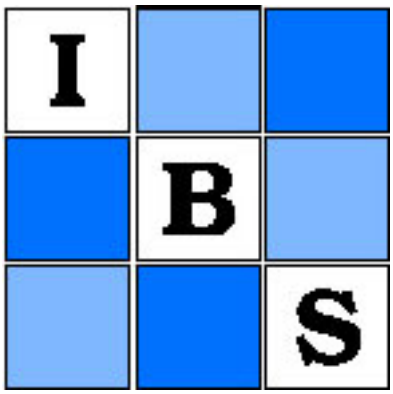

Adequacy of Approximations to Distributions of Test Statistics in Complex Mixed Linear Models

G. Bruce SCHAALJE, Justin B. MCBRIDE, and Gilbert W. FELLINGHAM

A recent study of lady beetle antennae was a small sample repeated measures design involving a complex covariance structure. Distributions of test statistics based on mixed models fitted to such data are unknown, but two recently developed methods for approximating the distributions of test statistics in mixed linear models have been included as options in the latest release of the MIXED procedure of SAS®. One method (FC, from Fai and Cornelius) computes degrees of freedom of an approximating F distribution for the test statistic using spectral decomposition of the hypothesis matrix together with repeated application of a method for single-degree-of-freedom tests. The other method (KR, from Kenward and Roger) adjusts the estimated covariance matrix of the parameter estimates, computes a scale adjustment to the test statistic, and computes the degrees of freedom of an approximating F distribution. Using the two methods, p values for a hypothesis of interest in the lady beetle study were quite different. Simulation studies on the Proc MIXED implementation of these methods showed that Type I error rates of both methods are affected by covariance structure complexity, sample size, and imbalance. Nonetheless, the KR method performs well in situations with fairly complicated covariance structures when sample sizes are moderate to small and the design is reasonably balanced. The KR method should be used in preference to the FC method, although it had inflated Type I error rates for complex covariance structures combined with small sample sizes.

Key Words: Ante-dependence; Covariance structures; Degrees of freedom; Residual maximum likelihood; Satterthwaite approximation; Simulation; Type I error rates.

1. INTRODUCTION

In a recent study of olfactory responses of male and female lady beetles, Hamilton, Dogan, Schaalje, and Booth (1999) examined the antennae of a sample of lady beetles using electron microscopy. Because of the time requirements and expense of electron mi-

G. Bruce Schaalje is Associate Professor, Department of Statistics, Brigham Young University, Provo, UT 84602 (E-mail: schaalje@byu.edu). Justin B. McBride is Division Statistician, 3M Electronic Handling and Protection Division, Building A141-4N-02, 6801 River Place Boulevard, Austin, TX 78726. Gilbert W. Fellingham is Professor, Department of Statistics, Brigham Young University, Provo, UT 84602.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 7, Number 4, Pages 512–524
DOI: 10.1198/108571102726

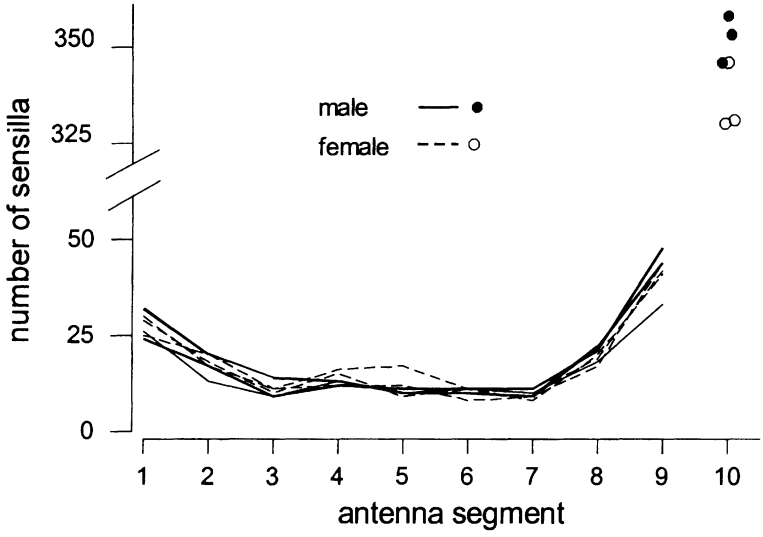

Figure 1. Numbers of sensilla on antenna segments for three male and three female lady beetles, with sequences of data for the first nine segments of each beetle shown as connected lines.

croscopy, the sample was necessarily small, consisting of three males and three females. The researchers counted the number of "sensilla" [small hair-like attachments; see Hamilton et al. (1999) for photographs] on each of 10 segments of an antenna of each beetle (Figure 1). The design was a small sample repeated measures design, and the covariance structure for the repeated measurements apparently had heterogeneous variances. A mixed linear model was fitted to these data, using the heterogeneous AR(1) covariance structure. Particular attention was focused on the mean number of sensilla on the last segment, and its relationship to the other segments and sex. Even though the design was balanced, the test statistics associated with the mixed model followed unknown null distributions because of the complex covariance structure (McCulloch and Searle 2001).

Small sample repeated measures designs like the foregoing are not uncommon in biology, agriculture, or environmental studies. Hence the attendant problem of approximating null distributions of test statistics for such situations must be addressed. Satterthwaite (1941) introduced a method-of-moments approximation to the degrees of freedom of an approximate t distribution for a two-sample t test with heterogeneous variances. This simple idea has been the basis of many modern general approaches to approximating distributions of test statistics, such as those proposed by Giesbrecht and Burns (1985) and Kackar and Harville (1984) for mixed linear models. Jeske and Harville (1988) and McLean and Sanders (1988) extended these ideas. Recently, two very general and promising proposals for approximating distributions of test statistics in mixed linear models have been made. Fai and Cornelius (1996) proposed a method for multi-degree-of-freedom tests in unbalanced split plot designs, but the development of the method is not specific to split-plot designs. Kenward and Roger (1997) proposed a method for tests in mixed linear models based on any covariance structure.

Table 1. Proc MIXED Tests of the Sex* (last segment) Term in a Mixed Linear Model with Heterogeneous ar(1) Covariance Structure

Approx. method	*Numerator D. F.*	*Denominator D. F.*	*F-statistic*	*p value*
SAS default	1	44	6.77	.0126
Fai-Cornelius	1	3.91	6.77	.0613
Kenward-Roger	1	3.91	4.61	.1000

Fai and Cornelius (1996) tested their method using simulation studies based on split-plot models. They varied the degree of imbalance and the value of the intra-class correlation, but did not vary the sample size. Kenward and Roger (1997) used simulation to investigate performance of their method under a variety of complex variance-covariance structures. In three of their four studies, they varied ratios of variance components. In a fourth simulation, they used a nonlinear variance-covariance structure without varying any of the components. They also did not address the small sample question, though they used moderately small samples in some of their simulations.

The MIXED procedure (Proc MIXED) of SAS® (Littell, Milliken, Stroup, and Wolfinger 1996) has made linear mixed model calculations accessible to researchers in a wide variety of fields and with a wide range of statistical training. Proc MIXED includes both the Fai Cornelius (FC) method and the Kenward Roger (KR) method as options in its latest release, but the documentation warns that full information about the performance of these methods is not available. For example, it states that properties of the F approximation using the FC method have not been fully examined for the various covariance structures available in the procedure, especially for small samples (SAS Institute Inc. 1999).

Proc MIXED was used to analyze the lady beetle sensilla data. The variances associated with the number of sensilla at each of the segments were estimated as 9.88, 6.16, 3.25, 3.09, 8.79, 1.55, 1.12, 3.40, 27.39, and 67.58, for segments 1 through 10, respectively. The autocorrelation coefficient was estimated as 0.29. Test statistics and associated p values for the interaction of sex with the contrast of the last segment to the other segments varied greatly, depending on whether the default, FC or KR method was used to approximate the distribution (Table 1). Comparing the default method to the FC and KR methods, it seems clear that the default method produced a p value that was too small. Using it, one would erroneously conclude that the evidence against the null hypothesis of no interaction was very strong. But the p values associated with the FC and KR methods were also quite different from each other. Because this is a small sample situation, it is unknown which p value is more reliable in this case, or indeed if either is reliable.

Because of the widespread use of Proc MIXED, often by nonstatisticians, it is essential to more fully examine properties of the FC and KR methods as implemented in Proc MIXED. The purpose of this article is to extend the simulation results of Fai and Cornelius (1996) and Kenward and Roger (1997), and to examine the Proc MIXED implementation of the FC and KR methods for small samples using some of the more complex covariance structures available. This article investigates the suitability of the approximation to the F distribution under the null hypothesis for both of these methods.

2. APPROXIMATION METHODS

A mixed model can be represented as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the vector of responses, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of unknown fixed effects, and $\boldsymbol{\varepsilon}$ is the vector of deviations from the expected value of the responses. It is assumed that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}))$, where $\boldsymbol{\theta}$ is the vector of variance parameters and $\mathbf{V}(\boldsymbol{\theta})$ is any proper covariance matrix. Associated with this model are several relevant estimators. The restricted maximum likelihood estimator (Patterson and Thompson 1971) of $\boldsymbol{\theta}$ is denoted as $\hat{\boldsymbol{\theta}}$, and therefore an estimator of the covariance matrix of $\mathbf{y}$ is $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$. $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$, which denotes the estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$, is the inverse of the negative Hessian of the restricted log-likelihood function. Given $\hat{\mathbf{V}}$, estimated generalized least squares estimators of $\boldsymbol{\beta}$ can be obtained as $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$, where the usual approximate covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$.

The FC and KR methods were proposed to approximate the distribution of multi-degree of freedom tests of $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ in mixed models, where $\mathbf{C}$ is a $(q \times p)$ matrix of contrasts of full row rank. A commonly used test statistic for this hypothesis is

$$F = \frac{1}{q}\left[(\mathbf{C}\hat{\boldsymbol{\beta}})'(\mathbf{C}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{C}')^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}})\right]. \tag{2.1}$$

Even though the statistic is typically called F, it usually does not follow an F distribution. In fact, it has an exact F distribution only on rare occasions. When the distribution is not exactly an F, however, it is often approximately an F if a suitable value of the denominator degrees of freedom can be calculated.

The FC method uses the test statistic in Equation (2.1). The method involves the spectral decomposition of $(\mathbf{C}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{C}')^{-1}$ to yield $\mathbf{P}'(\mathbf{C}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{C}')^{-1}\mathbf{P} = \text{diag}(\lambda_m)$, where columns of $\mathbf{P}$ are normalized eigenvectors and the λ_m are the corresponding eigenvalues of $(\mathbf{C}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{C}')^{-1}$. Using this decomposition, $Q = qF$ can be written as a sum of q approximate squared t variables,

$$Q = \sum_{m=1}^{q} \frac{(\mathbf{p}_m'\mathbf{C}\hat{\boldsymbol{\beta}})^2}{\lambda_m} = \sum_{m=1}^{q} t_{\nu_m}^2, \tag{2.2}$$

where $\mathbf{p}_m'$ is the mth eigenvector and ν_m is the approximate degrees of freedom for the mth independent single degree of freedom t test. Once the decomposition is performed, the method computes ν_m values by repeatedly applying the method for single degree of freedom contrasts proposed by Giesbrecht and Burns (1985).

Giesbrecht and Burns (1985) investigated the distribution of the approximate t statistic

$$t = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{c}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{c}}}$$

for testing $H_0 : \mathbf{c}'\boldsymbol{\beta} = 0$ where $\mathbf{c}$ is a vector of known constants. Following Satterthwaite's

(1941) premise, Giesbrecht and Burns (1985) assumed that the quantity

$$\frac{df(\mathbf{c}'\hat{\Sigma}_{\hat{\beta}}\mathbf{c})}{(\mathbf{c}'(\mathbf{X}'(\mathbf{V}(\boldsymbol{\theta}))^{-1}\mathbf{X})^{-1}\mathbf{c})}$$

approximately follows a chi-square distribution. Equating the second moment of (2.1) to the second moment of a chi-square distribution results in the method-of-moments approximation to the degrees of freedom

$$df = \frac{2(\mathbf{c}'\hat{\Sigma}_{\hat{\beta}}\mathbf{c})^2}{[\text{var}(\mathbf{c}'\hat{\Sigma}_{\hat{\beta}}\mathbf{c})]}.$$

Taking $f(\boldsymbol{\theta}) = \mathbf{c}'\Sigma_{\hat{\beta}}\mathbf{c} = \mathbf{c}'(\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X})^{-1}\mathbf{c}$, $\text{var}(f(\boldsymbol{\theta}))$ can be approximated using the multivariate delta method (Lehmann 1998) as

$$\text{var}(f(\boldsymbol{\theta})) \cong [\nabla_{f(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}})]'\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}[\nabla_{f(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}})],$$

where $\nabla_{f(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}})$ is a vector of partial derivatives of $f(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, evaluated at $\hat{\boldsymbol{\theta}}$.

For the test with $q > 1$, Fai and Cornelius (1996) noted that each ν_m in Equation (2.2) can be approximated by the Giesbrecht-Burns single degree of freedom method. Using the relationship $E(F_{\text{FC}_{q,v}}) = \nu/\nu - 2$ for $\nu > 2$, they then find ν such that $q^{-1}Q \sim F_{q,v}$ approximately. Since the t_{ν_m} can be regarded as having independent Student's t distributions with ν_m degrees of freedom,

$$\begin{aligned} E(Q) &= \sum_{m=1}^{q} E\left(t^2_{\nu_m}\right) = \sum_{m=1}^{q} E(F_{q,\nu}) \\ &= \sum_{m=1}^{q} \frac{\nu_m}{\nu_m - 2} \\ &= E_Q \quad \text{(say)}. \end{aligned}$$

Now, since

$$\frac{1}{q}E_Q = \frac{\nu}{\nu - 2}$$

it can be shown that

$$\nu = \frac{2E_Q}{E_Q - q}.$$

The KR method was designed to yield test statistics and denominator degrees of freedom for approximate F distributions when exact tests are not available, and exact F distributions when exact tests are available, namely for Hotelling T^2 type statistics and for many standard analysis of variance F ratios. The method first implements an adjustment to $\hat{\Sigma}_{\hat{\beta}}$ to account for small sample bias and incorporate the variability in $\hat{\boldsymbol{\theta}}$. The adjusted estimator is denoted $\hat{\Sigma}^*_{\hat{\beta}}$. Kenward and Roger pointed out that there are two sources of bias in $\hat{\Sigma}_{\hat{\beta}}$ when it is used to estimate the true variance of $\hat{\boldsymbol{\beta}}$ in small samples: (1) the estimand of $\hat{\Sigma}_{\hat{\beta}}$, $\Sigma_{\hat{\beta}} =$

$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, does not account for the impact of the variability in $\hat{\boldsymbol{\theta}}$ on the true variance of $\hat{\boldsymbol{\beta}}$ and (2) $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ is a biased estimator of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$. Kackar and Harville (1984) gave an approximation to the first source of bias, and Kenward and Roger proposed a method to adjust for the second source of bias. Both approximations are based on a Taylor series expansion about $\boldsymbol{\theta}$. Kenward and Roger combined their adjustment with Kackar and Harville's approximation, and proposed an adjusted approximate covariance matrix of $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\Sigma}}^{*}_{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} + 2\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \left\{ \sum_{i=1}^{r} \sum_{j=1}^{r} S_{ij} \left(\mathbf{Q}_{ij} - \mathbf{P}_i \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \mathbf{P}_j \right) \right\} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}},$$

where S_{ij} is the (i,j)th element of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$, $\mathbf{Q}_{ij} = \mathbf{X}' \frac{\partial \hat{\mathbf{V}}^{-1}}{\partial \theta_i} \hat{\mathbf{V}} \frac{\partial \hat{\mathbf{V}}}{\partial \theta_j} \mathbf{X}$, and $\mathbf{P} = \mathbf{X}' \frac{\partial \hat{\mathbf{V}}^{-1}}{\partial \theta_i} \mathbf{X}$.

Once $\hat{\boldsymbol{\Sigma}}^{*}_{\hat{\boldsymbol{\beta}}}$ is obtained, Kenward and Roger calculated a scale factor, δ, and the approximate denominator degrees of freedom, ν. The test statistic is

$$F^{*} = \delta F_{\mathrm{KR}} = \frac{\delta}{q} (\mathbf{C}\hat{\boldsymbol{\beta}})' (\mathbf{C}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \mathbf{C}')^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}}).$$

Kenward and Roger used a second order Taylor series expansion of $(\mathbf{C}'\hat{\boldsymbol{\Sigma}}^{*}_{\hat{\boldsymbol{\beta}}}\mathbf{C})^{-1}$ about $\boldsymbol{\theta}$ and conditional expectation relationships to yield $E(F_{\mathrm{KR}})$ and $V(F_{\mathrm{KR}})$ approximately. The approximate moments of F^{*} are then generated and equated to the moments of an F distribution to solve for δ and ν. This yields

$$\nu = 4 + \frac{q+2}{q\gamma - 1},$$

and

$$\delta = \frac{\nu}{\tilde{E}[F_{\mathrm{KR}}](\nu - 2)},$$

where

$$\gamma = \frac{\tilde{V}[F_{\mathrm{KR}}]}{2\tilde{E}[F_{\mathrm{KR}}]^2}.$$

3. SIMULATION STUDY

For each of several fully specified linear mixed models, 10,000 datasets were generated and then analyzed using Proc MIXED of SAS® v8. F statistics and accompanying p values were calculated by both the FC and KR methods for true null hypotheses. Distributions of the p values were examined and simulated Type I error rates computed.

All linear mixed models were split-plot or repeated measures designs. Whole plots were each assigned to one of three levels of a fixed factor, and subplots were each assigned to one of three levels of another fixed factor. Datasets were generated using the 20 combinations of four sample size options and five covariance structures.

Table 2. Parameter Values for Covariance Structures Used in the Simulations in Variance-Correlation Form

Covariance structure	*Parameter values*
Compound symmetry	$\begin{bmatrix} 1 & & & & \\ 0.50 & 1 & & & \\ 0.50 & 0.50 & 1 & & \\ 0.50 & 0.50 & 0.50 & 1 & \\ 0.50 & 0.50 & 0.50 & 0.50 & 1 \end{bmatrix}$
Toeplitz	$\begin{bmatrix} 1 & & & & \\ 0.50 & 1 & & & \\ 0.30 & 0.50 & 1 & & \\ 0.20 & 0.30 & 0.50 & 1 & \\ 0.10 & 0.20 & 0.30 & 0.50 & 1 \end{bmatrix}$
Heterogeneous compound symmetry	$\begin{bmatrix} 1.00 & & & & \\ 0.50 & 2.81 & & & \\ 0.50 & 0.50 & & 4.80 & \\ 0.50 & 0.50 & 0.50 & 6.35 & \\ 0.50 & 0.50 & 0.50 & 0.50 & 6.79 \end{bmatrix}$
First-order heterogeneous autoregressive	$\begin{bmatrix} 1.00 & & & & \\ 0.70 & 2.81 & & & \\ 0.49 & 0.70 & 4.80 & & \\ 0.343 & 0.49 & 0.70 & 6.35 & \\ 0.24 & 0.343 & 0.49 & 0.70 & 6.79 \end{bmatrix}$
First-order ante-dependence	$\begin{bmatrix} 1.00 & & & & \\ 0.54 & 2.81 & & & \\ 0.33 & 0.61 & 4.80 & & \\ 0.20 & 0.37 & 0.61 & 6.35 & \\ 0.12 & 0.22 & 0.35 & 0.58 & 6.79 \end{bmatrix}$

The four sample size options consisted of combinations of two numbers of whole plots per treatment, namely three and five, and two numbers of subplots per whole plot, also three and five. Because the subplot factor had three levels, when there were five subplots per whole plot the design was unbalanced. In such cases, each subplot treatment appeared at least once, but two of the three appeared twice within each whole plot.

The five covariance structures were compound symmetry, Toeplitz, heterogeneous compound symmetry, first-order heterogeneous autoregressive, and first-order ante-dependence (SAS Institute Inc. 1999). Parameter values used for the covariance structures in the simulations are given in Table 2 in variance-correlation form for the five subplot case. When there were three subplots, the upper 3×3 submatrices of the covariance structures in Table 2 were used to generate the data.

The first-order ante-dependence parameter values were taken from the Kenward and Roger (1997) simulation study so that results from this simulation could be compared to their results. Diagonal elements for the first-order heterogeneous autoregressive and heterogeneous compound symmetry structures were the same as in the first-order ante-dependence structure in order to be able to evaluate the effect of covariance complexity on effectiveness of the methods.

To generate data with a specific covariance structure, the method used by Ripley (1987)

was employed. This is a two-step method where (1) a random vector is generated using the independent standard multivariate normal distribution and then (2) this vector is premultiplied by the Cholesky decomposition of the specified covariance matrix under the null hypothesis.

The data were analyzed using Proc MIXED. The five covariance structures were specified in the REPEATED statement using the TYPE option as CS, TOEP, CSH, ARH(1), and ANTE(1). The FC method was specified as the DDFM = SATTERTH option of the MODEL statement, and the KR method was specified as DDFM = KENWARDROGER. The NOBOUNDS option was not used. Because there is controversy about the propriety of testing for main effects when there are interaction terms in the model, and because it was desired to evaluate the performance of the FC and KR methods for testing effects subject to whole plot and subplot error to varying degrees, the model in this study was fitted without an interaction term. P values for main effects of the whole plot and subplot factors were examined separately.

The data generation and analysis for each of the 20 combinations of covariance structures and sample size options was accomplished using a SAS® macro in which Proc IML was used to generate the data, Proc MIXED was used for the analysis, and an ODS (output delivery system) statement was used to capture the p values. Four p values were calculated for each dataset generated: one per approximation method (FC or KR) per factor tested (whole or subplot).

A chi-square lack-of-fit test with 100 bins was performed to determine if the p values in each simulation scenario followed the uniform(0,1) distribution. The percentages of p values less than or equal to $\alpha = 0.05$ and $\alpha = 0.01$ were calculated as the simulated Type I error rates. For 10,000 observations, standard errors for the simulated $\alpha = 0.05$ and $\alpha = 0.01$ Type I error rates are about 0.002 and 0.001, respectively. Therefore, if the methods work perfectly the simulated $\alpha = 0.05$ error rates should be between about 0.046 and 0.054, and the simulated $\alpha = 0.01$ error rates should be between about 0.008 and 0.012.

In a few cases, PROC MIXED did not converge. In such cases, missing values were generated for the observed p values, and the simulated Type I error rates were calculated based on the nonmissing values.

To directly address the lady beetle example, one additional simulation study was carried out. This involved a 2×10 factorial arrangement of treatments having an ARH(1) covariance structure with parameters equal to those which were estimated for the lady beetle example. As in the other simulation studies, 10,000 datasets were generated. Using the FC and KR methods, p value distributions were examined for the test of the interaction of sex with the contrast of the last segment to the other segments.

4. RESULTS

4.1 CONVERGENCE

Proc MIXED converged for almost all of the simulated datasets in the study. It converged for all of the simulated datasets using the CS, TOEP, and ANTE(1) covariance structures. For the CSH structure, Proc MIXED converged in all simulations with five whole plots per

Table 3. Results of Chi-Square Lack-of-Fit Tests Comparing the uniform(0,1) Distribution to the Simulated *P* Value Distributions

				p values for lack-of-fit tests				
Method	*Treatment Tested*	*Whole Plots*	*Sub-Plots*	*CS*	*TOEP*	*CSH*	*ARH(1)*	*ANTE(1)*
FC	Whole	5	5	0.97	<0.01	<0.01	<0.01	<0.01
			3	0.81	0.50	<0.01	<0.01	<0.01
		3	5	0.99	<0.01	<0.01	<0.01	<0.01
			3	0.26	<0.01	<0.01	<0.01	<0.01
	Sub	5	5	0.85	<0.01	<0.01	<0.01	<0.01
			3	0.90	<0.01	<0.01	<0.01	<0.01
		3	5	0.57	<0.01	<0.01	<0.01	<0.01
			3	0.07	<0.01	<0.01	<0.01	<0.01
KR	Whole	5	5	0.97	0.77	0.19	0.14	<0.01
			3	0.81	0.67	<0.01	0.38	0.03
		3	5	0.99	0.51	<0.01	<0.01	<0.01
			3	0.26	0.53	<0.01	<0.01	<0.01
	Sub	5	5	0.54	<0.01	0.76	0.91	<0.01
			3	0.90	0.16	0.66	0.38	<0.01
		3	5	0.92	<0.01	<0.01	<0.01	<0.01
			3	0.07	<0.01	<0.01	<0.01	<0.01

treatment and five subplots per whole plot, but it did not converge for 4, 27, and 180 of the 10,000 simulations with five whole plots and three subplots, three whole plots and five subplots, and three whole plots and three subplots, respectively. For the ARH(1) structure, Proc MIXED did not converge for 7, 4, 179, and 184 of the simulations with five whole plots and five subplots, five whole plots and three subplots, three whole plots and five subplots, and three whole plots and three subplots, respectively.

4.2 FC METHOD

For the FC method, distributions of p values followed the uniform(0,1) distribution for the CS structure regardless of sample size, type of treatment being tested, or balance (Table 3). There was a hint that the distribution of p values did not exactly follow the uniform(0,1) distribution for the smallest design (three whole plots and three subplots) based on the CS structure, but the simulated Type I error rates (Table 4) were close enough to 0.05 and 0.01 that the lack of fit would not cause practical problems. These results agree with previous simulation results of Fai and Cornelius (1996) and act as a check that the simulations were set up properly.

The FC method gave dramatically different results when other covariance structures were used. In fact, there was only one scenario in which the FC method produced p values following the uniform(0,1) distribution and Type I error rates close to 0.05 and 0.01, namely the balanced design with five whole plots per treatment and three subplots per whole plot based on the TOEP structure. For covariance structures more complex than CS, the FC method apparently produces approximate degrees of freedom that are too large because in all other cases, lack-of-fit was significant and Type I error rates were inflated. In all

Table 4. Simulated Type I Error Rates

					Proportion of p values less than α				
α	*Method*	*Factor tested*	*Whole plots*	*Sub-plots*	*CS*	*TOEP*	*CSH*	*ARH(1)*	*ANTE(1)*
0.05	FC	Whole	5	5	0.0520	0.0630	0.0900	0.0825	0.1145
				3	0.0445	0.0528	0.0889	0.0842	0.0934
			3	5	0.0489	0.0867	0.1418	0.1275	0.2239
				3	0.0474	0.0702	0.1304	0.1225	0.1641
		Sub	5	5	0.0520	0.1011	0.0914	0.0885	0.1415
				3	0.0515	0.0789	0.0815	0.0895	0.1206
			3	5	0.0491	0.1506	0.1405	0.1415	0.2607
				3	0.0486	0.1022	0.1262	0.1325	0.2013
	KR	Whole	5	5	0.0520	0.0497	0.0580	0.0557	0.0654
				3	0.0445	0.0455	0.0576	0.0551	0.0602
			3	5	0.0489	0.0544	0.0820	0.0712	0.1165
				3	0.0474	0.0505	0.0711	0.0699	0.0825
		Sub	5	5	0.0518	0.0610	0.0557	0.0537	0.0655
				3	0.0515	0.0546	0.0512	0.0551	0.0645
			3	5	0.0480	0.0704	0.0698	0.0653	0.1091
				3	0.0486	0.0635	0.0632	0.0686	0.0938
0.01	FC	Whole	5	5	0.0090	0.0166	0.0241	0.0202	0.0331
				3	0.0089	0.0123	0.0232	0.0232	0.0259
			3	5	0.0102	0.0289	0.0468	0.0369	0.1003
				3	0.0097	0.0197	0.0365	0.0387	0.0607
		Sub	5	5	0.0109	0.0309	0.0227	0.0252	0.0469
				3	0.0084	0.0222	0.0237	0.0238	0.0395
			3	5	0.0092	0.0536	0.0487	0.0470	0.1149
				3	0.0095	0.0279	0.0418	0.0460	0.0788
	KR	Whole	5	5	0.0090	0.0108	0.0134	0.0120	0.0132
				3	0.0089	0.0092	0.0132	0.0119	0.0124
			3	5	0.0102	0.0130	0.0216	0.0180	0.0408
				3	0.0097	0.0115	0.0148	0.0198	0.0255
		Sub	5	5	0.0106	0.0153	0.0115	0.0118	0.0166
				3	0.0084	0.0141	0.0126	0.0119	0.0168
			3	5	0.0088	0.0215	0.0198	0.0182	0.0352
				3	0.0095	0.0148	0.0172	0.0170	0.0290

NOTE: The proportions were calculated for cases in which Proc MIXED converged. It failed to converge with the CSH structure for .04%, .27%, and 1.8% for the studies with five whole plots and three subplots, three whole plots and five subplots, and three whole plots and three subplots, respectively. It failed to converge with the ARH(1) structure for .07%, .04%, 1.79%, and 1.84% for the studies with five whole plots and five subplots, five whole plots and three subplots, three whole plots and five subplots, and three whole plots and three subplots, respectively.

cases when lack-of-fit to the uniform(0,1) distribution was significant, deviation from the uniform(0,1) was such that there were increasingly more p values close to zero than expected and increasingly fewer p values close to one than expected.

Although the FC method produced valid tests for unbalanced designs using the CS structure, imbalance seems to affect the FC method when the TOEP structure is used because a larger unbalanced design exhibited significant lack-of-fit while a smaller balanced design (five whole plots per treatment and five subplots per whole plot) did not. Even though Type I error target values were not achieved for most simulations using the TOEP, CSH, ARH(1), and ANTE(1) structures, simulated Type I error rates were always closer to target values for balanced designs than for unbalanced designs. Also, when comparing designs of similar

balance, Type I error rates were always closer to target values for designs with larger sample sizes.

As mentioned previously, these results agree with the simulations reported by Fai and Cornelius (1996). Sample sizes in the Fai-Cornelius study were comparable to those used in this study, but all of their simulations used the CS structure. Some of their simulations involved a much higher degree of imbalance than this study. The FC method produced Type I error rates close to the target values for all of their simulations.

4.3 KR METHOD

For almost every simulation, the KR method produced a smaller lack-of-fit statistic than the FC method. Thus, the distributions of p values produced by the KR method followed the uniform(0,1) distribution at least as closely and usually much more closely than distributions of p values from corresponding simulations using the FC method (Table 3). Nonetheless, p values produced by the KR method still exhibited significant lack-of-fit in many cases. In addition to all tests involving the CS structure, distributions of p values were uniform(0,1) for most tests involving the TOEP structure and for the larger sample size simulations for the CSH and ARH(1) structures. None of the distributions of p values involving the ANTE(1) structure were uniform(0,1).

The KR method produced simulated Type I error rates very close to the target values for all simulations involving the CS structure, for most simulations involving the TOEP structure, and for the larger sample size simulations involving the CSH and ARH(1) structures (Table 4). If one were willing to accept error rates as high as 0.07 as reasonable practical approximations to the target value 0.05 (and error rates as high as 0.015 as reasonable practical approximations to the target value 0.01), many of the smaller sample size simulations involving the CSH and ARH(1) structures could also be considered to produce acceptable Type I error rates. Even the larger sample size simulations involving the ANTE(1) structure produced simulated Type I error rates that would be acceptable in this sense.

Imbalance apparently also affects the KR method because simulated Type I error rates were usually closer to target values for balanced designs than for unbalanced designs based on the TOEP, CSH, ARH(1), and ANTE(1) structures.

In previous simulations, Kenward and Roger (1997) found that the KR method produced acceptable Type I error rates using the ANTE(1) structure. However, their datasets were generally larger than those used in our study, and the design was different as well. The simulations involved only one fixed effect with two levels, and they had up to six repeated measures per experimental unit.

4.4 LADY BEETLE EXAMPLE

For the simulation study using the ARH(1) covariance structure with parameter estimates from the lady beetle example, Proc MIXED did not converge for 117 of the 10,000 simulations. For the simulations which converged, the simulated $\alpha = 0.05$ and $\alpha = 0.01$ Type I error rates were 0.0602 and 0.0153, respectively, for the FC method. The simulated

$\alpha = 0.05$ and $\alpha = 0.01$ Type I error rates were 0.0311 and 0.0063, respectively, for the KR method.

5. CONCLUSIONS

Agreement with results of previous simulation studies suggests that these simulations can be seen as extensions of those results. This agreement also suggests that the methods were implemented adequately on Proc MIXED and therefore the implications of this study apply not only to the Proc MIXED implementation of the FC and KR methods, but to the methods in general.

Complexity of the covariance structure, sample size, and imbalance affect the performance of both methods. However, these factors affect the FC method much more than the KR method. The FC method can only be recommended for use with the CS covariance structure or with simple structures such as TOEP when sample sizes are moderately large. The KR method works reasonably well with more complicated covariance structures when sample sizes are moderate to small and the design is reasonably balanced. Even the KR method had problems, however, with structures as complex as the ANTE(1) structure when sample sizes were small. Even though it worked well in connection with the CS structure, there seems little reason to use the FC method. The KR method works as well as or better than the FC method in all situations. When using the KR method, however, it must be kept in mind that it does not perform perfectly in all situations. When the covariance structure is complex and the sample size is small, the KR method produces inflated Type I error rates. Hence, the p value produced by the KR method in such situations should be treated as a lower bound.

The simulation study using the estimated covariance structure from the lady beetle example produced somewhat different results than those for the other simulation studies. Type I error rates using both methods were reasonably close to target values. However, error rates using the FC method were slightly inflated and Type I error rates using the KR method were slightly deflated. This could be due to any of several factors. Variances in the ARH(1) structure for this example were much more heterogeneous than in the other simulation studies, the design had a large number (10) of subplots per whole plot, and the autocorrelation coefficient was much closer to 0 than those used in the other simulation studies. The default p value produced by Proc MIXED for the lady beetle data (Table 1) is clearly too small. The p values for the FC and KR methods are more reliable, but based on the simulation study neither is preferable to the other in this situation. A better p value would appear to be between those produced by the FC and KR methods.

Research should continue on methods of approximating the distributions of test statistics in mixed models, and on the implementation and use of the FC and KR methods. For example, in this study as well as in the simulation studies of Fai and Cornelius (1996) and Kenward and Roger (1997) it was assumed that the true covariance structure was known. It would be interesting to evaluate the performance of the FC and KR methods when the data are used to select the appropriate covariance structure.

[Received June 2001. Revised December 2001.]

REFERENCES

Fai, A. H. T., and Cornelius, P. L. (1996), "Approximate F-tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-plot Experiments," *Journal of Statistical Computing and Simulation*, 54, 363–378.

Giesbrecht, F. G., and Burns, J. C. (1985), "Two-Stage Analysis Based on a Mixed Model: Large-sample Asymptotic Theory and Small-Sample Simulation Results," *Biometrics*, 41, 853–862.

Hamilton, R. M., Dogan, E. B., Schaalje, G. B., and Booth, G. M. (1999), "Olfactory Response Of The Lady Beetle Hippodamia Convergens (*Coleoptera: coccinellidae*) To Prey Related Odors, Including A SEM Study Of The Antennal Sensilla," *Environmental Entomology*, 28, 812–822.

Jeske, D. R., and Harville, D. A. (1988), "Prediction Interval Procedures and (Fixed-Random) Confidence-Inverval Procedures For Mixed Linear Model," *Communications in Statistics A. Theory and Methods*, 17, 1053–1087.

Kackar, R. N., and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association*, 79, 853–862.

Kenward, M. G., and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, 53, 983–997.

Lehmann, E. (1998), *Elements of Large Sample Theory*, New York: Springer-Verlag.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, New York: SAS Institute Inc.

McCulloch, C. E., and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, New York: Wiley.

McLean, R. A., and Sanders, W. L. (1988), "Approximating Degrees of Freedom for Standard Errors in Mixed Linear Models," in *Proceedings of the Statistical Computing Section*, Alexandria, VA: American Statistical Association, pp. 50–59.

Patterson, H. D. and Thompson, R. (1971), "Recovery of Inter-Block Information When Block Sizes are Unequal," Biometrika 58: 545-554.

Ripley, B. E. (1987), *Stochastic Simulation*, New York: Wiley.

SAS Institute, Inc. (1999), SAS version 8, Online Help, Cary, NC: SAS Institute.

Satterthwaite, F. E. (1941), "Synthesis of Variance," *Pyschometrika*, 6, 309–316.