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 $var(\varepsilon) = V$. 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 $\hat{\beta} = (X^t V^{-l} X)^{-l} X^t V^{-l} Y = BY$ and at the minimum the sum of squares is $Y^t(V^{-l} - V^{-l}(X^t V^{-l} X)^{-l} X^t V^{-l})Y = Y^t QY$

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.

We note that Robinson (1991) stated "A convention has 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* λ_z^t, *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* $var(Z - \lambda_z^t Y) + \gamma_z^t (X^t \lambda_z - A^t)$, *where* γ_z *is a* $p \times 1$ *vector of Lagrange multipliers, where* p *is the length of the parameter vector* β. *Setting to zero the derivatives with respect to* λ_z *and* γ_z *yields the system*

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

If the 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

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

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 cov(Z, Y) and can be speci- fied for whatever Z is being considered. It is this feature of system (1) that makes a generic approach to estimation possible.

Generalized linear models

1 Generalized Linear model

In statistics, the generalized linear model (GzLM) is a flexible generalization of ordinary least squares regression. The GzLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Mixed Effects Models offer a flexible framework by which to model the sources of variation and correlation that arise from grouped data. This grouping can arise when data collection is undertaken in a hierarchical manner, when a number of observations are taken on the same observational unit over time, or when observational units are in some other way related, violating assumptions of independence.

2 Generalized Model(GzLM)

Nelder and Wedderburn (1972) integrated the previously disparate and separate approaches to models for non-normal cases in a framework called "generalized linear models." The key elements of their approach is to describe any given model in terms of it's link function and it's variance function.

2.1 What is a GzLM

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{1}$$

where $E(Y)$ is the expected value of Y, $X\beta$ is the linear predictor, a linear combination of unknown parameters,β and g is the link function.

$\mathrm{Var}(\mathbf{Y}) = \mathrm{V}(\boldsymbol{\mu}) = \mathrm{V}(g^{-1}(\mathbf{X}\boldsymbol{\beta}))$

2.2 GzLM Structure

The GzLM consists of three elements.
1. A probability distribution from the exponential family.
2. A linear predictor $\eta = X\beta$.
3. A link function g such that $E(Y) = \mu = g^{-1}(eta)$.

2.3 Link Function

Definition 1 : The link function provides the relationship between the linear predictor and the mean of the distribution function. There are many commonly used link functions, and their choice can be somewhat arbitrary. It can be convenient to match the domain of the link function to the range of the distribution function's mean.
Definition 2 : A link function is the function that links the linear model specified in the design matrix, where columns represent the beta parameters and rows the real parameters.

2.4 Canonical parameter

θ, called the dispersion parameter,

2.5 Dispersion parameter

τ, called the dispersion parameter, typically is known and is usually related to the variance of the distribution.

2.6 Iteratively weighted least square

IWLS is used to find the maximum likelihood estimates of a generalized linear model. Definition: An iterative algorithm for fitting a linear model in the case where the data may contain outliers that would distort the parameter estimates if other estimation procedures were used. The procedure uses weighted least squares, the influence of an outlier being reduced by giving that observation a small weight. The weights chosen in one iteration are related to the magnitudes of the residuals in the previous iteration with a large residual earning a small weight.

2.7 Residual Components

In GzLMS the deviance is the sum of the deviance components

$$D = \sum d_i \tag{2}$$

In GzLMS the deviance is the sum of the deviance components

3 Generalized linear mixed models

[pawitan section 17.8]

The Generalized linear mixed model (GLMM) extend classical mixed models to non-normal outcome data.

In statistics, a generalized linear mixed model (GLMM) is a particular type of mixed model. It is an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects. These random effects are usually assumed to have a normal distribution.

Fitting such models by maximum likelihood involves integrating over these random effects.

4 Assessment of Agreements in Linear and Generalized Linear Mixed Models

- Study of measuring agreement is intend to evaluate whether the readings from one rater/ measurement agree with those from other raters/measurements. In this dissertation, we are going to present a general method to assess agreement for a large variety of data with repeated measurements using linear and generalized linear mixed models.

- In the first place, a set of agreement statistics, including mean square deviation, concordance correlation coefficient, precision and accuracy coefficients, is presented for evaluating the intra-, inter-, and total-rater agreement in the multiple-rater and multiple-replications cases.

- Secondly, likelihood-based approaches are developed to estimate all the agreement statistics. Asymptotic properties of these estimates are also discussed for different data structures.

- Furthermore, our method has the merit of handling missing values and covariates naturally, and a new set of restricted agreement statistics is proposed in order to capture the true random variations and between-instrument effects adjusted for the covariate effects.

- Simulations for both linear and generalized linear mixed models are conducted to show the accuracy and effectiveness of our approaches. In the end, two industry datasets are evaluated using our approach.

- One is the cardiac function measurements used to determine the agreement between impedance cardiography and radionuclide ventriculography estimates, and the other one is an antihypertensive patch dataset given by FDA for assessing individual bioequivalence.

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements.

Maximum likelihood estimation is used to estimate the parameters. The REML estimation is not considered since it does not lead to a joint distribution of the estimates of fixed effects and random effects parameters, upon which the assessment of agreement is based.

General Linear model

5 Haslett Dillane Hayes

Haslett and Dillane (2004) offers an procedure to assess the influences for the variance components within the linear model, complementing the existing methods for the fixed components. The essential problem is that there is no useful updating procedures for $\hat{V}$, or for $\hat{V}^{-1}$.

Haslett and Dillane (2004) remark that linear mixed effects models didn't experience a corresponding growth in the use of deletion diagnostics, adding that McCullough and Searle (2001) makes no mention of diagnostics whatsoever.

Haslett and Dillane (2004) propose an alternative, and computationally inexpensive approach, making use of the '`delete=replace`' identity.

Haslett (1999) considers the effect of 'leave k out' calculations on the parameters β and σ^2, using several key results from Haslett and Hayes (1998) on partioned matrices.

6 General Linear model

Mixed Effects Models are seen as especially robust in the analysis of unbalanced data when compared to similar analyses done under the General Linear Model framework (Pinheiro and Bates, 2000).

A Mixed Effects Model is an extension of the General Linear Model that can specify additional random effects terms

6.1 Equivalence of LME model

Henderson's mixed model equations are presented on page 147 of Youngjo et al. Youngjo et al demonstrate that this formulation is equivalent to an augmented general linear model.

Youngjo et al show that the linear mixed effects model can be shown to be the augmented classical linear model involving fixed effects parameters only.

7 The LME model as a general linear model

Henderson's equations in (**??**) can be rewritten $(T'W^{-1}T)\delta = T'W^{-1}y_a$ using

$$\delta = \begin{pmatrix} \beta \\ b \end{pmatrix},\ y_a = \begin{pmatrix} y \\ \psi \end{pmatrix},\ T = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix},\ \text{and } W = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix},$$

where Lee et al. (2006) describe $\psi = 0$ as quasi-data with mean $\mathrm{E}(\psi) = b$. Their formulation suggests that the joint estimation of the coefficients β and b of the linear mixed effects model can be derived via a classical augmented general linear model $y_a = T\delta + \varepsilon$ where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = W$, with *both* β and b appearing as fixed parameters. The usefulness of this reformulation of an LME as a general linear model will be revisited.

8 Simplifying GLS (K Hayes)

8.1 Introduction

Hayes and Haslett (1998) 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. The crux of the proposed approach is that one system of equations is sufficiently versatile, and provides for

- the estimation of new observations,
- estimation of fixed parameters in regression
- estimation of fixed and random effects in mixed models,
- the diagnostics associated with conditional and marginal residuals
- and of subset deletion.

8.2 Overview

Hayes and Haslett (1998) 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 the following equation.

$$\begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_z \\ \boldsymbol{\gamma}_z \end{pmatrix} = \begin{pmatrix} \mathrm{cov}(Y, Z) \\ A^t \end{pmatrix} \tag{3}$$

Hence BLUE and BLUP can be considered as the estimation of two different variables from Y. This equation 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.

9 Generalized Least Squares

generalized least squares (GLS) is a technique for estimating the unknown parameters in a linear regression model. The GLS is applied when the variances of the observations are unequal (heteroscedasticity), or when there is a certain degree of correlation between the observations. In these cases ordinary least squares can be statistically inefficient, or even give misleading inferences.

$$Y = X\beta + \varepsilon, \qquad \mathrm{E}[\varepsilon|X] = 0,\ \mathrm{Var}[\varepsilon|X] = \Omega.$$

9.1 Introduction to Generalized Least Squares

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{4}$$

Estimation under this model has been studied extensively in the linear regression model.

10 Hierarchical likelihood

Inferential method was developed for the mixed linear model via Lee and Nelder's (1996) hierarchical-likelihood (h-likelihood).

11 Importance-Weighted Least-Squares (IWLS)

11.1 Introduction

11.1.1 Robinson's (1991) review

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* $var(\varepsilon) = V$*. 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 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 $\hat{\beta} = (X^t V^{-l} X)^{-l} X^t V^{-l} Y = BY$ and at the minimum the sum of squares is $Y^t(V^{-l} - V^{-l}(X^t V^{-l} X)^{-l} X^t V^{-l})Y = Y^t QY$.

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, Haslett Hayes ,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.

11.2 Predictors and Estimators

We note that Robinson (1991) stated "A convention has 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 λ_z^t, 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 $var(Z - \lambda_z^t Y) + \gamma_z^t(X^t\lambda_z - A^t)$, where γ_z is a $p \times 1$ vector of Lagrange multipliers, where p is the length of the parameter vector β. Setting to zero the derivatives with respect to λ_z and γ_z yields the system.

$$\begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix} \begin{pmatrix} \lambda_z \\ \gamma_z \end{pmatrix} = \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix} \tag{5}$$

If the inverse exists we have that

$$\begin{pmatrix} \lambda_z \\ \gamma_z \end{pmatrix} = \begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix}^{-1} \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix} \tag{6}$$

so that

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

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 $cov(Z, Y)$ 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.

11.3 Two Options

- Wald Type CIs
- PL Type CIs

11.4 Profile Likelihood Confidence Intervals

The Profile-likelihood based confidence intervals methods is described in Venzon and Moolgavkar, Journal of the Royal Statistical Society, Series C vol 37, no.1, 1988, pp. 87-94.

Profile likelihood confidence intervals can be computed for real parameter estimates.

The default confidence intervals for real parameter estimates in the 0-1 interval are based on the standard error and the logit transformation. That is, a 95% confidence interval is computed on the logit estimate, and then these intervals are transformed to the real scale.

12 Simplifying GLS

It is straightforward to show that $\hat{\beta} = (X^t V^{-l} X)^{-l} X^t V^{-l} Y = BY$ and at the minimum the sum of squares is $Y^t(V^{-l} - V^{-l}(X^t V^{-l} X)^{-l} X^t V^{-l})Y = Y^t QY$.

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, Haslett Hayes ,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 $\mathrm{cov}(Z, Y)$ and achieve particular tasks by

identification of these. Our presentation is for a scalar Z, but the notation facilitates generalization to vector Z.

Haslett and Hayes - Residuals

Haslett and Hayes (1998) and Haslett (1999) considered the case of an LME model with correlated covariance structure.

13 Haslett's Analysis

For fixed effect linear models with correlated error structure Haslett (1999) showed that the effects on the fixed effects estimate of deleting each observation in turn could be cheaply computed from the fixed effects model predicted residuals.

14 Computation and Notation

with $\boldsymbol{V}$ unknown, a standard practice for estimating $\boldsymbol{X\beta}$ is the estime the variance components σ_j^2, compute an estimate for $\boldsymbol{V}$ and then compute the projector matrix A, $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{AY}$.

Zewotir and Galpin (2005) remarks that $\boldsymbol{D}$ is a block diagonal with the $i-$th block being $u\boldsymbol{I}$

15 Haslett Hayes

For fixed effect linear models with correlated error structure Haslett (1999) showed that the effects on the fixed effects estimate of deleting each observation in turn could be cheaply computed from the fixed effects model predicted residuals.

A general theory is presented for residuals from the general linear model with correlated errors. It is demonstrated that there are two fundamental types of residual associated with this model, referred to here as the marginal and the conditional residual. These measure respectively the distance to the global aspects of the model as represented by the expected value and the local aspects as represented by the conditional expected value. These residuals may be multivariate.

In contrast to classical linear models, diagnostics for LME are difficult to perform and interpret, because of the increased complexity of the model

16 Haslett's Analysis

For fixed effect linear models with correlated error structure Haslett (1999) showed that the effects on the fixed effects estimate of deleting each observation in turn could be cheaply computed from the fixed effects model predicted residuals.

A general theory is presented for residuals from the general linear model with correlated errors. It is demonstrated that there are two fundamental types of residual

associated with this model, referred to here as the marginal and the conditional residual.

These measure respectively the distance to the global aspects of the model as represented by the expected value and the local aspects as represented by the conditional expected value.

These residuals may be multivariate.

Haslett and Hayes (1998) developes some important dualities which have simple implications for diagnostics.

Haslett & Dillane (199X) offers an procedure to assess the influences for the variance components within the linear model, complementing the existing methods for the fixed components.

The essential problem is that there is no useful updating procedures for $\hat{V}$, or for $\hat{V}^{-1}$. Haslett & Dillane (199X) propose an alternative , and computationally inexpensive approach, making use of the '`delete=replace`' identity.

Haslett (1999) considers the effect of 'leave k out' calculations on the parameters β and σ^2, using several key results from Haslett and Hayes (1998) on partioned matrices.

Haslett & Dillane (19XX) remark that linear mixed effects models didn't experience a corresponding growth in the use of deletion diagnostics, adding that McCullough and Searle (2001) makes no mention of diagnostics whatsoever.

17 Haslett's Analysis

For fixed effect linear models with correlated error structure Haslett (1999) showed that the effects on the fixed effects estimate of deleting each observation in turn could be cheaply computed from the fixed effects model predicted residuals.

A general theory is presented for residuals from the general linear model with correlated errors. It is demonstrated that there are two fundamental types of residual associated with this model, referred to here as the marginal and the conditional residual.

These measure respectively the distance to the global aspects of the model as represented by the expected value and the local aspects as represented by the conditional expected value.

These residuals may be multivariate.

Haslett and Hayes (1998) developes some important dualities which have simple implications for diagnostics.

Augmented GLMs

Generalized linear models are a generalization of classical linear models.

18 Augmented GLMs

With the use of h-likihood, a random effected model of the form can be viewed as an 'augmented GLM' with the response varaibkes $(y^t, \phi_m^t)^t$, (with $\mu = E(y)$,$u = E(\phi)$, $var(y) = \theta V(\mu)$. The augmented linear predictor is

$$\eta_{ma} = (\eta^t, \eta_m^t)^t) = T\omega.$$

.

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{7}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{8}$$

$y_a = T\delta + e^*$
Weighted least squares equation

18.1 The Augmented Model Matrix

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{9}$$

19 Augmented GLMs

With the use of h-likihood, a random effected model of the form can be viewed as an 'augmented GLM' with the response varaibkes $(y^t, \phi_m^t)^t$, (with $\mu = E(y)$,$u = E(\phi)$, $var(y) = \theta V(\mu)$. The augmented linear predictor is

$$\eta_{ma} = (\eta^t, \eta_m^t)^t) = T\omega.$$

.

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{10}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{11}$$

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{12}$$

$$y_a = T\delta + e^* \tag{13}$$

Weighted least squares equation
$y_a = T\delta + e^*$
Weighted least squares equation
Generalized linear models are a generalization of classical linear models.
The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{14}$$

20 Augmented GLMs

Generalized linear models are a generalization of classical linear models.

With the use of h-likihood, a random effected model of the form can be viewed as an 'augmented GLM' with the response varaibkes $(y^t, \phi_m^t)^t$, (with $\mu = E(y)$,$u = E(\phi)$, $var(y) = \theta V(\mu)$. The augmented linear predictor is

$$\eta_{ma} = (\eta^t, \eta_m^t)^t) = T\omega.$$

.

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{15}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{16}$$

Weighted least squares equation

21 Augmented GLMs

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{17}$$

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{18}$$

$$y_a = T\delta + e^* \tag{19}$$

Weighted least squares equation

Generalized linear models are a generalization of classical linear models.

22 Generalized Least Squares

22.1 Introduction to Generalized Least Squares

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{20}$$

Estimation under this model has been studied extensively in the linear regression model.

23 Augmented GLMs

Generalized linear models are a generalization of classical linear models.

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{21}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{22}$$

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{23}$$

$$y_a = T\delta + e^* \tag{24}$$

Weighted least squares equation

24 Augmented GLMs

With the use of h-likihood, a random effected model of the form can be viewed as an 'augmented GLM' with the response varaibkes $(y^t, \phi_m^t)^t$, (with $\mu = E(y)$,$u = E(\phi)$, $var(y) = \theta V(\mu)$. The augmented linear predictor is

$$\eta_{ma} = (\eta^t, \eta_m^t)^t) = T\omega.$$

.

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{25}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{26}$$

$y_a = T\delta + e^*$
Weighted least squares equation

24.1 The Augmented Model Matrix

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{27}$$

24.2 Importance-Weighted Least-Squares (IWLS)

24.3 H-Likelihood

Generalized linear models are a generalization of classical linear models.

25 Algorithms : ML v REML

Maximum likelihood estimation is a method of obtaining estimates of unknown parameters by optimizing a likelihood function. The ML parameter estimates are the values of the argument that maximise the likelihood function, i.e. the estimates that make the observed values of the dependent variable most likely, given the distributional assumptions

The most common iterative algorithms used for the optimization problem in the context of LMEs are the EM algoritm, fisher scoring algorithm and NR algorithm, which [cite:West] commends as the preferred method.

A mixed model is an extension of the general linear models that can specify additional random effects terms.

Parameter of the mixed model can be estimated using either ML or REML, while the AIC and the BIC can be used as measures of "goodness of fit" for particular models, where smaller values are considered preferable.

(***Wikipedia***)The restricted (or residual, or reduced) maximum likelihood (REML) approach is a particular form of maximum likelihood estimation which does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function calculated from a transformed set of data, so that nuisance parameters have no effect.

In contrast to the earlier maximum likelihood estimation, REML can produce unbiased estimates of variance and covariance parameters.

ML procedures for LME

The maximum likelihood procedure of Hartley and Rao yields simultaneous estimates for both the fixed effects and the random effect, by maximising the likelihood of $\boldsymbol{y}$ with respect to each element of $\boldsymbol{\beta}$ and $\boldsymbol{b}$.

26 Estimation of random effects

Estimation of random effects for LME models in the NLME package is accomplished through use of both EM (Expectation-Maximization) algorithms and Newton-Raphson algorithms.

- EM iterations bring estimates of the parameters into the region of the optimum very quickly, but convergence to the optimum is slow when near the optimum.
- Newton-Raphson iterations are computationally intensive and can be unstable when far from the optimum. However, close to the optimum they converge quickly.
- The LME function implements a hybrid approach, using 25 EM iterations to quickly get near the optimum, then switching to Newton-Raphson iterations to quickly converge to the optimum.
- If convergence problems occur, the "controlargument in LME can be used to change the way the model arrives at the optimum.

27 Covariance Parameters

The unknown variance elements are referred to as the covariance parameters and collected in the vector θ.

27.1 Methods and Measures

The key to making deletion diagnostics useable is the development of efficient computational formulas, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model.

Zewotir and Galpin (2005) lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,
- the variance (information) ration,
- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

References

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 D. Dillane (2004). Application of 'delete = replace' to deletion diagnostics for variance component estimation. *Journal of the Royal Statistical Society (Series B) 66*, 131–143.

Haslett, J. and K. Hayes (1998). Residuals for the linear model with general covariance structure. *Journal of the Royal Statistical Society (Series B) 60*, 201–215.

Lee, Y., J. A. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall Ltd.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

McCullough, C. and S. Searle (2001). *Generalized , Linear and Mixed Models*. Wiley Interscience.

Zewotir, T. and J. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of Data Science 3*, 153–177.