

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.

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 $\text{var}(\varepsilon) = V$. These include linear mixed models, geo-statistics, 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 β 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 $Y^t (V^{-1} - V^{-1} (X^t V^{-1} X)^{-1} X^t V^{-1}) Y = Y^t QY$

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 (CPJ, HaslettHayes ,Martin) 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 $var(Z - Z(Y))$. 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 $cov(Z, Y)$. 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 $cov(Z, Y)$ and achieve particular tasks by identification of these. Our presentation is for a scalar Z , but the notation facilitates generalization to vector Z .