Comparing Methods of Small Sample Inference for Linear Mixed Models

Min Zhu[1] and Guixian Lin[1]

[1]SAS Institute, 600 Research Drive, Cary, NC 27513

Abstract

Linear mixed models have become a popular framework for analyzing data that have complicated designs and structured covariances. No exact test is available for fixed effects except for balanced data and special covariance structures. Test statistics based on asymptotic distribution of parameter estimates are useful for large samples. However, these tests can be unreliable in applications with small sample sizes. Various methods have been proposed to better approximate small sample distributions of these test statistics. This paper compares the performance of these methods through simulation studies that vary in experiment design, covariance structure, sample size, and so on. It also investigates robustness to misspecification of covariance structures and nonnormal random effects.

Keywords: mixed model; small sample inference; Kenward-Roger method; approximate *F* test; degrees-of-freedom method

1 Introduction

Mixed models have become an established tool for analyzing correlated data that arise in a wide range of settings, including hierarchical models, blocked design, crossover models, and repeated measurements studies. The parameters of a linear mixed model are usually estimated using restricted maximum likelihood (REML). The estimation of the precision and the inference for fixed effects are usually based on their asymptotic distributions. These asymptotic results do not take into account the variability in the estimates of covariance parameters, which for certain combinations of design and covariance structure has a significant impact on the small sample inference of fixed effects. Small samples are commonly encountered in analyses of veterinary studies, clinical trials, and so on. This paper reviews several approaches to small sample inference and compares their performance in a number of settings through simulation studies.

2 Comparison of Methods through an Example

The multicenter clinical trial that is discussed in Beitler and Landis (1985) [1] is an unbalanced design in which two treatments are randomly assigned to patients at eight randomly sampled clinics. The outcome, the number of favorable and unfavorable responses, is recorded for each clinic. The SAS procedure GLIMMIX can be used to model the relationship of the outcome to the fixed effect treatment and the random effects—clinic and

clinic treatment interaction. This model has two random effects and independent residuals. The default degrees-of-freedom (df) method is the containment method. For comparison, the χ^2 test and the t tests for treatment effect are also carried out based on the Satterthwaite [2] [5] and Kenward-Roger [7] methods.

From the output of PROC GLIMMIX, you can see that the p-value increases from 0.0244 of the χ^2 test to 0.0987 of the Kenward-Roger t test. The standard errors of the containment method and the Satterthwaite method are the same, and they are smaller than that of the Kenward-Roger method. The degrees of freedom of the Satterthwaite and Kenward-Roger methods are the same, and they are smaller than that of the containment method. The χ^2 test, the containment method, and the Satterthwaite method all use the asymptotic standard errors, whereas the Kenward-Roger method uses a biased-adjusted standard error. The χ^2 test is equivalent to an F test with infinite degrees of freedom; the containment method computes ANOVA-type degrees of freedom, and both the Satterthwaite and Kenward-Roger methods compute approximated degrees of freedom. From this example you can conclude that the choices of standard error estimators and degrees of freedoms are critical in the small sample inferences.

To review the computation of standard error estimates and degrees-of-freedoms, consider a Gaussian linear model

$$Y \sim N(X\beta, V(\sigma))$$

where parameters β and σ are estimated using REML. A linear mixed model

$$Y|\gamma \sim N(X\beta + Z\gamma, R(\sigma)), \gamma \sim N(0, G(\sigma))$$

can be written as a Gaussian linear model, where $V(\sigma) = ZG(\sigma)Z' + R(\sigma)$. Other than for special cases, an exact F test does not exist for the inference of fixed effect β. In the absence of exact results, the asymptotic covariance estimator $\phi(\hat{\sigma}) = (X^T V(\hat{\sigma})^{-1} X)^{-1}$ and ANOVA-type degrees of freedom are computed. The asymptotic estimator has a significant bias in small samples, whereas the ANOVA-type degrees of freedom do not use the interblock information in linear mixed models. Therefore, to ensure the validity of small sample inference of fixed effects, you need methods to adjust the bias in the standard error estimators and to appropriately approximate the degrees of freedom.

3 The Kenward-Roger Method

Kenward and Roger (1997) [6] propose a scaled Wald statistic that uses a bias-adjusted covariance estimator and derive an F approximation to its sampling distribution. For the inference of the contrast $L\beta$, where L is an $(l \times p)$ fixed matrix, they construct a scaled Wald statistic,

$$\begin{aligned} F^* &= \lambda F \\ &= \frac{\lambda}{l}(\hat{\beta} - \beta)^T L(L^T \hat{\Phi}_A L)^{-1} L^T (\hat{\beta} - \beta) \end{aligned}$$

where $\hat{\Phi}_A$ is a bias-adjusted estimator of the covariance of $\hat{\beta}$ and $0 < \lambda < 1$. An appropriate $F_{l,m}$ approximation to the sampling distribution of F^* is derived through matching the first two moments of F^* with those from the approximating F distribution. This approximation produces values for the scale factor λ and the denominator degrees of freedom m.

Note that a bias-adjusted covariance estimator $\hat{\Phi}_A$ is used instead of the conventional covariance estimator $\Phi(\hat{\sigma}) = (X'V(\hat{\sigma})^{-1}X)^{-1}$. To find the breakdown of the bias, denote the estimator of β for known fixed σ as $\tilde{\beta}$. Then you can write $\text{var}(\hat{\beta})$ as

$$\text{var}(\hat{\beta}) = \text{bias}_1 + \text{bias}_2 + \Phi(\hat{\sigma})$$

where $\text{bias}_1 = E(\hat{\beta} - \tilde{\beta})^2$ and $\text{bias}_2 = \Phi(\sigma) - \Phi(\hat{\sigma})$. bias_2 comes from the fact that $\Phi(\hat{\sigma})$ is a biased estimator of $\Phi(\sigma)$; bias_1 comes from the fact that $\Phi(\sigma)$ itself underestimates $\text{var}(\hat{\beta})$ when σ is unknown. bias_1 is nonnegative and thus always leads to underestimation, whereas bias_2 can lead to underestimation or overestimation. So the overall effect of the two biases is either underestimation or overestimation. Approximations to the two biases are proposed by Kackar and Harville (1984) [3], Harville and Jeske (1992) [4], and several other authors. The bias-adjusted covariance estimator is then $\hat{\Phi}_A = \text{adj}_1 + \text{adj}_2 + \Phi(\hat{\sigma})$, where adj_1 and adj_2 are approximations to bias_1 and bias_2, respectively. Kenward and Roger (2009) [7] introduce an extra term in the approximation to bias_2. This leads to an improved covariance estimator $\hat{\Phi}^*_A$. With a better approximation to bias_2, $\hat{\Phi}^*_A$ works well with nonlinear covariance structures.

One useful property of the improved estimator $\hat{\Phi}^*_A$ is that it is invariant under re-parameterization within the classes of intrinsically linear and intrinsically linear inverse covariance structures. For some choice of λ and fixed matrices $A_1, \ldots, A_r$ of the same dimension as V, the covariance matrix V with an intrinsically linear structure can be written as

$$V = \sum_{i=1}^{r} \lambda_i A_i$$

and the inverse of a covariance matrix V with an intrinsically linear inverse structure can be written as

$$V^{-1} = \sum_{i=1}^{r} \lambda_i A_i$$

To see the benefit of the invariance property, consider an unstructured covariance structure. This is an intrinsically linear structure. With the invariance property, you can parameterize an unstructured covariance matrix as its Cholesky decomposition to ensure that the estimated covariance matrix is positive semidefinite. Another useful feature of the Kenward-Roger method is that it reproduces the results of the univariate or multivariate approach to repeated measures when the $F_{l,m}$ distribution is exact.

4 Two Simulation Studies

These two simulation studies are adapted from Kenward-Roger (1997) [6]. The first simulation study is a four-treatment, two-period crossover trial. To study the effect of four treatments A, B, C, and D, all 12 pairs of treatments are randomly assigned to 12 subjects. The model is

$$Y_{ijk} = \mu + s_{k(i)} + p_j + \tau_{d[i,j]} + e_{ijk}$$

where $e_{ijk} \sim N(0, \sigma^2)$ is the random error and p_j, $\tau_{d[i,j]}$, $s_{k(i)} \sim N(0, \sigma_s^2)$ are the period, treatment, and subject effects, respectively.

In the simulation, let the ratio $\rho = \sigma_s^2/\sigma^2 = 0.25, 0.5, 1, 2, 4$. For each setting of ρ, 10,000 sets of data are simulated. To compare the biases, the percentage relative bias is

computed for both $\hat{\Phi}$ and $\hat{\Phi}_A^*$:

$$100\left(\frac{E_s[l^T\hat{\Phi}l]}{V_s[l^T\hat{\beta}]}-1\right)$$

Table 1 shows the simulation results.

Table 1: Simulation Results for Crossover Clinical Trial

	% Relative bias variance estimates			Observed size Nominal 5% F tests		
ρ	Asy.	KR	KR df	Asy.	Cont.	KR
0.25	−13	−3.8	14.3	12.8	4.6	5.5
0.5	−14	−3.8	13.0	13.9	5.2	5.8
1.0	−8	2	11.5	13.1	4.9	5.3
2.0	−8	0	10.1	13.6	4.9	5.0
4.0	−8	−2	9.1	13.6	5.1	5.1

Table 1 reveals that the asymptotic variance estimates have large negative percentages of relative bias. These percentages are reduced to acceptable levels by the Kenward-Roger method. The observed sizes of the asymptotic tests are significantly inflated, whereas the observed sizes of the containment and Kenward-Roger methods are rather close to the nominal value. Note that with increasing ρ, the contribution of between-subject information increases. As a result, the percentage of bias in variance estimate decreases, the observed test sizes get closer to nominal value, and the Kenward-Roger effective degrees of freedom decrease. For this crossover design, between-subject information contributes far less to the estimation of the treatment effect contrast than does the within-subject information. In such a case, the type I errors of the containment method are comparable to those from the Kenward-Roger method, despite significant small sample bias in the asymptotic covariance estimator and the large difference in degrees of freedoms.

The second simulation study is a repeated measurements experiment. This is a balanced design in which each of three treatments is randomly assigned to eight patients. Measurements at the end of four periods are recorded for each patient. The model is

$$Y_{ijk} = \mu + trt_i + p_j + e_{ijk}$$

where trt_i and p_j are the treatment and period effects, respectively, and $e_{ij} \sim N(0,\Sigma)$ is the random error. In the simulation, the residual covariance has an ANTE(1) structure,

$$\Sigma = \begin{pmatrix} 1 & & & \\ 0.54 & 2.81 & & \\ 0.33 & 0.61 & 4.8 & \\ 0.20 & 0.37 & 0.61 & 6.35 \end{pmatrix}$$

and $10,000$ sets of data are simulated. The simulation results are shown in Table 2.

The percentage of relative bias of the asymptotic estimator of the treatment effect is -23%; the Kenward-Roger method reduces this bias to -9%. The degrees of freedom from the between-within method and the Kenward-Roger method are not far apart. The observed test size of the treatment effect decreases from 11.7% of the asymptotic test to 6.4% of the

Table 2: Simulation Results for Repeated Measurements Study

	% Relative bias variance estimates		df		Observed size Nominal 5% F tests			
Parameter	Asy.	KR	BW	KR	Asy.	BW	Sat.	KR
trt	−23	−9	21	28.7	11.7	9.7	9.5	6.4
period	2	2	69	28.1	7.8	6.8	5.4	4.8

Kenward-Roger method. For the period effect, both the asymptotic test and the Kenward-Roger method produce small relative bias. The between-within degrees of freedom are 69, which is much larger than the Kenward-Roger degrees of freedom, 28. The observed test size decreases from 7.8% of the asymptotic test to 4.8% of the Kenward-Roger method. Type I error from the Kenward-Roger method is reduced but still inflated for treatment effect, because significant bias still exists in the adjusted variance estimator. In contrast, the biases in $\hat{\Phi}$ and $\hat{\Phi}^*$ are both small for the period effect, leading to a Kenward-Roger type I error that is very close to the nominal value.

5 Summary

The Kenward-Roger method tends to produce type I errors that are closer to nominal values when the bias adjustment of the covariance estimator has a significant impact. This is the case when significant small sample bias exists in the asymptotic estimator and when the information contribution of the interblock stratum to the contrast is dominant. You can expect to see significant small sample bias when the covariance of the contrast is estimated using information from multiple strata and the effective sample size is small. In practice, the Kenward-Roger method usually outperforms other methods in balanced incomplete block designs and repeated measures. Also, in hierarchic design it shows an advantage as soon as inference runs across the stratum levels.

References

[1] Beitler, P. J. and Landis, J. R., A Mixed-Effects Model for Categorical Data, Biometrics, 1985, 41, 991–1000.

[2] Satterthwaite, F. E., An Approximate Distribution of Estimates of Variance Components, Biometrics Bulletin, 1946, 2, 110–114.

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

[4] Harville, D. A. and Jeske, D. R., Mean Squared Error of Estimation or Prediction under a General Linear Model, Journal of the American Statistical Association, 1992, 87, 724–731.

[5] Fai, A. H. T. and Cornelius, P. L., Approximate F-Tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-Plot Experiments, Journal of Statistical Computation and Simulation, 1996, 54, 363–378.

[6] Kenward, M. G. and Roger, J. H., Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood, Biometrics, 1997, 53, 983–997.

[7] Kenward, M. G. and Roger, J. H., An Improved Approximation to the Precision of Fixed Effects from Restricted Maximum Likelihood, Computational Statistics and Data Analysis, 2009, 53, 2583–2595.