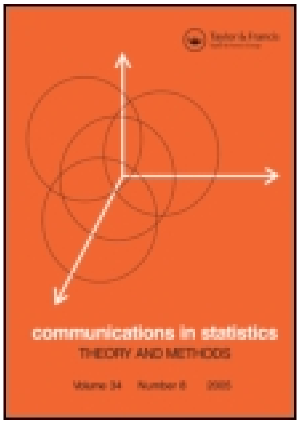

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

The Use of Permutation Tests for Variance Components in Linear Mixed Models

Monjed H. Samuh [a], Leonardo Grilli [b], Carla Rampichini [b], Luigi Salmaso [c] & Nicola Lunardon [a]

[a] Department of Statistical Sciences, University of Padua, Padua, Italy

[b] Department of Statistics, University of Florence, Florence, Italy

[c] Department of Management and Engineering, University of Padua, Italy

Version of record first published: 17 Jul 2012

Communications in Statistics—Theory and Methods, 41: 3020–3029, 2012

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2011.587933

The Use of Permutation Tests for Variance Components in Linear Mixed Models

MONJED H. SAMUH[1], LEONARDO GRILLI[2], CARLA RAMPICHINI[2], LUIGI SALMASO[3], AND NICOLA LUNARDON[1]

[1]Department of Statistical Sciences, University of Padua, Padua, Italy
[2]Department of Statistics, University of Florence, Florence, Italy
[3]Department of Management and Engineering, University of Padua, Italy

Standard asymptotic chi-square distribution of the likelihood ratio and score statistics under the null hypothesis does not hold when the parameter value is on the boundary of the parameter space. In mixed models it is of interest to test for a zero random effect variance component. Some available tests for the variance component are reviewed and a new test within the permutation framework is presented. The power and significance level of the different tests are investigated by means of a Monte Carlo simulation study. The proposed test has a significance level closer to the nominal one and it is more powerful.

Keywords Bootstrap test; Likelihood ratio test; Linear mixed model; Permutation test; Variance component.

Mathematics Subject Classification Primary 62G09; Secondary 62J12.

1. Introduction

Mixed models (e.g., Verbeke and Molenbeghs, 2000), hierarchical models (e.g., Raudenbush and Bryk, 2002), or multilevel regression models (e.g., Snijders and Bosker, 1999) are an extension of regression models in which data have a hierarchical structure with units nested in clusters. A common application is on individuals nested in institutions or organizations, for example students in schools, employees in firms, or patients in hospitals. Another kind of application is on repeated measures where occasions are nested in individuals.

Mixed models are widely used in many research fields such as social sciences (see Afshartous, 2004), econometrics (see Swamy, 1970), and political science (see Garner and Raudenbush, 1991).

To facilitate calculations and clarify ideas, the simplest case of linear mixed models, i.e., the random-intercept model involving two levels of analysis, is

Received December 10, 2010; Accepted April 15, 2012
Address correspondence to Monjed H. Samuh, College of Applied Sciences, Palestine Polytechnic University, Hebron, Palestine; E-mail: monjedsamuh@ppu.edu

considered as a guide. Level one units are referred to as *subjects* and level two units as *clusters*. We consider a model with one level-1 predictor, which is observable and has a linear relationship with the level-1 dependent variable.

Let the random variable Y_{ij} denote the response of interest for the ith subject in the jth cluster, X_{ij} denote the related observed covariate, β_1 is a fixed parameter or regression coefficient, γ_{0j} is the cluster intercept, β_0 is the average intercept across the clusters, ε_{ij} is the level-1 residual, and u_j is the level-2 residual. The level-1 model, which relates the response variable to the covariate, is written as

$$Y_{ij} = \gamma_{0j} + \beta_1 X_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n_j; \;\; j = 1, \ldots, J, \tag{1}$$

while the level-2 model, describing the variation between clusters, is written as

$$\gamma_{0j} = \beta_0 + u_j, \quad j = 1, \ldots, J \tag{2}$$

When we combine Eqs. (1) and (2) into a single equation, we get one that looks like a common regression equation with an extra error term u_j. This error term indicates that the mean intercepts can randomly differ across clusters. The combined model is written as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij}, \quad i = 1, \ldots, n_j; \;\; j = 1, \ldots, J. \tag{3}$$

For fixed X_{ij}, the essential assumptions for the random-intercept model are that: (1) the u_j are independently and identically normally distributed with $\mathbb{E}(u_j) = 0$ and $Var(u_j) = \sigma_u^2$; (2) ε_{ij} are independently and identically normally distributed with $\mathbb{E}(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \sigma_\varepsilon^2$; and (3) u_j and ε_{ij} are independent. Under these assumptions it follows that

$$\mathbb{E}(Y_{ij}) = \beta_0 + \beta_1 X_{ij}$$

and

$$Cov(Y_{ij}, Y_{i'j'}) = \begin{cases} \sigma_u^2 + \sigma_\varepsilon^2, & \text{for } i = i' \text{ and } j = j' \\ \sigma_u^2, & \text{for } i \neq i' \text{ and } j = j' \\ 0, & \text{otherwise} \end{cases}$$

It is of interest to test whether the random effects should be included in the model. This is equivalent to testing if the between-cluster σ_u^2 is zero:

$$H_0 : \sigma_u^2 = 0 \quad \text{versus} \quad H_1 : \sigma_u^2 > 0. \tag{4}$$

This problem is nonstandard because the parameter value under H_0 is on the boundary of the parameter space $[0, \infty)$. Therefore, the likelihood ratio and score statistics no longer have the standard asymptotic chi-squared distribution (Self and Liang, 1987; Stram and Lee, 1994; Verbeke and Molenberghs, 2003).

This article is organized as follows. Section 2 is devoted to reviewing the likelihood ratio tests and their asymptotic distributions. Simulation-based tests are reviewed in Sec. 3, in particular exact likelihood ratio tests, parametric bootstrap tests, and permutation tests. A new permutation test is proposed in Sec. 4. A simulation study comparing the alternative tests is described in Sec. 5. Section 6 concludes with some remarks and directions for future work.

2. Likelihood Ratio Tests

Suppose we wish to test

$$H_0 : \sigma_u^2 \in \Theta_0 \quad \text{versus} \quad H_1 : \sigma_u^2 \in \Theta_1, \quad \Theta = \Theta_0 \cup \Theta_1.$$

Let $\ell_{\Theta_0}^{ML}$ and ℓ_{Θ}^{ML} be the log-likelihood functions maximised over Θ_0 and Θ, respectively. Then the likelihood ratio test (LRT) statistic is given by

$$LRT = -2\left[\ell_{\Theta_0}^{ML} - \ell_{\Theta}^{ML}\right].$$

Using the restricted likelihood functions, the restricted likelihood ratio test ($RLRT$) statistic is given by

$$RLRT = -2\left[\ell_{\Theta_0}^{REML} - \ell_{\Theta}^{REML}\right].$$

It follows from classical likelihood theory (see, e.g., Pace and Salvan, 1997, Sec. 3.4) that under some regularity conditions LRT and $RLRT$ follow, asymptotically under H_0, a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in Θ and Θ_0. One of the regularity conditions under which the chi-squared approximation is valid is that the parameter value under the null hypothesis is not on the boundary of the parameter space Θ, such as in hypothesis (4). Self and Liang (1987) and Stram and Lee (1994) showed that the LRT statistic in this case has an asymptotic null distribution that is a mixture of χ_0^2 and χ_1^2 distributions, each having an equal weight of 0.5. χ_0^2 denotes the distribution with all probability mass at zero, so the correct p-value is obtained by halving the p-value obtained from the χ_1^2 distribution. This result also applies for $RLRT$, as shown by Morrell (1998) (see also Verbeke and Molenbeghs, 2000).

3. Simulation-Based Tests in the Literature

3.1. *Finite Sample Distribution of LRT and RLRT*

In linear mixed models with one variance component, finite sample distributions of the LRT and $RLRT$ are derived by Crainiceanu and Ruppert (2004). They considered the spectral representations of the LRT and $RLRT$ as the basis of efficient simulation algorithms of their null distributions. They provide an algorithm for simulating the null finite distribution of LRT (and $RLRT$). For more details, see Crainiceanu and Ruppert (2004, p. 168).

Crainiceanu and Ruppert's algorithm is implemented in R by Scheipl (2010) in the package "RLRsim". The Function "exactLRT" is used for finite sample LRT, and "exactRLRT" for finite sample $RLRT$.

In R, the function "lmer" in the package "lme4" produced by Bates (2010) can be used to fit the linear mixed models. It is worth to observe that the "exactLRT" function is not working properly with "lmer" function. This is due to some later modifications on the lmer function.

3.2. *Parametric Bootstrap Tests*

A parametric bootstrap test (Efron and Tibshirani, 1993; Davison and Hinkley, 1997) based on the score test for variance components in generalized linear mixed models is proposed by Sinha (2009). Via simulation he showed that the significance level of the parametric bootstrap test is much closer to the nominal level and it is more powerful. Bootstrap tests are more commonly based on LRT or $RLRT$ (Faraway, 2006, Sec. 8.4).

To obtain a parametric bootstrap estimate of the LRT statistic's p-value, the following steps are required.

1. For a given data set, calculate the LRT statistic, denoted by LRT_0.
2. Generate a bootstrap sample from the model under H_0 and calculate the corresponding bootstrap LRT^* statistic.
3. Independently, repeat step 2 many times, say B times, giving B test statistics, say $\{LRT_b^*, b = 1, \ldots, B\}$.
4. The bootstrap p-value is obtained as the proportion of samples with LRT_b^* greater than or equal to LRT_0.

3.3. *Permutation Tests*

Permutation tests are a subclass of nonparametric tests (Lehmann and Romano, 2005; Pesarin and Salmaso, 2010). They are computationally intensive, but modern computational power makes permutation tests feasible. Nonparametric test statistics do not rely on a specific probability distribution (e.g., normal, chi-square, or binomial) that describes the underlying population. Permutation tests are always distribution free provided that data are a sufficient statistic in the null hypothesis (see Pesarin and Salmaso, 2010, Sec. 2.1.3). Some assumptions are required in relation to the samples (e.g., exchangeability). The exchangeability assumption is generally assured by random allocation of treatments to units in experimental work.

Fitzmaurice et al. (2007) proposed a permutation test for variance components in generalized linear mixed models based on the LRT statistic. Their results are compared with the asymptotic 50:50 chi-square distribution of the LRT and with the LRT distribution proposed by Crainiceanu and Ruppert (2004). The proposed permutation test has the correct nominal level under the null hypothesis, and it is more powerful. Although their results were obtained for the case of LRT, the same procedure can be used for $RLRT$.

The following is an algorithm for obtaining a permutation estimate of the p-value of the LRT statistic.

1. For a given data set, calculate the LRT, denoted by LRT_0.
2. Randomly permute the cluster indices while maintaining a fixed number of subjects within a cluster and calculate the corresponding permutation LRT^* statistic.
3. Independently, repeat step 2 many times, say B times, giving B test statistics, say $\{LRT_b^*, b = 1, \ldots, B\}$.
4. The permutation p-value is obtained as the proportion of samples with LRT_b^* greater than or equal to LRT_0.

4. A New Permutation Test

Let us consider random-intercept model (3), repeated here as a guide:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij}, \quad i = 1, \ldots, n_j; \;\; j = 1, \ldots, J.$$

Normality assumptions for the random error components are not required. The hypotheses of interest are given by

$$H_0 : \sigma_u^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_u^2 > 0.$$

Under H_1, the cluster-specific regression lines have different intercepts but the same slope. The testing problem can be treated as permutation ANOVA by removing the effect of the covariate. To this end we compute the least square estimators of β_0 and β_1 under H_0 in order to obtain the empirical deviates $R_{ij} = Y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{ij}$. The R_{ij} are exchangeable, so the resulting problem is equivalent to permutation ANOVA. In terms of the population deviates $(u_j + e_{ij})$, the testing problem is:

$$H_0 : \{u_1 = \cdots = u_J\} \equiv \{\sigma_u^2 = 0\} \quad \text{vs.} \quad H_1 : \{H_0 \text{ is false}\}.$$

The usual F-test statistic is

$$F = \frac{N-J}{J-1} \frac{\sum_{j=1}^{J} n_j (\overline{R}_j - \overline{R})^2}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (R_{ij} - \overline{R}_j)^2}, \tag{5}$$

where $\overline{R}_j = \frac{\sum_i R_{ij}}{n_j}$ and $\overline{R} = \frac{1}{N} \sum_j^J n_j \overline{R}_j$. The F statistic (5) is permutationally equivalent to the following T statistic (see Pesarin and Salmaso, 2010, Sec. 2.4)

$$T = \sum_{j=1}^{J} n_j \overline{R}_j^2.$$

The algorithm for obtaining a conditional Monte Carlo (*CMC*) estimate of the permutation p-value is as follows.

1. For a given data set, under H_0, compute the least square estimates of β_0 and β_1 and calculate the empirical deviates $R_{ij} = Y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{ij}$.
2. Calculate the observed test statistic $T_0 = \sum_{j=1}^{J} n_j \overline{R}_j^2$.
3. Randomly permute the cluster indices while maintaining the same number of subjects within a cluster and calculate the corresponding permutation estimate $T^* = \sum_{j=1}^{J} n_j \overline{R}_j^{*2}$.
4. Independently, repeat step 3 many times, say B times, giving B test statistics, say $\{T_b^*, b = 1, \ldots, B\}$.
5. The permutation p-value is obtained as the proportion of samples with T_b^* greater than or equal to T_0.

5. Simulation Study

A simulation study is conducted to assess the significance level and the power of the proposed permutation test for variance components and to compare it with the aforementioned available tests. The design considered different configurations. For each combination of $J = \{10, 50\}$ clusters and $n_j = n = \{5, 25, 100\}$ observations within a cluster. We also performed several other combinations, not reported here, and the results follow the same behavior. We performed a simulation study based on 2,000 datasets. The permutations and the bootstrap are based on $B = 500$ replications. To examine the significance level of the tests we set

Table 1
Times (in seconds) for a single computation of the test calculated using a PC with a single CPU, considering a design where $n = 100$ and $J = 50$

Test	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
Time	0.18	0.25	35.00	38.00	0.30

LRT: Likelihood Ratio Test; *ERLRT*: Exact Restricted *LRT*; *Boot*: Bootstrap Test; *Fitz*: Fitzmaurice et al. (2007) Test; *PT*: our proposed Permutation Test.

$\sigma_u^2 = 0$, while to investigate the power behavior we select values of σ_u^2 in the set $\{0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.60, 0.80, 1.00\}$. The nominal significance level was set to $\alpha = 0.05$. In the simulation we consider the model in Eq. (3), where $u_j \sim N(0, \sigma_u^2)$, $\varepsilon_{ij} \sim N(0, 1)$, $X_{ij} \sim N(0, 1)$, $\beta_0 = 0$, and $\beta_1 = 1$.

Table 1 reports the time taken for a single computation of each test, using a PC with a single CPU and considering a design where $n = 100$ and $J = 50$. Of course, the *LRT* and *ERLRT* methods are faster than the others because they do not require resampling process. Our proposed permutation test *PT* is largely the fastest among the resampling methods.

Table 2 reports the empirical significance level for all the tests. The empirical level of significance of the bootstrap approach in the simulation configurations is between 0.049 and 0.055, which is much closer to the nominal 0.05 level than the other tests. Our proposed *PT* is the second preferable test in terms of empirical level of significance.

To investigate the power of the proposed permutation test, some configurations are reported in Table 3. It is clear that *PT* is more powerful than the *LRT* and *ERLRT* methods and it is a good competitor of the *Boot* and *Fitz* methods.

One configuration with an unbalanced design is investigated, $J = 10$ clusters with average cluster size equal to 25 (half clusters of size 10 and half clusters of size 40). The empirical significance level and power of the tests are reported in Table 4.

Table 2
Empirical significance level times 100 from the simulation study, nominal significance level $\alpha = 5\%$ – balanced designs

(J, n)	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
(10, 5)	3.1	3.6	5.0	4.6	5.1
(10, 25)	2.6	4.3	5.1	4.6	4.6
(10, 100)	2.3	4.7	4.9	4.9	5.1
(50, 5)	3.6	4.3	4.9	3.9	5.3
(50, 25)	3.8	5.0	5.5	5.5	5.2
(50, 100)	3.6	4.8	5.0	5.0	5.1

LRT: Likelihood Ratio Test; *ERLRT*: Exact Restricted *LRT*; *Boot*: Bootstrap Test; *Fitz*: Fitzmaurice et al. (2007) Test; *PT*: our proposed Permutation Test.

Table 3
Empirical power times 100 from the simulation study – balanced designs

(J, n)	σ_u^2	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
(10, 5)	0.05	8.6	9.8	11.9	12.4	12.8
	0.10	16.1	18.6	21.6	21.9	21.9
	0.15	24.6	27.8	31.3	31.5	30.8
	0.20	31.8	34.8	40.0	39.6	40.2
	0.30	49.0	53.2	57.5	58.0	56.2
	0.40	60.8	64.8	68.8	67.8	68.0
	0.60	77.9	80.2	83.4	83.5	82.2
	0.80	88.2	89.8	91.0	91.2	91.5
	1.00	93.5	94.5	95.3	95.1	95.3
(10, 25)	0.05	48.0	58.8	59.1	59.0	59.1
	0.10	77.6	83.4	83.7	83.3	83.4
	0.15	91.2	93.4	93.7	93.7	94.0
	0.20	94.6	96.2	96.4	96.4	96.4
	0.30	98.9	99.2	99.2	99.2	99.2
	0.40	99.6	99.9	99.9	99.9	99.9
	0.60	99.8	99.9	99.9	99.9	99.9
	0.80	99.9	99.9	99.9	99.9	99.9
	1.00	99.9	99.9	99.9	99.9	99.9
(50, 5)	0.05	25.2	27.5	28.9	26.2	29.8
	0.10	54.0	57.0	58.6	55.2	58.4
	0.15	78.0	79.3	81.1	78.5	80.7
	0.20	91.3	91.8	92.7	91.8	93.0
	0.30	98.7	98.8	99.1	98.7	99.1
	0.40	99.8	99.8	99.8	99.8	99.9
	0.60	99.9	99.9	99.9	99.9	99.9
	0.80	99.9	99.9	99.9	99.9	99.9
	1.00	99.9	99.9	99.9	99.9	99.9

The power of the *LRT* method is the worst. The *PT* method is a good competitor of the *ERLRT*, *Boot*, and *Fitz*. In addition, *Boot* and *Fitz* have an empirical level of significance much closer to the nominal level than the others.

The power of the proposed permutation test when the distributions of the random error components are misspecified is investigated. Specifically, the model of Eq. (3) is considered but a Gamma distribution is assumed for the random error components u_j and ε_{ij}, i.e., $u_j = \sigma_u(U_j^* - 1)$ where U_j^* is distributed as Gamma with location and scale parameters equal to 1. A similar distribution is used to generate the level 1 errors ε_{ij}. Table 5 reports the empirical significance level. The proposed permutation test *PT* and the *Boot* test have an empirical level of significance between 0.045 and 0.051 which are much closer to the nominal level than the other tests. In terms of power, Table 6 reports some configurations. The proposed *PT* is more powerful than the *LRT* and *ERLRT* and it is a very good competitor of the *Boot* and *Fitz* methods.

Table 4
Empirical power times 100 from the simulation study of unbalanced design $J = 10$, $n_1 = \cdots = n_5 = 10$ and $n_6 = \cdots = n_{10} = 40$

σ_u^2	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
0.00	2.3	4.6	5.0	5.0	5.5
0.05	45.1	56.0	56.8	56.6	52.4
0.10	73.5	81.0	81.1	81.2	80.4
0.15	87.2	92.2	92.3	92.2	92.4
0.20	91.9	94.8	94.8	94.8	95.3
0.30	97.5	98.1	98.2	98.0	98.7
0.40	98.8	99.1	99.2	99.1	99.2
0.60	99.8	99.9	99.9	99.9	99.9
0.80	99.7	99.9	99.9	99.9	99.9
1.00	99.9	99.9	99.9	99.9	99.9

Table 5
Empirical significance level times 100 from the simulation study when both level 1 and level 2 errors follow a Gamma distribution

(J, n)	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
(10, 5)	2.6	3.5	4.9	4.7	5.2
(10, 25)	2.5	4.8	5.1	5.0	5.1
(10, 100)	2.2	4.7	4.8	5.2	5.0
(50, 5)	3.9	4.2	5.0	4.1	4.8
(50, 25)	3.5	4.5	5.1	5.0	4.7
(50, 100)	2.9	4.5	4.5	4.8	4.5

6. Concluding Remarks

To test variance components in a linear mixed model with balanced design, the proposed permutation test has a significance level close to the nominal level and it is more powerful than the tests based on the 50:50 mixture chi-square distribution and the exact restricted likelihood ratio method given by Crainiceanu and Ruppert (2004). In terms of speed, our proposed permutation test is the fastest method among the resampling-based methods. This is due to the way the distribution of the test statistic is obtained; in our approach the algorithm requires the fitted model under the null hypothesis only once, while the other algorithms require the fitted model under at least the null hypothesis for every iteration. The proposed permutation test is also fully nonparametric while the other approaches rely on distributional assumptions.

With unbalanced designs, the proposed permutation test still has a significance level close to the nominal level and it is more powerful than the likelihood ratio test based on the 50:50 mixture chi-square distribution and the approximate exact

Table 6
Empirical power times 100 from the simulation study when both level 1 and level 2 errors follow a Gamma distribution

(J, n)	σ_u^2	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
(10, 5)	0.05	7.9	10.5	12.7	13.4	13.1
	0.10	17.0	20.4	23.6	24.4	23.2
	0.15	23.5	26.2	29.9	30.6	31.3
	0.20	33.0	37.0	40.6	40.5	41.7
	0.30	43.2	46.5	50.2	51.4	51.0
	0.40	54.1	56.6	59.0	59.4	59.9
	0.60	68.0	70.2	72.8	73.4	73.8
	0.80	76.2	78.5	80.8	81.2	81.0
	1.00	83.0	84.5	86.6	86.5	86.3
(10, 25)	0.05	41.2	49.6	50.3	50.7	50.5
	0.10	66.2	73.2	73.9	73.4	73.5
	0.15	77.8	82.7	83.0	83.1	83.2
	0.20	87.1	89.8	90.1	90.3	90.2
	0.30	93.6	95.3	95.5	95.6	95.5
	0.40	95.7	97.2	97.2	97.4	97.2
	0.60	98.2	98.6	98.7	98.8	98.7
	0.80	99.2	99.3	99.3	99.2	99.3
	1.00	99.5	99.6	99.7	99.7	99.7
(50, 5)	0.05	23.4	25.2	27.2	24.1	26.9
	0.10	49.8	52.9	55.1	51.4	54.9
	0.15	73.7	75.4	76.8	74.6	76.6
	0.20	85.2	86.6	87.0	85.9	86.9
	0.30	96.4	96.9	97.0	96.7	96.9
	0.40	98.3	98.5	98.8	98.5	98.9
	0.60	99.7	99.8	99.9	99.8	99.8
	0.80	99.9	99.9	99.9	99.9	99.9
	1.00	99.9	99.9	99.9	99.9	99.9

restricted likelihood ratio method. It is worthwhile to observe that all tests discussed in this article are more powerful for balanced designs than for unbalanced designs.

When the distributions of the model errors are misspecified all the tests under consideration loose power. Also in this case, the three resampling-based tests, which have similar performances, are clearly preferable to the standard *LRT* and the exact *RLRT*.

Acknowledgments

The authors would like to thank Dr. Paolo Emilio Mazzon and Professor Gianfranco Bilardi of Department of Information Engineering – Univeristy of Padua for their helping to use their own computer clusters which facilitate and accelerate our computations. Moreover, the authors wish to thank the University of Padova (CPDA092350/09) and the Italian Ministry for University and Research

MIUR project PRIN2008-CUP number C91J10000000001 (2008WKHJPK/002) for providing the financial support for this research.

References

Afshartous, D., de Leeuw, J. (2004). An application of multilevel model prediction to nels:88. *Behaviormetrika* 31(1):43–66.

Bates, D. M. (2010). *Mixed-Effects Modeling with R.* New York: Springer.

Crainiceanu, C. M., Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Statist. Soc. Ser. B* 66(1):165–185.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability, 57. New York: Chapman and Hall/CRC.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. New York: Chapman and Hall/CRC.

Fitzmaurice, G. M., Lipsitz, S. R., Ibrahim, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* 63(3):942–946.

Garner, C. L., Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: a multilevel analysis. *Sociol. Educ.* 64(4):251–262.

Lehmann, E. L., Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer Texts in Statistics. New York: Springer.

Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* 54(4):1560–1568.

Pace, L., Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific Publishing Company.

Pesarin, F., Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Application and Software*. Wiley Series in Probability and Statistics. Chichester, UK: Wiley.

Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. London: Sage Publications.

Scheipl, F. (2010). RLRsim/CAP/R package v.2.0=5. Exact (Restricted) Likelihood Ratio Tests for Mixed and Additive Models.

Self, S. G., Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* 82:605–610.

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Can. J. Statist.* 37:219–234.

Snijders, T., Bosker, R. (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*. London/Thousand Oaks, CA: Sage.

Stram, D. O., Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* 50:1171–1177.

Swamy, P. A. V. B. (1970). Efficient inference in a random coefficient regression model. *Econometrica* 38(2):311–323.

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

Verbeke, G., Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* 59:254–262.