A Note on Bootstrap Tests for Variance Components in Generalized Linear Mixed Models

Sanjoy K. Sinha

Carleton University

School of Mathematics and Statistics

Ottawa, ON, K1S 5B6 Canada

e-mail: sinha@math.carleton.ca

December 8, 2008

Abstract

In many applications of generalized linear mixed models to clustered correlated or longitudinal data, often we are interested in testing whether a random effects variance component is zero. The usual asymptotic mixture of chi-square distributions of the score statistic for testing constrained variance components does not necessarily hold. In this paper, we propose and explore a parametric bootstrap test that appears to be valid based on its estimated level of significance under the null hypothesis. Results from a simulation study indicate that the bootstrap test has a level much closer to the nominal one while the asymptotic test is conservative, and is more powerful than the usual asymptotic score test based on a mixture of chi-squares. The proposed bootstrap test is illustrated using three sets of real-life data obtained from clinical trials.

KEY WORDS: Bootstrap test; Generalized linear model; Likelihood ratio test; Mixed model; Score test; Variance component.

1 INTRODUCTION

Generalized linear mixed models (GLMMs) are commonly used in the analysis of clustered data including longitudinal data or repeated measurements (see, for example, Breslow and Clayton 1993). GLMMs are useful for accommodating the overdispersion often observed among nonnormally distributed responses and for modeling the dependence among responses inherent in longitudinal or repeated measures data by incorporating random effects (Stiratelli, Laird, and Ware 1984; Zeger, Liang, and Albert 1988). It is usually assumed that the random effects have a multivariate normal distribution with mean vector zero and a covariance matrix depending on some variance components.

In many applied statistical problems, often there is a need for inference on variance components. Many authors considered testing for variance components in a variety of models, which include overdispersion in binomial and Poisson regression models (Cox 1983; Breslow 1984; Dean and Lawless 1989; Dean 1992), linear mixed models (Verbeke and Molenberghs 2000, 2003), longitudinal mixed models (Stram and Lee 1994; Diggle, Heagerty, Liang, and Zeger 2002), and generalized linear mixed models (McCulloch and Searle 2000; Fitzmaurice, Lipsitz, and Ibrahim 2007). While there is much discussion on testing the variance components in the context of a likelihood ratio test, only a few authors studied score tests for testing the variance components in generalized linear mixed models, which include Jacqmin-Gadda and Commenges (1995); Lin (1997); and Hall and Praestgaard (2001). Jacqmin-Gadda and Commenges (1995) proposed a score test for testing homogeneity among clustered data adjusting for the affects of covariates. Lin (1997) developed a score test based on an integrated quasi likelihood function for the marginal distribution of the response vector. Lin (1997) did not specify the alternative hypothesis for the variance components explicitly, thereby implicitly assumed two-sided alternatives while being in the setting of a one-sided test. Hall and Praestgaard (2001) suggested the use of restricted score tests to improve upon the earlier work of Lin (1997) in terms of efficiency.

In this paper, we explore a score test that is derived from a Taylor series expansion of the likelihood function about the mean of the random effects. The score test is attractive in that it only requires estimation of the model parameters under the null hypothesis

that the variance component is zero. In finite-sample inference, it is well-known that the usual one-sided score tests and likelihood ratio tests based on mixtures of chi-squares often result in incorrect estimates of the level of significance (see, for example, Shephard and Harvey 1990; Shephard 1993; Pinheiro and Bates 2000; Crainiceanu, Ruppert, and Vogelsang 2003; Crainiceanu and Ruppert 2004; Fitzmaurice, Lipsitz, and Ibrahim 2007). In this note we propose a bootstrap test that approximates the p-value of a one-sided score test for the variance component. The proposed test provides a level of significance for finite samples that is closer to the nominal level, and is also more powerful than tests based on mixtures of chi-square distributions.

The paper is organized as follows. Section 2 introduces the score test for the variance component in the context of a generalized linear mixed model, and also describes the proposed bootstrap method for approximating the p-value of the score test. Section 3 presents two examples from binary and Poisson mixed models, and illustrates the calculation of the score statistic for the variance components. Section 4 presents results from a simulation study, which was conducted to explore the finite-sample properties of the proposed bootstrap test, and the score test based on asymptotic mixtures of chi-squares. Section 5 discusses applications of the bootstrap test to some actual binary and count data obtained from clinical experiments. Section 6 describes an extension of the score test for testing a subset of the variance components in generalized linear mixed models. Section 7 concludes the paper with some discussion.

2 SCORE TESTS

For simplicity, we first describe the score test using a generalized linear mixed model with a single intercept random effect. Let y_{ij} be the outcome for the jth member of cluster i $(i = 1, \ldots, k; j = 1, \ldots, n_i)$. Given a random effect u_i, assume that the observations y_{ij} from ith cluster are independent and follow a distribution in the exponential family:

$$f_{y_{ij}|u_i}(y_{ij}|u_i, \boldsymbol{\beta}, \phi) = \exp\{(y_{ij}\theta_{ij} - b(\theta_{ij}))/a(\phi) + c(y_{ij}, \phi)\} \tag{1}$$

for some functions a, b and c. Here the canonical parameter $\theta_{ij} = \mathbf{x}_{ij}^t\boldsymbol{\beta} + u_i$, with $\mathbf{x}_{ij}$ being the covariate vector for the jth member of the ith cluster. Also assume that the random

effects u_i are independently and normally distributed with mean 0 and variance σ^2. Since the mean responses within a cluster share the same random effect, the observations within clusters are not independent. In the limit as $\sigma^2 \to 0$, the observations tend to be independent.

In many practical situations including binary and Poisson models, the dispersion parameter ϕ is fixed at unity. So we consider $\phi = 1$, for simplicity. In some situations, however, it may be necessary to estimate the nuisance parameter ϕ together with other nuisance parameters $\boldsymbol{\beta}$ when calculating the score statistic for testing the composite hypothesis $H_0 : \sigma^2 = 0$.

From (1), the marginal density of the ith response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^t$ can be obtained as $f_{y_i}(\mathbf{y}_i) = E_{u_i}[f_{y_i|u_i}(\mathbf{y}_i|u_i)]$, where E_{u_i} denotes the expectation with respect to the distribution of u_i. Following Cox (1983), we can expand $f_{y_i|u_i}(\mathbf{y}_i|u_i)$ using a Taylor series expansion about $E_{u_i}(u_i) = 0$, and take expectations with respect to u_i to obtain the marginal density of $\mathbf{y}_i$ as

$$f_{y_i}(\mathbf{y}_i) = f_{y_i|u_i}(\mathbf{y}_i|u_i = 0, \boldsymbol{\beta}) + \sum_{r=2}^{\infty} \frac{\alpha_r}{r!} \left\{ \frac{\partial^r}{\partial u_i^r} f_{y_i|u_i}(\mathbf{y}_i|u_i, \boldsymbol{\beta}) \Big|_{u_i=0} \right\}, \tag{2}$$

where $\alpha_r = E_{u_i}[u_i^r]$.

The score test for testing $H_0 : \sigma^2 = 0$ against the one-sided alternative $H_1 : \sigma^2 > 0$ is based on the score function

$$U_0 = \sum_{i=1}^{k} \frac{\partial l_i}{\partial \sigma^2} \Big|_{\sigma^2=0} = \sum_{i=1}^{k} S_i(\tilde{\boldsymbol{\beta}}), \tag{3}$$

where $l_i = \log f_{y_i}(\mathbf{y}_i)$ is the marginal log-likelihood for the ith response vector $\mathbf{y}_i$, $\tilde{\boldsymbol{\beta}}$ is the ML estimator of the nuisance parameter $\boldsymbol{\beta}$ under $H_0 : \sigma^2 = 0$, and $S_i(\boldsymbol{\beta})$ is considered to be the score function of σ^2 for the ith cluster evaluated at the null hypothesis that the

variance component is zero:

$$\begin{aligned}
S_i(\boldsymbol{\beta}) &= \left.\frac{\partial \log f_{y_i}(\mathbf{y}_i)}{\partial \sigma^2}\right|_{\sigma^2=0} \\
&= \frac{1}{2}\left\{\left.\frac{\partial^2}{\partial u_i^2} f_{y_i|u_i}(\mathbf{y}_i|u_i,\boldsymbol{\beta})\right|_{u_i=0}\right\}\left\{f_{y_i|u_i}(\mathbf{y}_i|u_i=0,\boldsymbol{\beta})\right\}^{-1} \\
&= \frac{1}{2}\left\{\left(\sum_{j=1}^{n_i}\frac{\partial}{\partial u_i}\log f_{y_{ij}|u_i}(y_{ij}|u_i,\boldsymbol{\beta})\right)^2 + \sum_{j=1}^{n_i}\frac{\partial^2}{\partial u_i^2}\log f_{y_{ij}|u_i}(y_{ij}|u_i,\boldsymbol{\beta})\right\}\Bigg|_{u_i=0} \\
&= \frac{1}{2}\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}\right)^2 - \sum_{j=1}^{n_i} b''(\theta_{ij})\right\} \qquad (4)
\end{aligned}$$

with θ_{ij} being evaluated at $u_i = 0$. Note that in the case of independent binomial and Poisson regression models, (4) corresponds to the overdispersion test statistic of Cox (1983) and Dean (1992). The asymptotic variance of the score function U_0 can be derived from the Fisher information matrix:

$$I(\boldsymbol{\beta}) = \begin{pmatrix} I_{\beta\beta} & I_{\beta\sigma} \\ I_{\sigma\beta} & I_{\sigma\sigma} \end{pmatrix},$$

evaluated at $\sigma^2 = 0$, where

$$\begin{aligned}
I_{\beta\beta} &= \sum_{i=1}^{k} E\left\{-\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\right\}\Bigg|_{\sigma^2=0} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} E\left\{-\frac{\partial^2 \log f(y_{ij}|u_i=0,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\right\} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} b''(\theta_{ij})\mathbf{x}_{ij}\mathbf{x}_{ij}^t, \qquad (5)
\end{aligned}$$

$$\begin{aligned}
I_{\beta\sigma} = I_{\sigma\beta}^t &= \sum_{i=1}^{k} E\left\{-\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\sigma^2}\right\}\Bigg|_{\sigma^2=0} \\
&= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i} b'''(\theta_{ij})\mathbf{x}_{ij}, \qquad (6)
\end{aligned}$$

and

$$\begin{aligned}
I_{\sigma\sigma} &= \sum_{i=1}^{k} E\left\{-\frac{\partial^2 l_i}{\partial(\sigma^2)^2}\right\}\Bigg|_{\sigma^2=0} \\
&= \frac{1}{4}\sum_{i=1}^{k}\left\{2\left(\sum_{j=1}^{n_i} b''(\theta_{ij})\right)^2 + \sum_{j=1}^{n_i} b''''(\theta_{ij})\right\}, \qquad (7)
\end{aligned}$$

with θ_{ij} in (5)-(7) being evaluated at $u_i = 0$.

The asymptotic variance of the score function U_0 is obtained as

$$D(\boldsymbol{\beta}) = I_{\sigma\sigma} - I_{\sigma\beta} I_{\beta\beta}^{-1} I_{\beta\sigma}. \tag{8}$$

Similarly to Silvapulle and Silvapulle (1995) (see also Kudo 1963), for testing $H_0 : \sigma^2 = 0$ against the one-sided alternative $H_1 : \sigma^2 > 0$ we use the score statistic:

$$T = \frac{U_0^2}{\tilde{D}} - \inf\left\{ \frac{(U_0 - d)^2}{\tilde{D}} : d > 0 \right\}, \tag{9}$$

where $\tilde{D} = D(\tilde{\boldsymbol{\beta}})$ with $\tilde{\boldsymbol{\beta}}$ being the ML estimator of the nuisance parameter $\boldsymbol{\beta}$ under the null $H_0 : \sigma^2 = 0$. The construction of this score statistic is motivated by the fact that in the limit as $k \to \infty$, the score statistic becomes the likelihood ratio statistic (see Silvapulle and Silvapulle 1995). When $\hat{\sigma}^2$ is positive, the score at zero is positive, and so in $\{d > 0\}$ the infimum in (9) becomes zero. For negative $\hat{\sigma}^2$, the score at zero is also negative, so the infimum in (9) is attained at $d = 0$ and the score statistic becomes $T = 0$. Verbeke and Molenberghs (2003) pointed out that it was crucial to have valid models for sufficiently small but negative values of σ^2, even under constrained setting.

Under the null hypothesis $H_0 : \sigma^2 = 0$, the score statistic T for the one-sided test $H_1 : \sigma^2 > 0$ does not have the usual chi-square distribution since the value of σ^2 under the null is on the boundary of the parameter space. As the number of clusters $k \to \infty$, the score statistic has the mixture distribution:

$$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2, \tag{10}$$

where χ_0^2 is a point mass at 0 and χ_1^2 is a chi-square distribution with one degree of freedom. However, for fixed cluster size k, this mixture of chi-square may lead to incorrect nominal level of significance (Type I error). For a linear mixed model with a fixed intercept and a random cluster effect, Crainiceanu and Ruppert (2004) showed that for fixed k and $n_i \to \infty$, the likelihood ratio statistic has the mixture distribution:

$$(1 - a_k)\chi_0^2 + a_k\chi_1^2,$$

where a_k is determined by cluster size k. Fitzmaurice, Lipsitz and Ibrahim (2007) point out that the appropriate mixture of chi-squares for likelihood ratio (or score) tests in GLMMs with nonidentity link functions is not straightforward to obtain.

In this paper, we propose and explore a parametric bootstrap test to overcome the aforementioned problems. This bootstrap test produces a one-sided p-value of the score test and preserves approximately the correct Type I error under the null hypothesis even for small sample sizes. To obtain a Monte Carlo estimate of the p-value of the score test, we adopt the following algorithm:

(i) For a given sample, estimate the regression coefficient $\boldsymbol{\beta}$ under the null $H_0 : \sigma^2 = 0$, denoted by $\tilde{\boldsymbol{\beta}}$. Then compute the score statistic T in (9), denoted by T_{obs}.

(ii) For given $\tilde{\boldsymbol{\beta}}$ and under the null $H_0 : \sigma^2 = 0$, generate a bootstrap sample of size $N = \sum_{i=1}^{k} n_i$ from an ordinary generalized linear model with only the fixed effects. Calculate the corresponding bootstrap estimate $\tilde{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}$ under $H_0 : \sigma^2 = 0$, and hence the bootstrap score statistic T^*.

(iii) Repeat the above process to generate B bootstrap samples, and hence obtain the corresponding bootstrap score statistics $T_b^*, b = 1, \ldots, B$.

(iv) The bootstrap p-value is obtained as the proportion of samples with T_b^* greater than or equal to T_{obs}.

To maintain good accuracy of the bootstrap estimates of the p-values, one should use a large number of bootstrap samples. We suggest $B = 1000$ bootstrap samples as a minimum for a test at the 5% level of significance. However, the size of the bootstrap samples may vary depending upon the complexity of the models considered. Although this bootstrap method is computationally intensive, it provides feasible solutions for a variety of data configurations.

The score test discussed above extends in a natural way to tests for multiple variance components. One can consider that the canonical parameter θ_{ij} in (1) has the linear form:

$$\theta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{z}_{ij}^t \mathbf{u},$$

where $\mathbf{z}_{ij}$ corresponds to the jth member of the ith cluster and is a known vector from the design matrix $\mathbf{Z}$ for the random effects. The vector $\mathbf{u}$ of random effects can be assumed to have a distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a $q \times 1$ vector of unknown variance components. Without loss of generality, we can also

postulate that each component of $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$ is a linear function of $\boldsymbol{\alpha}$, and that $\boldsymbol{\Sigma}(\boldsymbol{\alpha}) = \mathbf{0}$ if $\boldsymbol{\alpha} = \mathbf{0}$. Here testing a general hypothesis of the form

$$H_0 : \boldsymbol{\alpha} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\alpha} \in C$$

may be of interest, where the alternative parameter space C equals the nonnegative real numbers (for example, when testing $H_0 : \sigma^2 = 0$ versus $H_1 : \sigma^2 > 0$ in mixed models with a single intercept random effect), or the set of positive semidefinite covariance matrices $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$ (for example, when testing the variance components in mixed models with both intercept and slope random effects). Similar to the test statistic (9), a one-sided score statistic for testing $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ can be obtained in the form:

$$T = U_0^t \tilde{D}^{-1} U_0 - \inf \left\{ (U_0 - d)^t \tilde{D}^{-1} (U_0 - d) : d \in C \right\},$$

where U_0 is the score function defined for the multivariate random effects $\mathbf{u}$, and $\tilde{D}$ is the corresponding covariance matrix of the score function (see Lin 1997, for details). Here also under the null, generalized linear mixed models become ordinary generalized linear models. Silvapulle and Silvapulle (1995) showed that the above score statistic T has the asymptotic mixture distribution:

$$\frac{1}{2}\chi^2_{q-1} + \frac{1}{2}\chi^2_q.$$

However, as for testing a single variance component discussed earlier, a finite-sample score test for multiple variance components based on the asymptotic mixture of chi-squares may also lead to an incorrect level of significance. The proposed bootstrap test can be extended here to approximate the p-value of the score test. As before, we only need to generate the bootstrap samples from an ordinary generalized linear model under the null that the variance components are zero.

3 ILLUSTRATIVE EXAMPLES

Here we provide some details for calculating the score statistics in simple binary and Poisson mixed models.

3.1 A Binary Mixed Model

Consider a binary mixed model with a single intercept random effect:

$$\begin{aligned} y_{ij}|u_i \sim \text{independent Bernoulli}(p_{ij}),\ i=1,\ldots,k;\ j=1,\ldots,n, \\ \theta_{ij} = \log\{p_{ij}/(1-p_{ij})\} = \beta_0 + u_i, \\ u_i \sim \text{independent } N(0,\sigma^2). \end{aligned} \tag{11}$$

Since the mean responses within a cluster share a common random effect, the observations within clusters are correlated. Here the conditional means and variances are $E[Y_{ij}|u_i] = \mu_{ij}(\beta_0, u_i) = \exp(\beta_0 + u_i)/\{1+\exp(\beta_0+u_i)\}$ and $\text{var}[Y_{ij}|u_i] = V_{ij}(\beta_0,u_i) = \exp(\beta_0 + u_i)/\{1+\exp(\beta_0+u_i)\}^2$.

For model (11), and the score function in (3) becomes

$$U_0 = \frac{1}{2}\sum_{i=1}^{k}\left\{\left(\sum_{j=1}^{n}\{y_{ij}-\mu_{ij}(\beta_0,u_i)\}\right)^2 - \sum_{j=1}^{n}V_{ij}(\beta_0,u_i)\right\}, \tag{12}$$

evaluated at $u_i = 0$. Also, for the Fisher information matrix we have under $H_0 : \sigma^2 = 0$:

$$I_{\beta\beta} = \sum_{i=1}^{k}\sum_{j=1}^{n}V_{ij}(\beta_0,u_i),$$

$$I_{\beta\sigma} = I_{\sigma\beta}^t = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n}\{1-2\mu_{ij}(\beta_0,u_i)\}V_{ij}(\beta_0,u_i),$$

and

$$I_{\sigma\sigma} = \frac{1}{4}\sum_{i=1}^{k}\left\{2\left(\sum_{j=1}^{n}V_{ij}(\beta_0,u_i)\right)^2 + \sum_{j=1}^{n}b''''(\theta_{ij})\right\},$$

with

$$b''''(\theta_{ij}) = -2\{V_{ij}(\beta_0,u_i)\}^2 + \{1-2\mu_{ij}(\beta_0,u_i)\}^2 V_{ij}(\beta_0,u_i).$$

Here $b(\theta_{ij}) = \log\{1+\exp(\theta_{ij})\}$, and all terms in the Fisher information are evaluated at $u_i = 0$.

3.2 A Poisson Mixed Model

Consider a Poisson mixed model with two covariates and a single intercept random effect:

$$\begin{aligned} y_{ij}|u_i \sim \text{independent Poisson}(\lambda_{ij}),\ i = 1, \ldots, k;\ j = 1, \ldots, n, \\ \theta_{ij} = \log(\lambda_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i, \\ u_i \sim \text{independent } N(0, \sigma^2). \end{aligned} \tag{13}$$

Here the conditional means and variances are $E[Y_{ij}|u_i] = \mu_{ij}(\boldsymbol{\beta}, u_i) = \exp(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i)$ and $\text{var}[Y_{ij}|u_i] = V_{ij}(\boldsymbol{\beta}, u_i) = \exp(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i)$.

Also, the score function for testing $H_0 : \sigma^2 = 0$ has the form:

$$U_0 = \frac{1}{2} \sum_{i=1}^{k} \left\{ \left(\sum_{j=1}^{n} \{y_{ij} - \mu_{ij}(\boldsymbol{\beta}, u_i)\} \right)^2 - \sum_{j=1}^{n} \mu_{ij}(\boldsymbol{\beta}, u_i) \right\}, \tag{14}$$

evaluated at $u_i = 0$. For the Fisher information matrix, we have under $H_0 : \sigma^2 = 0$:

$$I_{\beta\beta} = \sum_{i=1}^{k} \sum_{j=1}^{n} \mu_{ij}(\boldsymbol{\beta}, u_i) \mathbf{x}_{ij} \mathbf{x}_{ij}^t,$$

$$I_{\beta\sigma} = I_{\sigma\beta}^t = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{n} \mu_{ij}(\boldsymbol{\beta}, u_i) \mathbf{x}_{ij},$$

and

$$I_{\sigma\sigma} = \frac{1}{4} \sum_{i=1}^{k} \left\{ 2 \left(\sum_{j=1}^{n} \mu_{ij}(\boldsymbol{\beta}, u_i) \right)^2 + \sum_{j=1}^{n} \mu_{ij}(\boldsymbol{\beta}, u_i) \right\},$$

where $\mathbf{x}_{ij} = (1, x_{1ij}, x_{2ij})^t$. All terms in the Fisher information are evaluated at $u_i = 0$.

4 SIMULATION STUDY

To explore the finite-sample properties of the proposed bootstrap test, we ran a series of simulations using clustered correlated data with a random cluster effect. First, data were generated from the binary mixed model (11) with the values of the parameter fixed at $\beta_0 = -1, 1$. For each combination of $k = 20, 30, 40$ clusters and $n = 2, 5, 10$ observations within a cluster, we performed a simulation study based on 1000 replicates of data sets. We set $\sigma^2 = 0$ to examine the level of significance of the proposed test. We compared

Table 1: Empirical level of significance for a binary mixed model (standard error in parenthesis).

	Number of		Cluster size (n_i)		
β_0	clusters (k)	Test	2	5	10
-1	20	Bootstrap	$.040_{(.0062)}$	$.047_{(.0067)}$	$.045_{(.0066)}$
		Mixture	$.017_{(.0041)}$	$.036_{(.0059)}$	$.038_{(.0060)}$
	30	Bootstrap	$.053_{(.0071)}$	$.045_{(.0066)}$	$.047_{(.0067)}$
		Mixture	$.025_{(.0049)}$	$.035_{(.0058)}$	$.046_{(.0066)}$
	40	Bootstrap	$.042_{(.0063)}$	$.051_{(.0070)}$	$.047_{(.0067)}$
		Mixture	$.010_{(.0031)}$	$.038_{(.0060)}$	$.048_{(.0068)}$
1	20	Bootstrap	$.041_{(.0063)}$	$.048_{(.0068)}$	$.044_{(.0065)}$
		Mixture	$.020_{(.0044)}$	$.037_{(.0060)}$	$.035_{(.0058)}$
	30	Bootstrap	$.047_{(.0067)}$	$.042_{(.0063)}$	$.051_{(.0070)}$
		Mixture	$.019_{(.0043)}$	$.031_{(.0055)}$	$.048_{(.0068)}$
	40	Bootstrap	$.050_{(.0069)}$	$.051_{(.0070)}$	$.057_{(.0073)}$
		Mixture	$.020_{(.0044)}$	$.044_{(.0065)}$	$.056_{(.0073)}$

the bootstrap p-value to the p-value obtained from the approximate $(0.5, 0.5)$ mixture of chi-square distributions (10) for the score statistic T in (9). The bootstrap p-value was based on $B = 1000$ bootstrap samples. The empirical level of significance was obtained as the proportion of samples in which a given p-value was less than $\alpha = 0.05$.

Table 1 presents the empirical levels of significance of the tests for the binary mixed model. We note from the table that the level of significance of the bootstrap test is generally much closer to the nominal 0.05 level of significance than a level obtained from the approximate mixture of chi-square distributions. For the latter case, the empirical levels get closer to the nominal 0.05 level only when the cluster size n and the number of clusters k increase. But the bootstrap test provides approximately the correct level of significance in each of the simulation configurations considered. Also, as pointed out by an Associate Editor, from the standard errors of the estimated levels, we can find 95% confidence intervals for the true levels of both asymptotic and bootstrap tests. For example, for an estimated level of 0.045, a 95% confidence interval for the true level should be $[0.0322, 0.0578]$, which implies that all of our bootstrap levels are not significantly different from the nominal 0.05 level whereas many of the mixture levels are significantly different.

Table 2: Empirical level of significance for a Poisson mixed model (standard error in parenthesis).

	Number of		Cluster size (n_i)		
β_0	clusters (k)	Test	2	4	8
-.5	20	Bootstrap	$.045_{(.0066)}$	$.047_{(.0067)}$	$.048_{(.0068)}$
		Mixture	$.019_{(.0043)}$	$.027_{(.0051)}$	$.030_{(.0054)}$
	30	Bootstrap	$.061_{(.0076)}$	$.051_{(.0070)}$	$.054_{(.0071)}$
		Mixture	$.026_{(.0050)}$	$.035_{(.0058)}$	$.045_{(.0066)}$
	40	Bootstrap	$.046_{(.0066)}$	$.055_{(.0072)}$	$.047_{(.0067)}$
		Mixture	$.025_{(.0049)}$	$.036_{(.0059)}$	$.034_{(.0057)}$
.5	20	Bootstrap	$.046_{(.0066)}$	$.048_{(.0068)}$	$.053_{(.0071)}$
		Mixture	$.018_{(.0042)}$	$.029_{(.0053)}$	$.039_{(.0061)}$
	30	Bootstrap	$.052_{(.0070)}$	$.051_{(.0070)}$	$.047_{(.0067)}$
		Mixture	$.027_{(.0051)}$	$.028_{(.0052)}$	$.033_{(.0056)}$
	40	Bootstrap	$.062_{(.0076)}$	$.046_{(.0066)}$	$.054_{(.0071)}$
		Mixture	$.030_{(.0054)}$	$.029_{(.0053)}$	$.040_{(.0062)}$

We then performed a simulation study based on the Poisson mixed model (13) with the values of the regression parameters fixed at $\beta_0 = -0.5, 0.5$, $\beta_1 = .5$, and $\beta_2 = -.5$. The value $\sigma^2 = 0$ was used to examine the empirical level of significance of the proposed bootstrap test. The covariates x_1 and x_2 were generated independently from two distributions $N(0.5, 2^2)$ and uniform$(0, 1)$, respectively. As before, for each simulation configuration, we performed a simulation study based on 1000 replicates of data sets. We used $B = 1000$ bootstrap samples to compute the bootstrap p-value of the score test. Table 2 presents the empirical levels of significance of both the bootstrap test and the test based on the $(0.5, 0.5)$ mixture of chi-square distributions. Here also we observe that the proposed bootstrap test provides levels of significance that are generally close to the nominal level. On the other hand, the mixture of chi-squares gives small estimates of the level of significance, and in many cases, these empirical levels are smaller than 0.040. Also, as noted in the binary case, we observe that when the cluster size is small (here $n = 2$), the empirical level based on this asymptotic score test is conservative.

To investigate the power of the bootstrap test, we used the same simulation configurations as above. To calculate the power of the test, for the binary mixed model (11), we considered $\sigma^2 = 0.25, 0.50$, and for the Poisson mixed model (13), we considered

Table 3: Empirical power for a binary mixed model with a correctly specified random effect: $u_i \sim N(0, \sigma^2)$.

	Number of		Cluster size (n_i)		
(β_0, σ^2)	clusters (k)	Test	2	5	10
(-1, .25)	20	Bootstrap	.049	.158	.373
		Mixture	.026	.129	.348
	30	Bootstrap	.102	.195	.452
		Mixture	.051	.168	.437
	40	Bootstrap	.078	.195	.539
		Mixture	.047	.176	.519
(1, .25)	20	Bootstrap	.061	.173	.354
		Mixture	.030	.141	.324
	30	Bootstrap	.082	.182	.437
		Mixture	.039	.160	.423
	40	Bootstrap	.068	.211	.535
		Mixture	.037	.190	.529
(1, .50)	20	Bootstrap	.079	.305	.657
		Mixture	.045	.262	.629
	30	Bootstrap	.114	.407	.792
		Mixture	.072	.369	.786
	40	Bootstrap	.143	.488	.879
		Mixture	.075	.458	.873

$\sigma^2 = 0.025, 0.050$. We used 1000 simulation replications for each simulation configuration, and also used 1000 bootstrap samples to find the bootstrap p-value of the score test. Table 3 presents the empirical powers of the two tests for the binary mixed model. It is clear from the table that the proposed bootstrap test is uniformly more powerful than the test based on the mixture of chi-square distributions in all simulation configurations considered.

Table 4 presents the empirical powers of the two tests for the Poisson mixed model. Clearly, the bootstrap test is much more powerful than the score test based on the mixture of chi-square distributions. The power advantage of the bootstrap test is more evident in situations where the number of clusters k and cluster size n are relatively small.

We also examined the power of the proposed bootstrap test when the distribution of the random effects is misspecified. Specifically, we generated the data from both the

Table 4: Empirical power for a Poisson mixed model with a correctly specified random effect: $u_i \sim N(0, \sigma^2)$.

	Number of		Cluster size (n_i)		
(β_0, σ^2)	clusters (k)	Test	2	4	8
(-.5, .025)	20	Bootstrap	.081	.098	.190
		Mixture	.026	.061	.151
	30	Bootstrap	.077	.108	.206
		Mixture	.032	.064	.171
	40	Bootstrap	.085	.125	.237
		Mixture	.048	.087	.203
(.5, .025)	20	Bootstrap	.135	.207	.433
		Mixture	.073	.142	.374
	30	Bootstrap	.150	.275	.565
		Mixture	.091	.217	.512
	40	Bootstrap	.168	.349	.661
		Mixture	.119	.293	.621
(.5, .050)	20	Bootstrap	.228	.421	.732
		Mixture	.127	.316	.680
	30	Bootstrap	.278	.556	.855
		Mixture	.192	.463	.838
	40	Bootstrap	.344	.658	.940
		Mixture	.253	.599	.922

Table 5: Empirical power for a binary mixed model with a misspecified random effect: $u_i = \exp(u_i^*) - \exp(\sigma^2/2)$, where $u_i^* \sim N(0, \sigma^2)$.

	Number of		Cluster size (n_i)		
(β_0, σ^2)	clusters (k)	Test	2	5	10
(-1, .25)	20	Bootstrap	.072	.273	.560
		Mixture	.048	.238	.542
	30	Bootstrap	.106	.379	.715
		Mixture	.052	.342	.704
	40	Bootstrap	.111	.423	.764
		Mixture	.064	.397	.754
(1, .25)	20	Bootstrap	.067	.158	.321
		Mixture	.040	.131	.305
	30	Bootstrap	.077	.169	.466
		Mixture	.049	.149	.445
	40	Bootstrap	.072	.222	.540
		Mixture	.037	.199	.540
(1, .50)	20	Bootstrap	.082	.298	.678
		Mixture	.059	.266	.656
	30	Bootstrap	.113	.394	.808
		Mixture	.066	.367	.802
	40	Bootstrap	.134	.471	.884
		Mixture	.089	.438	.880

binary and the Poisson mixed models (11) and (13), respectively, but by considering a log-normal distribution for the random effects u_i with $u_i = \exp(u_i^*) - \exp(\sigma^2/2)$ and $u_i^* \sim N(0, \sigma^2)$. Here the distribution of u_i is log-normal, but centered at zero. We used $\sigma^2 = 0.25, 0.50$ for the binary mixed model, and $\sigma^2 = 0.025, 0.050$ for the Poisson mixed model. As before, for each simulation configuration, 1000 data sets were generated from the two mixed models with the misspecified log-normal random effects. The bootstrap test was based on 1000 bootstrap samples. Tables 5 and 6 present the empirical powers of the two tests for the binary and Poisson mixed models, respectively. It is evident from the results in the tables that the proposed bootstrap test is uniformly more powerful than the test based on the mixture of chi-squares for all configurations considered.

The overall results from the simulation study demonstrate that the proposed bootstrap test has generally the correct level of significance and is more powerful than the

Table 6: Empirical power for a Poisson mixed model with a misspecified random effect: $u_i = \exp(u_i^*) - \exp(\sigma^2/2)$, where $u_i^* \sim N(0, \sigma^2)$.

	Number of		Cluster size (n_i)		
(β_0, σ^2)	clusters (k)	Test	2	4	8
(-.5, .025)	20	Bootstrap	.083	.112	.174
		Mixture	.035	.066	.148
	30	Bootstrap	.085	.130	.240
		Mixture	.039	.083	.200
	40	Bootstrap	.096	.145	.255
		Mixture	.050	.112	.217
(.5, .025)	20	Bootstrap	.145	.214	.440
		Mixture	.066	.153	.380
	30	Bootstrap	.169	.321	.590
		Mixture	.090	.248	.541
	40	Bootstrap	.193	.377	.678
		Mixture	.119	.311	.647
(.5, .050)	20	Bootstrap	.265	.462	.756
		Mixture	.169	.396	.718
	30	Bootstrap	.323	.604	.889
		Mixture	.231	.537	.868
	40	Bootstrap	.410	.702	.960
		Mixture	.307	.651	.946

approximate test based on a mixture of chi-square distributions. Also, the bootstrap test maintains a power advantage even when the distribution of the random effects is misspecified.

A referee pointed out that since the empirical level of a score test based on the mixture of chi-squares is typically smaller than the nominal 0.05 level, a size-corrected test might be more appropriate when estimating the power of the test. The size-adjusted version, however, cannot be used in practice. Moreover, as observed in Tables 1 and 2, when the cluster size n and the number of clusters k increase, the two test procedures provide estimated levels that are close to the nominal 0.05 level. And in these cases too, the proposed bootstrap test appears to be more powerful than the asymptotic test, although there is no dramatic increase in power as we have observed for smaller sample sizes.

5 APPLICATIONS

5.1 GUIDE Data

Preisser and Qaqish (1999) introduced an interesting set of data obtained from the Guidelines for Urinary Incontinence Discussion and Evaluation (GUIDE) study. The purpose of the study was to identify factors among urinary incontinent men and women of age 76 or above that were predictive of the responses to the question of whether individuals in that age group consider this accidental loss of urine a problem that interferes with their day to day activities or bothers them in other ways. The data were collected from 137 patients in 38 medical practices, and the binary response y_{ij} was defined as 1 if the jth patient from the ith medical practice was "bothered" by the urinary incontinence, and 0 if not. The predictors considered were gender, age, weekacc = how many leaking accidents the patients have in an average week, severe = 1 if it just creates some moisture when they lose urine, 2 if it wets their underwear (or pad), 3 if it trickles down their thigh, 4 if it wets the floor, and toilet = how many times during the day they usually to go the toilet to urinate.

Preisser and Qaqish (1999) used a marginal logistic model to describe the GUIDE data, and proposed a resistant generalized estimating equation (REGEE) approach to predict the mean response $\mu_{ij} = E(Y_{ij})$ based on the covariates age = (age in years - 76)/10, dayacc = weekacc/7, severe, toilet and an indicator variable female. Sinha (2004) reanalyzed the data using a binary logistic mixed model. Similarly to Sinha (2004), here we consider a mixed effect logistic model for the conditional mean response $\mu_{ij} = E(y_{ij}|u_i)$:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \text{female}_{ij} + \beta_2 \text{age}_{ij} + \beta_3 \text{dayacc}_{ij} + \beta_4 \text{severe}_{ij} + \beta_5 \text{toilet}_{ij} + u_i,$$

where u_i is the ith cluster (medical practice) effect, and is assumed to be independent normal with mean 0 and variance component σ^2.

To assess the significance of the variance component σ^2, we use the proposed score test in which we first fit the model under the null $H_0 : \sigma^2 = 0$ using the maximum likelihood method. The ML estimates of the model parameters are presented in Table 7. The score test produced a value of 0.6788 for the test statistic. Assuming that the score

Table 7: Logistic model fit to the GUIDE data under $H_0 : \sigma^2 = 0$.

Coefficient	Estimate	STD error	z value
INTERCEPT	-3.293	1.108	-2.971
FEMALE	-0.672	0.612	-1.099
AGE	-0.641	0.585	-1.095
DAYACC	0.415	0.096	4.337
SEVERE	0.829	0.365	2.273
TOILET	0.111	0.086	1.296

statistic has the mixture distribution $0.5\chi_0^2 + 0.5\chi_1^2$, the p-value of the test is obtained as 0.2050. We then applied the proposed bootstrap test based on 2000 bootstrap samples, which produced a p-value of 0.1565. Both methods suggest that there is little evidence of a cluster (medical practice) effect. However, different p-values from the two methods indicate that the $(0.5, 0.5)$ mixture of chi-square distributions for the score statistic may lead to an incorrect inference in certain situations.

5.2 Leprosy Data

Snedecor and Cochran (1980) presented and analyzed some count data from a clinical trial of 30 patients with leprosy at the Eversley Childs Sanitarium in the Philippines. Participants in the study were randomized to either of two antibiotics (treatment drugs A and B) or a placebo (treatment drug C). Ten patients were selected for each treatment. For each patient, baseline data on the number of leprosy bacilli at six sites of the body where the bacilli tend to congregate were recorded prior to receiving any treatment. The number of leprosy bacilli at those six sites were recorded again after several months of treatment. The purpose of the study was to investigate whether treatment with antibiotics (drugs A and B) reduces the abundance of leprosy bacilli at the six sites of the body as compared to placebo (drug C). Snedecor and Cochran (1980) studied the data using a simple analysis a covariance method in a completely randomized design. Fitzmaurice, Laird, and Ware (2004) reanalyzed the data, and studied both overdispersion and within-subject association using a generalized estimating equations approach for marginal models with longitudinal count data.

Here we revisit the data, and consider analyzing them using a mixed effect log-linear model for the conditional mean response $E(y_{ij}|u_i) = \mu_{ij}$:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{time}_{ij} \times \text{trt}_{1i} + \beta_3 \text{time}_{ij} \times \text{trt}_{2i} + u_i,$$

where y_{ij} represents the number of leprosy bacilli for the ith patient observed in the jth period ($j = 1, 2$). The predictors trt_1 and trt_2 are indicator variables for treatment drugs A and B respectively, with $\text{trt}_1 = 1$ if a patient was randomized to drug A and $\text{trt}_1 = 0$ if not, and $\text{trt}_2 = 1$ if a patient was randomized to drug B and $\text{trt}_1 = 0$ if not. The predictor time is also an indicator variable representing the baseline and post-treatment follow-up periods, with time = 0 for the baseline period and time = 1 for the follow-up period. The variable u_i is a subject-specific random effect, and is assumed to be independent normal with mean 0 and variance component σ^2. The model does not include the main effects of treatment since the mean number of leprosy bacilli can be assumed to be equal in the three treatment groups.

To assess the significance of the variance component σ^2, we apply the score test to the leprosy data using both the exact method based on the $(0.5, 0.5)$ mixture of chi-squares and the proposed bootstrap method based on 2000 bootstrap samples. The value of the score statistic is obtained as 17.983. The assumption of the mixture of chi-squares produces a p-value of 1.114248e-05, whereas the bootstrap method produced a p-value of 0.000. Here both methods indicate that there is a very strong evidence against the null, that is, the subject-specific variance component σ^2 may be considered highly significant. Then we fit the above Poisson mixed effect model with an intercept random effect to the leprosy data. The maximum likelihood estimates of the model parameters, and their corresponding asymptotic standard errors are presented in Table 8.

6 SCORE TESTS FOR SUBSETS OF VARIANCE COMPONENTS

In the previous sections, we have discussed the score test for assessing the significance of all variance components in a GLMM. An advantage of using such a score test is that we

Table 8: Poisson mixed model fit to the Leprosy data with an intercept random effect.

Coefficient	Estimate	STD error	z value
INTERCEPT	2.2411	0.1147	19.5388
TIME	0.0031	0.1235	0.0251
TIME $\times$ TRT1	-0.6056	0.2036	-2.9745
TIME $\times$ TRT2	-0.5228	0.1963	-2.6633
σ^2	0.2812	0.0953	2.9507

only need to fit a simple generalized linear model for calculating the score statistic under the null hypothesis that the variance components are zero. In certain situations, however, one may wish to test a subset of the variance components in mixed models. In such cases, the likelihood score test can still be used, but such tests will require somewhat extensive computation, since under the null we still have a mixed model, and the likelihood function cannot be defined in a closed form in general. If the dimension of the random effects is small, we can still evaluate the likelihood using some numerical quadrature method, and hence compute the ML estimators and the corresponding Fisher information matrix. Here I describe the calculation of the score test for a subset of variance components in a GLMM with two variance components assuming that the link function θ_{ij} in (1) has the form:

$$\theta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta} + u_{0i} + z_{ij} u_{1i}, \tag{15}$$

where z_{ij} are elements of the design matrix Z for random effects. The random effects u_{0i} and u_{1i} are assumed to be independent, and follow independent normal distributions with means 0 and variance components σ_0^2 and σ_1^2, respectively.

In the next section, we develop the score test for testing the variance component σ_1^2 only. Here under the null, we still need to fit a generalized linear mixed model with the intercept random effect u_{0i}. The ML estimators of the regression coefficients $\boldsymbol{\beta}$, the variance component σ_0^2 and the Fisher information matrix can be calculated using a numerical integration technique.

6.1 The Score Test for $H_0 : \sigma_1^2 = 0$

Here the score test can be developed based on the score function:

$$U_0^* = \sum_{i=1}^{k} \frac{\partial l_i}{\partial \sigma_1^2}\bigg|_{\sigma_1^2=0} = \sum_{i=1}^{k} S_i^*(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}_0^2), \tag{16}$$

where $l_i = \log f_{y_i}(\mathbf{y}_i)$, with $f_{y_i}(\mathbf{y}_i)$ being the density function for the ith response vector $\mathbf{y}_i$:

$$f_{y_i}(\mathbf{y}_i) = \int f_{y_i|u_i}(\mathbf{y}_i|\mathbf{u}_i) f_{u_i}(\mathbf{u}_i) d\mathbf{u}_i. \tag{17}$$

The integral in (17) is with respect to the two-dimensional random effect $\mathbf{u}_i = (u_{0i}, u_{1i})^t$ with its density function $f_{u_i}(\mathbf{u}_i)$. Also, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}_0^2$ in (16) are the ML estimators of the regression coefficients $\boldsymbol{\beta}$ and the variance component σ_0^2 under $H_0 : \sigma_1^2 = 0$.

Following similar arguments as used in Section 2, we can expand the conditional density $f_{y_i|u_i}(\mathbf{y}_i|\mathbf{u}_i)$ in (17) about $E_{u_{1i}}(u_{1i}) = 0$, and show that the score function $S_i^*(\boldsymbol{\beta}, \sigma_0^2)$ in (16) has the form

$$S_i^*(\boldsymbol{\beta}, \sigma_0^2) = \frac{1}{2}\left[E\left\{ \left(\sum_{j=1}^{n_i} \{y_{ij} - b'(\theta_{ij})\} z_{ij} \right)^2 \middle| \mathbf{y}_i \right\} - E\left\{ \sum_{j=1}^{n_i} b''(\theta_{ij}) z_{ij}^2 \middle| \mathbf{y}_i \right\} \right] \tag{18}$$

with θ_{ij} being evaluated at $u_{1i} = 0$. The expectation in (18) is with respect to the conditional distribution of the random effects u_{0i} given the response vector $\mathbf{y}_i$. Since the random effect u_{0i} is one-dimensional, we can evaluate the expectations in (18) using a numerical integration technique.

An approximate variance of the score function U_0^* can be obtained from the observed Fisher information matrix:

$$I^*(\boldsymbol{\beta}, \sigma_0^2) = \begin{pmatrix} I_{\beta\beta}^* & I_{\beta\sigma_0}^* & I_{\beta\sigma_1}^* \\ I_{\sigma_0\beta}^* & I_{\sigma_0\sigma_0}^* & I_{\sigma_0\sigma_1}^* \\ I_{\sigma_1\beta}^* & I_{\sigma_1\sigma_0}^* & I_{\sigma_1\sigma_1}^* \end{pmatrix} \tag{19}$$

evaluated at $\sigma_1^2 = 0$, where

$$\begin{aligned} I_{\beta\beta}^* = & \sum_{i=1}^{k}\sum_{j=1}^{n_i} E\left\{ b''(\theta_{ij}) | \mathbf{y}_i \right\} \mathbf{x}_{ij}\mathbf{x}_{ij}^t - \sum_{i=1}^{k}\sum_{j=1}^{n_i} E[\{y_{ij} - b'(\theta_{ij})\}^2 | \mathbf{y}_i] \mathbf{x}_{ij}\mathbf{x}_{ij}^t \\ & + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left\{ E[\{y_{ij} - b'(\theta_{ij})\} | \mathbf{y}_i] \right\}^2 \mathbf{x}_{ij}\mathbf{x}_{ij}^t, \end{aligned} \tag{20}$$

$$I^*_{\beta\sigma_0} = I^{*^t}_{\sigma_0\beta} = -\frac{1}{2\sigma_0^4}\sum_{i=1}^{k} E\left\{\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}\mathbf{x}_{ij}\left(u_{0i}^2-\sigma_0^2\right)\middle|\,\mathbf{y}_i\right\} + \frac{1}{2\sigma_0^4}\sum_{i=1}^{k} E\left\{\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}\mathbf{x}_{ij}\middle|\,\mathbf{y}_i\right\} E\left\{\left(u_{0i}^2-\sigma_0^2\right)\middle|\,\mathbf{y}_i\right\}, \quad (21)$$

$$I^*_{\sigma_0\sigma_0} = \frac{1}{2\sigma_0^6}\sum_{i=1}^{k} E\left\{\left(2u_{0i}^2-\sigma_0^2\right)\middle|\,\mathbf{y}_i\right\} - \left(\frac{1}{2\sigma_0^4}\right)^2\sum_{i=1}^{k}\left[E\left\{\left(u_{0i}^2-\sigma_0^2\right)^2\middle|\,\mathbf{y}_i\right\}\right] + \left(\frac{1}{2\sigma_0^4}\right)^2\sum_{i=1}^{k}\left[E\left\{\left(u_{0i}^2-\sigma_0^2\right)\middle|\,\mathbf{y}_i\right\}\right]^2, \quad (22)$$

$$\begin{aligned} I^*_{\beta\sigma_1} = & -\frac{1}{2}\sum_{i=1}^{k} E\left[\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}\mathbf{x}_{ij}\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\right)^2-\sum_{j=1}^{n_i}b''(\theta_{ij})z_{ij}^2\right\}\middle|\,\mathbf{y}_i\right] \\ & +\frac{1}{2}\sum_{i=1}^{k} E\left\{\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}\mathbf{x}_{ij}\middle|\,\mathbf{y}_i\right\} E\left[\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\right)^2-\sum_{j=1}^{n_i}b''(\theta_{ij})z_{ij}^2\right\}\middle|\,\mathbf{y}_i\right] \\ & +\sum_{i=1}^{k} E\left\{\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\sum_{j=1}^{n_i}b''(\theta_{ij})z_{ij}\mathbf{x}_{ij}\middle|\,\mathbf{y}_i\right\}+\frac{1}{2}\sum_{i=1}^{k} E\left[\sum_{j=1}^{n_i}b'''(\theta_{ij})z_{ij}^2\mathbf{x}_{ij}\middle|\,\mathbf{y}_i\right], \quad (23) \end{aligned}$$

$$\begin{aligned} I^*_{\sigma_0\sigma_1} = & -\frac{1}{4\sigma_0^4}\sum_{i=1}^{k} E\left[\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\right)^2-\sum_{j=1}^{n_i}b''(\theta_{ij})z_{ij}^2\right\}\left(u_{0i}^2-\sigma_0^2\right)\middle|\,\mathbf{y}_i\right] \\ & +\frac{1}{4\sigma_0^4}\sum_{i=1}^{k} E\left[\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\right)^2-\sum_{j=1}^{n_i}b''(\theta_{ij})z_{ij}^2\right\}\middle|\,\mathbf{y}_i\right] E\left\{\left(u_{0i}^2-\sigma_0^2\right)\middle|\,\mathbf{y}_i\right\}, \quad (24) \end{aligned}$$

and

$$
\begin{aligned}
I^*_{\sigma_1\sigma_1} &= -\frac{1}{3}\sum_{i=1}^{k} E\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\right)^4\middle|\mathbf{y}_i\right\} \\
&\quad +2\sum_{i=1}^{k} E\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\right)^2\sum_{j=1}^{n_i} b''(\theta_{ij})z_{ij}^2\middle|\mathbf{y}_i\right\} \\
&\quad +2\sum_{i=1}^{k} E\left\{\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\sum_{j=1}^{n_i} b'''(\theta_{ij})z_{ij}^3\middle|\mathbf{y}_i\right\} \\
&\quad -\sum_{i=1}^{k} E\left[\left\{\left(\sum_{j=1}^{n_i} b''(\theta_{ij})z_{ij}^2\right)^2-\frac{1}{3}\sum_{j=1}^{n_i} b''''(\theta_{ij})z_{ij}^4\right\}\middle|\mathbf{y}_i\right] \\
&\quad +\frac{1}{4}\sum_{i=1}^{k}\left[E\left\{\left\{\left(\sum_{j=1}^{n_i}\{y_{ij}-b'(\theta_{ij})\}z_{ij}\right)^2-\sum_{j=1}^{n_i} b''(\theta_{ij})z_{ij}^2\right\}\middle|\mathbf{y}_i\right\}\right]^2. \quad (25)
\end{aligned}
$$

The elements θ_{ij} in (20)-(25) are evaluated at $u_{1i} = 0$. The variance of the score function U_0^* can be approximated as

$$D^*(\boldsymbol{\beta}, \sigma_0^2) = I^*_{\sigma_1\sigma_1} - I^*_{\sigma_1\gamma} I^{*^{-1}}_{\gamma\gamma} I^*_{\gamma\sigma_1}, \quad (26)$$

where $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \sigma_0^2)$ with its corresponding Fisher information $I^*_{\gamma\gamma}$ being obtained from (19).

Using U_0^* and $D^*(\tilde{\boldsymbol{\beta}}, \tilde{\sigma_0^2})$, we can develop a score statistic in a form similarly to (9). To find a bootstrap p-value of the score test, we can adopt the same algorithm as described in Section 2, except that instead of generating the bootstrap samples from an ordinary generalized linear model, here we need to generate them from a GLMM with the intercept random effect u_{0i}, and then find the ML estimators of $\boldsymbol{\beta}$ and σ_0^2 for each bootstrap sample.

The next section presents an application of this score test to some actual count data obtained from a clinical experiment.

6.2 Epilepsy Data

Thall and Vail (1990) presented and analyzed some count data obtained from a clinical trial of 59 epileptics who were randomized to either the antiepileptic drug progabide (trt = 1) or a placebo (trt = 0), as an adjuvant to standard chemotherapy. The number of seizures during the 2-weeks before each of the four clinic visits (visit, coded

$\text{visit}_1 = -3, -1, 1$, $\text{visit}_4 = 3$) was recorded. The predictors considered in the study were the logarithm of $\frac{1}{4}$ the number of baseline seizure count recorded in the preceding 8-week period (base), logarithm of age in years at entry into the trial (age), the binary indicators trt for the progabide group, and visit for the four clinic visits. The seizure counts of the epileptic patients demonstrated a high degree of extra-Poisson variation, heteroscedasticity, and within-patient dependence.

Thall and Vail (1990) analyzed the data by fitting log-linear models for the marginal event rates and the covariance parameters in various forms of covariance matrices based on the generalized quasi-likelihood (GQL) approach similar to the formulation by Liang and Zeger (1986). Breslow and Clayton (1993) reanalyzed these data using a generalized linear mixed model, but used an approximate penalized quasi-likelihood (PQL) method for estimating the regression parameters.

Similarly to Breslow and Clayton (1993), here we used a GLMM to describe the epilepsy data. It is assumed that the conditional mean $E(y_{ij}|\mathbf{u}_i) = \lambda_{ij}$ is related to the linear predictor θ_{ij} by the link function

$$\theta_{ij} = \ln(\lambda_{ij}) = \mathbf{x}_{ij}^t\boldsymbol{\beta} + u_{0i} + u_{1i}\text{visit}_j/10,$$

where $\mathbf{x}_{ij}$ is a vector of the predictors base, trt, age, visit, and the interaction between base and trt. The random effects (u_{0i}, u_{1i}), which represent the residual level and rate of change in the event rate for the ith subject, are assumed to be independent and normally distributed each with mean zero, and with variance components $\text{var}(u_{0i}) = \sigma_0^2$ and $\text{var}(u_{1i}) = \sigma_1^2$.

To test the variance component σ_1^2, we first find the ML estimates of the regression coefficients $\boldsymbol{\beta}$ and the variance component σ_0^2. The ML estimates and their estimated standard errors are presented in Table 9. The score test for $H_0 : \sigma_1^2 = 0$ against $H_1 : \sigma_1^2 > 0$ produced a value of 8.3368 for the score statistic. The p-value of the test based on the (0.5, 0.5) mixture of chi-squares is obtained as 0.0019, whereas the proposed bootstrap test based on 2000 bootstrap samples produced an estimated p-value 0.0025. Clearly, both asymptotic and bootstrap p-values indicate strong evidence against the null.

Table 9: Poisson mixed model fit to the Epilepsy data with an intercept random effect.

Coefficient	Estimate	STD error	z value
INTERCEPT	-1.3775	1.1823	-1.1651
BASE	0.8844	0.1312	6.7409
TRT	-0.9330	0.4008	-2.3278
BASE $\times$ TRT	0.3383	0.2033	1.6640
AGE	0.4842	0.3473	1.3942
VISIT/10	-0.2936	0.1014	-2.8955
σ_0^2	0.2528	0.0590	4.2847

7 DISCUSSION

For testing the variance components in generalized linear mixed models, the proposed bootstrap test is a simple alternative to computing approximate p-values based on a mixture of chi-square distributions. We have demonstrated in the simulations that the bootstrap test has a level of significance close to the nominal level. In addition, the simulation results indicate that the bootstrap test is more powerful than tests based on mixtures of chi-square distributions, and also maintains higher powers in cases of misspecified (non-Gaussian) random effects. The bootstrap test is easy to implement; one only needs to generate the bootstrap samples from simple generalized linear models under the null that the variance components are zero. Also, the bootstrap test requires estimation of only the fixed effects regression coefficients under the null hypothesis.

For testing a subset of variance components in generalized linear mixed models, we have demonstrated that the proposed bootstrap test can still be used, but in such cases, the bootstrap method will require somewhat extensive computation for evaluating the marginal likelihood by integrating out the random effects. If the dimension of the random effects is relatively small (two or three), then one can still evaluate the likelihood using some numerical method. For high-dimensional random effects, however, the calculation of the score test would be tedious, and in such cases, some approximate methods may be investigated in a future study.

REFERENCES

Breslow, N. E. (1984), "Extra-Poisson Variation in Log-Linear Models," *Applied Statistics*, 33, 38-44.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models", *Journal of the American Statistical Association*, 88, 9-25.

Cox, D. R. (1983), "Some Remarks on Overdispersion," *Biometrika*, 70, 269-74.

Crainiceanu, C., Ruppert, D., and Vogelsang (2003), "Probability that the MLE of a Variance Component is Zero with Applications to Likelihood Ratio Tests," Unpublished manuscript (available from www.orie.cornell.edu/ daviddr/papers).

Crainiceanu, C. and Ruppert, D. (2004), "Likelihood Ratio Tests in Linear Mixed Models with One Variance Component," *Journal of the Royal Statistical Society*, Series B, 66, 165-185.

Dean, C. B. and Lawless, J. F. (1989), "Tests for Detecting Overdispersion in Poisson Regression Models," *Journal of the American Statistical Association*, 84, 467-472.

Dean, C. B. (1992), "Testing for Overdispersion in Poisson and Binomial Regression Models," *Journal of the American Statistical Association*, 87, 451-457.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data*, Oxford University Press, 2nd Edition.

Fitzmaurice, G. M., Lipsitz, S. R., and Ibrahim, J. G. (2007), "A Note on Permutation Tests for Variance Components in Multilevel Generalized Linear Mixed Models," *Biometrics*, 63, 942-946.

Hall, D. B. and Praestgaard, J. T. (2001), "Order-Restricted Score Tests for Homogeneity in Generalised Linear and Nonlinear Mixed Models," *Biometrika*, 88, 739-751.

Jacqmin-Gadda, H. and Commenges, D. (1995), "Tests of Homogeneity for Generalized Linear Models," *Journal of the American Statistical Association*, 90, 1237-1246.

Kudo, A. (1963), "A Multivariate Analogue of the One-Sided Test," *Biometrika*, 50, 403-418.

Lin, X. (1997), "Variance Component Testing in Generalised Linear Models with Random Effects," *Biometrika*, 84, 309-326.

McCulloch, C. E. and Searle, S. R. (2000), *Generalized, Linear, and Mixed Models*, New York: Wiley .

Pinheiro, J. C. and Bates, D. M. (2000), *Mixed-Effects Models in S and S-Plus*, Springer-Verlag, New York.

Preisser, J. S. and Qaqish, B. F. (1999), "Robust Regression to Clustered Data With Application to Binary Responses," *Biometrics*, 55, 574-579.

Shephard, N. G. and Harvey, A. C. (1990), "On the Probability of Estimating a Deterministic Component in the Local Level Model," *Journal of Time Series Analysis*, 4, 339-347.

Shephard, N. G. (1993), "Maximum Likelihood Estimation of Regression Models with Stochastic Trend components," *Journal of the American Statistical Association*, 88, 590-595.

Silvapulle, M. J. and Silvapulle, P. (1995), "A Score Test Against One-Sided Alternatives," *Journal of the American Statistical Association*, 90, 342-349.

Sinha, S. K. (2004), "Robust Analysis of Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 99, 451-460

Snedecor, G. W. and Cochran, W. G. (1980), *Statistical Methods*, 7th ed. Ames, Iowa: Iowa State University Press.

Stiratelli, R., Laird, N. and Ware, J. (1984), "Random Effects Models for Serial Observations With Binary Responses," *Biometrics*, 40, 961-971.

Stram, D. O. and Lee, J. W. (1994), "Variance Components Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 50, 1171-1177.

Thall, P. F. and Vail, S. C. (1990), "Some Covariance Models for Longitudinal Count Data With Overdispersion," *Biometrics*, 46, 657-671.

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

Verbeke, G. and Molenberghs, G. (2003), "The Use of Score Tests for Inference on Variance Components," *Biometrics*, 59, 254-262.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049-1060.