

Analysis of Unbalanced Mixed Model Data: A Case Study Comparison of ANOVA versus REML/GLS

Ramon C. LITTELL

Major transition has occurred in recent years in statistical methods for analysis of linear mixed model data from analysis of variance (ANOVA) to likelihood-based methods. Prior to the early 1990s, most applications used some version of analysis of variance because computer software was either not available or not easy to use for likelihood-based methods. ANOVA is based on ordinary least squares computations, with adoptions for mixed models. Computer programs for such methodology were plagued with technical problems of estimability, weighting, and handling missing data. Likelihood-based methods mainly use a combination of residual maximum likelihood (REML) estimation of covariance parameters and generalized least squares (GLS) estimation of mean parameters. Software for REML/GLS methods became readily available early in the 1990s, but the methodology still is not universally embraced. Although many of the computational inadequacies have been overcome, conceptual problems remain. Also, technical problems with REML/GLS have emerged, such as the need for adjustments for effects due to estimating covariance parameters. This article attempts to identify the major problems with ANOVA, describe the problems which remain with REML/GLS, and discuss new problems with REML/GLS.

Key Words: Likelihood; Linear models; Random effects.

1. INTRODUCTION

Linear mixed models provide the theoretical basis for a large segment of statistical data analysis. This is one of the most developed and frequently researched topics in statistics. The concept of linear models has greatly expanded in recent decades, so that one must carefully specify the particular reference arena. The arena for this article is linear mixed models with normally distributed errors, referring to models in which the expectation of the response variable is linear in a set of parameters, and the errors are normally distributed

Ramon C. Littell is Professor, Department of Statistics, Institute of Food and Agricultural Sciences, University of Florida, Box 110339, Gainesville, FL 32611 (E-mail: littell@stat.ulf.edu).

©2002 American Statistical Association and the International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics, Volume 7, Number 4, Pages 472–490
DOI: 10.1198/108571102816

and correlated. Correlated errors occur, for example, when they can be expressed as sums of random effects, making the response a sum of fixed and random effects; hence, the expression “mixed” model. Correlated errors are often the result of multiple measurements made on the same sampling or experimental unit.

The past two decades have seen great advances surrounding the use of linear mixed models in data analysis. Prior to that time, most research in linear mixed models was on variance component estimation. Attention to variance components came from disciplines such as animal breeding and industrial quality control, where the focus is on ascription and quantification of variation in data. Variance component estimation is still an important and active area of research, but in typical linear model applications, the inference is on fixed effects, that is, on the parameters defining the mean structure. Most of the attention to variance components and covariance structure in recent years has been on their use in making appropriate inference about fixed effects. Thus, covariance parameters frequently assume the role of “nuisance” parameters.

Developments in linear mixed models in the past two decades resulted in large part from research in agricultural and health science applications, combined with advances in computing hardware and software. The primary purpose of this article is to review and summarize some of the methodological advances that affect everyday data analysis using linear mixed models. These relate to likelihood- and generalized least squares-based inference implemented in computer programs. A secondary objective is to describe analysis of variance (ANOVA) methods as they relate to the contemporary statistical practice. Some of the problems and inadequacies of ANOVA have been overcome by the newer methods, but others remain. In addition, certain new problems occur with likelihood methods that do not occur in ANOVA. The difficult problems in linear models usually occur in the context of “unbalanced data.” The term “unbalanced data” is not precisely defined in statistics, but usually refers to multi-factor data classifications that have different numbers of observations in the factor combination categories. It should be noted, though, that “balanced incomplete block designs” are not “balanced” in the sense of the previous sentence. The essence of unbalanced data is that measures of effects of one factor depend upon other factors in the model.

A common example of a situation using mixed linear models with unbalanced data is a multi-center trial. In this article, such an example is used to illustrate data analysis methods. Two drugs were compared in a multi-center study at 53 research centers. At each center, drugs were assigned to multiple subjects. Numbers of subjects assigned to the drugs varied from 0 to 14. Nine centers had subjects on only one of the drugs. One of the objectives was to assess side effects of the drugs. At triweekly intervals following treatment, subjects returned to the clinics for measurement of several critical signs, including sitting heart rate. Some subjects did not return on all scheduled occasions, thus creating an additional aspect of unbalanced data.

Although this example involved human subjects, the same statistical issues occur in numerous areas of research. For instance, in the Animal Sciences Department at the University of Florida, scientists measured the expression of a protein that they suspected to be related to cancer prevalence. They used female and male piglets from several litters, and

extracted tissues from four areas of the gut from each piglet. The objective was to compare expression between sex and tissue types. Thus, in relation to the multi-center trial, litters are analogous to centers, sex is analogous to treatments, piglets are analogous to subjects, and tissues are analogous to time intervals.

The linear mixed model is described in Section 2, and illustrated in the multi-center trial example. Likelihood-based and ANOVA methods are described using the multi-center trial data. In Sections 3 and 4 issues are discussed that have come to the forefront of statistical data analysis only in recent years, and are not commonly understood.

2. THE LINEAR MIXED MODEL

The linear mixed model (LMM) considered in this article has equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \mathbf{e}, \quad (2.1)$$

where $\mathbf{Y}$ is the $n \times 1$ observation vector, $\mathbf{X}$ is a matrix of covariates with fixed coefficients and $\mathbf{Z}$ is a matrix of covariates with random coefficients, $\boldsymbol{\beta}$ is the vector of fixed coefficients, and $\mathbf{U}$ is the vector of random coefficients. The random vector $\mathbf{U}$ has multivariate normal (MVN) distribution with expectation $\mathbf{E}(\mathbf{U}) = 0$ and variance $\mathbf{V}(\mathbf{U}) = \mathbf{G}$, and the vector of random errors $\mathbf{e}$ has MVN distribution with $\mathbf{E}(\mathbf{e}) = 0$ and $\mathbf{V}(\mathbf{e}) = \mathbf{R}$. In addition, $\mathbf{U}$ and $\mathbf{e}$ are independent. This is the model presented by Henderson (1975), and studied extensively by him and his students and colleagues in genetics applications. It is now the standard form of the mixed model referenced in most statistical research and implemented in computer programs. From model (2.1) it follows that

$$\mathbf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathbf{V}(\mathbf{Y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}. \quad (2.2)$$

The term $\mathbf{Z}\mathbf{U}$ in (2.1) does not add to the generality of the model. It could be absorbed into $\mathbf{e}$ because, theoretically, $\mathbf{R}$ can accommodate any possible covariance structure. Instead, the term $\mathbf{Z}\mathbf{U}$ adds convenience of model formulation and data analysis because the covariance matrix structure $\mathbf{V}(\mathbf{Y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ commonly occurs with straightforward forms of $\mathbf{G}$ and $\mathbf{R}$. In many instances, between-unit variability can be incorporated into $\mathbf{Z}\mathbf{U}$ and within-unit variability into $\mathbf{e}$. This enables one to model between- and within-unit variability separately and distinctly. It can also be exploited in computations.

2.1 MODEL FORMULATION FOR MULTI-CENTER TRIAL WITH REPEATED MEASURES

In the clinical trial described in Section 1, there were n_{ij} patients in center j assigned to drug i . Sitting heart rate was measured at five visits on each patient. Let y_{ijkl} denote the measurement at visit l on patient k in center j assigned to drug i . A model will be developed in two phases, corresponding to the random and fixed effects, respectively. The random effects portion of a model generally refers to the random variation incurred at the various stages of sampling or data collection. When conceptualizing random effects, it is

helpful to imagine a model of data with no fixed effects. In the multi-center trial example, think of the sources of variation that would be present in the data if drugs had been assigned but not administered. In the present case, that also means no time effects, because time trends would be induced only by the administration of drugs. Thus, y_{ijkl} would contain random variation attributable to centers, patients within centers, and samples within patients within centers, which is a hierarchical classification of random effects. A model is then

$$y_{ijkl} = \mu + b_j + c_{ijk} + e_{ijkl} \quad (2.3)$$

where $E(y_{ijkl}) = \mu$ is an overall mean heart rate for the patient population, b_j is the random effect of center j , c_{ijk} is the random effect of the k th patient in center j assigned to drug i , and e_{ijkl} is the random error at visit l for this patient. Now imagine that drugs have been administered to patients, and consider the consequences of treatment on heart rate measurements: Each drug may affect heart rate, and the effect may change with time. The mean heart rate measured at visit l , for all patients at all centers who receive drug i , can be represented

$$\mu_{il} = E(y_{ijkl}) = \mu + \alpha_i + \tau_l + (\alpha\tau)_{il}, \quad (2.4)$$

where α_i , τ_l , and $(\alpha\tau)_{il}$ are effects of drug, visit, and drug*visit interaction, respectively. This is the mean response averaged over centers and patients. Also, there will be a possible interaction effect between drug i and center j , $(\alpha b)_{ij}$. This is a random effect because it would change with center. Thus, combining (2.3) and (2.4), and adding a random drug*center interaction term, one obtains a linear mixed model (LMM)

$$y_{ijkl} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + c_{ijk} + \tau_l + (\alpha\tau)_{il} + e_{ijkl}. \quad (2.5)$$

It might be tempting to include a $(b\tau)_{jl}$ term to the model, representing center-by-time interaction. Before doing so, the meaning of such a term should be considered along with its potential for existence. Arbitrarily including all possible terms leads to cumbersome models and uninterpretable outcomes.

It is reasonable to assume the b_j terms are distributed $\text{NID}(0, \sigma_{\text{center}}^2)$, the $(\alpha b)_{ij}$ terms are distributed $\text{NID}(0, \sigma_{\text{drug*center}}^2)$, and the c_{ijk} terms are distributed $\text{NID}(0, \sigma_{\text{patient}}^2)$. It would be convenient to assume $V(e_{ijkl}) = \sigma_{\text{error}}^2$, and that e_{ijkl} values on the same patient are independent, conditional on the patient. However, it is possible that correlations between measurements on the *same* patient decrease with the length of time between the measurements; for example, that adjacent measurements are more highly correlated than nonadjacent measurements. So one might wish to assume that the vector $\mathbf{e}_{ijk} = (e_{ijk1}, \dots, e_{ijk5})'$ has covariance matrix Σ for all ijk . Such a phenomenon frequently occurs in repeated measures data and, in fact, occurs to a small degree in this example. However, as will be explained in Section 2.3.3, for all practical purposes, it is reasonable to assume the elements of $\mathbf{e}_{ijk}$ are independent, conditional on patient. This is almost equivalent to deleting c_{ijk} from the model and assuming the matrix Σ has *compound symmetric* structure; that is, that elements of e_{ijkl} have equal variances and equal covariances. The difference between the two formulations is that the latter allows for negative covariance between observations within the same

patient. See Littell, Pendergast, and Natarajan (2000) for methods of modeling covariance structure of repeated measures.

Assumptions of the previous paragraph result in the variance

$$V(y_{ijkl}) = \sigma_{\text{center}}^2 + \sigma_{\text{drug*center}}^2 + \sigma_{\text{patient}}^2 + \sigma_{\text{error}}^2. \quad (2.6)$$

Covariances between two observations depend on the factors they have in common:

Factors in common:

$$\begin{array}{ll} \text{Center, drug, patient} & \text{cov}(y_{ijkl}, y_{ijkl'}) = \sigma_{\text{center}}^2 + \sigma_{\text{drug*center}}^2 + \sigma_{\text{patient}}^2 \\ \text{Center, drug} & \text{cov}(y_{ijkl}, y_{ijk'l'}) = \sigma_{\text{center}}^2 + \sigma_{\text{drug*center}}^2 \\ \text{Center} & \text{cov}(y_{ijkl}, y_{i'jk'l'}) = \sigma_{\text{center}}^2 \end{array} \quad (2.7)$$

Let $\mathbf{Y}$ be the vector of data, and assume it is sorted according to center, drug, patient, and visit. This is the hierarchy of random effects. Then the model can be written in the form (2.1), with columns of the matrix $\mathbf{X}$ containing indicator variables for the fixed effect parameters μ, α_i, τ_l , and $(\alpha\tau)_{il}$, and β containing those parameters. The columns of $\mathbf{Z}$ contain indicator variables for random effects $b_j, (\alpha b)_{ij}$, and c_{ijk} , and $\mathbf{U}$ contains these elements. Finally, the vector $\mathbf{e}$ contains the random errors e_{ijkl} for all observations.

The expected value of the observation vector also can be expressed

$$\mathbf{E}(\mathbf{Y}) = \mathbf{X}_2 \boldsymbol{\mu}, \quad (2.8)$$

where $\mathbf{X}_2$ is a matrix of that contains the columns of indicator variables corresponding to combinations of drug and visit, and $\boldsymbol{\mu}$ is a vector containing the parameters μ_{il} . This is the so-called “mean-model” parameterization. It has the advantage that the matrix $\mathbf{X}_2$ is of full rank and, equivalently, the parameters of $\boldsymbol{\mu}$ are unambiguously defined.

The covariance matrix $\mathbf{V}(\mathbf{Y})$ assumes a nested block-diagonal form with elements given by expressions in (2.6) and (2.7). The largest blocks correspond to the centers. Within each center are blocks corresponding to drug*center combinations, and within each of these are blocks corresponding to individual patients.

2.2 ISSUES RELATED TO ANALYSIS OF TWO-WAY CROSSED CLASSIFICATIONS

The multi-center trial example has three factors (centers, drug, and visits) that are pairwise crossed with each other. Interaction terms between drugs and centers and between drugs and visits are in the model. One of the interactions (drug*center) involves a fixed and a random factor, which makes it random. The other (drug*visit) involves two fixed factors, and is fixed. The two types of interactions have different implications for comparison of drugs and must be dealt with differently. Of particular concern is inference about main effects when interaction is in the model. Following is a brief review of some of the issues related to analysis of data from two-way crossed factors.

Consider two generic crossed factors, say A and B , with a levels of A and b levels of B . An example is a factorial experiment with factors A and B , in which the treatment consisting of the combination of the i th level of A and j th level of B is assigned to m_{ij}

experimental units. Or, perhaps each treatment was assigned to the same number of units initially, but only m_{ij} units yield useful data for the combination of level i of A and level j of B . Still another example is an observational study in which m_{ij} subjects are classified as belonging to levels i and j , respectively, of factors A and B , and x_{ijk} is measured on the k th of these units. In this section, we review some of the issues surrounding construction of tests and estimates for main effects of a factor. If both A and B are fixed, a model for the data is

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \quad (2.9)$$

where x_{ijk} is an “observation” in level i of A and level j of B , and μ, α_i, β_j and $(\alpha\beta)_{ij}$ are fixed but unknown parameters, $k = 1, \dots, m_{ij}$. The e_{ijk} are random variables distributed $MVN(0, \text{diag}(\sigma_{ijk}^2))$. Thus, $E(x_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ is the population mean for the i th level of A and j th level of B .

The primary concern is to make inference about main effects of factor A ; that is, about means and differences between means for the levels of factor A . A considerable literature exists on construction of statistical tests and estimates of main effects in this situation, culminating with the book by Searle (1987). Basically, the problem is to determine meaningful linear combinations of the means for one factor averaged over levels of the other factor.

In terms of a statistical test, Speed, Hocking, and Hackney (1978) described the problem as specifying the weighted averages for the null hypothesis

$$H_0 : w_{11}\mu_{11} + \dots + w_{1b}\mu_{1b} = \dots = w_{a1}\mu_{a1} + \dots + w_{ab}\mu_{ab}, \quad (2.10)$$

where $\sum_{j=1,b} w_{ij} = 1$, for $i = 1, \dots, t$. This is manifest, for example, in the types of sums of squares in the GLM procedure or the types of hypotheses in the MIXED procedures of SAS[®]. Different situations lead to different specifications of the weights. In many cases, equal weights ($w_{ij} = 1/b$) are appropriate, which relates roughly to SAS Type III. If population sizes N_{ij} are known, then choosing w_{ij} proportional to N_{ij} might be appropriate. In any event, the weights in (2.10) refer to a *hypothesis*, rather than the numbers of observations in a category, or even the variance of x_{ijk} . The hypothesis should be framed in terms of the meaningful scientific questions, rather than characteristics of the dataset. It is often necessary to return to the objectives of a study to ascertain the meaningful weights. While there might not be complete agreement on the choice of the weights, at least the problem should be clearly formulated.

The goal might be to estimate the linear combinations in (2.10), rather than test the hypothesis of their equality. Estimation entails the same problems as hypothesis testing in terms of choosing weights. This is manifest in the so-called least squares means computed by the GLM and MIXED procedure, using equal weights, but other weights might also be meaningful. Whether testing hypotheses or estimating linear combinations of main effects of a factor, one of the challenges is to specify the weights for averaging across the levels of the other factors. Error terms (i.e., denominators of test statistics or standard errors of estimates) result automatically from the variance of the linear combinations of x_{ijk} .

Now consider the two-way model with factor B random;

$$y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk}, \quad (2.11)$$

where μ and α_i are fixed parameters, the b_j are random effects for factor B , and the $(\alpha b)_{ij}$ are random effects for $A \times B$ interaction. Thus, $E(y_{ijk}) = \mu_i = \mu + \alpha_i$. In this relatively simple two-way mixed model setting, the hypothesis of equality of factor A means is easy to formulate because B is random. Ordinarily, one would not assign weights according to the levels of B . Instead, the relevant hypothesis would usually be

$$H_0 : \mu_1 = \cdots = \mu_a, \quad (2.12)$$

or equivalently

$$H_0 : \alpha_1 = \cdots = \alpha_a.$$

In principle, the hypothesis (2.12) is meaningful even if $A \times B$ interaction is large, at least if one is interested in inference in the “broad sense.” That is, if one wishes to make statements that apply to the entire population of levels of factor B . But in practice, the presence of a very large interaction might point to the necessity to examine and explain effects of factor A at specific levels of B even if B is random. In multi-center studies, some inferential statements might apply to the entire population of possible centers; that is, *marginal* inference about A . But large interactions might indicate the need to investigate effects of A at specific centers; that is, *conditional* inference about A .

It is easy to prescribe linear combinations of the factor A means that one might wish to estimate, such as those implied in (2.12) or more specific ones like

$$(\mu_1 + \mu_2)/2 \quad \text{or} \quad \mu_1 - (\mu_2 + \mu_3)/2. \quad (2.13)$$

But, it is not such a simple matter to specify a test statistic of (2.12) or an estimate of linear combinations as in (2.13). There are many linear combinations of the data that could be used to test (2.12) or estimate (2.13). The difficulty is to find those that provide powerful tests and precise estimates. When using a computer program such as SAS GLM, the problem translates into unclear choices among available sums of squares, or to specifying coefficients in an ESTIMATE statement in lieu of the built-in “lsmeans.” Generalized least squares would yield a test statistic for (2.12) or the BLUE of (2.13) if $\mathbf{V}$ were known (Searle, Casella, and McCulloch 1992). But $\mathbf{V}$ is not known and must be estimated, which introduces other complex problems discussed in Section 3.

2.3 INFERENCE METHODS

There are basically two approaches for inference about fixed effects in the linear mixed model (2.1). First, and still most familiar to many readers, is analysis of variance (ANOVA), which is based on linear functions $\mathbf{LY}$ and certain types of quadratic forms $\mathbf{Y}'\mathbf{QY}$ of the data vector $\mathbf{Y}$. The linear functions and quadratic forms are constructed to assess functions

of fixed and random effect parameters. Second is the family of methods using variations of generalized least squares and maximum likelihood. Included in the family are so-called estimated generalized least squares (EGLS), maximum likelihood (ML), and residual maximum likelihood (REML). The acronym REML traditionally refers to a method of estimating variance components, but recently (e.g., Kenward and Roger 1997) has been adopted to refer to a combination of REML estimation of the random effect parameters, and EGLS estimation of fixed effect parameters. Statistical inference is obtained by using the GLS methods with an estimate $\hat{\mathbf{V}}$ of $\mathbf{V}$ substituted for $\mathbf{V}$.

2.3.1 ANOVA Methods

The residual sum of squares after fitting a regression model $\mathbf{Y} = \mathbf{M}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$ is the quadratic form $\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}')\mathbf{Y}$. Sums of squares used in analysis of variance can be expressed as $R(\mathbf{M}_2|\mathbf{M}_1) = \text{SSE}_1 - \text{SSE}_2$, where SSE_1 is the residual sum of squares after a least squares fit of the “reduced” model with $\mathbf{E}(\mathbf{Y}|\boldsymbol{\eta}_1) = \mathbf{M}_1\boldsymbol{\eta}_1$, and SSE_2 is the residual sum of squares after a least squares fit of the “full” model with $\mathbf{E}(\mathbf{Y}|\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \mathbf{M}_1\boldsymbol{\eta}_1 + \mathbf{M}_2\boldsymbol{\eta}_2$. When applied to the mixed model (2.1), $\mathbf{M}_1$ and $\mathbf{M}_2$ are matrices whose columns are linear combinations of the columns of $\mathbf{X}$ and $\mathbf{Z}$. The quadratic form $R(\mathbf{M}_2|\mathbf{M}_1)$ is the “reduction in error sum of squares” due to adding variables in $\mathbf{M}_2$ to a model that already contains variables in $\mathbf{M}_1$. It is a measure of the effect of variables in $\mathbf{M}_2$ on $\mathbf{E}(\mathbf{Y})$, after the effects of variables in $\mathbf{M}_1$ have been removed. Sometimes $R(\mathbf{M}_2|\mathbf{M}_1)$ is called the “effect of $\mathbf{M}_2$ adjusted for $\mathbf{M}_1$.” It would be used to test the null hypothesis $H_0 : \boldsymbol{\eta}_2 = \mathbf{0}$ in the model $\mathbf{E}(\mathbf{Y}|\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \mathbf{M}_1\boldsymbol{\eta}_1 + \mathbf{M}_2\boldsymbol{\eta}_2$. Incidentally, this notation is valid for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ representing either fixed or random effects, or combinations thereof. In “balanced data,” the value of $R(\mathbf{M}_2|\mathbf{M}_1)$ depends only on $\mathbf{M}_2$, and not $\mathbf{M}_1$.

Classical ANOVA tables for balanced data appear in numerous statistical texts, along with instruction for computing and interpreting them. With few exceptions, though, details of how to perform analysis of variance with unbalanced mixed model data are woefully inadequate. One explanation is that ANOVA methods use computations derived from fixed effect models, rather than mixed models, so there are no model-based principles to form guidelines in mixed model applications. The ANOVA computations adapt readily to mixed model applications in balanced, but not unbalanced, situations. Essentially, the difficulty is in deciding what columns to include in $\mathbf{M}_1$ in the model with $\mathbf{E}(\mathbf{Y}|\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \mathbf{M}_1\boldsymbol{\eta}_1 + \mathbf{M}_2\boldsymbol{\eta}_2$. There are two parts to this question; one pertains to fixed effects and the other to random effects. Partition the matrix $\mathbf{M}_1$ into $\mathbf{M}_{1F}$ and $\mathbf{M}_{1R}$; $\mathbf{M}_1 = [\mathbf{M}_{1F} : \mathbf{M}_{1R}]$, where columns of $\mathbf{M}_{1R}$ are linear combinations of columns of $\mathbf{X}$ and columns of $\mathbf{M}_{1F}$ are linear combinations of columns of $\mathbf{Z}$. Choice of $\mathbf{M}_{1F}$ is the fixed effects part of the question and choice of $\mathbf{M}_{1R}$ is the random effects part. Choice of $\mathbf{M}_{1F}$ is dictated by the hypothesis one wishes to test, as in (2.10). But choice of $\mathbf{M}_{1R}$ does not affect the hypothesis. Instead, it affects the statistic used to *test* the hypothesis.

Hypotheses about fixed effects are tested using ratios of quadratic forms to create statistics that approximately follow F distributions. Conceptually, the ratios have the form (error + signal)/error, where “signal” is a parametric function that measures departure from the null hypothesis, and “error” is a measure of random variation. Expected means squares

can be used to separate the signal and expected error, and thus determine the “hypothesis tested” and the appropriate “error term” for the denominator.

Typically, an expected mean square looks like

$$\mathbf{E}(\mathbf{Y}'\mathbf{Q}\mathbf{Y}) = \sum_l c_l \sigma_l^2 + \varphi^2, \quad (2.14)$$

where the c_l are known constants, the σ_l^2 are variance components for random effects, and $\varphi^2 = \beta' \mathbf{X}' \mathbf{Q} \mathbf{X} \beta$ is a quadratic form of fixed effect parameters. The sum $\sum_l c_l \sigma_l^2$ is the “expected error” and φ^2 is the “signal.” The ratio is constructed to test $H_0 : \varphi^2 = 0$. In balanced cases, the procedure usually results in a unique statistic that follows a noncentral F distribution with noncentrality parameter equal to a multiple of φ^2 . But in unbalanced cases the procedure is ad hoc and does not necessarily yield unique, much less optimal solutions. In fact, several different quadratic forms may be used to test the same null hypothesis H_0 , but it is not always clear how to choose among them. That is, several quadratic forms, $\mathbf{Y}'\mathbf{M}_k\mathbf{Y}$, $k = 1, \dots, K$, could have expectations $\mathbf{E}(\mathbf{Y}'\mathbf{Q}_k\mathbf{Y}) = \sum_l c_{kl} \sigma_l^2 + \varphi_k^2$, such that $\varphi_k^2 = 0$ when H_0 is true, for $k = 1, \dots, K$. In fact, $\varphi_k^2 = \beta' \mathbf{X}' \mathbf{Q}_k \mathbf{X} \beta$. Then any one of these is a candidate for the numerator of an approximate F statistic for testing H_0 . In particular, PROC GLM in SAS provides Types I, II, III, and IV sums of squares. These have specific interpretations in fixed effect settings, but there are no definitive guidelines for testing a hypothesis about fixed effects in the presence of random effects. Results of the tests from the three types can vary greatly. Milliken and Johnson (1992), Damon and Harvey (1987), and Rawlings, Pantula, and Dickey (1998), all present examples of analysis that basically use different “types” of sums of squares for this type of problem, but they do not present logical rationale for their chosen methods. Littell, Stroup, and Freund (2002) suggested approximating noncentrality parameters associated with test statistics based on the various types of sums of squares and choosing the type that maximizes this parameter. But in general, the problem is without evident solution. Related to this problem is the choice of quadratic form for ANOVA estimation of variance components, as noted by Harvey (1982).

In addition to the complication of choosing a numerator mean square for an ANOVA test statistic, one must select an “error term,” that is a denominator “mean square” for the F statistic. This ordinarily entails using expected means squares to determine a linear combination of mean squares that has the desired expectation. In other words, if $\mathbf{Y}'\mathbf{Q}\mathbf{Y}$, with expectation (2.14), is used as a mean square for the numerator, then a linear combination of mean squares $\sum_h a_h \mathbf{Y}'\mathbf{Q}_h\mathbf{Y}$ with expectation $\sum_l c_l \sigma_l^2$ must be found to use in the denominator. Then Satterthwaite’s (1946) formula is used to estimate degrees of freedom associated with the approximate distribution of $\sum_h a_h \mathbf{Y}'\mathbf{Q}_h\mathbf{Y}$. The process entails several levels of approximation and estimation, and tends to have good performance under well-behaved conditions, such as when all a_h are non-negative. But under other conditions the procedure is unreliable (Guerin and Stroup 2000, Remenga and Johnson 1995). One of the most troublesome problems occurs when some of the a_h are negative. Then the distribution of $\sum_h a_h \mathbf{Y}'\mathbf{Q}_h\mathbf{Y}$ has support on the negative axis, and thus cannot be approximated very well by the chi-squared distribution. By the same token, the distribution of a ratio with $\sum_h a_h \mathbf{Y}'\mathbf{Q}_h\mathbf{Y}$ in its denominator cannot be approximated very well with an F distribution.

2.3.2 EGLS Methods

Denoting $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, the generalized least squares (GLS) estimator of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \quad (2.15)$$

provided all indicated inverses exist. If not, then generalized inverses are used. The GLS estimate provides best unbiased estimates $\mathbf{L}\hat{\beta}$ of estimable linear combinations $\mathbf{L}\beta$ of the elements of the fixed effect parameter vector.

The covariance matrix of $\hat{\beta}$ is

$$\mathbf{V}(\hat{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}, \quad (2.16)$$

and the covariance matrix of a set of linear combinations $\mathbf{L}\hat{\beta}$ is

$$\mathbf{V}(\mathbf{L}\hat{\beta}) = \mathbf{L}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{L}'. \quad (2.17)$$

Equation (2.17) provides the machinery for statistical inference regarding $\mathbf{L}\beta$. Specifically, a test of the null hypothesis $H_0 : \mathbf{L}\beta = \mathbf{0}$ is given by the statistic

$$\hat{\beta}'\mathbf{L}'[\mathbf{L}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{L}]^{-1}\mathbf{L}\hat{\beta}, \quad (2.18)$$

which has a chi-square distribution with degrees of freedom equal to $r(\mathbf{L})$, the rank of $\mathbf{L}$. A $100(1 - \alpha)\%$ confidence interval for a single linear combination $\mathbf{l}\beta$ is given by

$$\mathbf{l}\hat{\beta} \pm z_{\alpha/2}\mathbf{l}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{l} \quad (2.19)$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ percentile of the standard normal distribution. Generalizations exist for simultaneous confidence regions for sets of linear combinations.

Methods shown in (2.15) through (2.19) provide basic tools for data analysis for the linear mixed model. They are quite simple and concise, so one might wonder, “Why all the fuss about linear mixed models.” The answer is that methods (2.15)–(2.19) all rely on the assumption that $\mathbf{V}$ is known. However, the contrary is usually true, so that (2.15)–(2.19) cannot be used in their basic forms.

Computation of $\hat{\beta}$ requires knowledge of $\mathbf{V}$, except in special circumstances in which $\hat{\beta}$ is invariant to $\mathbf{V}$ (see Puntanen and Styan 1989). Moreover, even if $\hat{\beta}$ is invariant with respect to $\mathbf{V}$, $\mathbf{V}(\hat{\beta})$ might not be. In reality, $\hat{\beta}$ hardly ever is invariant to $\mathbf{V}$ with unbalanced data. Thus, an estimate $\hat{\mathbf{V}}$ of $\mathbf{V}$ must be used in the computation, providing the “estimated” generalized least squares (EGLS) estimate of β ,

$$\tilde{\beta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}. \quad (2.20)$$

Likewise, $\hat{\mathbf{V}}$ is used in place of $\mathbf{V}$ to estimate the covariance matrix of $\tilde{\beta}$,

$$\hat{\mathbf{V}}(\tilde{\beta}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}. \quad (2.21)$$

However, the true variance of $\tilde{\beta}$ is not $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ because $\tilde{\beta}$ contains variation due to $\hat{\mathbf{V}}$, so $\hat{\mathbf{V}}(\tilde{\beta})$ in (2.21) is not always a good estimate of $\mathbf{V}(\tilde{\beta})$, as will be discussed in Section 3.

In most cases, $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ is a function of a vector of covariance parameters, θ . The EGLS approach first estimates parameters of the covariance matrices $\mathbf{G}$ and $\mathbf{R}$, then substitutes these estimates for the parameters into the functional form of $\mathbf{G}(\theta)$ and $\mathbf{R}(\theta)$ to obtain $\hat{\mathbf{G}} = \mathbf{G}(\hat{\theta})$ and $\hat{\mathbf{R}} = \mathbf{R}(\hat{\theta})$. Then $\hat{\mathbf{V}}$ is computed as $\hat{\mathbf{V}} = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}' + \hat{\mathbf{R}}$. Residual maximum likelihood has become the favored method of estimation (Searle, Casella, and McCulloch 1992) of random effect parameters. Hypotheses about fixed effects can be tested using the statistic

$$\tilde{\beta}'\mathbf{L}' [\mathbf{L}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{L}']^{-1} \mathbf{L}\tilde{\beta}. \quad (2.22)$$

In balanced data cases, the statistic (2.22) divided by the rank of $\mathbf{L}$ yields the ANOVA F statistic, except for caveats concerning estimates of variance components that are negative or zero. With unbalanced data, the EGLS approximate F is not directly related to any particular ANOVA F statistic. It has certain desirable attributes not possessed by the ANOVA statistic. For example, there is no guesswork in formulating (2.22) corresponding to the murky choice of choosing a numerator for an ANOVA F statistic. The statistic does not have an exact F distribution, but is asymptotically distributed as a constant times a chi-square as the *number of levels of random effects* becomes large. In addition, “error” terms do not have to be manufactured. They result naturally by substituting estimates of random effect parameters into the GLS chi-squared statistics. Based on the fact that (2.22) is an exact F in ideal circumstances and is asymptotically chi-square otherwise, it is reasonable to use F as an approximating distribution. Giesbrecht and Burns (1985), Fai and Cornelius (1996), and Kenward and Roger (1997), constructed approximate denominator degrees of freedom by generalizing the Satterthwaite method for ANOVA mean squares. These are described in Section 4.

2.3.3 Comparison of EGLS and ANOVA for Multi-Center Trial

The MIXED procedure in SAS can be used to illustrate EGLS methodology. The first step in the EGLS analysis is to estimate the parameters of the covariance structure. Following methods described by Littell, Pendergast, and Natarajan (2000), three covariance structures for the within-patient error vectors e_{ijkl} in model (2.9) were evaluated: compound symmetric (CS), autoregressive order one (AR(1)), and “unstructured” (UN). (Note that the random effect for patient in model (2.9) can be absorbed into the CS and UN structures, but not into AR(1). For this reason, in order to discuss the three models in comparable terms, we shall refer to AR(1) as AR(1) + RE.) These structures are nested, so models may be compared with likelihood ratio tests. Table 1 shows $-2\log$ -likelihood and related fit measures. According to likelihood ratio tests, UN does not fit better than AR(1) + RE, but AR(1) + RE fits better than CS. Also, AR(1)+RE is the preferred structure according to the AICC and BIC criteria. The “statistically” significant differences between the models occur because of the relatively large number of centers. But the three models produce almost indistinguishable results for inference about fixed effects. Therefore, we elect to use the simplest structure, CS, and conclude with a variance components-type model. Based on model (2.9) with CS structure, the REML algorithm yields $\hat{\sigma}_{\text{center}}^2 = 0.45$, $\hat{\sigma}_{\text{drug*center}}^2 = 4.76$, $\hat{\sigma}_{\text{patient}}^2 = 30.83$, and $\hat{\sigma}_{\text{error}}^2 = 62.65$.

Table 1. -2 Log-Likelihood Values for Three Covariance Structures

Structure	CS	AR(1) + RE	UN
-2 log-Likelihood	8855.9	8839.3	8824.0
Number of parameters	4	5	17
AICC	8863.9	8849.4	8858.5
BIC	8871.8	8859.2	8891.5

The next step is to make inference about fixed effect parameters. Substitution of the variance component estimates into the expression for the EGLS F statistic (2.22), using Types 1, 2, and 3 coefficients for estimable functions in $\mathbf{L}$, gives results in Table 2. The ultimate goal is to compare drugs. Before constructing a test of “main effect” of drug, interaction between drug and visit should be analyzed. The tests in Table 2 indicate little, if any, evidence of drug*visit interaction. Thus, a test for main effects of drugs is appropriate.

Now we are in the situation of wanting to test for the main effect of one factor (drug) in the presence of another factor (visit) that is crossed with the first. So, we must choose weights for the means as in (2.10) with $A = \text{drug}$ and $B = \text{visit}$. These weights are shown in Table 3 for Types 1, 2, and 3 EGLS tests from the MIXED procedure (and Types I, II, and III ANOVA tests from the GLM procedure). For this application, we choose Type 3 coefficients because they weight the means equally across visits. That gives the hypothesis we would wish to test in the original plan. Therefore, we report the Type 3 $F = 4.29$ from Table 2. This EGLS approximate F has 1 degree of freedom for drug and 29.3 “denominator” degrees of freedom from a Satterthwaite-type approximation that is discussed in Section 4. The p value is 0.047, indicating a statistically significant difference between the drug means, averaged over visits.

The EGLS estimates (2.17) of the drug means and difference ($\pm$ s.e.) are $\hat{\mu}_1 = 77.20(\pm 0.73)$ and $\hat{\mu}_2 = 75.08(\pm 0.70)$, and $\hat{\mu}_1 - \hat{\mu}_2 = 2.12(\pm 1.00)$. These were obtained from the lsmeans statement in PROC MIXED, and use equal coefficients as in the Type 3 EGLS F statistics. The drug lsmeans are estimable from MIXED because the fixed effects in the model contain only drug and visit, and there are data for all combinations of the factors. However, it may be a concern that data are more likely to be missing for later visits than early ones. In general and overly simplified terms, the methods used in PROC MIXED are valid when data are missing at random (see Little and Rubin 1987). Failure for the condition to hold may seriously bias estimates and invalidate results, but further discussion is beyond the scope of this article.

Analysis of variance concepts of the multi-center trial data can be illustrated using

Table 2. Denominator Degrees of Freedom, F Statistics, and p Values for Types 1, 2, and 3 EGLS Tests of Fixed Effects

	Type 1			Type 2			Type 3		
	DDF	F	P	DDF	F	P	DDF	F	P
drug	28.5	3.76	.062	28.6	3.72	.064	29.3	4.29	.047
visit	947	0.36	.839	947	0.36	.839	947	0.34	.853
drug*visit	947	1.45	.214	947	1.45	.214	947	1.45	.214

Table 3. Coefficients on Means for GLM and MIXED Tests of Drug Main Effect

<i>Visit</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>MIXED</i>					
<i>Drug 1</i>					
Type 1	.2572	.2164	.2027	.1635	.1601
Type 2	.2540	.2255	.2060	.1629	.1516
Type 3	.2	.2	.2	.2	.2
<i>Drug 2</i>					
Type 1	-.2500	-.2358	-.2091	-.1619	-.1431
Type 2	-.2540	-.2255	-.2060	-.1629	-.1516
Type 3	-.2	-.2	-.2	-.2	-.2
<i>GLM</i>					
<i>Drug 1</i>					
Type I	.2572	.2164	.2027	.1635	.1601
Type II	.2611	.2280	.2040	.1592	.1477
Type III	.2	.2	.2	.2	.2
<i>Drug 2</i>					
Type I	-.2500	-.2358	-.2091	-.1619	-.1431
Type II	-.2611	-.2280	-.2040	-.1592	-.1477
Type III	-.2	-.2	-.2	-.2	-.2

the GLM procedure in SAS. The sources of variation are drug, center, drug*center, visit, and drug*visit. Table 4 shows degrees of freedom and Types I, II, and III expected mean squares, and coefficients of the “synthetic” mean squares for the denominator. Notice that all three expected mean squares for drug contain the variance component for drug*center, with different coefficients. This is because columns of dummy variables for drug in $\mathbf{X}$ are spanned by columns for drug*center in $\mathbf{U}$. However, only the Type I expected mean square for drug contains the variance component for center, because the Types II and III sums of squares for drug have been “adjusted” for center, but the Type I has not. In terms of the reduction in sum of squares notation $R(\mathbf{M}_2|\mathbf{M}_1)$, the matrix $\mathbf{M}_1$ for the Types II and III sums of squares contains columns for center, but it does not for the Type I sum of squares.

It should be noted that the statements concerning Type I sums of squares and expected mean squares are true for a PROC GLM model statement with terms in the order as in Equation (2.5) because Type I sums of squares are model-order dependent. Each effect is “adjusted” for all effects that precede it in the GLM MODEL statement. Some authors, such as Milliken and Johnson (1992), recommend that all fixed effects precede all random effects in the GLM MODEL statement to make each random effect adjusted for each fixed effect. This is one of the few universal guidelines of analysis of unbalanced mixed model data. Then the expected mean squares of random effects would not contain any fixed effect parameters. In this article, we use the order as in (2.5) for illustrative purposes, only.

According to the ANOVA principles based on expected mean squares in Section 2.3.1, any one of the Types I, II, or III mean squares could be used to construct a test for differences between drug means, but they use slightly different weights in (2.12), where $A = \text{drug}$ and $B = \text{visits}$. (See Table 3). The three tests are shown in Table 5, including values of the approximate F statistic, degrees of freedom, and p values. Results of these tests

Table 4. Coefficients of Variance Components in Expected Mean Squares and Linear Combinations of Means Squares Used in Denominator for Drug F Statistics

Variance component	σ^2_{error}	$\sigma^2_{patient}$	$\sigma^2_{drug*center}$	σ^2_{center}
<i>Type I coefficients</i>				
drug	1	4.42	26.93	1.41
center	1	4.29	12.05	22.48
drug*center	1	4.02	12.59	
patient	1	3.82		
Error for drug: $0.06 MS_{center} + 2.08 MS_{drug*center} - 1.01 MS_{patient} - 0.04 MS_{error}$				
<i>Type II coefficients</i>				
drug	1	4.28	26.76	
center	1	4.29	12.04	22.45
drug*center	1	4.02	12.59	
patient	1	3.82		
Error for drug: $2.12 MS_{drug*center} - 1.11 MS_{patient} - 0.01 MS_{error}$				
<i>Type III Coefficients</i>				
drug	1	3.12	7.16	
center	1	3.65	9.72	18.75
drug*center	1	3.62	10.85	
patient	1	3.82		
Error for drug: $0.67 MS_{drug*center} + 0.19 MS_{patient} + 0.15 MS_{error}$				

are disconcerting because their p values of 0.066, 0.105, and 0.019 for Types I, II, and III, respectively, convey different shades of conclusion, if not contradiction, regarding the difference between drug means. Discrepancies in p values result for two reasons: One, the hypotheses being tested are not exactly the same; that is, the weights in (2.13) are not the same for the three tests. For the present example, it is doubtful that this reason is influential because the weights are similar. Also, the Types 1, 2, and 3 EGLS tests for drug in Table 2 from the MIXED procedure, which use almost the same coefficients, produce p values similar to each other. In other applications, the weights could differ substantially. Two, the types of sums of squares incorporate random effects differently. This second reason is the likely culprit, because the expected mean squares for the three types are quite different. In addition, Types I and II approximate F statistics use denominator linear combinations of mean squares with negative weights.

Estimates of drug or drug*visit means are not available from the GLM procedure because it performs a least squares fit of the model containing center and drug*center as fixed effects, and there are no data for some drugs in some centers. Thus, overall drug

Table 5. Denominator Degrees of Freedom, F Statistics, and p Values for Types I, II, and III ANOVA Tests of Fixed Effects

	<i>Type I</i>			<i>Type II</i>			<i>Type III</i>		
	DDF	F	P	DDF	F	P	DDF	F	P
drug	18.0	3.84	.066	16.6	2.95	.105	75.4	5.79	.019
visit	904	0.29	.886	904	0.29	.886	904	0.25	.909
drug*visit	904	1.64	.163	904	1.64	.163	904	1.64	.163

Table 6. EGLS Variance Estimates of With and Without Correction for Complete Dataset and Three Subsets

No. centers	53	15	10	5
No. Patients	311	127	75	36
$\hat{\mathbf{V}} = \hat{\mathbf{V}}(\hat{\mu}_1 - \hat{\mu}_2)$	0.99	2.07	3.79	4.31
$\tilde{\mathbf{V}} = \hat{\mathbf{V}}(\tilde{\mu}_1 - \tilde{\mu}_2)$	1.00	2.16	4.23	7.27
$(\tilde{\mathbf{V}} - \hat{\mathbf{V}})/\hat{\mathbf{V}}$	.005	.042	.116	.687

means are considered non-estimable by GLM. This would be true regardless of the computer program used if it fitted the random effects as if they were fixed. Modeling random effects into the covariance structure, where they belong, is a huge advancement in linear model computing technology because it overcomes the troublesome, and sometimes disabling, spurious problems that result when random effects are modeled as fixed.

3. VARIANCE PROPAGATION IN EGLS

The estimator $\tilde{\beta}$ is not a linear combination of the data vector $\mathbf{Y}$, and has a complicated variance expression. Kackar and Harville (1984) showed that $\tilde{\beta}$ is an unbiased estimate of β , but the variance of any linear combination $\mathbf{L}\tilde{\beta}$ is greater than the variance of the corresponding linear combination $\mathbf{L}\hat{\beta}$; that is, $\mathbf{V}(\mathbf{L}\tilde{\beta}) > \mathbf{V}(\mathbf{L}\hat{\beta})$. Jeske and Harville (1988), Harville and Jeske (1992), and other authors identified expressions for the difference between $\mathbf{V}(\mathbf{L}\tilde{\beta}) - \mathbf{V}(\mathbf{L}\hat{\beta})$ in special cases. Prasad and Rao (1990) and Kenward and Roger (1997) further developed the expressions. These are implemented as options into computations of standard error and test statistics in the MIXED procedure of SAS[®] software whenever the KR option is specified.

The difference $\mathbf{V}(\mathbf{L}\tilde{\beta}) - \mathbf{V}(\mathbf{L}\hat{\beta})$ results from variation in $\hat{\mathbf{V}}$ as an estimator of $\mathbf{V}$; that is, variation in $\hat{\mathbf{V}}$ is *propagated* into variation in $\mathbf{L}\tilde{\beta}$. Therefore, it makes sense that the difference would decrease as the amount of data increases. However, it is not simply the total amount of data that matters. Rather, the numbers of centers and patients affect the difference between $\mathbf{V}(\mathbf{L}\tilde{\beta})$ and $\mathbf{V}(\mathbf{L}\hat{\beta})$. The example is a large study with 53 centers and 311 patients. The difference between $\mathbf{V}(\mathbf{L}\tilde{\beta})$ and $\mathbf{V}(\mathbf{L}\hat{\beta})$ is almost imperceptible for the estimate of $\mathbf{1}\beta + \mu_1 - \mu_2$ based on the entire dataset. The standard error estimates were computed for the complete dataset, for the first 5, 10, and 15 centers, and are shown in Table 6. The last line in Table 6 shows, for this example, that the increase in variance is only 0.5% for the full dataset, 4% for the first 15 centers, 12% for the first 10, and 69% for the first 5. This illustrates the critical need for the adjustment if the number of levels of a random effect is small. More research is needed on this topic to better understand the nature of EGLS variance propagation and the properties of estimators.

There is no variance propagation in OLS estimators used in ANOVA applications because those are linear combinations with known weights. The OLS estimate $\mathbf{L}\hat{\beta}$ of $\mathbf{L}\beta$ has covariance matrix $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}$. Strictly speaking, estimates from the GLM procedure are not OLS estimates of fixed effects $\mathbf{L}\beta$ in the LMM (2.1), because GLM fits (2.1) treating $\mathbf{U}$ as fixed. Thus, estimates from GLM come from solving the normal

equations $(\mathbf{M}'\mathbf{M})[\check{\boldsymbol{\beta}}' : \mathbf{U}']' = \mathbf{M}'\mathbf{Y}$, where $\mathbf{M}$ is the partitioned matrix $\mathbf{M} = [\mathbf{X} : \mathbf{Z}]$. Linear combinations $\mathbf{L}\check{\boldsymbol{\beta}}$ are unbiased, but *not* minimum variance, for $\mathbf{L}\boldsymbol{\beta}$.

4. DEGREES OF FREEDOM

A mundane but important issue in mixed model applications is determining degrees of freedom for approximate F statistics and standard errors of estimates of linear combinations of fixed effects parameters. In the ANOVA setting, the Satterthwaite approximation for linear combinations of mean squares is well known. But in EGLS applications, there is no such thing as “mean squares.” Instead, for a single estimable function $\mathbf{l}\boldsymbol{\beta}$, one devises the variance of a linear combination of fixed effect estimates as

$$\mathbf{V}(\mathbf{l}\hat{\boldsymbol{\beta}}) = \mathbf{l}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{l}'. \quad (4.1)$$

Then one estimates the variance as

$$\hat{\mathbf{V}}(\mathbf{l}\hat{\boldsymbol{\beta}}) = \mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}'. \quad (4.2)$$

The chi-square distribution can be used to approximate the distribution of $\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}'$. That is, one assumes

$$v\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}' / E(\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}') \quad (4.3)$$

is approximately distributed as chi-square. Then

$$E[v\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}' / E(\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}')] = v \quad (4.4)$$

and

$$\mathbf{V} [v\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}' / E(\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}')] = 2v. \quad (4.5)$$

Substitution and solving for v gives

$$v = 2 \left[E [\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}']^2 / \mathbf{V}(\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}') \right], \quad (4.6)$$

which leads to the estimated approximate degrees of freedom

$$\hat{v} = 2 [\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}']^2 / \hat{\mathbf{V}}(\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}'). \quad (4.7)$$

Calculation of this expression is not trivial because of the denominator. It is implemented in the SAS procedure MIXED through the Satterthwaite option, using the second derivative matrix and gradients to approximate $\hat{\mathbf{V}}(\mathbf{l}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{l}')$. In fact, (4.6) contains the machinery for Satterthwaite’s df computation in ANOVA, but direct computation is simpler.

Fai and Cornelius (1996) expanded the concept for multiple linear combinations. In addition, Kenward and Roger (1997) incorporated the variance inflation phenomenon of Section 3 into the computations. These are implemented in the MIXED procedure whenever the KR (for Kenward–Roger) option is specified. Guerin and Stroup (2000) showed that the KR method substantially improves the degrees of freedom approximation.

5. INFERENCES ABOUT RANDOM EFFECT PARAMETERS

Sometimes the variance components are of inferential interest instead of being nuisance parameters. Tests of hypotheses and confidence intervals can be obtained in both the ANOVA and EGLS settings, although the procedures are not exact in either setting. The EGLS methods rely on asymptotic distributions of the REML estimates and likelihood ratio statistics, and therefore are of questionable value for small to medium sized numbers of levels of the random effect in question. Also, the validity of asymptotic standard errors for estimates and asymptotic distributions of likelihood ratio statistics for parameters require an open interval containing the parameter within the parameter space. This condition is not met when testing a hypothesis that a variance component is zero, because all variances are non-negative and thus on the boundary of the parameter space when zero. There are ways to skirt this technical difficulty, but they entail some risk. Self and Liang (1987), showed that, under certain conditions, the asymptotic distribution of $-2\log$ -likelihood ratio is a 0.5 and 0.5 mixture of a discrete distribution with mass at zero and a chi-square with $df = 1$. The mass at zero corresponds to the MLE of the variance component being zero, that is, of the likelihood having a local maximum over the parameter space on the boundary. Thus, to use this result, one would simply cut the usual p value in half. In addition, Self and Liang (1987) presented related results concerning the asymptotic distributions of ML estimates when the true values of parameters are on the boundary. The practical consequence for testing the null hypothesis that a variance component is zero using a Wald statistic is essentially the same as that of using a likelihood ratio statistic. In both cases, the mixing proportions may be quite different from 0.5 with finite samples, and validity of the continuous portions of the asymptotic distributions are subject to the same perils as in regular likelihood-based inference. For all of these reasons, such inference is not default output from the MIXED procedure.

In the clinical trials example, it might be of interest to assess the statistical significance of the drug*center variance component. In a broader sense, this is treatment-by-environment interaction. It measures covariance between observations on subjects who received the same treatment within the same clinic. If the data were balanced, then ANOVA would provide an exact test with $F = MS(\text{drug*center})/MS(\text{patient}(\text{drug*center}))$. Since the data are unbalanced, one could use either the Type II or III mean squares in the numerator. Type I would not be appropriate because it is not adjusted for the fixed effects of visit and drug*visit and is therefore confounded with the effects. The choice between II and III is not immediately clear, and they produce quite different results, with p values of 0.0735 and 0.2552, respectively. The disagreement arises from substantially different values of $MS(\text{drug*center})$. These are issues similar to those discussed in Section 2 regarding tests of fixed effects, and have not been thoroughly researched.

The REML Wald statistic, converted to a one-sided test, has $p = 0.0669$. With relatively large number of centers in this study, the assumption of equal mass in the continuous and discrete portions of the distribution of the REML estimate may be reasonable. But with smaller numbers of centers, the Wald test based on the REML estimate is highly unreliable, *even with balanced data*. This test is *not* equivalent to an ANOVA F test. Experience has

shown many cases in which an ANOVA test produced overwhelming evidence of a positive variance component, yet the REML-Wald test was not significant.

6. SUMMARY AND CONCLUSIONS

Analysis of variance has been a mainstay of statistical data analysis for several decades. Its application in mixed models uses OLS computations based on a model in which random effects are treated as fixed. Then those computations are adapted for mixed models with no central theoretical basis. Nonetheless, ANOVA methods have served us well. Perhaps the greatest shortcoming has been the lack of computer programs that fully extend ANOVA computations to mixed model applications, rather than inadequacies of the methods. However, ANOVA methods can be expected to decline in usage as more refined methodology becomes available. Only in the most recent decade or so has computer software been available to use more contemporary likelihood-based methods. The basic concepts of REML and GLS have been available for many years. But when computer technology made such programs feasible, it became apparent that there were many unanswered questions, such as the effects of estimating the covariance matrix in EGLS and degrees of freedom counts. Developments have been dramatic, but many questions remain on the basic behavior of REML/GLS statistics. Moreover, competent modeling is critical with such methods. Computer programs are in the hands of users who have no way of making sound decisions on model selection. Likelihood-based methods are not commonly taught in applied courses at many universities, wherein students frequently do not have the background to understand them. Finally, likelihood methods are less intuitive to data analysts than the appealing ANOVA concept of partitioning variation.

Nonetheless, advantages of REML/GLS methods compared to ANOVA far exceed the disadvantages. Foremost, the methods match the models, and therefore provide the basis for optimal results. If the models are formulated properly, appropriate inference should follow. Idiosyncrasies and inadequacies of OLS-based computer programs will not plague the user. Technical problems exist in REML/GLS programs, but they are being solved. Also, new approaches will lead to better solutions. For example, resampling techniques may prove superior to the Taylor series approximations that have been used to estimate $V(L\tilde{\beta})$.

Likelihood-based methods in mixed linear models are here to stay. In some fields of study, such as animal science, mixed model analyses are often required. Continuing education programs are underway within universities, corporations, government agencies and professional societies to train personnel on the use of modern methodology. It is only a matter of time until computing technology and education make the methods as familiar as ANOVA.

ACKNOWLEDGMENTS

This work was supported by the Florida Agricultural Experiment Station.

[Received July 2002. Revised October 2002.]

REFERENCES

- Damon, R. A., Jr., and Harvey, W. R. (1987), *Experimental Design, ANOVA, and Regression*, New York: Harper & Row.
- Fai, A. H. T., and Cornelius, P. L. (1996), "Approximate F -tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analysis of Unbalanced Split-Plot Experiments," *Journal of Statistical Computation and Simulation*, 54, 363–378.
- Giesbrecht, F. G., and Burns, J. C. (1985), "Two-Stage Analysis Based on a Mixed Model: Large-Sample Theory and Small-Sample Simulation Results," *Biometrics*, 41, 477–486.
- Guerin, L., and Stroup, W. W. (2000), "A Simulation Study to Evaluate PROC MIXED Analysis of Repeated Measures Data," in *Proceedings of the 12th Annual Conference on Applied Statistics in Agriculture*, Manhattan, KS: Kansas State University.
- Harvey, W. R. (1982), "Mixed Model Capabilities of LSML76," *Journal of Animal Science*, 54, 1279–1285.
- Harville, D. A., and Jeske, D. R. (1992), "Mean Squared Error of Prediction Under a General Linear Model," *Journal of the American Statistical Association*, 87, 724–731.
- Henderson, C. R. (1975), "Best Linear Estimation and Prediction Under a Selection Model," *Biometrics*, 31, 423–449.
- (1984), "Applications of Linear Models in Animal Breeding," Guelph, Ontario: University of Guelph.
- Jeske, D. R., and Harville, D. A. (1988), "Prediction-Interval Procedures and (Fixed Effect) Confidence-Interval Procedures for Mixed Linear Models," *Communications in Statistics: Theory and Methods*, 17, 1053–1088.
- Kackar, R. N., and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association*, 79, 853–862.
- Kenward, M. G., and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, 53, 983–997.
- Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.
- Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute, Inc.
- Littell, R. C., Pendergast, J., and Natarajan, R. (2000), "Modelling Covariance Structure in the Analysis of Repeated Measures Data," *Statistics in Medicine*, 19, 1793–1819.
- Littell, R. C., Stroup, W. W., and Freund, R. J. (2002), *SAS for Linear Models* (4th ed.), Cary, NC: SAS Institute, Inc.
- Milliken, G. A., and Johnson, D. E. (1992), *Analysis of Messy Data, Volume 1: Designed Experiments*, New York: Chapman and Hall.
- Prasad, N. G. N., and Rao, J. N. K. (1990), "The Estimation of Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163–171.
- 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, 153–164.
- Rawlings, J. O., Pantula, D. A., and Dickey, D. A. (1998), *Applied Regression Analysis*, New York: Springer.
- Remenga, M. D., and Johnson, D. E. (1995), "A Comparison of Inference Procedures in Unbalanced Split-Plot Designs," *Journal of Statistical Computation and Simulation*, 51, 353–367.
- Satterthwaite, F. W. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114.
- Self, S. G., and Liang, K.-Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.
- Speed, F. M., Hocking, R. R., and Hackney, O. P. (1978), "Methods for Analysis of Linear Models with Unbalanced Data," *Journal of the American Statistical Association*, 73, 105–112.
- Searle, S. R. (1987), *Linear Models for Unbalanced Data*, New York: Wiley.
- Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley.