Statistical Practice

Selecting the Best Linear Mixed Model Under REML

Matthew J. GURKA

Restricted maximum likelihood (REML) estimation of the parameters of the mixed model has become commonplace, even becoming the default option in many statistical software packages. However, a review of the literature indicates a need to update and clarify model selection techniques under REML, as ambiguities exist on the appropriateness of existing information criteria in this setting. A simulation study as well as an application assisted in gaining an understanding of the performance of information criteria in selecting the best model when using REML estimation.

KEY WORDS: Information criteria; Longitudinal data; Model selection; Random effects; Restricted likelihood.

1. INTRODUCTION

Linear mixed model theory has expanded greatly over the past few decades, resulting in its widespread application in many areas of research. This in turn has led to the development of procedures in multiple statistical packages for the analysis of linear mixed models, such as SAS's `proc mixed` (SAS 2003). Linear mixed models are especially useful in longitudinal data settings, because one not only models the mean (referred to as the fixed effects), but also the covariance (in terms of the random effects and pure error term). Restricted, or residual, maximum likelihood (REML) estimation of mixed models is recommended when interest lies in accurate estimators of the variance components of the mixed model (Verbeke and Molenberghs 2000). In fact, many current mixed model-fitting procedures (e.g., `proc mixed` in SAS; `lme` in S-Plus, MathSoft 2002) include REML estimation as the default option.

Additionally, the area of model selection has received increased attention in recent years as datasets and the models that analyze them have become more and more complex (Burnham and Anderson 2002). Model selection tools such as Akaike's information criterion, or AIC (Akaike 1974), the corrected AIC, or AICC (Hurvich and Tsai 1989), the consistent AIC, or CAIC (Bozdogan 1987), and Schwarz's Bayesian information criterion, or BIC (Schwarz 1978), are also computed automatically when fitting a linear mixed model with some of the very same procedures that use REML as a default. The documentation of these software packages indicate that the criteria can then be used to compare a set of models with varying covariance structures. However, in practice one usually wishes to select the best mixed model with respect to not only its covariance, but its mean as well.

Verbeke and Molenberghs (2000) noted that the likelihood-ratio test based on the REML log-likelihood function is not valid when interest lies in the comparison of models with different sets of fixed effects. Welham and Thompson (1997) introduced an adjusted likelihood-ratio test for the fixed effects under REML. In more complex model selection scenarios, though, likelihood-ratio tests may not be very useful; it is in these cases that information criteria are often employed. The comparison of mixed models with different mean structures using information criteria such as the AIC or BIC is generally seen as inappropriate under REML (Verbeke and Molenberghs 2000). For such a comparison, the employment of information criteria calculated from maximum likelihood (ML) parameter estimates is recommended, even after first fitting the model using REML estimation (Wolfinger 1993). Careful examination of the documentation of the cited mixed model fitting procedures relays this opinion as well. However, these procedures calculate the criteria under REML automatically without a clear indication in the resulting output that the criteria should not be used in this way. Only the most diligent and resourceful reader of the software documentation will then employ the model selection techniques in the recommended manner.

Even though it is evident that likelihood ratio tests in their true form cannot be computed using the restricted log-likelihood, it is this author's opinion that it is not apparent as to why model selection criteria computed under REML should not be used to select the best set of fixed effects. Additionally, Shi and Tsai (2002) raised the point that information criteria under ML are calculated from biased estimators and hence may not be suitable. The appropriateness of information criteria for both ML and REML estimation when interest lies in mixed model selection will be discussed. Likewise, the various formulas of these criteria will be described, as inconsistencies exist across the literature and software. Moreover, empirical results presented aid in the understanding of the validity of model selection criteria under REML when attempting to choose the best possible mixed model. These results assist in gaining more general knowledge about the relatively unknown effectiveness of information cri-

Matthew J. Gurka is Assistant Professor, Division of Biostatistics and Epidemiology, Department of Public Health Sciences, University of Virginia School of Medicine, P.O. Box 800717, Charlottesville, VA 22908-0717 (E-mail: mgurka@virginia.edu). The author is grateful to the editor, the associate editor, and the referee for their valuable comments.

 DOI: 10.1198/000313006X90396

teria in mixed model selection. Finally, an application of the discussed criteria and their formulas provides additional insight.

2. ELABORATION OF THE PROBLEM

2.1 Likelihoods of the Linear Mixed Model

Before discussion of the related issues of model selection when using REML estimation, some preliminary notation must be introduced. In the context of repeated measures data, the linear mixed model for $i \in \{1, \ldots, m\}$, m the number of independent sampling units (subjects), is written in the following form

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i. \tag{1}$$

Here, $\mathbf{y}_i$ is a $(n_i \times 1)$ vector of observations on the ith subject, $\mathbf{X}_i$ is a $(n_i \times p)$ known, constant design matrix for the ith subject with rank p, and $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown, constant population parameters. Also, $\mathbf{Z}_i$ is a $(n_i \times q)$ known, constant design matrix for the ith subject with rank q corresponding to $\mathbf{b}_i$, a $(q \times 1)$ vector of unknown, random individual-specific parameters, and $\mathbf{e}_i$ is a $(n_i \times 1)$ vector of random within-subject, or "pure" error terms. Additionally, let $\boldsymbol{\epsilon}_i = \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$ denote the "total" error term of the model, and let $N = \sum_{i=1}^{m} n_i$ signify the total number of observations in the dataset.

For the above mixed model (1), we make the following distributional assumptions: $\mathbf{b}_i$ is normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{D}$, and $\mathbf{e}_i$ is distributed normally with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{R}_i$, independent of $\mathbf{b}_i$. The covariance matrices $\mathbf{D}$ and $\mathbf{R}_i$ are characterized by unique parameters contained in the $(k \times 1)$ vector $\boldsymbol{\theta}$. The total variance for the response vector is $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$. The marginal log-likelihood function for (1) is

$$l_{\mathrm{ML}}(\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{2}$$

Maximization of $l_{\mathrm{ML}}(\boldsymbol{\theta})$ produces ML estimators (MLE's) of the unknown parameters. When $\boldsymbol{\theta}$ is known, the MLE of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{y}_i. \tag{3}$$

In the usual case when $\boldsymbol{\theta}$ is unknown, $\boldsymbol{\Sigma}_i$ is simply replaced with its estimate, $\widehat{\boldsymbol{\Sigma}}_i$. However, the ML estimator of $\boldsymbol{\theta}$ is biased, and thus REML estimators of $\boldsymbol{\theta}$, and hence $\boldsymbol{\beta}$, are typically sought. The REML estimator of $\boldsymbol{\theta}$ is calculated by maximizing the likelihood function of a set of error contrasts that stem from the fixed effects design matrix. The resulting function, not dependent on $\boldsymbol{\beta}$, is based on a transformation of the original observations that lead to a new set of $N - p$ observations. Harville (1974) showed that the restricted log-likelihood function can be written in the following form based on the original observations

$$\begin{aligned} l_{\mathrm{REML}}(\boldsymbol{\theta}) &= -\frac{N-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i\right| \\ &\quad -\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| \\ &\quad -\frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)'\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right), \end{aligned} \tag{4}$$

where $\widehat{\boldsymbol{\beta}}$ is of the form given in (3).

2.2 Information Criteria

The complexity of the mixed model has led to the increased use of model selection tools over traditional inference techniques. This is especially the case when comparing mixed models that are non-nested, such as models with different covariance structures. Information criteria such as the AIC, AICC, CAIC, and BIC, and many other variations, are often used for these purposes. In general, these information criteria are functions of the calculated likelihood for a given model and a penalty term based on the number of parameters in the model. The use of these criteria is strictly subjective; no formal inference based on their values can be made. Comparison of the values of the criteria for a set of models simply indicates if a superior model in that set exists.

When discussing model selection criteria, one should introduce the large-sample notions of efficiency and consistency. Efficient criteria target the best model of finite dimension when the "true model" (which is unknown) is of infinite dimension. In contrast, consistent criteria choose the correct model with probability approaching 1 when a true model of finite dimension is assumed to exist. Selection criteria usually fall into one of the two categories; for instance, the AIC and AICC are efficient criteria, while the BIC and CAIC are considered to be consistent criteria. Debate has ensued as to which characteristic is preferred, as opinions are largely driven by the field of application in which one is interested in applying model selection techniques. For further discussion, see Burnham and Anderson (2002) or Shi and Tsai (2002).

In their original forms, a larger value of the criteria for a given model indicates a better fit of the data. However, it is common to see them presented in a "smaller-is-better" form when they are calculated directly from the $-2 \times$ log-likelihood. Table 1 displays the formulas for the AIC, AICC, CAIC, and BIC from both angles, based on formulas familiar to readers of Vonesh and Chinchilli (1997). Here, l is either $l_{\mathrm{REML}}(\boldsymbol{\theta})$ or $l_{\mathrm{ML}}(\boldsymbol{\theta})$, s

Table 1. General Formulas for Commonly Used Information Criteria

Criteria	*Larger-is-better formula*	*Smaller-is-better formula*
AIC	$l - s$	$-2l + 2s$
AICC	$l - s\left(\frac{N^*}{N^* - s - 1}\right)$	$-2l + 2s\left(\frac{N^*}{N^* - s - 1}\right)$
CAIC	$l - s(\log N^* + 1)/2$	$-2l + s(\log N^* + 1)$
BIC	$l - s(\log N^*)/2$	$-2l + s(\log N^*)$

NOTE: Here, l is either $l_{REML}(\boldsymbol{\theta})$ or $l_{ML}(\boldsymbol{\theta})$, s refers to the number of parameters of the model, and N^* is a function of the number of observations.

refers to the number of parameters of the model, and N^* is a function of the number of observations. When using ML estimation, most often $s = p + k$, the total number of parameters in the model. However, under REML, it is stated that information criteria based on the restricted likelihood cannot be used to compare models with varying sets of fixed effects, as the contrast used to develop the restricted likelihood is dependent on the fixed effects design matrix. Thus, models with different fixed effects have likelihoods based on different observations and are no longer comparable. So, in the case when information criteria are computed under REML for procedures such as `proc mixed`, $s = k$, the number of covariance parameters only. However, as stated previously, the resulting output from such a procedure does not explicitly state when the criteria should be employed; this caveat is only discussed in the documentation of the procedure.

Other inconsistencies exist that display the confusion over the formulas used to compute the aforementioned criteria. One such issue is the use of the complete REML likelihood. The formula for $l_{\mathrm{REML}}(\boldsymbol{\theta})$ has a constant term relative to estimation of the parameters, $\frac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i\right|$, and hence this term is not included in the computation of the REML likelihood in SAS `proc mixed` (2003). To clarify, in computing the restricted log-likelihood value from the estimators, SAS applies the formula

$$\begin{aligned} l_{\mathrm{REML}_2}(\boldsymbol{\theta}) \;=\; & -\frac{N-p}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| \\ & -\frac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| \\ & -\frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)'\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right). \quad (5) \end{aligned}$$

Consequently, the computations of the criteria from $l_{\mathrm{REML}_2}(\boldsymbol{\theta})$ do not include the constant term as well. It is clear that this term does not affect estimation of the model parameters; it is not evident if it should be included in the REML likelihood formula when computing the selection criteria for the purposes of assessing the fixed effects portion of the mixed model. This constant term has appeared in a previous comparison of different versions of the BIC (Neath and Cavanaugh 1997), and its inclusion will be assessed here as well.

The AIC works well in settings in which the sample size is fairly large, but it is biased when the sample size is small (Hurvich and Tsai 1989). Thus, "corrected" versions of the AIC have been proposed, such as the AICC. Essentially, the corrections are some function of the number of observations in the dataset. Like corrected forms of the AIC, the BIC accounts for the total number of observations in the dataset and has generally been known to perform relatively better for small sample size settings. One point of view (Vonesh and Chinchilli 1997) is that under ML, $N^* = N$, the total number of observations, and under REML, $N^* = N - p$, given that the restricted likelihood is based on $N - p$ observations. However, this recommendation has not been consistently employed, as SAS `proc mixed` uses $N^* = m$ under both ML and REML when computing the BIC and CAIC, where m is the total number of subjects, or independent sampling units. To further add to the confusion, SAS `proc mixed` uses $N^* = N$ and $N^* = N - p$ (under ML and REML, respectively) for the correction term of the AICC only. Kass and Raftery (1995) explained that the sample size in the penalty term of the BIC should be "the rate at which the Hessian matrix of the log-likelihood function grows." Hence, at least for the BIC, m is suggested in the correction rather than N (or $N - p$). It is clear that a comparison of various sample size corrections for all of the discussed criteria is warranted.

The motivation for the problem to be analyzed and discussed can now be summarized. It would prove useful to know if criteria under REML are truly inappropriate in choosing from a set of mixed models with different fixed effects. It is understandable that likelihood ratio tests (LRT's) are invalid when using restricted likelihoods. The true forms of restricted likelihoods, based on error contrasts, cannot be compared for models with different mean structures because they will use different sets of observations. When viewed in the formulation of Harville (1974), the restricted likelihood (3) depends on the terms from the originally noted model (1), rather than a model of a transformation of the original observations. This version of the restricted likelihood still allows one to conclude with ease that a LRT is not appropriate under REML, as terms arise not present in the marginal likelihood that do not allow for the assumption of a chi-square distribution of the LRT. However, viewing the REML function in the manner described by Harville (1974) complicates the notion that information criteria should not be subjectively employed for mean model selection.

Because it has been argued that ML estimators are biased, and because most procedures use REML as a default, it is worth comparing the performance of criteria under both estimation methods. Even though the benefit of the application of information criteria in comparing models with varying means is not apparent, information criteria can be especially useful when comparing models with nonnested mean structures and/or distinct covariance models.

3. A SIMULATION STUDY

A straightforward Monte Carlo simulation study that examines the performance of information criteria in selecting the correct linear mixed model should begin to answer the questions listed above. In assessing linear mixed model selection, three scenarios were considered: (1) selection of the correct set of fixed effects when the covariance structure is known; (2) selection of the correct set of random effects when the fixed effects are known; and (3) simultaneous selection of the correct set of fixed and random effects.

This performance of the AIC, AICC, CAIC, and the BIC was assessed. These criteria were chosen due to their popularity in statistical research as well as software packages. Because this simulation is assuming the correct model exists and is of finite dimension, it should be expected that the consistent criteria will perform better in this case. Because the ability of the criteria to choose the proper mean structure is also of interest here, the formulas listed in Table 1 use $s = p + k$. The criteria were evaluated first under ML and then under REML estimation, using both the likelihood formula with the term $\frac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i\right|$ (coined REML_1) as well as without it (referred to as REML_2). All of the criteria involving a sample size correction were evalu-

Table 2. Monte Carlo Assessment of Fixed Effects Model Selection: 10,000 datasets, 100 subjects each, 6 observations per subject

				Percentage of correct model selection					
				$\sigma^2 = 1$			$\sigma^2 = 4$		
Criterion	Estimation method	N^*	$\rho =$	0.25	0.50	0.75	0.25	0.50	0.75
AIC	ML	–		84.0	84.1	84.1	83.8	81.6	77.7
	$REML_1$	–		50.7	65.2	78.5	0	0	29.8
	$REML_2$	–		92.7	94.3	96.1	82.5	83.2	83.4
AICC	ML	N		84.4	84.5	84.5	84.2	81.8	77.8
		m		86.5	86.5	86.5	86.1	83.4	78.9
	$REML_1$	$N-p$		52.5	66.2	79.1	0	0	33.7
		m		60.7	71.1	82.0	0	0	47.9
	$REML_2$	$N-p$		92.9	94.5	96.3	83.0	83.4	83.3
		m		93.7	95.2	96.6	85.2	84.5	83.2
CAIC	ML	N		99.3	99.3	99.1	90.1	70.0	52.6
		m		98.1	98.1	98.1	93.9	80.4	66.0
	$REML_1$	$N-p$		98.3	98.6	99.1	95.2	91.5	87.1
		m		95.4	96.3	97.5	89.3	90.7	92.0
	$REML_2$	$N-p$		99.6	99.7	99.2	86.7	66.2	50.5
		m		99.1	99.3	99.4	91.3	77.0	64.2
BIC	ML	N		98.8	98.8	98.7	92.6	76.1	60.2
		m		96.8	96.4	96.4	93.8	84.2	72.5
	$REML_1$	$N-p$		97.2	97.6	98.4	92.9	92.2	90.8
		m		91.5	93.4	95.6	80.1	84.9	90.4
	$REML_2$	$N-p$		99.3	99.5	99.3	89.6	72.5	58.3
		m		98.4	98.7	99.1	92.2	81.9	71.7

NOTE: "$REML_1$" denotes the criteria were computed using the REML function l_{REML} (that includes the constant term $\frac{1}{2}\log|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i|$). "$REML_2$" signifies the criteria were computed using the REML function l_{REML_2} (that excludes the constant term $\frac{1}{2}\log|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i|$). The estimated simulation standard error for each of the reported percentages, $\widehat{p}$, is equal to $\sqrt{\{\widehat{p}(1-\widehat{p})/10{,}000\}}$

ated using both $N^* = N$ and $N^* = m$ under ML ($N^* = N - p$ and $N^* = m$ under REML).

In all three scenarios of interest, a repeated measures study was simulated that assumes compound symmetry of the data; that is, the data generated were based on a model with only a random intercept and an iid within-unit error term. The compound symmetry structure was simulated such that $\sigma^2 = \sigma_b^2 + \sigma_0^2$, where σ^2 is the total variance, σ_b^2 is the variance of the random intercepts, and σ_0^2 is the variance of the within-unit error term. Datasets were simulated using varying values of the total variance, namely $\sigma^2 = 1$ and 4, to get a sense of the effect of the true variance on the performance of the criteria. Corresponding to the standard model assumptions, the random intercepts and within-unit error terms were generated as independent normal random variables with means zero and variances σ_b^2 and σ_0^2, respectively. To get an idea of the impact of the within-unit correlation on the performance of the criteria, data were simulated using $\rho = 0.25$, 0.50, and 0.75, where $\rho = \sigma_b^2/\sigma^2$. To begin with, large sample performance was assessed; simulated datasets consisted of 100 subjects with 6 observations each. This particular simulation study considered only the complete data case. Each simulation for the varying sample sizes, variances, and correlation values consisted of 10,000 realizations.

For Scenario 1, data were simulated from a true linear mixed model consisting of the following fixed effects: an intercept, a dummy variable indicating membership in one of two groups, and a continuous covariate. The continuous covariate took equally spaced values in $(0, 1)$. More formally, the true linear mixed model has $\boldsymbol{\beta}_0 = (2,\ 1,\ 0.5)'$ corresponding to an intercept, group, and slope. In order to assess the performance of the criteria in choosing the proper set of fixed effects based on this simulation study, a set of candidate models was fit for each generated dataset, and the number of times the criteria chose the correct model from this set of 10,000 was tallied. The set of candidate models consisted of three models each having the same compound symmetric covariance structure and fixed effects corresponding to: (1) a model with common intercept and common slope; (2) a model with a common intercept, a slope, and an additional group covariate (corresponding to the true model); and (3) a model with intercept, slope, group, and group $\times$ slope interaction. The number of times out of the 10,000 possibilities that the criterion in question chose model 2 as the best model was recorded.

For Scenario 2, data were again simulated from a true model with the same fixed and random effects as in Scenario 1. In order to assess the performance of the criteria in choosing the proper random effects structure, three candidate models were fit for each generated dataset, and the number of times the criteria chose the correct model from this set of 10,000 was tallied. The set of candidate models consisted of three models each having the same correct set of fixed effects and iid within-unit error term, but random effects structures corresponding to: (1) no random effects (equivalent to a univariate linear model); (2) a random intercept only (corresponding to the true model); and (3) a random intercept and a random slope (unstructured covariance). The number of times out of the 10,000 possibilities that the criterion in question chose model 2 as the best model was recorded.

Table 3. Monte Carlo Assessment of Simultaneous Fixed Effects and Random Effects Selection: 10,000 datasets, 100 subjects each, 6 observations per subject

				Percentage of correct model selection					
				$\sigma^2 = 1$			$\sigma^2 = 4$		
Criterion	*Estimation method*	*N**	$\rho =$	*0.25*	*0.50*	*0.75*	*0.25*	*0.50*	*0.75*
AIC	ML	–		76.8	76.7	77.2	77.2	75.0	71.0
	$REML_1$	–		45.0	58.6	72.0	0	0	27.8
	$REML_2$	–		84.9	85.9	87.8	75.9	76.3	75.8
AICC	ML	N		77.6	77.4	77.7	77.7	75.6	71.7
		m		81.0	80.8	81.4	81.1	78.5	74.1
	$REML_1$	$N - p$		47.0	59.9	72.9	0	0	31.8
		m		56.3	66.2	77.1	0	0	45.6
	$REML_2$	$N - p$		85.4	86.4	88.3	76.6	76.8	76.2
		m		88.3	88.7	90.7	79.8	79.2	77.7
CAIC	ML	N		99.3	99.4	99.1	89.9	69.7	52.6
		m		98.0	97.9	98.0	93.7	79.6	65.9
	$REML_1$	$N - p$		98.4	98.7	99.2	95.5	91.6	86.9
		m		95.6	96.0	97.6	89.4	90.9	91.9
	$REML_2$	$N - p$		99.6	99.7	99.3	86.3	66.0	50.7
		m		98.9	99.2	99.1	91.3	76.2	64.2
BIC	ML	N		98.8	98.8	98.7	92.5	75.3	59.9
		m		96.3	95.9	96.4	93.8	83.4	72.2
	$REML_1$	$N - p$		97.3	97.6	98.4	93.1	92.4	91.1
		m		91.5	92.8	95.4	79.9	84.5	89.8
	$REML_2$	$N - p$		99.4	99.6	99.2	89.4	72.1	58.1
		m		98.0	98.2	98.5	91.8	80.9	71.1

NOTE: "$REML_1$" denotes the criteria were computed using the REML function l_{REML} (that includes the constant term $\frac{1}{2}\log|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i|$). "$REML_2$" signifies the criteria were computed using the REML function l_{REML_2} (that excludes the constant term $\frac{1}{2}\log|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i|$). The estimated simulation standard error for each of the reported percentages, $\widehat{p}$, is equal to $\sqrt{\{\widehat{p}(1-\widehat{p})/10{,}000\}}$

Finally, for Scenario 3, data were again simulated from a true model with the same fixed and random effects as in the previous two scenarios. In order to assess the performance of the criteria in choosing the proper fixed and random effects structure simultaneously, six candidate models were fit for each generated dataset, and the number of times the criteria chose the correct model from this set of 10,000 was tallied. Each of the three fixed effects models listed in Scenario 1 were combined with the last two random effects structures in Scenario 2 (a random-intercept-only model, and a model with a random intercept and slope), resulting in six possible candidate models. The number of times out of the 10,000 possibilities that the criterion in question chose the true model (fixed intercept, group, and slope effects, and a random intercept) as the best model was recorded.

In calculating the proportion of correct model selection out of a set of candidate models for each of the three scenarios, it is necessary to discuss what is considered an acceptable level of performance. Unlike most other simulation studies, when one is interested in comparing simulated rejection rates to some nominal significance level (e.g., $\alpha = 0.05$), a similar number for model selection performance is not obvious. In Scenarios 1 and 2, in which a finite set of candidate models for only one aspect of the linear mixed model are compared, and it is assumed the other portion of the model is known, it is reasonable to expect at least a 90–95% rate of correct model selection. Even though such rates would obviously be acceptable in Scenario 3, the very fact that both primary portions of the mixed model are allowed to vary should naturally decrease the expectations of the performance of the criteria. In this case, a rate of correct model selection of 80% would probably be acceptable. Of course, such a determination is subjective and can and should vary based on the analyst and the context of the data analysis.

Tables 2 and 3 display the results of the simulations regarding Scenarios 1 and 3, respectively. The results of the Scenario 2 simulations are not displayed, as there were not many distinguishing features between the four criteria and their examined variations. All versions of the four criteria selected the proper random effects structure over 90% of the time, no matter the true variance or correlation. A factor in the excellent performance of the criteria in this simulation is the fact that in roughly half of the simulated datasets, the covariance matrix estimate of the random intercept and random slope in model 3 was not positive definite, implying no variance of the one of the components. In this case, the random slope in the mixed model was found to have zero variance, a phenomenon that makes sense given the true model from which the data were generated did not contain a random slope. Not surprisingly, the consistent criteria (CAIC and BIC) performed slightly better than their efficient counterparts (AIC and AICC), selecting the correct model around 99% of the time (compared to approximately 90–93% selection rate). Large true variance values or within-unit correlations did not alter performance of the criteria. For this particular scenario, REML selection was roughly equivalent to ML selection.

Results of the other two scenarios proved to be more insightful, particularly when examining selection performances for the larger variance value. Some characteristics and trends are noticed regarding the four criteria examined in the Monte Carlo

study. First, it is necessary to address the questions brought to light earlier in this discussion of linear mixed model selection. Specifically, despite the common belief, can model selection criteria be employed in choosing the best mean model when using REML methods? Based on the results of this simple simulation comparison, the answer is yes, with the caveat that this notion needs to be studied further before a definitive conclusion can be made. The performance of the criteria under REML certainly contradicts the contention that mean model selection for the mixed model is inappropriate using REML estimation. In many cases, the criteria actually performed better in choosing the proper set of fixed effects under REML compared to when using ML estimation methods. Without making a definitive conclusion (yet), it can certainly be stated that REML mixed model selection of the fixed effects is not inappropriate.

The next level of assessment involves which version of the REML likelihood should be used when computing these criteria. From the two simulation scenarios, some initial conclusions can be made. The consistent criteria, the BIC and the CAIC, performed better overall when using the full residual likelihood, including the constant term (denoted as REML_1). Performance of the BIC and CAIC was better under REML_1 than under ML in both scenarios. It is very apparent that REML_1 should not be used, however, for the efficient criteria (AIC and AICC). For these two efficient criteria, REML_2 selection was superior to REML_1 as well as ML. If one were to use the performance of the criteria under ML as the basis for comparison, the performance of all four criteria under REML_2 more closely resembles that under ML. To be expected, the AICC performed slightly better than the AIC.

Another comparison to be made regarding the formulas is the assessment of the "correction term" involving the sample size. Specifically, should the total number of observations be used, or should the total number of independent sampling units be applied instead? Under ML, evidence points towards the use of m in the correction terms of the CAIC and BIC, as well as the AICC. Under REML, we should first keep in mind which of the two examined REML functions was indicated to be used for the three criteria. For the consistent BIC and CAIC, use of the complete REML function (REML_1) is suggested; in this case, the total number of observations minus the number of fixed parameters, $N - p$, seems to be more effective than using m. However, the performance of the BIC and CAIC when using m is not terrible, and thus cannot be excluded from consideration. In fact, performance of the criteria under REML_2 using m most closely resembles their performance under ML using m.

Finally, how do the four criteria perform overall when examining varying variance, correlation, and sample size? All four criteria perform reasonably well, but are very sensitive to large total variances. Not surprisingly, simulations repeated for a smaller sample size situation ($m = 25$, $n_i = 3$) demonstrated decreasing performance for all criteria (not shown). For the fixed effects-only selection scenario, when the examined criteria did not choose the proper model, they almost always selected the full model for $\sigma^2 = 1$. For the larger total variance value, the criteria, when incorrect, usually still preferred the full model, but the frequency of selection of the smaller models increased. In fact, the consistent criteria (CAIC and BIC) chose the smaller incorrect model more frequently in these cases, a type of selection error that is in the author's opinion more severe than selecting a model with too many parameters.

In the more general scenario—when both the fixed and random effects models had to be selected—a similar pattern emerged. Most of the time, when the criteria did not select the correct model, they selected the correct random effects structure, but the full fixed effects model (that included the group $\times$ slope interaction). However, as the variance increased, the percentage of selection of the model with too few fixed effects (again with the correct random effects structure) increased for the consistent criteria (particularly under REML_2). Within-subject correlation somewhat affects the performance of all four criteria in both scenarios, but the extent of this influence is relatively minor compared to sample size and variance. It is difficult to determine the best criterion to employ based on this limited simulation study. Surprisingly, though, the consistent criteria did not outperform their efficient counterparts as much as was hypothesized prior to running simulations that would seem to favor consistency.

4. AN EXAMPLE

Even though the simulation study provided much insight into the performance of multiple model selection criteria computed in a variety of ways and under many different conditions, use of these criteria in an application would also be extremely informative. The data to be analyzed were introduced by Verbeke and Molenberghs (2000, pp. 7–9). This longitudinal study was interested in the impact of testosterone inhibition on the craniofacial growth of rats. The "rat study" involved the application of a testosterone-inhibiting drug in two different doses to two groups of rats, in addition to a control group (50 rats total). The response of interest in the book is one type of measurement used to characterize the height of the rat skull; this response was measured repeatedly on each rat. The primary purpose of the data analysis was to estimate changes over time in this response, and to test if these changes vary across the treatment levels. Thus, the full model of interest is as follows (Verbeke and Molenberghs 2000):

$$y_{ij} = \beta_0 + \beta_1 t_{ij} L_i + \beta_2 t_{ij} H_i + \beta_3 t_{ij} C_i + b_{0i} + e_{ij}; \quad (6)$$

$i = 1, \ldots, m; j = 1, \ldots, n_i$. Here, m is the number of rats, n_i is the number of observations on rat i, t_{ij} =ln[1+$\left(\text{Age}_{ij} - 45\right)/10$], and L_i, H_i, and C_i are indicator variables equal to one if rat i belongs to the low-dose group, the high-dose group, or the control group (equal to 0 otherwise), respectively. Thus, the three treatment groups share a common intercept, but have different slopes in the full model.

An approximate Wald test under REML (Verbeke and Molenberghs 2000, p. 75) led to the conclusion that the three treatment-specific slopes were not significantly different ($p = 0.0987$). The authors then proceeded to demonstrate the application of the AIC and BIC for this particular example, computing the two criteria under ML for the full model and for the model with a common slope for the three treatment groups. Table 4 contains these values, as well as values for all four discussed criteria in all of their forms.

Verbeke and Molenberghs (2000, p. 76) noted that the AIC and BIC under ML point to different final models; the AIC leads to the full model, while the BIC prefers the reduced model. Ex-

Table 4. Rat Study Example: Model Selection Criteria Values for Two Models of Interest

Criterion	Estimation Method	N^*	Smaller-is-better values: Full model (distinct slopes)	Smaller-is-better values: Reduced model (common slopes)
AIC	ML	–	**940.7**	941.3
	$REML_1$	–	**924.0**	933.4
	$REML_2$	–	944.4	**943.8**
AICC	ML	N	**941.0**	941.4
		m	942.6	**942.1**
	$REML_1$	$N-p$	**924.3**	933.5
		m	**925.9**	934.2
	$REML_2$	$N-p$	944.8	**944.0**
		m	946.4	**944.7**
CAIC	ML	N	967.8	**959.4**
		m	958.1	**952.9**
	$REML_1$	$N-p$	**951.1**	951.4
		m	**941.5**	945.0
	$REML_2$	$N-p$	971.5	**961.9**
		m	961.9	**955.4**
BIC	ML	N	961.8	**955.4**
		m	952.1	**948.9**
	$REML_1$	$N-p$	**945.1**	947.4
		m	**935.5**	941.0
	$REML_2$	$N-p$	965.5	**957.9**
		m	955.9	**951.4**

NOTE: Values in bold indicate which of the two models is preferred by the criterion.

amination of Table 4 further demonstrates the validity of these very same criteria when computed under REML. However, as seen in the simulations, the REML function without the constant, $\frac{1}{2}\log|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{X}_i|$ (denoted as $REML_2$), again displays the most consistency across the four criteria. In fact, all four criteria under $REML_2$, no matter what correction is used, prefer the reduced model, which is consistent with the conclusion based on the approximate Wald test. None of the criteria in any of their examined forms indicated a random slope is needed in the model (6); these values are not shown.

One aspect of the use of model selection criteria becomes evident from this example. Comparing models on the basis of criteria values is strictly subjective and relative to the magnitude of those values. Thus, in examining the above numbers, even for the ML AIC value that technically "prefers" the full model, many practicing statisticians would ultimately decide on the reduced model since such a simplification of the final model does not translate to a dramatic change in value of the AIC.

5. CONCLUSIONS AND DISCUSSION

So what judgments can be made based on this concise yet informative Monte Carlo study, as well as from the example? Indeed, some of the questions raised in the beginning can be partially answered by this analysis. This study overall was able to shed light into the performance of information criteria in selecting the best linear model at a very basic level. More importantly, this discussion highlights the need for a thorough, unified methodological examination of model selection techniques under REML.

The use of information criteria is commonplace in model selection for the mixed model, as exhibited by their automatic computation in popular statistical software packages. Despite their popularity, the combination of statistical literature and software documentation has led to discrepancies in the formulas and uncertainties of the proper use of information criteria for the purpose of mixed model selection. The aim of this study was to clearly document these inconsistencies and potentially provide answers to the questions that arise.

The primary question that needed to be addressed is the appropriateness of restricted likelihood-based model selection tools for the fixed effects portion of the model. Although not proved beyond a reasonable doubt, this study was able to act as a counterexample to the notion that information criteria under REML cannot be used to select the best mean model. It could be argued that this examination is not important, as the employment of information criteria are probably not necessary in selecting the best mean model out of a set of nested models. In this case, one could use existing inference techniques that are valid under REML. However, the presented simulations demonstrate in this instance the suitability of REML-based criteria in selecting the proper mean model, even when the true covariance structure is unknown. It is in this situation where information criteria can be especially useful, and this study clearly leads one to believe that they are apt for such a comparison, even under REML.

The study also sought to provide some numerical evidence of the performance of criteria in selecting the best linear mixed model. All four examined criteria perform exceptionally well in selecting the proper set of random effects. However, the discussed simulation was admittedly limited, as a more sophisticated covariance model selection evaluation (including different models for the within-unit error term) is worth an article by itself. The results of this simulation study whose primary focus is on mean model selection show that information criteria such as the AIC, AICC, CAIC, and BIC can be valuable, but they also indicate that characteristics of the data, such as variance and sample size, can greatly impact the criteria's performance. No one criteria clearly stands above and beyond the others in terms of selection performance in this simulation study.

Even though use of these criteria under REML is supported here, the results clearly indicate that the correct form of the REML function needs to be studied further. These initial findings indicate the superiority of the REML function without the constant in question, denoted as l_{REML_2}, for the efficient criteria, while the full REML function, l_{REML_1}, seems to be advantageous for the consistent criteria. Further methodological and analytical research on this possible trend is needed. Likewise, additional study is necessary in determining the proper correction factor in the context of longitudinal or clustered data. Some conclusions can be made based on this research, namely the dismissal of the notion that REML-based information criteria are not appropriate for selection of the fixed effects of the mixed model. But, it is clearly displayed here that more work needs to be done, both theoretically and numerically, in understanding the role of information criteria in mixed model selection.

[Received March 2005. Revised September 2005.]

REFERENCES

Akaike, H. (1974), "A New Look At The Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Bozdogan, H. (1987), "Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions," *Psychometrika*, 52, 345–370.

Burnham, K. P., and Anderson, D. R. (2002), *Model Selection And Multimodel Inference: A Practical Information-Theoretic Approach*, New York: Springer.

Harville, D. A. (1974), "Bayesian Inference for Variance Components Using Only Error Contrasts," *Biometrika*, 61, 383–385.

Hurvich, C. M., and Tsai, C. L. (1989), "Regression And Time Series Model Selection In Small Samples," *Biometrika*, 76, 297–307.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

MathSoft, Inc. (2002), *S-Plus* (release 6), Seattle: MathSoft Inc.

Neath, A. A., and Cavanaugh, J. E. (1997), "Regression and Time Series Model Selection using Variants of the Schwarz Information Criterion," *Communications in Statistics—Theory and Methods*, 26, 559–580.

SAS Institute Inc. (2003), *SAS* (release 9.1), Cary, NC: SAS Institute Inc.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shi, P., and Tsai, C. L. (2002), "Regression Model Selection—A Residual Likelihood Approach," *Journal of the Royal Statistical Society*, Series B, 64, 237–252.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models For Longitudinal Data*, New York: Springer-Verlag.

Vonesh, E. F., and Chinchilli, V. M. (1997), *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, New York: Marcel Dekker.

Welham, S. J., and Thompson, R. (1997), "A Likelihood Ratio Test For Fixed Model Terms Using Residual Maximum Likelihood," *Journal of the Royal Statistical Society*, Series B, 59, 701–714.

Wolfinger, R. (1993), "Covariance Structure Selection In General Mixed Models," *Communications in Statistics—Simulation and Computation*, 22, 1079–1106.