Abstract

Motivated by a number of drawbacks of classical methods of point estimation, we generalize the definitions of point estimation, and address such notions as unbiasedness and estimation under constraints. The utility of the extension is shown by deriving more reliable estimates for small coefficients of regression models, and for variance components and random effects of mixed models. The extension is in the spirit of generalized confidence intervals introduced by Weerahandi (1993) and should encourage much needed further research in point estimation in unbalanced models, multi-variate models, non-normal models, and non-linear models.

Keywords: Mixed Models, Generalized P-Values, Generalized Confidence Intervals, Bayesian Inference, Unbiased Estimation

Generalized Point Estimation with Application to Small Response Estimation (to appear in Communications in Statistics (T&M))

Samaradasa Weerahandi

November 9, 2010

1 Introduction

This article is motivated by author's unpleasant experience in applying widely used estimation methods in two types of applications in marketing. First of all, when one has to estimate a large number of parameters in a Mixed Model setting, while SAS Proc Mixed often fail or yield unreliable estimates, the Bayesian approach could take as much as one month to estimate just one model. Secondly, widely used classical methods such as the Least Squares Estimation (LSE), and the Maximum Likelihood Estimation (MLE), often yield negative estimates for positive parameters even when the model involve just one covariate. The drawbacks persist even in econometrics, marketing, public health, and other applications where one gets millions of data. For example, in estimating the effect of an advertising campaign, one often gets highly unreliable estimates for the effect. The problem is not limited to econometrics

and marketing applications where the noise tends to be high - in designed experiments, say in clinical trials, one gets observations with less noise, but the problem is equally serious since the sample size in such applications tends to be much smaller compared to marketing and public health related applications.

In recent years, practitioners have alleviated the problem of getting estimates with wrong signs using the Best Linear Unbiased Predictor (BLUP) in Mixed Models. For example, if one attempts to estimate the effect of an advertising campaign by region, say a TV campaign with ad-stocked GRP as the variable of interest, using a simple regression model with only fixed effects, it is quite typical to get as much as 40% of the regional coefficients negative even when there is no reason for such campaign to alienate any demographic segment. As we demonstrate below with some simulated data, it is just an artifact of the nature of normal data and inability of LSE and MLE to take advantage of known signs of parameters. If one models the same data using a mixed model in a hierarchical model setting, where one would treat regional effects as random effects around the national average effect, and estimate parameters using a tool such as SAS Proc Mixed, or S-Plus lme, then most of the estimates would become tight and reliable due to the shrinkage nature of predictors of random effects. Even in this situation, when the consumer response to such campaigns is small, one still gets negative estimates for the fixed effects and hence predictors of random effects also could become unreliable. Moreover, in Mixed Models, it is well known that the standard MLE and unbiased estimates of variance components frequently become zero or negative thus making BLUP unreliable. Use of REML to tackle variance components alleviates the glitch somewhat, but unless the number of factor levels is very large, they also yield zero variance components and unreliable results (cf. Chapter 3 of Weerahandi (2004)). In all these situations, practitioners end up applying ad hoc methods to fix the problems and cherry picking estimates, thus giving a bad name to the Statistical Science.

The Bayesian methods do provide a promising approach to overcome such drawbacks, but they also have their own drawbacks, especially due to treating all model parameters, including nuisance parameters as random variables. Moreover, they tend to be cumbersome and computationally intensive. Some practicing Bayesians do cherry picking via the prior distributions and setting values for hyper parameters. Moreover, when one has to work with millions of data and hundreds of thousands of parameters, as in the case of estimating the individual physician response to detailing (calling on doctors and dropping samples) by representatives of a pharmaceutical company, widely used methods either fail or run for weeks to estimate parameters. In a recent application, having failed to estimate all the parameters simultaneously, the author first estimated the parameters by region, and then estimated the response of individual physicians in each region, one by one, by using a hierarchical Bayesian model and the MCMC algorithm to estimate model parameters. With all the hype about the efficacy of latest algorithms and computer power, it took over one month to complete the estimation of just one model. At the end of that long process, when doing a sanity check one gets an uncomfortable feeling when she sees the sensitivity of estimates when a slightly different prior distribution, say log normal distribution instead of gamma distribution, is used as the prior for positive parameters, or when different values are placed on hyper parameters of no importance.

The main drawbacks of the classical treatment to point estimation are (i) except in problems involving location parameters and in simple academic problems, the MLE based methods are the only systematic method available to tackle any parameter such as a function of variance components, but ML and REML could yield unreliable estimates or fail to estimate BLUP unless the sample size is large, (ii) the classical approach to point estimation does not provide a systematic method to incorporate the knowledge that one may have about the parameter space.

In view of this unsatisfactory state of point estimation of parameters, the purpose of this article is to generalize the classical point estimation methods to provide reliable estimates without resorting to Bayesian methods or resorting to ad hoc methods having no sound theory behind. As a result, one will be able to tackle estimation problems, in hours if not minutes, that may take days or weeks to run by Bayesian methods. The extensions of the classical extensions are in the spirit of Generalized P-values and Generalized Confidence Intervals introduced by Tsui and Weerahandi (1989), and Weerahandi (1993), respectively. We will illustrate the utility of the generalized definitions in point estimation by applying the notions to resolve all the problems discussed above involving estimation of small fixed effects, variance components of Linear Mixed Models, and predictors of random effects of such mixed models.

1.1 Motivating Example

In the applications section below, we will address a variety of situations where LSE and MLE have various drawbacks. To demonstrate one situation involving the problem of estimating small parameters, consider a perfect regression model with just one covariate, say X. Assume that the response variable Y is normal perhaps after a logarithmic transformation. Assume the exact model $Y = \alpha + \beta X + \epsilon$ and that β is a small parameter of known sign, where ϵ is the residual error. Such parameters with known signs arise in practice in Physics, Chemistry, Engineering, and even in econometric and marketing applications. For example, if Y is the gas consumption when X represents the outside temperature of a home, or when we analyze consumer response to a TV advertisement we need to deal with parameters of known signs. To be specific, let us assume that β is supposed to be positive, the assumed linear structure is perfect, and that ϵ is exactly normally distributed, perhaps after a suitable transformation such as the logarithmic transformation, with 0 mean and variance 1. In a simulated experiment, a set of random numbers are gen-

erated from the exact model when $\alpha = 10$ and $\beta = .05$, a typical value when X represents a variable such as a promotional variable in a marketing campaign. In our simulation, ten equally spaced values between 0 and 1 are repeated 50 times for the X variable thus producing a sample of size 500. Shown in table below are the LSEs (also MLEs) of β and its Generalized Estimates (GE) derived in Section 4 obtained in 10 runs from the assumed model.

Table 1.

Performance of GE Vs. LSE in Estimating Small Response of $\beta = .05$

Run #:	1	2	3	4	5	6	7	8	9	10
MLE/LSE of β:	.041	-.263	.232	.069	.153	-.097	.221	.049	.113	.041
GE of β:	.058	.037	.179	.074	.123	.048	.170	.065	.099	.119

Notice that despite the fairly large sample of size 500, LSE yields unreliable estimates with large variance, sometimes even yielding negative estimates for a parameter which is positive. This is partly due to the fact that LSE does not depend on the error variance. In practical applications the error variance tends to be larger than what is assumed here and the sample size could be smaller than that assumed in this illustration and hence could yield highly unreliable estimates. As we will discuss in Section 3, the generalized estimate we derive later in Section 4.4 depends on both the LSE and its variance, and takes advantage of known signs or the range of possible values of the parameters thus producing more reliable estimates as evident from the above table. Both sets of estimators are consistent in that they tend to the true value of the parameter as the sample size becomes large.

Of course LSE is unbiased and hence if the above experiment is repeated for a large number of runs, then the average value of estimates will be the actual value of $\beta = .05$, but in real world applications we get only one set of data and hence we do not get the benefit of average performance. Of course unbiasedness is an important property to have since it assures that by using an

unbiased estimator, practitioners will not over estimate or under estimate on average. In Section 4 we will provide further simulated studies for situations where the MLE is highly bias, classical unbiased estimators do not exist, and the proposed GE is close to being unbiased.

2 Generalized Estimation

Motivated by the Generalized Chow Test derived by Weerahandi (1987), Tsui and Weerahandi (1989) and Weerahandi (1993) extended the classical definitions of statistical tests and confidence intervals so that one could obtain reasonable interval estimates for situations where the classical approach fails or yield results lacking small sample accuracy. While these methods are referred to as exact methods since they are based on exact probability statements, Hanning, Iyer and Patterson (2006) showed that usually they are also asymptotically exact in a classical sense. Roy and Bose (2007) further refined the generalized intervals to make them perform better with small sample sizes. Extended definitions have been successfully applied by Gamage et al (2004) in solving the Multi-Variate Behrens-Fisher problem, by Lee and Lin (2004) in constructing intervals for the ratio of two normal means, by Roy and Mathew (2005) in two parameter exponential distributions, Krishnamoorthy et al (2006), by Chen and Zhou (2006) in dealing with log normal means, and by Bebu et al (2009) in dealing with ratios of two regression coefficients, and so on. The generalized confidence intervals are obtained by means of generalized pivotal quantities, which are counterparts of classical pivotal quantities, when such quantities are allowed to be functions of observed values and nuisance parameters as well. Here we take a similar approach by letting the extended version of classical estimators to be functions of the observable random vectors, observed values, and underlying parameters, just like posteriors of parameters in a Bayesian treatment of the problem, but without having to

treat constant parameters as random variables thus avoiding the burden of too many unknown parameters and too many random variables to deal with.

2.1 Definitions and Notions

Consider an observable random vector $\mathbf{X}$ representing a certain population. Since the concept of *Minimum Sufficient Statistics* allows us to simplify inference problems to base on a few summary statistics, we can even assume that $\mathbf{X}$ is a set of minimum sufficient statistics. The form of the distribution of $\mathbf{X}$ is assumed to be known except for certain unknown parameters, say $\boldsymbol{\zeta} = (\theta, \boldsymbol{\delta})$ where θ is one or a vector of parameters of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters. Let Ψ be the sample space of possible values of $\mathbf{X}$, let Ξ be the parameter space of $\boldsymbol{\zeta}$, and let Θ be the parameter space of θ. The observed value of the random vector $\mathbf{X}$ is denoted by the lower case $\mathbf{x}$. Consider the problem of estimating the parameter θ. In the classical approach to estimation, an estimator is required to be an observable random quantity that is free of all unknown parameters. As in the generalized approach to hypotheses testing and interval estimation, and in the Bayesian approach, we relax this requirement so that we will have a greater class of estimators to choose from. This is accomplished by allowing an estimator-like quantity to depend on all unknown parameters and even the currently observed sample, but requiring certain properties.

Definition 1 *A random quantity $Q = Q(\mathbf{X}, \mathbf{x}, \boldsymbol{\zeta})$, a map of $\Psi \times \Psi \times \Xi$ to a Euclidean space, is said to be a Generalized Estimator for θ if it satisfies the following properties:*

1. The cumulative distribution of Q is a monotonic function of θ,
2. $Q(x, x, \zeta) = c$, where c is a constant free of nuisance parameters, but possibly

could depend on observed x and θ.

An optional, but desirable additional property to have is

3. The distribution of Q is free of nuisance parameters.

Property 1 will assure that the location parameters (mean, median, etc.) of Q or its monotonic transformations will increase as θ increases. Property 2 is imposed to assure that such quantities and even other inferences based on Q will dot depend on nuisance parameters. Property 3 above is essential when we need to make additional inferences such as statistical tests and interval estimation based on Q . In fact, it is easily verified that if Q is a generalized estimator satisfying Property 3, then it is also a *Generalized Pivotal Quantity*, as defined by Weerahandi (1993). Moreover, $T = Q - q$ is a *Generalized Test Variable*, where q is the observed value of Q. When such additional inferences are of secondary importance, it is desirable to have at least the milder property that the expected value of Q is free of nuisance parameters. For example, in sampling from a normal population, say $N(\mu, \sigma^2)$, the random variable $Q_1 = \overline{X} - \mu$ is a generalized estimator having just the first two properties above, whereas the generalized estimator $Q_2 = \overline{x} - s(\overline{X} - \mu)/S$ has all three properties, where $\overline{X}$ is the sample mean and S is the sample standard deviation.

Definition 2 *Let Q be a generalized estimator for a parameter θ. Suppose $t(Y)$ is a monotonic real-valued function free of unknown parameters, where Y could possibly depend on random variables free of unknown parameters. Then, $t()$ is said to be an e-transformation (abbreviated for estimator transformation) of Q if $t(Q)$ is also a generalized estimator.*

For example, $Q_1' = \exp(\overline{X} - \mu)$ obtained by transforming Q_1 above is also generalized estimator having just the first two properties, whereas the generalized estimator $Q_2' = \exp(\overline{x} - s(\overline{X} - \mu)/S)$ has all three properties.

Proposition 3 *If Q is a generalized estimator for a parameter θ and if $t()$ is an e-transformation, then for any given confidence level γ, $t(Q)$ and Q can generate identical one-sided confidence intervals.*

The proof of the proposition trivial, because the monotonicity $t()$ leads to identical probability statements such as $\Pr(t(Q) \leq t(q)) = \Pr(Q \leq q)$ in generalized interval estimation. The proposition 1 implies that e-transformations of generalized estimators can produce equivalent tests as well. However, as in the classical estimation, given one generalized estimator, by means of e-transformations one can find a class of generalized estimators yielding different estimates. In the development below we assume that the generalized estimator of interest has already been appropriately transformed to achieve certain desirable properties such as the unbiasedness, which only certain transformations will have.

Without loss of generality, we can take the quantity c to be 0 in Definition 1. When necessary, it can also be set to be equal to such quantities as a function of $\mathbf{x}$ or θ, the parameter of interest. Obviously any classical estimator $\widehat{\theta}(\mathbf{x})$ leads to a generalized estimator, because $Q = \widehat{\theta}(\mathbf{X}) - \widehat{\theta}(\mathbf{x})$ is a random variable satisfying the first two properties of Definition 1. Moreover, as in the classical approach to point estimation, one can construct a variety of estimates based on Q, including the MLE, by imposing additional requirements. To a minimum, one would require the property of *Consistency* that such estimators will tend to θ in probability as the sample size goes to ∞. In practical applications one can reduce the search for good estimates to those based on sufficient statistics. We provide a method below to further facilitate the selection of a suitable estimator based on such sufficient statistics. Yet, as we will clarify in applications below, still one will have a choice of estimators, and so she will have to impose further requirements of practical importance such as minimizing the *Mean Squared Error* (MSE), $MSE = E((\widehat{\theta} - \theta)^2)$ or similar other metrics based on appropriate loss functions.

Being an extension of classical estimators, the Generalized Estimators can produce better or worse estimators than a given classical estimator. In providing better estimators for small samples, there are a few simple ways to

establish such asymptotic properties as Consistency and unbiasedness. For instance, one can try showing that under a certain diffuse prior, the proposed estimator is equivalent to a Bayesian estimator possessing such properties. Even better, one can simply show that the MSE of the proposed estimator is no bigger (and strictly smaller with smaller samples) than that of the MLE, the estimator that we will improve upon in each of the applications undertaken in the article. Then, the following proposition would guarantee that the estimator is consistent, asymptotically unbiased, and more efficient.

Proposition 4 *Let $\widetilde{\theta}$ be the MLE or any other asymptotically unbiased and consistent estimator of θ. If $\widehat{\theta}$ is a generalized estimator such that $E((\widehat{\theta}-\theta)^2) \leq E((\widetilde{\theta}-\theta)^2)$ for all $\theta \in \Theta_0$, then $\widehat{\theta}$ is an asymptotically unbiased and more efficient consistent estimator of $\theta \in \Theta_0$.*

The proof of the proposition trivially follows from the decomposition of each MSE into two non-negative quantities as in, $E((\widetilde{\theta}-\theta)^2) = Var(\widetilde{\theta}) + (E(\widetilde{\theta})-\theta)^2$, and from the Chebyshev inequality.

Of particular interest are the estimates that are based on the expected value of Q or that of an e-transformation, the median of Q when the observed value of Q is θ, and so on. In the treatment below, we assume that the generalized estimator has been already transformed to achieve desirable properties or objectives.

Definition 5 *Let $Q = Q(\mathbf{X}, \mathbf{x}, \boldsymbol{\zeta})$ be a generalized estimator with $c = Q(\mathbf{x}, \mathbf{x}, \boldsymbol{\zeta}) = 0$. Then, a solution of the equation*

$$E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})) = 0 \tag{1}$$

for θ is said to be a Generalized Estimate of θ based on the expected value of Q.

Note that any conventional unbiased estimator is a generalized estimator, because if $\widehat{\theta}(\mathbf{X})$ is an unbiased estimator of θ, then the solution of the equation

$$E(\widehat{\theta}(\mathbf{X}) - \widehat{\theta}(\mathbf{x})) = \theta - \widehat{\theta}(\mathbf{x}) = 0 \tag{2}$$

for θ itself is a classical unbiased estimator. However, a generalized estimate, say $\widehat{\theta}(\mathbf{x})$ obtained by solving the above extension of the classical definition does not necessarily have the property that $E(\widehat{\theta}(\mathbf{X})) = \theta$. Nevertheless, in many applications it is possible to achieve unbiasedness (or close unbiasedness when unbiased estimators do not exist) by means of a suitable e-transformation, as done in most applications undertaken in this paper. In dealing with location parameters it is customary to derive unbiased estimates with additional properties by such methods as BLUE and MVUE. In complicated estimation problems, perhaps the most convenient way to establish desirable properties, such as unbiasedness and minimum MSE, in this computer age is via simulation, as we will do in our illustrations.

The generalized approach will prove to be useful especially when MLEs and LSEs have undesirable properties or when unbiased estimators do not exist. In particular, the generalized definition of point estimators allow us to take advantage of the knowledge that a practitioner may have concerning the parameter space, or feasible values of the parameter, without having to treat non-random parameters as random variables, as required by the Bayesian treatment of the problem, or taking ad hoc approaches. For example, in Mixed models, variance components are positive parameters, but the classical approach frequently yield negative estimates, or one has to artificially truncate them to zero. In many other applications such as that of studying the effect of a TV advertisement, the parameters of interest are supposed to be positive, and yet LSE frequently yields negative estimates. When we have such knowledge about the parameters, some of the drawbacks of the classical estimators could be alleviated via conditional estimation.

2.2 Estimation Under Constrains

In some applications, the practitioners may have specific knowledge about the parameter space. For example, variance components are supposed to be

positive quantities and certain variance ratios are supposed to be bounded by 0 and 1. Such articles as Mandelkern (2002) deal with bounded parameters, and deal with such inference problems as interval estimation, and here we will address the problem of point estimation.

When dealing with problems involving bounded parameters, it is desirable to assure that estimates we get with any set of data observed from a model with such bounded parameters will also fall within the same boundary. When the classical interval estimators fails to satisfy such properties, Kiefer (1977) showed how one could avoid such drawbacks via conditioning. Weerahandi (1994, Chapter 9) showed how the conditioning approach could be employed in Generalized Interval Estimation to avoid negative confidence bounds for variance components. By taking a parallel approach we can take advantage of the knowledge of the parameter space by conditioning on the generalized estimators.

Definition 6 *Let Θ_0 be a subset of the parameter space of a parameter θ, and X_0 be a subset of the sample space, and let $Q(\mathbf{X}, \mathbf{x}, \boldsymbol{\zeta})$ be generalized estimator of θ satisfying all three properties of Definition 1. Assume without loss of generality that the observed value of Q is 0. If the solution, $\widehat{\theta}(\mathbf{x})$, of the equation*

$$E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|\mathbf{X} \in X_0) = 0 \tag{3}$$

falls in Θ_0 for all values of $\theta \in \Theta_0$, then $\widehat{\theta}(\mathbf{x})$ is said to be a Conditional Generalized Estimate given $\theta \in \Theta_0$.

It should be noted that, as in the case of Bayesian estimation and in decision theoretic approach to estimation under loss functions, such conditional estimation with small samples could produce bias estimates in a classical sense. However, such conditioning could produce highly superior consistent estimates in terms of more important metrics such as the Mean Squared Error (MSE). To see why smaller MSE is more important than the unbiasedness, when we

have a random sample from a normal population if we throw out half of the data and estimate the population mean by the sample mean of the remaining sample, the resulting estimate will still be unbiased and consistent. However, in terms of the MSE that estimate will be much inferior compared to the mean of the whole sample. Therefore, when dealing with conditional estimators below, in the spirit of the decision theoretic approach to point estimation, we will use such metrics as the MSE to measure how tight the estimate from the true value of the parameter is. In some applications, of course it is desirable to have both properties, an area requiring further research as we discuss in the last section.

A special version of conditional generalized estimators when the observed value of Q is θ is of particular interest. When we have some information about the parameter, as evident from the proposition below, Definition 5 yields estimates with a highly desirable property of generalized estimators.

Proposition 7 *Suppose that $Q(\mathbf{X}, \mathbf{x}, \boldsymbol{\zeta})$ is a generalized estimator with distribution free of unknown parameters, its observed value is θ, and that θ is known to belong to the interval $[a, b]$. Then, the solution of the equation $E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|Q \in [a, b]) = \theta$, say $\widehat{\theta}(\mathbf{x})$, is a conditional generalized estimate that also falls on the interval $[a, b]$.*

Proof. *Note that $Q' = Q - \theta$ is a generalized estimator with the observed value 0 and that*

$$E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|Q \in [a, b]) = \frac{\int_a^b qf(q)dq}{\int_a^b f(q)dq} \leq \frac{\int_a^b bf(q)dq}{\int_a^b f(q)dq} = b.$$

Similarly $E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|Q \in [a, b]) \geq a$.Moreover, since Q is free of unknown parameters, the generalized estimate of θ given by Q as well as Q' is $\widehat{\theta}(\mathbf{x}) = E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|Q \in [a, b])$. Therefore, $\widehat{\theta}(\mathbf{x}) \in [a, b]$. ■

While the above proposition is convenient in deriving conditional estimates when it permits, except in dealing with location parameters, it is too restrictive

to have the requirement that the observed value of the generalized estimator as θ. Then, as illustrated in applications undertaken in later sections of this article, one can use variations of the proposition without imposing the requirement that the observed value of Q is θ. The following proposition was found to be adequate in the applications undertaken in this article.

Proposition 8 *Suppose that $Q(\mathbf{X}, \mathbf{x}, \boldsymbol{\zeta})$ is a generalized estimator with its distribution free of unknown parameters, its observed value 0, and that θ is known to belong to the interval $[a, b]$. Suppose Q can be expressed in the form $Q = Q_1(Q_2 - \theta)$, where Q_1 is a random variable free of unknown parameters and $E(Q_1) > 0$. Then, the solution of the equation $E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|Q_2 \in [a, b]) = 0$, say $\widehat{\theta}(\mathbf{x})$, is a conditional generalized estimate that also falls on the interval $[a, b]$.*

Proof. *Note that*

$$\begin{aligned} E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|Q_2 \in [a, b]) &= E(Q_1 \frac{\int_a^b (q_2 - \theta) f(q_2|q_1) dq_2}{\int_a^b f(q_2|q_1) dq_2}) \\ &\leq E(Q_1 \frac{\int_a^b (b - \theta) f(q_2|q_1) dq_2}{\int_a^b f(q_2|q_1) dq_2}) \\ &= E(Q_1(b - \theta)) = (b - \theta) E(Q_1). \end{aligned}$$

Hence, the solution of the equation $E(Q(\mathbf{X}, \mathbf{x}, \theta, \boldsymbol{\delta})|Q_2 \in [a, b]) = 0$ satisfies the inequality $\widehat{\theta}(\mathbf{x}) \leq b$. Similarly $\widehat{\theta}(\mathbf{x}) \geq a$, and hence, $\widehat{\theta}(\mathbf{x}) \in [a, b]$. ■

2.3 Finding Generalized Estimators

Once the inference problem has been reduced by means of a set of minimum sufficient statistics, one can easily deduce generalized estimators from generalized pivotal quantities that researchers have now found for a variety of practical

applications, such as that by Lee and Lin (2004) in constructing intervals for the ratio of two normal means, by Roy and Mathew (2005) in two parameter exponential distributions, by Ananda (2003) and by Chen and Zhou (2006) in dealing with log normal means, and by Mu et al (2008) in unbalanced two-way ANOVA, and so on. Nevertheless, it is desirable to provide a self-contained approach to find such variables to tackle many problems. It should be emphasized that, except in dealing with location parameters (for which, such methods as BLUE and LSE are available), not even the classical approach to unbiased estimation provides a systematic method to find potential estimators. In the context of interval estimation Weerahandi (2004) provided a method called the substitution method to facilitate the problem of finding generalized pivotal quantities. By taking a similar approach we can modify the substitution method in the current context to find generalized estimators and hence generalized estimators of parameters satisfying all three properties.

Let $\mathbf{V} = \mathbf{V}(\boldsymbol{\zeta})$ be a set of random variables with distributions free of unknown parameters obtained by transforming the sufficient statistics $\mathbf{X}$; the reader is referred to articles in generalized inference discussed by Weerahandi (2004) to get details on how to find such variables. It is assumed that the joint distribution of the random vector $\mathbf{V}$ is known. The *Substitution Method* in the context of point estimation is then carried out in the following steps:

- Express the parameters ζ, and then θ in terms $\boldsymbol{X}$ and the random variables $\boldsymbol{V}$,

- Define a potential generalized estimator, say Q_1, by replacing the statistics $\boldsymbol{X}$ by the observed value $\boldsymbol{x}$ and argue that its distribution is free of unknown parameters,

- Rewrite V terms appearing in Q_1 in terms of $\boldsymbol{X}$ and and show that when $\boldsymbol{X} = \boldsymbol{x}$, the observed value of the quantity $Q_1(\boldsymbol{x}; \boldsymbol{x}, \boldsymbol{\zeta})$ does not depend on nuisance parameters,

- Define an equivalent generalized estimator as $Q_2 = \theta - Q_1(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\zeta})$,
- By rearranging the V terms in Q_2 by intuition or trial and error, find a suitable e-transformation, $t()$ to achieve such desirable properties as the unbiasedness and minimum mean squared error, and define an appropriate generalized estimator as $Q = t(Q_2)$.

The above approach usually yields a few estimators to choose from, a situation much better than having to guess by intuition when the parameter of interest is not a simple location parameter. When the choice is not unique, one can find a unique estimator by requiring additional desirable properties such as the consistency, the invariance, the unbiasedness, and the minimum mean squared error - the approach will be illustrated in our applications discussed below. It should also be noted that, although the above approach works in many applications, not in all situations that one can find a one-to-one correspondence between the number of unknown parameters and the number of $\mathbf{V}$ random variables. Moreover, in unbalanced models usually the substitution method does not really work. However, once we establish the structure of generalized estimators in special cases such as in balanced models, then usually it is easy to construct the counter parts for general models, as we will illustrate later in our applications.

3 Estimation Issues in Mixed Models

The first three applications of generalized estimation undertaken in the next section deal with mixed models, which have become very popular lately for estimating parameters by a large number of segments. For example, if a practitioner wishes to estimate the price elasticity of demand for a consumer product by region (or by individual store of a chain of stores), or estimate the effect of a marketing campaign by region, estimating parameters region by region will yield highly unreliable results. In a typical application of these kinds, one has

to estimate thousands of parameters using longitudinal data on the response variable by regressing on variables of interest or ad-stocked variables in the case of promotions. If one attempts to estimate such parameters by means of a simple regression model, LSE would basically yield highly unreliable estimates, including a large fraction of estimates for parameters with even the wrong sign. Therefore, in such situations practitioners now use Mixed Models to better structure the model.

To briefly introduce the estimation issues in mixed models, consider a mixed linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{y}$ is a vector of n observations of the response variable, $\mathbf{X}$ is a $n \times k$ matrix of covariates with fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}$ is $n \times r$ a matrix of coviatites with random effects $\boldsymbol{\alpha}$, and $\boldsymbol{\epsilon}$ is the vector of residuals. It is further assumed that $\boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_e)$, $\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$ are independently distributed. Usually the vector $\boldsymbol{\beta}$ includes the fixed effects, say $\boldsymbol{\mu}$ around which the original random effects, say $\boldsymbol{\alpha} = \boldsymbol{\mu}+\boldsymbol{u}$ are distributed, and $\mathbf{Z}$ is the corresponding design matrix. Either one has to treat $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Sigma}_\alpha$ as known covariance matrices or rather provide a structure to them with a few unknown parameters. Most widely used structure for $\boldsymbol{\Sigma}_e$ is $\boldsymbol{\Sigma}_e = \sigma_e^2\mathbf{I}_n$ and that for $\boldsymbol{\Sigma}_\alpha$ are $\sigma_\alpha^2\mathbf{I}_r$, $\mathbf{I}_k \otimes \sigma_k^2\mathbf{I}_l$, and $\sigma_\alpha^2\mathbf{A}$, where $\mathbf{A}$ is a known $r \times r$ matrix referred to as *a relationship matrix*, as introduced by Henderson (1975) in his pioneering research in the animal breeding context.

Since $\mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_e + \mathbf{Z}\boldsymbol{\Sigma}_\alpha\mathbf{Z}')$, the GLSE of $\boldsymbol{\beta}$ is easily obtained as $\widehat{\boldsymbol{\beta}} = (\ \mathbf{X}'\boldsymbol{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Delta}^{-1}\mathbf{y}$,where $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_e + \mathbf{Z}\boldsymbol{\Sigma}_\alpha\mathbf{Z}'$.With data from designed experiments yielding orthogonal design matrices, under independent errors with constant variance, the matrix $\boldsymbol{\Delta}$ will have the compound symmetric covariance structure, and then GLSE of $\boldsymbol{\beta}$ reduces to the ordinary LSE $\widehat{\boldsymbol{\beta}} = (\ \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. With other structures of the covariance, one will have to estimate such nuisance parameters and approximate the GLSE by plugging in

the estimates in place of the variance terms in $\mathbf{\Delta}$.

Henderson (1975) derived a predictor (estimator of random effects) of $\boldsymbol{u}$ as well as the estimate of $\boldsymbol{\beta}$ using the best linear unbiased estimation approach. In the applications undertaken below we need to know the distribution of the predictor as well, and therefore it is more convenient to derive the predictor using basics. From the joint distribution of $\boldsymbol{u}$ and $\mathbf{y}$, it is easy to see that, the conditional distribution $\boldsymbol{u}$ given $\mathbf{y}$ is

$$\boldsymbol{u}|\mathbf{y} \sim \mathbf{N}(\mathbf{\Sigma_\alpha Z'\Delta^{-1}(y - X\boldsymbol{\beta}),(\Sigma_\alpha^{-1} + Z'\Sigma_e^{-1}Z)^{-1}}). \tag{5}$$

Hence the predictor of the vector of random effects is

$$\widehat{\boldsymbol{u}} = E(\boldsymbol{u}|\mathbf{y}) = \mathbf{\Sigma}_\alpha \mathbf{Z}'\mathbf{\Delta}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\mathbf{.} \tag{6}$$

3.1 Estimating Segment Coefficients

In an application involving a linear model, consider a variable, say $\mathbf{Z}$, whose coefficient we wish to estimate by a number of segments, as in the problem of estimating the price elasticity of demand by region, as described above. Suppose one wishes to obtain more reliable estimators than those given by LSE by assuming a mixed model of the form

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i u_i + \boldsymbol{\epsilon}_i \tag{7}$$

where $\mathbf{Z}_i$ is the $\mathbf{Z}$ observations from segment i, $u_i \sim N(0, \sigma_a^2)$ is its random effect, and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_{n_i})$. One of the simplest versions of the segment level model is the one-way layout model

$$Y_{ij} = \beta + \alpha_i + \in_{ij}, \tag{8}$$

where β is a fixed effect and α_i is a random effect. A more general version of (7) is the situation where we need to randomize a number of variables, including interactions, by segment. Then, (7) can be easily extended as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_j \mathbf{Z}_{ij} u_{ij} + \boldsymbol{\epsilon}_i, \tag{9}$$

where $u_{ij} \sim N(0, \sigma^2_{a_j})$.

The parameter vector $\boldsymbol{\beta}$ contains the fixed effect of $\mathbf{Z}$, say $\boldsymbol{\mu}$, common to all segments. Assume that $\boldsymbol{\beta}$ has already been estimated by $\widehat{\boldsymbol{\beta}}$ using estimates of variance components, a problem we will further examine in Section 5. Then, suppressing the subscript i for the purpose of dealing with one segment, and letting $\mathbf{y} = \mathbf{y}_i$, $\mathbf{z} = \mathbf{Z}_i$ and $\mathbf{x} = \mathbf{X}_i$, the parameter $\alpha_i = \mu + u_i$ represents the total effect of the randomized variable of interest, where μ is the common effect the variable, usually one parameter of the vector $\boldsymbol{\beta}$. Since $\mathbf{z}$ is a $n_i \times 1$ vector, from the above results we can deduce the predictor of the random effect as

$$\begin{aligned} E(u|\mathbf{y}) &= \sigma^2_a \mathbf{z}'(\sigma^2_e \mathbf{I} + \boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\alpha}} \mathbf{z}\mathbf{z}')^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) = \boldsymbol{\varsigma}\mathbf{z}'(\mathbf{I} - \frac{\boldsymbol{\varsigma}\mathbf{z}\mathbf{z}'}{1 + \boldsymbol{\varsigma}\mathbf{z}'\mathbf{z}})\boldsymbol{\delta} \\ &= \boldsymbol{\varsigma}(\frac{\mathbf{z}'\boldsymbol{\delta}}{1 + \boldsymbol{\varsigma}\mathbf{z}'\mathbf{z}}) = \frac{\mathbf{z}'\boldsymbol{\delta}}{\mathbf{z}'\mathbf{z}}(1 - \frac{\sigma^2_e}{\sigma^2_e + \mathbf{z}'\mathbf{z}\sigma^2_a}), \end{aligned} \tag{10}$$

where $\boldsymbol{\varsigma} = \boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\alpha}}/\sigma^2_e$ and $\boldsymbol{\delta} = \mathbf{y} - \mathbf{x}\boldsymbol{\beta}$ is the vector of observed segment i deviations from the assumed model. Therefore, the the coefficient of the variable of interest estimated by the mixed model approach is

$$\widehat{\alpha}_i = \mu + \widehat{u}_i(1 - \frac{\sigma^2_e}{\sigma^2_e + \mathbf{z}'\mathbf{z}\sigma^2_a}), \tag{11}$$

where $\widehat{u}_i = \mathbf{z}'\boldsymbol{\delta}/\mathbf{z}'\mathbf{z}$. From the above representation it is evident that, unlike the LSE of the coefficient of $\mathbf{Z}_i$, $\widehat{\alpha}$ given by (11) is a shrinkage estimator that pulls the segment estimates towards the common effect μ. As a result, one gets more reliable tight estimates compared to the estimates she gets in segment by segment estimation by LSE. This is the main reason why mixed models have become very popular lately.

Now consider the problem of obtaining unbiased predictors for α when the parameters $\boldsymbol{\beta}$, σ^2_e, and σ^2_a are unknown. Assume that an unbiased estimator or an almost unbiased estimator of $\boldsymbol{\beta}$ is available. Since the role that the parameters $\boldsymbol{\beta}$ and μ play have a linear structure, we can simply replace the unknown parameters by their estimates $\widehat{\boldsymbol{\beta}}$ and maintain the close unbiasedness of the resulting predictor of α. Now it is clear that the problem of obtaining

unbiased estimators reduces to the problem of estimating parameters of the form $\rho = \sigma_e^2/(\sigma_e^2 + A\sigma_a^2)$.

For additional inferences concerning random effects we also need to estimate the variance of α. When we are dealing with one set of random effects, the variance of interest can be deduced from (5) as

$$\theta = Var(\boldsymbol{u}|\mathbf{y}) = \left(\frac{1}{\sigma_a^2} + \frac{1}{\mathbf{z}'\mathbf{z}/\sigma_e^2}\right)^{-1}. \tag{12}$$

One handles the variances appearing in above estimators by simply plugging in their MLEs. Such estimators and their variances based on asymptotic results are known (cf. Weerahandi (2004)) to be very unreliable. Moreover, they produce highly bias estimators even when the covariance matrices have just two unknown parameters and there is just one set of random effects. So, in the following applications of generalized estimation, we will derive estimators which are either unbiased or close to being unbiased when unbiased estimators do not exist.

4 Estimating Parameters of Mixed Models

In this section we consider a number of applications of immense practical importance in linear regression that arise in estimating the coefficients of a certain variable by a number of segments formulated as a mixed model, or coefficients of small magnitude of a regular liner model or a mixed linear model. We will deal with four types of estimation problems that arise in estimating parameters of Mixed Models. For simplicity in our illustrations, let us first consider the balanced models in canonical form (cf. Weerahandi (2004)), for which simple distributional results are available. Then, in the next section we will address the estimation problems in unbalanced models including those arise in regression models with random effects as discussed above. The first three applications in this section are useful in a variety of random effects

models and mixed models regardless of the number of fixed effects. And our last application below will prove useful even in a simple linear regression with just one covariate or in dealing with one parameter of a mixed model.

In our illustrations below involving variance components and location parameters, we derive and compare certain generalized estimates with maximum likelihood based methods. When we deal with functions of variance components, those estimates obtained by simple substitution of maximum likelihood estimates of σ_e^2 and $\sigma_e^2 + A\sigma_a^2$ will be referred to as MLEs. The estimates available from such software procedures as SAS Proc Mixed that truncate negative MLEs to zero and those obtained by restricted (or residual, or reduced) maximum likelihood method will be referred to as ML and REML, respectively. ML and REML do not produce predictors of random effects when MLE is negative, which frequently occur with small samples. Therefore, we defer discussion of ML and REML until Section 4.3.

4.1 Estimating Predictors of Random Effects

Consider the problem of obtaining predictors of random effects of a mixed model that arise in a balanced designed experiment. Since this is a problem involving location parameters, unbiased estimators are desirable. As illustrated in the previous section, the problem of unbiased estimation of the predictor of the random effect can be reduced to the unbiased estimation of ratios of variance components of the form

$$\rho = \frac{\sigma_e^2}{\sigma_e^2 + B\sigma_a^2} = \frac{\sigma_e^2}{\sigma_t^2},$$

a parameter of interest in other applications such as industrial quality control (cf. Hamada and Weerahandi (2000)) as well, where $\sigma_t^2 = \sigma_e^2 + B\sigma_a^2$, where B is a positive constant thus implying that $\rho \in (0, 1)$. Consider the problem of estimating ρ when we have unbiased estimators (also MLEs) of the variance components $\widehat{\sigma}_e^2 = S_e/e$ and $\widehat{\sigma}_t^2 = S_a/a$ distributed as

$$U = \frac{S_e}{\sigma_e^2} \sim \chi_e^2 \tag{13}$$

$$V = \frac{S_a}{\sigma_e^2 + A\sigma_a^2} \sim \chi_a^2, \tag{14}$$

where σ_e^2, σ_a^2 are the variance components of the problem. For example, in the one-way random effects model involving k groups with n observations from each group, $A = n$, $a = k-1$, $e = nk-k$. Although the generalized estimation allows us to obtain unbiased estimates of any ratio of variance components, in this first illustration let us assume for convenience that $B = A$ as in the canonical form.

When nuisance parameters σ_e^2 and σ_a^2 are unknown, then one may simply replace the unknown parameters by these estimates and obtain an estimate of α as well as ρ, the MLE of each parameter. Since random effects arose from a linear model it is reasonable to look for unbiased estimators when available.

In any case, let us consider the nature of generalized estimators. By means of the substitution method or otherwise, we can obtain a potential generalized estimator as

$$Q = \rho\frac{S_e/\sigma_e^2}{S_a/\sigma_t^2} - \frac{s_e}{s_a} = \rho\frac{U}{V} - \frac{s_e}{s_a} \tag{15}$$

$$= \frac{e}{a}(\rho W - \frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_t^2}), \tag{16}$$

where W is distributed as F with e and a degrees of freedom; i.e. $W \sim F_{e.a}$. From the second representation of Q in (15) it follows that the distribution of Q is free of nuisance parameters and that it is an increasing function of ρ. And from the first representation of Q it is clear the the observed value of Q is 0. Hence, Q is indeed a generalized estimator. Hence, the generalized estimator suggested by the expected value of Q is the solution of the equation

$$0 = E(\rho W - \frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_t^2}) = \rho\frac{a}{a-2} - \frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_t^2}.$$

Hence the generalized estimator of ρ based on the proposed Q is

$$\widehat{\rho} = \frac{(a-2)}{a}\frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_t^2}. \tag{17}$$

It is easy to verify that when $B = A$ this is the standard unbiased estimator that one can directly obtain using W, but we were able to obtain it by the substitution method rather than by guessing. As evident from the simulation comparison given by the table below, the estimator given by (17), say GE, has much smaller MSE than the MLE, and that MLE tends to be bias when the number of segments is not large.

Table 2. **MSE Performance of MLE and GE Under $(\sigma_\alpha^2, \sigma_e^2, e, a)$ Scenarios**

Scenario:	.1,1,20,4	.1,1,40,9	.1,2,20,4	.1,2,40,9	.5,2,20,4	.5,2,40,9	1,2,20,4	1,2,40,9
MLE:	3.53	0.460	14.8	0.095	58.1	0.181	3.22	0.077
GE:	0.78	0.249	3.51	0.055	14.4	0.100	0.77	0.042

Although the generalized estimator has superior MSE performance, it does not exists if $a < 3$; i.e. when $k < 4$ in the one-way layout. One can obtain alternative estimates based on e-transformations of Q. For example, $Q_1 = \rho - \frac{s_e}{s_a}\frac{S_a/\sigma_t^2}{S_e/\sigma_e^2} = \rho - \frac{\widehat{\sigma}_t^2}{\widehat{\sigma}_e^2}E(\frac{1}{W})$ and $Q_2 = \rho s_a \frac{\sigma_t^2}{S_a} - s_e\frac{\sigma_e^2}{S_e} = \rho\frac{s_a}{V} - \frac{s_e}{U}$ are also generalized estimators yielding the generalized estimators $\widehat{\rho}_1 = \frac{e}{(e-2)}\frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_t^2}$ and $\widehat{\rho}_2 = \frac{e(a-2)}{a(e-2)}\frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_t^2}$, respectively. Since e is a parameter depending on the total sample size, $\widehat{\rho}_1$ is closely agree with the MLE, whereas $\widehat{\rho}_2$ is an estimator that closely agrees with the unbiased estimator. The answer to the question of which estimator is better depends on the objective of the estimation, but in most applications when one does not have a well defined loss function, the clear choices are unbiased estimators with minimum MSE - such properties as BLUE and LSE are appropriate and apply only dealing with location parameters of linear models. Therefore, in our illustrations below we will seek estimators with superior MSE performance compared to widely used LSEs and MLEs. We will seek unbiasedness, when that property does not lead to estimators with wrong signs or possess other undesirable properties.

4.2 Estimating the Conditional Variance of Random Effects

Now consider the problem of estimating the conditional variance given by (12) of the form

$$\upsilon = \frac{1}{A/\sigma_e^2 + 1/\sigma_a^2} = \frac{\sigma_e^2\sigma_a^2}{\sigma_t^2}. \tag{18}$$

For this problem there is no unbiased estimator available and so let us use the generalized estimation method to find an estimator that is close to being unbiased than the MLE

$$\widehat{\upsilon} = \frac{\widehat{\sigma}_e^2(\widehat{\sigma}_t^2 - \widehat{\sigma}_e^2)}{A\widehat{\sigma}_t^2}. \tag{19}$$

as we demonstrate later in this section. Moreover, obtaining Bayesian estimates using typical priors in this type of applications is very cumbersome and computationally very intensive. So let us try to find a generalized estimator having better MSE performance without having to deal with any numerical integrations as the Bayesian approach requires. For now, let us do so with no regard to our knowledge that υ is supposed to be a positive random variable, an issue we will address in the next section in a context, in which the problem is more serious.

By applying the substitution method with an appropriate e-transformation, or otherwise, consider the generalized estimator

$$\begin{aligned} Q &= \upsilon A\frac{s_a}{s_e}\frac{S_e/\sigma_e^2}{S_a/\sigma_t^2} - \frac{s_a}{S_a/\sigma_t^2} + \frac{s_e}{S_e/\sigma_e^2} \\ &= \upsilon A\frac{s_a}{s_e}\frac{U}{V} - \frac{s_a}{V} + \frac{s_e}{U} \\ &= \upsilon A\frac{\widehat{\sigma}_t^2}{\widehat{\sigma}_e^2}W - a\frac{\widehat{\sigma}_t^2}{V} + e\frac{\widehat{\sigma}_e^2}{U} \end{aligned}$$

It is evident from the above representations of Q that it is a generalized estimator having all three properties of Definition 1. We can derive various

generalized estimators using its direct expected value or expected values of its e-transformations. When $a > 3$ as in the previous section, we can search for a close unbiased estimator by taking the direct expected value and solving the equation

$$\begin{aligned} 0 &= E(\upsilon A\frac{\widehat{\sigma}_t^2}{\widehat{\sigma}_e^2}W - a\frac{\widehat{\sigma}_t^2}{V} + e\frac{\widehat{\sigma}_e^2}{U}) \\ &= \upsilon A\frac{\widehat{\sigma}_t^2}{\widehat{\sigma}_e^2}\frac{a}{a-2} - \frac{a\widehat{\sigma}_t^2}{a-2} + \frac{e\widehat{\sigma}_e^2}{e-2}. \end{aligned}$$

Hence the generalized estimator given by the expected value of Q is

$$\widehat{\widehat{\upsilon}} = \frac{\widehat{\sigma}_e^2(\widehat{\sigma}_t^2 - \kappa\widehat{\sigma}_e^2)}{A\widehat{\sigma}_t^2}, \tag{20}$$

where $\kappa = e(a-2)/(a(e-2))$.

Superiority of the generalized estimator (GE) given by (20) over the MLE (19) is demonstrated by a simulated experiment with typical values of parameters - only the unconstrained estimators are considered since we have not yet addressed the issue of taking advantage of known signs of parameters. Since the residual error variance typically tends to be higher than the other variance components, in this study the variances are set at $\sigma_a^2 = 1$ and $\sigma_e^2 = 2$. The simulation is carried out by generating $100,000$ samples from the chi-squared distributions (13) and (14) and then evaluating the mean squared error (MS) and the mean of the two sets of estimators to measure the degree of unbiasedness. The number of groups, $k = e + 1$ is set at 3, 4, 5, and 10, which are typical values in clinical trials involving one-way layouts. The number of subjects from in each group, $n = A$, is set at typical values 5, 10, and 100. Results also apply to other designs with $a = k - 1$ and $e = n(k-1)$. The results of the study are summarized in the table below.

Table 3. **Bias and MSE Performance of MLE and GE**

(n,k)	e	$Actual\ \upsilon$	$Mean(MLE)$	$Mean(GE)$	$MSE(MLE)$	$MSE(GE)$
5, 3	10	.286	-.878	.401	1519	.046
5, 4	15	.286	.024	.257	5.84	.839
5, 5	20	.286	.149	.261	.389	.116
5, 10	45	.286	.246	.274	.013	.009
10, 3	20	.167	-.994	.200	3823	.004
10, 4	30	.167	.091	.162	1.756	.219
10, 5	40	.167	.131	.164	.032	.009
10, 10	90	.167	.156	.165	.0012	.0009
100, 3	200	.0196	.0115	.0200	.8436	.0000
100, 4	300	.0196	.0188	.0196	.0004	.0000
100, 5	400	.0196	.0192	.0196	.0000	.0000
100, 10	900	.0196	.0195	.0196	.0000	.0000

It seems that the MLE, which is widely used in software packages dealing with Mixed Models, systematically under estimate the parameter of interest, the variance of predictors, whereas the generalized estimator is almost unbiased except when n and k are both very small. The MSE performance of the generalized estimator seems to be much superior than the MLE except when the sample size is large. Moreover, the unconstrained MLE is highly sensitive when the total sample size is small and yield negative mean values even with 100,000 simulated samples. In some sample runs MLE yields estimators as small as -50,000. In applying MLE of course one would truncate the negative values to zero, but the point is that the actual value of the parameter υ in this case, 0.286, which is the largest value in our simulation, is not even close to 0. In the two next sections we will handle the problem of negative estimators more formerly.

4.3 Estimating Variance Components

In this section we consider the problem of estimating the variance components of mixed models, a well known problem where the classical unbiased estimators known to have a serious drawback, namely they frequently yield negative estimates for parameters which are supposed to be positive. The Bayesian approach does not suffer from the drawback when one assumes a prior distribution such as the inverse gamma distribution. However, the Bayesian estimation of variance components, fixed effects, and random effects, in applications involving large number of random effects is very cumbersome and computationally very intensive - in author's experience, it is typical for the process to take as much as one month to estimate parameters of a mixed model when the number of random effects is as large as one hundred thousand.

So, let us consider the problem of estimating a variance component without adding unnecessary complexity to the problem by taking the generalized approach. Considering the canonical form of the problem, we wish to estimate σ_α^2 based on the sufficient statistics S_e and S_a having the distributions

$$U = \frac{S_e}{\sigma_e^2} \sim \chi_e^2 \tag{21}$$

$$V = \frac{S_a}{\sigma_e^2 + A\sigma_a^2} \sim \chi_a^2. \tag{22}$$

The classical unbiased estimator of σ_α^2, which is also the same as the MLE, is

$$\widehat{\sigma}_\alpha^2 = \frac{1}{A}(\frac{S_a}{a} - \frac{S_e}{e}) = (\widehat{\sigma}_t^2 - \widehat{\sigma}_e^2)/A.$$

The main drawback of this estimator is that it frequently yields negative estimates. The generalized approach can produce the same estimate if we use the generalized estimator $Q' = S_a/a - S_e/e - A\theta$, but this variable does not satisfy Property 3 of Definition 1 to enable inferences beyond computing unbiased estimates. So, let us find another generalized estimator that does not yield negative estimates.

Before applying an e-transformation, the substitution method yields $Q' = \sigma_\alpha^2 - \left(\frac{s_a(\sigma_e^2+A\sigma_\alpha^2)}{S_a} - \frac{s_e\sigma_e^2}{S_e}\right)/A = \sigma_\alpha^2 - (\frac{s_a}{V} - \frac{s_e}{U})/A$, but obviously this is a bad choice for estimating σ_α^2, because as $s_e \longrightarrow 0$, the estimate given by this estimator does not quite tend to the estimator $\widehat{\sigma}_t^2/A$ as desired. So by applying the e-transformation $t(Q, UV) = AUVQ'$ we get the generalized estimator

$$Q = AUV\sigma_\alpha^2 - (s_aU - s_eV). \tag{23}$$

This estimator has the advantage that, while it produces the classical estimator if not conditioned, it allows conditional estimation. Moreover, it can be employed to produce generalized tests and confidence intervals. This is possible because the distribution of Q is free of unknown parameters and it is stochastically increasing in θ.

Now consider the problem of finding estimates under conditional inference so as to guarantee the sign of variance components. To avoid bias estimates, it is also desirable to have an upper bound for the estimates. A natural upper bound to have is the estimate $u_{b1} = \widehat{\sigma}_t^2 = s_a/a$ of σ_t^2 which bounds $A\sigma_\alpha^2$ as seen from the definition, $\sigma_t^2 = \sigma_e^2 + A\sigma_a^2$. Another appropriate upper bound is $u_{b2} = q_p$ the $(1-p)^{th}$ quantile of the distribution of $g(U, V) = s_a/V - s_e/U$, where $p = \Pr(g(U, V) < 0)$. It can be seen that while the former choice for the upper bound lead to estimates with much smaller MSE than the ML and REML, the latter choice avoids the bias. In dealing with estimators with skewed distributions (see plots below), a situation that get more extreme when negative estimates are truncated to zero as done by ML and REML, it does not seem as a good idea to insist on small sample unbiasedness, as done by REML. In any case, a compromise, is to use former for smaller samples and the latter for larger samples. This type of double conditions are frequently arise and derived in the context improving estimates under loss functions in the decision theoretic literature and this is a good topic for such research. While ML and GE with u_{b1} as the upper bound both require large a as well as e to attain the large sample unbiasedness, the GE with such double conditions

require only large e to attain the large sample unbiasedness.

The class of conditional estimators $\widetilde{\sigma}^2_\alpha$of σ^2_α based on Q are solutions of equations of the form

$$E(A\widetilde{\sigma}^2_\alpha UV - (s_aU - s_eV) \mid C) = E(UV(A\widetilde{\sigma}^2_\alpha - (\frac{s_a}{V} - \frac{s_e}{U})) \mid C) = 0 \ , \quad (24)$$

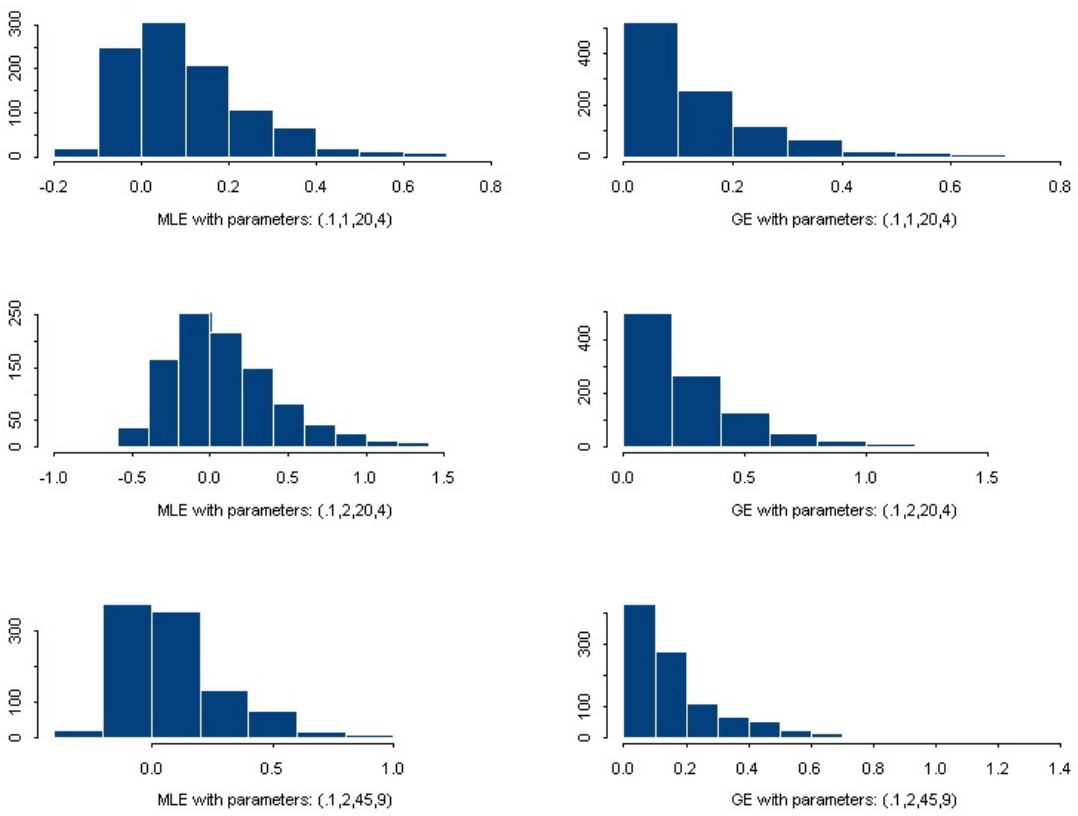
where C is the subset on which the random variables (U, V) is to be conditioned. Now it is evident, by applying Proposition 8 that the desired bounds of $\widetilde{\sigma}^2_\alpha$ for any data (s_a, s_e) can be assured by defining C based on the two bounds. Obviously, assuring the positivity of the estimate require the condition $s_aU > s_eV$ or equivalently $\delta = s_a/V - s_e/U > 0$. Imposing the above upper bound as well, the conditional generalized estimators $\widetilde{\sigma}^2_\alpha$of σ^2_α satisfying both conditions can be found as

$$\widetilde{\sigma}^2_\alpha = \frac{1}{A}\frac{E(s_aU - s_eV| \ 0 \leq s_a/V - s_e/U \leq u_b)}{E(UV| \ 0 \leq s_a/V - s_e/U \leq u_b)}. \quad (25)$$

Unlike the classical unbiased estimator (UE) or the MLE, the generalized estimator $\widetilde{\sigma}^2_\alpha$ will not become negative or zero with any set of (s_a, s_e). The expectations appearing in (25) can be easily evaluated by simple Monte Carlo integration. The details of the computation will be given below more generally in the regression setting.

To demonstrate the advantage of the generalized estimate over the MLE based methods, we carried out a simulation study using a set of simulated samples and generating 1,000 estimates under a number of scenarios on typical values of parameters. Obviously there is no point studying the biasedness because MLE is perfectly unbiased in repeated samples. Moreover, when one truncates the negative estimates to zero, ML becomes biased and REML becomes unbiased by construction, although the need to do so is arguable since its distribution is skewed for small samples.

The major problem with the likelihood based methods is the unreliability with a given sample. In the chart below we report the distribution of individual estimates obtained by MLE and the GE given by (25).



Distribution of Estimators Under $(\sigma_\alpha^2, \sigma_e^2, e, a)$ Scenarios

Notice that in each of the scenarios the MLE/UE generates negative estimates in a substantial fraction of runs. Of course, one can truncate the negative estimates to 0, but this will produce an artificial lump at 0. The situation is specially serious when the ratio $\sigma_\alpha^2/\sigma_e^2$ is small. In fact, when $\sigma_\alpha^2 = .1$ and $\sigma_e^2 = 1$, 46% of the MLEs generated by samples from the exact assumed model are negative estimates. As expected, the generalized estimators always produce positive estimates of the variance component around the true value.

Although the widely used REML are MLE are unbiased by construction, with small samples they fail to yield predictors of random effects. Therefore we study below the mean squared error (MSE) performance of the estimators under alternative scenarios on the parameters $(\sigma_\alpha^2, \sigma_e^2, e, a)$. Note that GE

obtained using u_{b1}as the upper bound in conditional estimation consistently performs substantially better than all MLE based estimators, especially when $\sigma_\alpha^2/\sigma_e^2$ ratio is small as usually the case in practice. This further establishes the superiority of GE not only in producing BLUP with any set of sample, but also in terms of MSE performance. Also notice that despite the small sample unbiasedness of REML, a property whose practical use is questionable with skewed distributions, the REML is worse than the ML in terms of MSE performance, which is considered a more important metric to assure tight reliable estimates.

Table 4. **MSE Performance of Estimators Under $(\sigma_\alpha^2, \sigma_e^2, e, a)$ Scenarios**

Scenario:	.1,1,20,4	.1,1,40,9	.1,2,20,4	.1,2,40,9	.5,2,20,4	.5,2,40,9	1,2,20,4	1,2,40,9
MLE/UE:	.06	.02	.14	.07	.40	.20	1.01	.45
ML:	.07	.01	.12	.03	.22	.17	.67	.42
REML:	.11	.02	.08	.05	.34	.24	.99	.95
GE:	.01	.01	.03	.02	.11	.08	.42	.29

4.4 Estimating Constrained Coefficients of a Regression Model

Next consider the problem of estimating a small parameter of a regression model when we know the sign of the parameter. To be specific we assume that the coefficients is supposed to be positive. When the parameter is small, LSE frequently yields negative estimates for positive parameters thus leaving practitioners to do cherry picking. Bayesians can easily tackle this problem by placing an appropriate prior such as a gamma prior or a log normal prior. This approach is fine except that in applications involving large number of parameters, the computation could take days or weeks, as pointed out in the introduction. So, let us tackle the problem by properly incorporating the known information rather than truncating the LSE as some practitioners do

when estimated parameters have wrong signs.

Let θ be the parameter of interest, a coefficients of linear model, including fixed effects of a mixed model. Assume that a normally distributed error structure is valid. Suppose the inference problem has been reduced by parameters estimated by LSE or MLE methods, say $\widehat{\theta}$, and their variances, say $c\widehat{\sigma}^2$ so that we can confine our attention to estimates based on sufficient statistics

$$\begin{aligned} \widehat{\theta} &\sim N(\theta, c\sigma^2) && (26) \\ Y &= \frac{S}{\sigma^2} = \frac{r\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_r. && (27) \end{aligned}$$

For example if θ is the i^{th} coefficient of a regression model of the form $W = X\beta + \epsilon$, then $\widehat{\theta} = (X'X)^{-1}W_i$, $c = x_i'(X'X)^{-1}x_i$ and $\widehat{\sigma}^2$ is the residual variance.

Now consider in general the problem of estimating θ under the constraint that θ belongs to the interval $[a, b]$. Even in the problem of estimating a parameter without a specific upper constraint, it is a good idea to constraint it by some reasonable upper bound. In the current application a natural upper bound is $b = q_p$, the $(1-p)^{th}$ quantile of the distribution of Q defined below, where $p = \Pr(Q < 0)$. . It is known that without such upper constraints, even the Bayesian approach produce bias estimates.

Kim (2008) provided close form solutions to the Bayesian estimation problem under constraints. Here we will take the generalized approach to tackle the problem. To do so, consider the random variable

$$\begin{aligned} Q &= \widehat{\theta} - \frac{\sqrt{s}\sigma}{\sqrt{S}}\left(\frac{\widehat{\Theta} - \theta}{\sigma}\right) = \widehat{\theta} + \frac{\sqrt{cs}}{\sqrt{Y}}Z, \\ &= \widehat{\theta} + \sqrt{c}\widehat{\sigma}T_r && (28) \end{aligned}$$

where $Z \sim N(0,1)$ and is a standard normal random variable and T_r is a students-t random variable with r degrees of freedom. Obviously the above representations of Q imply that Q has all three properties of a generalized estimator with the observed value at θ. If there are no constraints, the point estimate suggested by the expected value of Q is the LSE $\widehat{\theta}$. The conditional generalized estimator given that $\theta \in [a, b]$ is found using Proposition 7 as

$$\begin{aligned}\widetilde{\theta} &= \widehat{\theta} + \sqrt{c}\widehat{\sigma}E(T_r|Q \in [a,b]) = \widehat{\theta}\sqrt{c}\widehat{\sigma}E(T_r|Q \in [a,b]) \\ &= \widehat{\theta} + \sqrt{c}\widehat{\sigma}E(T_r|T_r \in [\frac{a-\widehat{\theta}}{\sqrt{c}\widehat{\sigma}}, \frac{b-\widehat{\theta}}{\sqrt{c}\widehat{\sigma}}]) = \widehat{\theta} + \sqrt{c}\widehat{\sigma}\frac{\int_{a'}^{b'} t f_T(t)dt}{\int_{a'}^{b'} f_T(t)dt},\end{aligned}$$

where $a' = \frac{a-\widehat{\theta}}{\sqrt{c}\widehat{\sigma}}$ and $b' = \frac{b-\widehat{\theta}}{\sqrt{c}\widehat{\sigma}}$. It is easily seen by integration of the numerator of the above expression that the generalized estimator can be explicitly evaluated as

$$\widetilde{\theta} = \widehat{\theta} + \frac{ks_\theta}{F_T(b') - F_T(a')}\frac{r}{(r-1)}\left((1+\frac{a'^2}{r})^{-(r-1)/2} - (1+\frac{b'^2}{r})^{-(r-1)/2}\right), \quad (29)$$

where F_T is the cdf of the t-distribution with r degrees of freedom, $s_\theta = \sqrt{c}\widehat{\sigma}$ is the standard error of $\widehat{\theta}$, and $k = \Gamma(r/2+.5)/(\sqrt{r\pi}\Gamma(r/2))$.

Unlike the LSE/MLE, the above estimator falls within the $[a,b]$ interval for all possible sample values. Moreover, it is a function of both sufficient statistics rather than relying just on $\widehat{\theta}$ regardless of the variance. It is clear from (29) that, unlike the LSE/MLE, it is a shrinkage estimator that pulls the estimate towards the $[a,b]$ interval, with the amount of shrinkage depending on the standard error of $\widehat{\theta}$. In the simulation study reported in the introduction we have already seen the advantage of the estimator in individual runs.

5 Estimating Parameters of Unbalanced Mixed Models

The results in the previous section do not quite apply to general mixed models of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{u}+\boldsymbol{\epsilon}$ that we discussed in Section 3. However, it is easy to obtain counter parts of the above estimates for the general mixed models with a number of variance components. Let us first consider the two variance components case so that the covariance matrices discussed in Section

3 are of the form $\mathbf{\Sigma}_e = \sigma_e^2 \mathbf{I}_n$ and $\mathbf{\Sigma}_\alpha = \sigma_\alpha^2 \mathbf{A}$, where $\mathbf{A}$ is a known relationship matrix. In this section we will illustrate how to extend the results of the previous section to the unbalanced case.

Most of the solutions in the literature on the unbalanced case are based on variations of distributional results given by Wald (1947) leading to two independent chi-squared random variates to tackle certain ratios of variance components and the error variance. Other methods are based on decomposing the residual sum of squares into a series of independent chi-squared variates. While we can obtain necessary extensions by taking either approach, in the illustrations in this section we confine to Wald's approach, which is computationally superior than other competitors, as we argue below.

Wald's results are derived for the case $\mathbf{A} = \mathbf{I}_k$, but that is not really a drawback because with under the mild assumption that $\mathbf{A}$ is a positive definite matrix, one can simply redefine $\mathbf{Z}$ and $\boldsymbol{u}$ as $\mathbf{Z} \rightarrow \mathbf{Z}\mathbf{A}^{1/2}$ and $\boldsymbol{u} \rightarrow \mathbf{A}^{-1/2}\boldsymbol{u}$, thus meeting the requirement. Wald (1974) basically obtained an extension of (13) and (14) by defining two residual sums of squares from an appropriate regression. In fact, in the problem of estimating random effects by group, Wald's approach works the model is of the form (9).

To outline Wald's simple derivation using familiar regression results and matrix notations, consider the ordinary least squares estimators we get when we regress $\widetilde{\mathbf{X}}$=($\mathbf{X}$ $\mathbf{Z}$) on $\mathbf{y}$ if $\boldsymbol{u}$ were constant. Letting $\boldsymbol{\gamma}'$= ($\boldsymbol{\beta}'$ $\boldsymbol{u}'$) and rewriting the model as $\mathbf{y} = \widetilde{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, let $\widehat{\boldsymbol{\gamma}}$ be the LSE of $\boldsymbol{\gamma}$ if $\boldsymbol{u}$ were fixed. Let $\mathbf{e} = \mathbf{y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\gamma}} = (\mathbf{I}_n - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}')\mathbf{y}$ be the resulting residual error. Now it is evident that the conditional distribution of the error given $\boldsymbol{u}$ is $\mathbf{e}|\boldsymbol{u} \sim \mathbf{N}(\mathbf{0}, \sigma_e^2\mathbf{P})$, which is free of $\boldsymbol{u}$, where $\mathbf{P} = (\mathbf{I}_n - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}')$ is an idempotent matrix. Hence $\mathbf{e}$ distributed independent of $\boldsymbol{u}$ and so from known results in regression analysis we get the counterpart of (13) as

$$U = \frac{S_e}{\sigma_e^2} \sim \chi_e^2, \qquad \text{where } S_e = \mathbf{y}'(\mathbf{I}_n - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}')\mathbf{y} \text{ and } e = n - k - r \quad (30)$$

To establish the counterpart of (14), let $\boldsymbol{\delta} = \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}')\boldsymbol{\epsilon}$ and ζ =

$\sigma_\alpha^2/\sigma_e^2$. Obviously $\boldsymbol{\delta}|\boldsymbol{u} \sim \mathbf{N}(\mathbf{0},\sigma_e^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-\mathbf{1}})$, which is also free of $\boldsymbol{u}$, and hence $\boldsymbol{\delta} \sim \mathbf{N}(\mathbf{0},\sigma_e^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-\mathbf{1}})$. Then, by partitioning $\boldsymbol{\delta}$ as $\boldsymbol{\delta}'=(\boldsymbol{\delta}_\beta' \;\; \boldsymbol{\delta}_u')$ we get $\boldsymbol{\delta}_u \sim \mathbf{N}(\mathbf{0},\sigma_e^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}})$, where $(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}}$ is a $r \times r$ matrix obtained by the corresponding partition of the $(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}$ matrix. Now it is evident that the LSE estimator $\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{\gamma}_u} = \boldsymbol{\delta}_u + \boldsymbol{u}$ of $\boldsymbol{u}$ is distributed as $\widehat{\boldsymbol{u}} \sim \mathbf{N}(\mathbf{0},\sigma_e^2\left((\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}} + \boldsymbol{\zeta}\mathbf{I}_r\right))$ and consequently we get

$$V = \frac{S_\alpha(\zeta)}{\sigma_e^2} \sim \chi_a^2; \quad \text{where } S_\alpha(\zeta) = \widehat{\boldsymbol{u}}'\left((\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}} + \boldsymbol{\zeta}\mathbf{I}_r\right)^{-1}\widehat{\boldsymbol{u}} \tag{31}$$

and $a = r$. The above results are based on the assumption that the random effects are not constrained. If m covariates are randomized by g groups and if each set random effects are required to sum to zero, then $a = r - m$. Moreover, Wald (1974) implicitly assumed that $\widetilde{\mathbf{X}}$ is of full rank, an assumption that can be relaxed as Seely and El-Bassiouni (1983) argued. They showed that, in general, the degrees of freedom of the above two chi-squared random variates are $e = n - rank(\widetilde{\mathbf{X}})$ and $a = rank(\widetilde{\mathbf{X}}) - rank(\mathbf{X})$, respectively. As they further clarified, Wald's full rank assumption can be maintained if the above argument is applied after an initial transformation.

Moreover, by diagonalizing $(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}}$ by means of an orthogonal matrix, say $\mathbf{P}$, the sum of squares term $S_\alpha(\zeta)$ can be further reduced to

$$\begin{aligned} S_\alpha(\zeta) &= \widetilde{\boldsymbol{u}}'\left(\mathbf{D} + \boldsymbol{\zeta}\mathbf{I}_r\right)^{-1}\widetilde{\boldsymbol{u}} && (32) \\ &= \sum \frac{u_i^2}{(\boldsymbol{\zeta} + \boldsymbol{\lambda}_i)}, && (33) \end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix formed by eigen values $\boldsymbol{\lambda}$ of $(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}}$ and $\widetilde{\boldsymbol{u}} = \mathbf{P}\widehat{\boldsymbol{u}}$. When the sample size n is large, the formula (32) is computationally much more efficient than the formula given by Olsen, et al (1976). Furthermore, Wald's argument works even if there are a number of variance components as the case when we need to estimate the effects of a number of covariates by segment, and we are dealing with one variance component at a time. To see this, assume a mixed model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_\mathbf{v}\boldsymbol{v} + \mathbf{Z}_u\boldsymbol{u} + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0},\sigma_e^2\mathbf{I}), \boldsymbol{v} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_v), \boldsymbol{u} \sim \mathbf{N}(\mathbf{0},\sigma_\alpha^2\mathbf{I}), \tag{34}$$

where $\boldsymbol{u}$ is the vector of random effects of interest, $\boldsymbol{v}$ is a vector formed by all other random effects with no interactions, and $\boldsymbol{\Sigma}_v$ is not specified in making inferences about $\boldsymbol{u}$. Now consider the regression of $\widetilde{\mathbf{X}}$=($\mathbf{X}$ $\mathbf{Z_v}$ $\mathbf{Z}_u$) on $\mathbf{y}$ given $\boldsymbol{u}$ and $\boldsymbol{v}$. Let $\mathbf{e} = (\mathbf{I}_n - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}')\mathbf{y}$ be the residual error from the regression. Then it is seen as before that the conditional distribution of the error given $\boldsymbol{u}$ and $\boldsymbol{v}$ is $\mathbf{e}|\boldsymbol{u}, \boldsymbol{v} \sim \mathbf{N}(\mathbf{0},\sigma_e^2\mathbf{P})$, which is free of $\boldsymbol{u}$, where $\mathbf{P} = (\mathbf{I}_n - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}')$ is an idempotent matrix. Hence, we get the unconditional distribution also as $\mathbf{e} \sim \mathbf{N}(\mathbf{0},\sigma_e^2\mathbf{P})$ and therefore, we still get the distributional result (30). Now, partitioning $\boldsymbol{\delta} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}')\boldsymbol{\epsilon}$, which is distributed as $\mathbf{N}(\mathbf{0},\sigma_e^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-\mathbf{1}})$, as before we get $\boldsymbol{\delta}_u \sim \mathbf{N}(\mathbf{0},\sigma_e^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}})$ for the sub vector of $\boldsymbol{\delta}$ corresponding to $\boldsymbol{u}$. Then it is evident that $\widehat{\boldsymbol{u}} = \boldsymbol{\delta}_u + \boldsymbol{u} \sim \mathbf{N}(\mathbf{0},\sigma_e^2\left((\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})_{\mathbf{u}}^{-\mathbf{1}} + \boldsymbol{\zeta}\mathbf{I}_r\right))$. From this result we still get (31) as claimed.

5.1 Estimating Predictors of Random Effects

As we showed in Section 3, the problem of estimating predictors of a set of random effects having a common variance component boils down to estimating variance ratios of the form

$$\rho = \frac{\sigma_e^2}{\sigma_e^2 + B\sigma_a^2} = \frac{1}{1 + B\zeta}, \tag{35}$$

where B is a known constant such as $\mathbf{z}'\mathbf{z}$ and $\zeta = \sigma_\alpha^2/\sigma_e^2$. Since generalized estimators of the form (15) promise almost unbiased (exactly unbiased in the balanced case) estimates, let us derive the counter part in the unbalanced case using (30) and (31). Unlike in the balanced case, it is not possible to apply the substitution method directly to find a generalized estimator. However, in view of the form of (15) in the balanced case, we can construct a generalized estimator as

$$Q = \frac{S_e/\sigma_e^2}{S_a(\rho)/\sigma_e^2} - \frac{s_e}{s_a(\rho)} \tag{36}$$

$$= \frac{U}{V} - \frac{s_e}{s_a(\rho)} = \frac{e}{a}W - \frac{s_e}{s_a(\rho)} \tag{37}$$

where $W \sim F_{e.a}$ and $s_a(\rho) = s_\alpha((\rho^{-1}-1)/B) = s_\alpha(\zeta)$ is an increasing function of ρ. From the second representation of Q in (37) it follows that the distribution of Q is free of nuisance parameters and that it is an increasing function of ρ. And from the first representation of Q it is clear the the observed value of Q is 0, thus implying that Q is a generalized estimator. Therefore, the generalized estimate suggested by the expected value of Q is the solution of the equation $E(\frac{e}{a}W - \frac{s_e}{s_a(\rho)}) = \frac{e}{a-2} - \frac{s_e}{s_a(\rho)} = 0$. Hence, the generalized estimator of ρ based on the proposed Q is

$$\widehat{\rho} = s_a^{-1}\left(\frac{(a-2)}{e}s_e\right), \tag{38}$$

where s_a^{-1} is the inverse function of the increasing function $s_a()$.

5.2 Estimating Variance Components

To establish the counterpart of (25), which has superior MSE performance than the MLE, we need to construct a generalized estimator similar to (23) using the distributional results (30) and (31). Noting that the generalized estimator leading to almost unbiased estimators in the balanced case is equivalent (i.e. they yield the same estimator) to

$$\widetilde{Q} = V\frac{(UA\sigma_\alpha^2 + s_e)}{s_a} - U$$

$$= V\frac{s_e}{s_a}\left(1 + A\sigma_\alpha^2\frac{U}{s_e}\right) - U, \tag{39}$$

by intuition, a potential generalized estimator cab be constructed as

$$Q = V\frac{s_e}{s_\alpha(\sigma_\alpha^2\frac{U}{s_e})} - U = U\left(\frac{V}{U}\frac{s_e}{s_\alpha(\sigma_\alpha^2\frac{U}{s_e})} - 1\right) \tag{40}$$

$$= U(\frac{S_\alpha(\zeta)s_e}{S_e s_\alpha(\zeta\frac{S_e}{s_e})} - 1). \tag{41}$$

From the first of the above two representations of Q it is clear that distribution of Q is free of nuisance parameters. And from the second representation it is clear the observed value of Q is 0. Hence, Q is indeed a generalized estimator that reduces to (39) in the balanced case. Moreover, it is stochastically increasing with respect to the parameter of interest, σ_α^2. Note that as $\sigma_\alpha^2 \to 0$, $\zeta \to 0$, and $S_\alpha(\zeta)$ reduces to the ordinary LSE residual sum of squares, say S_a, so we can employ results of the balanced case. Moreover, in view of the fact that the MLE is asymptotically unbiased, we can find an appropriate generalized conditional generalized estimator, that reduces to (25) in the balanced case by solving the equation

$$E\left(V\frac{s_e}{s_\alpha(\sigma_\alpha^2\frac{U}{s_e})} - U \mid 0 \le s_a/V - s_e/U \le u_b\right) = 0, \tag{42}$$

where u_b is an appropriate upper bound such as s_a/a and the conditional expectation is taken with respect to the chi-squared random variables, $U\sim\chi_e^2$ and $V\sim\chi_a^2$. Although the equation (42) has no close form solution, it can be easily evaluated by Monte Carlo integration:

- Estimate the error sums of squares s_e and the classical regression sum of squares s_a due to the inclusion of the term of interest
- Define a function $f(\sigma_\alpha^2) = mean(\mathbf{V}\frac{s_e}{s_\alpha(\sigma_\alpha^2\frac{U}{s_e})} - \mathbf{U})$ to be called by a root finding algorithm such as the Newton–Raphson method, where $(\mathbf{U}, \mathbf{V})$ is set of chi-squared random numbers to be passed by the calling function subject to any desired constraint
- Set up the initial value $\widetilde{\sigma}_\alpha^2$ as or 0 or any other initial estimate available as required by the root finding algorithm
- Set up $[0,\ u_b]$ as the interval within which the root is to be found
- Generate a large set of chi-squared random numbers from $U\sim\chi_e^2$ and $V\sim\chi_a^2$

- Drop the values that do not satisfy the condition $0 \leq s_a/V - s_e/U \leq u_b$ and define the rest as $(\mathbf{U}, \mathbf{V})$
- Evaluate $f(\sigma_\alpha^2)$ for a given value of σ_α^2, as the average of $(\mathbf{V}\frac{s_e}{s_\alpha(\sigma_\alpha^2\frac{U}{s_e})} - \mathbf{U})$
- Estimate $\widehat{\sigma}_\alpha^2$ as the root of the equation $f(\sigma_\alpha^2) = 0$ by solving it using the desired numerical method with the required accuracy

It is observed that above is a well behaved numerical method that converges rapidly. It can be also verified that the solution reduces to (25) in the balanced case.

A performance comparison of the Generalized Estimator compared with the maximum likelihood based estimators in the unbalanced case is conducted using a simulation of the mixed model involving g groups having random effects of of one covariate X and one fixed effect due to a variable T and one random effect of the form

$$Y_{ij} = 50 + .20T_j + (.25 + \alpha_i)X_{ij} + \epsilon_{ij},\ i = 1, \cdots, g,\ j = 1, \cdots J$$

where J is the number of Y and X observations available from each group. In our simulation $T_j = j$ for each group and X_{ij} values are generated from gamma distribution with scale parameter .5 and the shape parameter 2. Simulated values of Y_{ij} are computed from the above model with random effects generated from $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim N(0, \sigma_e^2)$. Table 5 displays the mean squared error performance of the alternative estimators. It is evident from the simulation that substantial MSE improvement can be obtained by using the proposed generalized estimator, GE. As in the balanced two-way layout, in this regression problem also GE is consistently better than any other likelihood based method, whereas REML is the worst estimator in terms of MSE performance. Of course in terms of classical unbiasedness, REML is the best choice for those who insists on that property despite the skewed nature of all estimators, whereas both ML and GE produce biased estimators with small samples as studied here.

Table 5. **MSE Performance of Estimators Under $(\sigma^2_\alpha, \sigma^2_e, g, J)$ Scenarios**

Scenario:	(1,1,5,5)	(1,1,5,10)	(1,4,5,10)	(5,10,4,1)	(10,10,1,1)	(10,10,4,1)	(10,10,1,4)
ML	.49	.45	.74	5.99	.22	3.58	.37
REML	.66	.62	1.12	8.14	.25	4.12	.43
GE:	.39	.37	.46	5.78	.21	3.23	.26

6 Concluding Remarks

The examples outlined in Section 4 demonstrate the utility of the generalized definitions in point estimation. For each application we were able to provide close form solutions requiring no complex numerical integration, but using simple Monte Carlo integrations minor computational burden. While the Bayesian approach could also provide satisfactory solutions to the problems undertaken in Sections 4 and 5, they involve computationally more intensive numerical integrations unless diffuse priors yielding basically the same results are imposed. Moreover, Bayes estimators typically tend to yield biased estimates due to use of prior distributions.

As evident from various applications reported by Weerahandi (2004), for instance, the Generalized definitions in testing of hypotheses and interval estimation have allowed statisticians to tackle a variety of solutions that have been found to be better than approximate and asymptotic solutions. The generalized definitions, methods, and illustrations given in this article should motivate statisticians to find better estimates than the MLEs in other applications such as those arise in unbalanced models, multivariate models, and non-normal models.

Another area requiring further research is to find systematic methods to find generalized estimators. For example, if one applies the substitution method proposed in this article, the task of finding a particular transformation to assure desirable properties such as close unbiasedness and minimum MSE is an

area requiring further research - in this article this is accomplished by intuition and by examining the performance of alternative estimates.

As we discussed in Section 2, the conditional estimation with small samples could produce bias estimates in the classical sense. However, since estimators with smaller MSE are more important, in our applications above, we have attempted to find estimates with superior MSE performance. Of course all such estimates are asymptotically unbiased. Although the importance of classical unbiasedness in such applications as variance estimation is arguable, in some applications such as those involving location parameters, it is desirable to have both properties. This should prove to be an interesting problem requiring much research. Also desirable is to prove MSE superiority of proposed classes of generalized analytically rather than by simulation.

References

[1] Ananda, M. M. A. (2003). Confidence intervals for steady state availability of a system with exponential operating time and lognormal repair time. *Applied Mathematics and Computation* , 137, 499-509.

[2] Bebu, I., and Mathew, T. (2009). Confidence intervals for limited moments and truncated moments in normal and lognormal models. *Statistics and Probability Letters*, 79, 375–380.

[3] Chen, Y., and Zhou, X. (2006). Generalized confidence intervals for the ratio or difference of two means for lognormal populations with zeros, Technical Report.

[4] Gamage, J., Mathew, T., and Weerahandi, S. (2004). Generalized p-values and generalized confidence Regions for the multivariate Behrens-Fisher Problem. *Journal of Multivariate Analysis*, 88, 177-189.

[5] Hamada, M., and Weerahandi, S. (2000). Measurement system sssessment via generalized inference, *Journal of Quality Technology*, 32, 241-253.

[6] Hanning, J., Iyer, H., and Patterson, P. (2006). Fiducial generalized confidence intervals. *Journal of the American Statistical Association*, 101, 254-269.

[7] Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31 423-447.

[8] Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 72, 789-808.

[9] Kim, H. (2008). Moments of truncated Student-t distribution. *Journal of the Korean Statistical Society*, 37, 81-87.

[10] Krishnamoorthy, K., Mathew, T., and Ramachandran, G. (2006). Generalized P-values and confidence intervals : A novel approach for analyzing lognormally distributed exposure data, *Journal of occupational and environmental hygiene* , 3, 642-650.

[11] Lee, J. C., and Lin, S. H. (2004). Generalized confidence intervals for the ratio of means of two normal populations. *Journal of Statistical Planning and Inference*, 123, 49-60.

[12] Mandelkern (2002), M. (2002). Setting confidence intervals for bounded parameters (with discussion), *Statistical Science*, 17, 149-172.

[13] Roy, A., and Mathew, T. (2005). A Generalized confidence limit for the reliability function of a two-parameter Exponential distribution. *Journal of Statistical Planning and Inference*, 128, 505-517.

[14] Mu, W., Xiong, S., and Xu, X. (2008). Generalized confidence regions of fixed effects in the two-way ANOVA, *Journal of Systems Science and Complexity*, 21, 276-282.

[15] Roy, A., and Bose, A. (2009). Edgeworth corrected generalized confidence intervals. *Journal of Multivariate Analysis*, 100, 1384-1397.

[16] Seely, J. F., and El-Bassiouni (1983). Applying Wald's variance component test, *Annals of Statistics*, 11, 197-201.

[17] Tsui, K., and Weerahandi, S. (1989). Generalized p-Values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, 18, 586-589.

[18] Wald, A. (1947). A Note on regression analysis, *Annals of Statistics,* 55, 1211-1215.

[19] Weerahandi, S. (1987). Testing regression equality with unequal variances. *Econometrica*, 55, 1211-1215.

[20] Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, 88, 899-905.

[21] Weerahandi, S. (1995). Exact Statistical Methods for Data Analysis, New York: Springer Verlag.

[22] Weerahandi, S. (2004). Generalized Inference in Repeated Measures, New York: Wiley.