameter vector $\beta$. Setting to zero the derivatives with respect to $\lambda_z$ and $\gamma_z$ yields the system

$$\begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix} \begin{pmatrix} \lambda_z \\ \gamma_z \end{pmatrix} = \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix}. \tag{1}$$

If an inverse exists we have that

$$\begin{pmatrix} \lambda_z \\ \gamma_z \end{pmatrix} = \begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix}^{-1} \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix},$$

so that

$$\hat{Z}(Y) = (\, \lambda_z^t, \quad \gamma_z^t \,) \begin{pmatrix} Y \\ 0 \end{pmatrix}.$$

In terms of the estimation problem being considered the square matrix on the left-hand side of (1) concerns "what we have," namely, the data plus constraints. The matrix does not depend on $Z$ and consequently need only be constructed once before application to a range of problems. The right-hand side contains the term $\text{cov}(Z, Y)$ and can be specified for whatever $Z$ is being considered. It is this feature of system (1) that makes a generic approach to estimation possible.

The matrices

$$F = \begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix},$$

$$F^{-1} = \begin{pmatrix} Q & B^t \\ B & C \end{pmatrix},$$

and

$$F^{-1} \begin{pmatrix} Y \\ 0 \end{pmatrix} = \begin{pmatrix} QY \\ BY \end{pmatrix},$$

are central to all subsequent calculations. The identification of the elements $Q, B$, and $C$ with $V^{-1} - V^{-1}X(X^tV^{-1}X)^{-1}X^tV^{-1}$, $(X^tV^{-1}X)^{-1}X^tV^{-1}$, and $-(X^tV^{-1}X)^{-1}$, is well known from standard inverse matrix decompositions (Graybill 1983, p. 184; Rao and Toutenburg 1995, p. 291), but is hardly intuitive to most students. The intuitive understanding of their roles is one of our objectives. When the augmented matrix $F$ is singular a generalized inverse (see Searle et al. 1992, p. 447) facilitates a solution to (1). Note that

$$\begin{aligned} \hat{Z}(Y) &= (\text{cov}(Z, Y), A) F^{-1} \begin{pmatrix} Y \\ 0 \end{pmatrix}, \\ &= (\text{cov}(Z, Y), A) \begin{pmatrix} QY \\ BY \end{pmatrix}. \end{aligned} \tag{2}$$

The prediction variance may also be expressed in terms of the matrix $F^{-1}$ as

$$\begin{aligned} \text{var}(Z - \hat{Z}(Y)) &= \text{var}(Z) + \lambda_z^t V \lambda_z - 2\lambda_z^t \text{cov}(Y, Z), \\ &= \text{var}(Z) - \begin{pmatrix} \lambda_z \\ \gamma_z \end{pmatrix}^t \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix}, \\ &= \text{var}(Z) - \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix}^t \\ &\quad \times F^{-1} \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix}. \end{aligned} \tag{3}$$

When a number of scalars are to be estimated, we may conveniently stack them as a vector $Z$ written now as $\hat{Z}(Y) = \Lambda_z^t Y$, where $\Lambda_z$ is an $n \times k$ matrix of estimation coefficients. A solution now requires element-wise minimization of diag $[\text{var}(Z - \Lambda_z^t Y) + \Gamma_z^t (X^t \Lambda_z - A^t)]$, where $\Gamma_z$ is now a $p \times k$ matrix of Lagrange multipliers. The solutions for a vector $Z$ are of the exact same form as in (2) and (3), except $A$ now represents an $n \times k$ matrix such that $E[Z] = A\beta$.

We conclude this section with a number of very simple examples.

*Example 1.* $Z = \beta$, so that $\text{cov}(\beta, Y) = 0$ and $A = I$.

$$\Rightarrow \hat{\beta} = (\, 0, \quad I \,) \begin{pmatrix} Q & B^t \\ B & C \end{pmatrix} \begin{pmatrix} Y \\ 0 \end{pmatrix} = BY,$$

and the matrix $B$ is seen to determine $\hat{\beta}$ from $Y$. From (2) we see that

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}(\beta - \hat{\beta}), \\ &= 0 - (\, 0, \quad I \,) \begin{pmatrix} Q & B^t \\ B & C \end{pmatrix} \begin{pmatrix} 0 \\ I \end{pmatrix} = -C. \end{aligned}$$

The variance of $\hat{\beta}$ is thus determined from the lower right quadrant of $F^{-1}$.

*Example 2.* $Z = \mu = EY$, with $\text{cov}(\mu, Y) = 0$ and $A = X$. This implies $\hat{\mu} = XBY = X\hat{\beta}$, as expected.

*Example 3.* $Z = Y$, with $\text{cov}(Y, Y) = V$ and $A = X$. From (1) it is apparent (with this right-hand side) that $\Lambda_z = 1$, and $\hat{Y} = Y$—that is, the best estimate of $Y$ is $Y$ itself.

*Example 4.* $Z = Y - \mu$, such that $\hat{Z} = \hat{Y} - \hat{\mu} = Y - \hat{\mu} = \hat{e}$. Then $\text{cov}(Z, Y) = \text{var}(Y) = V$ and $A = 0$, which yields $\hat{e} = VQY$. It follows that $V^{-1}\hat{e} = QY$ and one of the roles of Q is to determine $\hat{e}$ from $Y$. Other roles emerge during our subsequent discussion of subset deletion in Section 3.

Note that regressing the residuals $\hat{e}$ (instead of $Y$) against the covariates yields new residuals $\hat{\hat{e}} = VQ(VQY) = \hat{e}$, and idempotence of $Q$ (*wrt* $V$) follows as an intuitive result.

*Example 5.* Consider $Z = Y_0$ associated with "new" covariates $X_0$ such that $EY_0 = X_0\beta = \mu_0$. We distinguish between the estimation of $\mu_0$ and $Y_0$. Classical terminology refers to the former as BLUE and the latter as BLUP.

1. From Example 2, $\hat{\mu}_0 = X_0\hat{\beta}$.
2. From Equation (2), $\hat{Y}_0 = X_0BY + \text{cov}(Y_0, Y)V^{-1}\hat{e} = \hat{\mu}_0 + \text{cov}(Y_0, Y)V^{-1}\hat{e}$.

Finally, when Y is composed of several independent processes, say $Y = Y_s + Y_m$, then it is natural to consider estimations of the type $\hat{Y}(Y), \hat{Y}_m(Y)$ and $\hat{Y}_s(Y)$. Such superimpositions arise naturally in spatial and temporal work, where $Y_t$ may denote some underlying slowly changing phenomenom and $Y_m$ may represent measurement error. Haslett and Hayes (1998) considered one such example in climatology. Christensen, Johnson, and Pearson (1992) used such a decomposition for diagnostics purposes in spatial work. In the following section we focus on mixed models.

## 2. LINEAR PREDICTION AND LINEAR MIXED MODELS

We show here that the formulation of the estimation problem presented in Section (1) is particularly useful when applied to linear mixed models. Mixed models have applications in statistical genetics, repeated measures, interlaboratory studies, and quality control, among other areas. Define

$$Y = X\beta + Wu + \eta, \tag{4}$$

where $Y$ is a vector of observable random variables, $\beta$ is a vector of unknown but fixed parameters, $u$ is a vector of random parameters, $X$ and $W$ are known design matrices, and $\eta$ is (possibly correlated) measurement error, respectively, such that $E(u) = 0, E(\eta) = 0$ and

$$\text{var}\begin{pmatrix} u \\ \eta \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix},$$

where $G$ and $R$ are known positive definite matrices. Note that Equation (4) may be expressed as $Y = X\beta + \varepsilon$, where $\varepsilon = Wu + \eta$ and $\text{var}(\varepsilon) = WGW^t + R$.

Equation (1) can be used to make various types of estimations from the model $Y = X\beta + Wu + \eta$. Eight examples are considered in Table 1, most of which have been considered by Robinson (1991) as useful aspects of the model $Y = X\beta + Wu + \eta$. Substituting the appropriate values of $A$ and $\text{cov}(Z, Y)$ into system (1) yields the desired estimation problem. The exact form of $\hat{Z}(Y)$ in each case can be obtained from (2), and the variance of the associated estimation error is available from (3).

Consider the eight variations in Table 1. The first two processes considered, $Z = \beta$, and $Z = X_0\beta$, are composed purely of *fixed effects,* the third $Z = u$, fourth $Z = W_0u$, and the eight $Z = W_0u + \eta$, are composed purely of *random effects*, and the remaining three are mixtures of both fixed and random effects. The approach taken in Section 1 does not seek to distinguish between the estimation of random and fixed effects. Both can be obtained by simply varying the right side of system (1). We believe the approach here supports a similar argument made by Robinson (1991) that the term "estimator" should be used for both fixed and random effects. Finally, the detailed algebraic expression for the solution to each of these estimation problems in terms of $X, G$, and $W$, can be easily obtained by substitution into (2). For example, Robinson (1991) gave $\hat{\beta} = [X^t(WGW^t + R)^{-1}X]^{-1}X^t(WGW^t + R)^{-1}Y$ which is hardly intuitive.

Variations of the form $\hat{Z}(X\beta + Wu)$ and $\hat{Z}(X\beta + \eta)$ for each of the eight examples considered in Table 1 may also be obtained by altering the $\text{var}(Y)$ and $\text{cov}(Z, Y)$ terms in system (1).

*Table 1. Estimation from the process $Y = X\beta + Wu + \eta$*

| | $Z$ | $A$ | $\text{cov}(Z,Y)$ |
|---|---|---|---|
| 1 | $\beta$ | $I$ | 0 |
| 2 | $X_0\beta$ | $X_0$ | 0 |
| 3 | $u$ | 0 | $GW^t$ |
| 4 | $W_0u$ | 0 | $W_0GW^t$ |
| 5 | $X_0\beta + W_0u + \eta$ | $X_0$ | $W_0GW^t + R$ |
| 6 | $X_0\beta + \eta$ | $X_0$ | $R$ |
| 7 | $X_0\beta + W_0u$ | $X_0$ | $W_0GW^t$ |
| 8 | $W_0u + \eta$ | 0 | $W_0GW^t + R$ |

## 3. SUBSET DELETION, RESIDUALS, AND DIAGNOSTIC MEASURES

Many important diagnostic measures are based on the concept of "deleting" subsets of the available data. The extensions of GLS to classical OLS diagnostics (such as deletion residuals, DFBETA and Cook's distance) has been considered by many authors (Christensen, Pearson, and Johnson 1992; Christensen, Johnson, and Pearson 1992; Martin 1992; Haslett and Hayes, 1998). Without exception the derivation of these measures involves highly technical results on the inverses of partitioned matrices. This may be almost entirely avoided by basing the derivations on (1).

We begin by partitioning $Y$ into two subsets, $Y_a$ and $Y_b$, and consider the estimation of $Z = Y_a$ using only the elements of $Y_b$. We write this as $\hat{Y}_a(Y_b) = \Lambda_{(a)}^t Y_b$. We similarly decompose $X$ and $\text{var}(Y)$. From (1) we then have

$$\begin{pmatrix} V_{bb} & X_b \\ X_b^t & 0 \end{pmatrix}\begin{pmatrix} \Lambda_{(a)} \\ \Gamma_{(a)} \end{pmatrix} = \begin{pmatrix} V_{ba} \\ X_a^t \end{pmatrix}, \tag{5}$$

from which $D_{(a)} = \text{var}(Y_a - \hat{Y}_a(Y_b))$ is available from (3) as $D_{(a)} = V_{aa} - V_{ab}\Lambda_{(a)} - X_a\Gamma_{(a)}$ (where the subscript $(a)$ denotes a quantity computed without reference to subset $a$). Equation (5) may be re-expressed as

$$\begin{pmatrix} V_{aa} & V_{ab} & X_a \\ V_{ba} & V_{bb} & X_b \\ X_a^t & X_b^t & 0 \end{pmatrix}\begin{pmatrix} I \\ -\Lambda_{(a)} \\ -\Gamma_{(a)} \end{pmatrix} = \begin{pmatrix} D_{(a)} \\ 0 \\ 0 \end{pmatrix}.$$

We may thus write

$$\begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix}\begin{pmatrix} L_{(a)} \\ -\Gamma_{(a)} \end{pmatrix} = \begin{pmatrix} D_{(a)} \\ 0 \\ 0 \end{pmatrix}, \tag{6}$$

where $L_{(a)}^t Y = Y_a - \hat{Y}_a(Y_b) = \tilde{e}_{(a)}$ is a conditional residual (Haslett and Hayes 1998) or in other terminology a deletion-prediction residual (Martin 1992), or a cross-validation residual (e.g., in the geostatistical literature; Cressie 1993, p. 101). Note once again the central role of the augmented matrix $F$. This result is the starting point of many subset-deletion results. We use the remainder of this section to

demonstrate three examples of these. Note that

$$\begin{aligned}\text{var}(\tilde{e}_{(a)}) &= \begin{pmatrix} D_{(a)} \\ 0 \end{pmatrix}^t QVQ \begin{pmatrix} D_{(a)} \\ 0 \end{pmatrix}, \\ &= \begin{pmatrix} D_{(a)} \\ 0 \end{pmatrix}^t Q \begin{pmatrix} D_{(a)} \\ 0 \end{pmatrix}, \\ &= D_{(a)} Q_{(aa)} D_{(a)}.\end{aligned}$$

As $\text{var}(\tilde{e}_{(a)}) = D_{(a)}$ it follows that $D_{(a)} = Q^{-1}_{(aa)}$ and $\tilde{e}_{(a)} = (Q^{-1}_{(aa)}, 0)QY = Y_a - Q^{-1}_{(aa)} Q_{bb} Y_b$. A similar updating formula has been derived in the context of the geostatistical method of kriging by Dudrule (1983).

*Example 6.* Decomposition of the generalized sum of squares $\hat{e}^t V^{-1} \hat{e}$.

Let $P = \{a, b, c, \ldots\}$ be a complete partition of the indexes of $Y$; that is, $Y$ is partitioned into the stacked vector $\{Y_a, Y_b, Y_c, \ldots\}$. Then if $\tilde{e}_{(a)}, \tilde{e}_{(b)}, \tilde{e}_{(c)}, \ldots$ denotes the conditional residuals associated with this partition and $\tilde{e}_{(P)}$ the (stacked) vector of these residuals, then the application of (6) to each subset in turn can be conveniently presented as

$$\begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix} \begin{pmatrix} L_{(P)} \\ -\Gamma_{(P)} \end{pmatrix} = \begin{pmatrix} D_{(P)} \\ 0 \end{pmatrix}, \tag{7}$$

where $D_{(P)}$ is block diagonal such that $D_{(a)} = Q^{-1}_{aa}, D_{(b)} = Q^{-1}_{bb}, D_{(c)} = Q^{-1}_{cc}$ and so on. It follows that

$$D^{-1}_{(P)} \tilde{e}_{(P)} = V^{-1} \hat{e} = QY. \tag{8}$$

Thus, the sum of squares $\hat{e}^t V^{-1} \hat{e}$ may be decomposed as

$$\sum_{i \in P} \hat{e}_i^t D^{-1}_{(i)} \tilde{e}_{(i)},$$

where $\hat{e}_i$ is the $i$th block of the residual vector $\hat{e}$. The basic result is that the sum of squares can be decomposed into a weighted sum of the products of two fundamentally different types of residuals. Applications of this result are discussed at length in Haslett and Hayes (1998).

If $P = \{1, 2, 3, \ldots, n\}$ partitions $Y$ into $n$ individual components, then $\Delta^{-1} \tilde{e} = V^{-1} \hat{e}$, where $\Delta^{-1} = \text{diag}(Q)$ and $\tilde{e} = (\tilde{e}_{(1)}, \tilde{e}_{(2)}, \ldots, \tilde{e}_{(n)})^t$ is the vector of leave-one-out conditional residuals. As this can be re-expressed as $\tilde{e} = \text{diag}(Q)^{-1} QY$ we see that the natural role of $Q$ is in computing, and describing the properties of, the conditional residuals.

*Example 7.* Interpretation of $\tilde{e}^t_{(a)} D^{-1}_{(a)} \tilde{e}_{(a)}$.

Note that further insight into the role of conditional residuals in the sum of squares may be gained by considering the following artificial problem. What reduction of the sum of squares can be achieved by optimally adjusting the values of $Y_a$ for fixed $Y_b$? Specifically, writing $Y^*_a = Y_a - \delta_a$, and $Y^*_b = Y_b$, what values of $\delta_a$ will maximally reduce $Y^{*t} Q Y^*$? A moment's work shows that

$$\begin{aligned}&\left(Y - \begin{pmatrix} \delta_a \\ 0 \end{pmatrix}\right)^t Q \left(Y - \begin{pmatrix} \delta_a \\ 0 \end{pmatrix}\right) \\ &\qquad = Y^t QY - 2Y^t Q \begin{pmatrix} \delta_a \\ 0 \end{pmatrix} + \begin{pmatrix} \delta_a \\ 0 \end{pmatrix}^t Q \begin{pmatrix} \delta_a \\ 0 \end{pmatrix},\end{aligned}$$

is maximally reduced when $Q_{aa} \delta_a = (QY)_a$, the $\{a\}$ block of $QY$. But from above this shows that $\delta_a = \tilde{e}_{(a)}$ (in other words that $Y_a = \tilde{Y}_a(Y_b)$) and that the reduction achieved is $\tilde{e}^t_{(a)} D^{-1}_{(a)} \tilde{e}_{(a)}$. This is referred to by Haslett and Hayes (1998) as the Mahalanobis distance of $Y_a$ from $Y_b$ *under the model.*

*Example 8.* Impact of $\hat{\beta}$ on subset deletion.

Here we consider $\hat{\beta}_{(a)} = \hat{\beta}(Y_b)$, the estimated coefficients when the subset $Y_a$ is deleted. Consider the following sequence of steps:

1. Fit a model to the data $(Y_b, X_b)$ using $V_{bb}$, the $\{bb\}$ block of $V$, and estimate $\hat{\beta}(Y_b)$.
2. Use the fitted model to predict $Y_a$ as $\tilde{Y}_a(Y_b)$.
3. Form $Y_{\tilde{a}} = (\tilde{Y}_a(Y_b), Y_b)$.
4. Fit the same model to $(Y_{\tilde{a}}, X)$ using $V$; estimate $\hat{\beta}(Y_{\tilde{a}})$.

Then $\hat{\beta}(Y_{\tilde{a}}) = \hat{\beta}(Y_b) = \hat{\beta}_{(a)}$. The formal proof of this intuitive result is given in Haslett (1999). It follows then, from Example 1, that $\hat{\beta}_{(a)} = BY_{\tilde{a}}$. The classical DFBETA may thus be written as

$$\hat{\beta} - \hat{\beta}_{(a)} = B(Y - Y_{\tilde{a}}) = B_a \tilde{e}_{(a)},$$

where $B_a$ denotes the columns of $B$ corresponding to the subset $a$. The generalization of the classical Cook's distance follows immediately as

$$\begin{aligned}\text{Cook}_{(a)} &= \frac{(\hat{\beta} - \hat{\beta}_{(a)})^t \text{var}(\hat{\beta})^{-1} (\hat{\beta} - \hat{\beta}_{(a)})}{c}, \\ &= \frac{\tilde{e}^t_{(a)} P_{aa} \tilde{e}_{(a)}}{c},\end{aligned}$$

where $\beta$ is of length $c$, and $P_{aa}$ is the $a$th block of $V^{-1} - Q$. Note that the explicit computation of $V^{-1}$ is the first time that it has been necessary to use results not already available in the augmented matrix $F$ or its inverse.

## 4. CLASSROOM EXPERIENCE

This approach has been used for some years in Trinity College, Dublin, in a senior undergraduate mathematics course of about 25 hours. The overall emphasis of the course is prediction via linear models in space and/or time. It is our experience that students have problems at two levels. These are:

(a) even with known mean parameters, the distinction between estimation and prediction, or as we prefer, the distinction between predicting a constant and a random variable; and

(b) the additional issues raised by lack of knowledge of the mean parameters.

The Lagrange multiplier approach proposed in this article as the key to (b) is not relevant to (a). We illustrate in the following a teaching example based on having available three observations from an AR(1) process contrasting the known and unknown mean cases.

*Example 9.* Data $y_1$, $y_2$, and $y_3$ are available from observations of an AR(1) process with known auto-covariance parameter $\alpha = .7$. It is desired to predict $Y_t$ as $\tilde{Y}_t = \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3$ for $t = 4$, 5, and 100 for the cases where the process mean $\mu$ is known and is unknown. We discuss the interpretation of the coefficients $\lambda_i$ in each case.

(a) Known mean $\mu$.

Writing $Z_t = Y_t - \mu$ and using predictor $\tilde{Z}_t = \lambda_1 z_1 + \lambda_2 z_2 + \lambda_3 z_3$ we see that $\tilde{Y}_t = (1 - \lambda_1 - \lambda_2 - \lambda_3)\mu + \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3$ where the $\lambda_i$'s satisfy

$$\begin{pmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \alpha^{t-1} \\ \alpha^{t-2} \\ \alpha^{t-3} \end{pmatrix}.$$

We find

| $t$ | 4 | 5 | 100 |
|---|---|---|---|
| $\lambda_1$ | .000 | .000 | .000 |
| $\lambda_2$ | .000 | .000 | .000 |
| $\lambda_3$ | .700 | .490 | .000 |

$\tilde{Y}_t$ is just a weighted average of $\mu$ and the most recent observation. For large $t$ the weights on the data converge to zero and the prediction depends exclusively on $\mu$.

In this example no reference has been made to the use of the Lagrange multiplier approach as the estimation of the mean parameter $\mu$ is not an issue. This example instead demonstrates that forecasts (BLUP's) made sufficiently beyond the range of the available data "return" towards the process mean. In this case the process being considered, namely the AR(1) model $Y_t = \mu + \alpha Y_{t-1} + \varepsilon_t$, also shows the Markov property that the "future" $(Y_t, t = 4, 5, \ldots)$ is independent of the "past" $(Y_1, Y_2)$ given the separating set $Y_3$. This is reflected by the fact that the only nonzero weight belongs to the value $y_3$, the most recent observation in the observed data $(y_1, y_2, y_3)$. It is easy to show algebraically that for an AR(1) process the weight $\lambda_3 = \alpha^{t-3}$.

(b) Unknown mean $\mu$.

The weights satisfy

$$\begin{pmatrix} 1 & \alpha & \alpha^2 & 1 \\ \alpha & 1 & \alpha & 1 \\ \alpha^2 & \alpha & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \gamma \end{pmatrix} = \begin{pmatrix} \alpha^{t-1} \\ \alpha^{t-2} \\ \alpha^{t-3} \\ 1 \end{pmatrix}.$$

We now find

| $t$ | 4 | 5 | 100 |
|---|---|---|---|
| $\lambda_1$ | .130 | .222 | .435 |
| $\lambda_2$ | .039 | .067 | .130 |
| $\lambda_3$ | .830 | .712 | .435 |

$\tilde{Y}_t$ is thus a weighted average of *all* the data, with at least for small $t$, greater weight on the most recent observation. For large $t$ the weights converge to (.435, .130, .435).

The optimal predictor (estimator) of $\mu$ is $\hat{\mu} = .435y_1 + .130y_2 + .435y_3$ and not the usual sample mean! This estimate now involves all the data but more weight is given to the endpoints. This is well understood in any traditional treatment of time series, and arises because the observations most isolated in the time domain (i.e., the end points) are heavily weighted by least squares fitting. A similar feature is observed in spatial prediction, with isolated observations or observations near the edges of the domain more heavily weighted, while the weights assigned to observations belonging to clusters tending to be lower.

It is easy to confirm that $\tilde{Y}_t = (1 - \lambda_1 - \lambda_2 - \lambda_3)\hat{\mu} + \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3$, where the weights $\lambda_i$ are as in example (a). Although all three observations are used in the estimator $\hat{\mu}$, the only nonzero weight in the "extra" component of $\tilde{Y}_t$ again belongs to $y_3$. This is a feature of the Markov property demonstrated by the AR(1) process being considered. As was the case in example (a), when extrapolating sufficiently far into the future the forecast $\tilde{Y}_t$ (i.e., BLUP) converges, except this time to the estimated mean $\hat{\mu}$ (i.e., BLUE).

## 5. CONCLUDING REMARKS

We have demonstrated how the problem of best linear unbiased estimation can be posed in terms of Lagrange multipliers. Both BLUE and BLUP can be treated as distinct estimation problems from (1). Hence BLUE and BLUP can be considered as the estimation of two different variables from $Y$.

Equation (1) has a natural role in the derivation of leave-$k$-out residuals and diagnostic measures, and replaces the traditional approach of using a variety of clumsy updating formulas. Note that this approach may be used to determine the impact of deletion on any quantity computed from $Y$.

*[Received February 1998. Revised March 1999.]*

## REFERENCES


Christensen, R. (1991), *Linear Models for Multivariate, Time Series, and Spatial Data*, New York: Springer.

——— (1996), *Plane Answers to Complex Questions: The Theory of Linear Models* (2nd Ed.), New York: Springer.

Christensen, R., Johnson, W., and Pearson, L. M. (1992), "Prediction Diagnostics for Spatial Linear Models," *Biometrica*, 79, 583–591.

Christensen, R., Pearson, L. M., and Johnson, W. (1992), "Case-Deletion Diagnostics for Mixed Models," *Technometrics*, 34, 38–45.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.

Dubrule, O. (1983), "Cross Validation of Kriging in a Unique Neighborhood," *Mathematical Geology*, 15, 687–699.

Graybill, F. A. (1983), *Matrices With Applications in Statistics*, Belmont, CA: Wadsworth.

Journel, A. G., and Huijbregts, C. J. (1981), *Mining Geostatistics*, New York: Academic Press.

Haslett, J. (1999), "A Simple Derivation of Deletion Diagnostic Results for the General Linear Model with Correlated Errors," *Journal of the Royal Statistical Society*, Series B, 61, 603–609.

Haslett, J., and Hayes, K. (1998), "Residuals for the Linear Model with General Covariance Structure," *Journal of the Royal Statistical Society*, Series B, 60, 201–215.

Henderson, J. R. (1983), *Applications of Linear Models in Animal Breeding*, Guelph, Ontario: University of Guelph.

Hocking, R. R. (1996), *Methods and Applications of Linear Models*, New York: Wiley.

McClean, R. A., Sanders, W. L., and Stroup, W. W. (1991), "A Unified Approach to Mixed Linear Models," *The American Statistician*, 45, 54–64.

Martin, R. J. (1992), "Leverage, Influence and Residuals in Regression Models When Observations Are Correlated," *Communications in Statistics—Theory and Methods*, 21, 1183–1212.

Matheron, G. (1962), *Les Variables Regionalisés et Leur Estimation*, Paris: Mason.

Puntanen, S., and Styan, G. P. H. (1989), "The Equality of the Ordinary Least Squares Estimator and the best Linear Unbiased Estimator," *The American Statistician*, 43, 153–161.

Rao, C. R. (1971), "Unified Theory of Linear Estimation," *SANKHYA: The Indian Journal of Statistics*, Series A, 33, 371–394.

Rao, C. R., and Toutenburg, H. (1995), *Linear Models: Least Squares and Alternatives*, New York: Springer.

Ripley, B. D. (1981), *Spatial Statistics*, New York: Wiley.

Robinson, G. K. (1991), "That BLUP Is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, 15–51.

Searle, S. R. (1997), "The Matrix Handling of BLUE and BLUP in the Mixed Linear Model," *Linear Algebra and its Applications*, 264, 291–311.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley.