Permutation Tests for Random Effects in Mixed Models

by

Oliver E. Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2012

Doctoral Committee:

Associate Professor Thomas M. Braun, Chair
Research Professor Mousumi Banerjee
Associate Professor Lynda D. Lisabeth
Professor Jeremy M. G. Taylor

ACKNOWLEDGMENTS

I would like to thank my family for all of their support over the years. Without them and the sacrifices that they made I could not have accomplished this work.

I would also like to express my sincere thanks to Dr. Thomas M. Braun for your guidance and patience during the course of my dissertation work. Without your encouragement and help this dissertation would never have been completed. Your advice has truly helped to make me a better statistician and a better person.

I am also extremely grateful for the support of Dr. Jeremy M. G. Taylor whose numerous research projects that I worked on during my time at Michigan has defined my development as an applied statistician. Your vast knowledge and experience serves as a goal for me to work towards and try to emulate.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Permutation Tests for Random Effects in Mixed Models

by

Oliver E. Lee

Chair: Thomas M. Braun

Inference regarding the inclusion or exclusion of random effects in mixed models is challenging because the variance components are located on the boundary of their parameter space under the null hypothesis. As a result, the asymptotic null distribution of the Wald, score, and likelihood ratio tests will not have the typical chi-squared distribution. Although it has been proved that the correct asymptotic distribution is a mixture of chi-squared distributions, the appropriate mixture distribution is cumbersome and non-intuitive when the null and alternative hypotheses differ by more than one random effect. This dissertation addresses these challenges through the use of permutation methods.

For the first chapter, we focus on linear mixed models and present two permutation tests, one that is based on the Best Linear Unbiased Predictors (BLUPs), and one that is based on the restricted likelihood ratio test statistic. The null permutation distributions of our statistics are computed by permuting the residuals both within-

and among-subjects and are valid both asymptotically and in small samples. Through simulations we show that our permutation tests are valid for small sample sizes and is more powerful than the asymptotic likelihood ratio test. The proposed tests are also shown to be more robust to violations of distributional assumptions compared with the asymptotic likelihood ratio tests.

For the second chapter we extend the linear mixed model permutation methods to inference on random effects in generalized linear mixed models (GLMMs). We use the idea of working variates to approximate the GLMM with a linear mixed model. Through simulations we show that our permutation tests are valid and display power that is comparable to the most powerful score test.

For the final chapter we demonstrate the versatility of our permutation tests with an application to linear penalized spline models. By re-expressing the penalized spline model as a mixed model our permutation tests can test the spline model alternative against a linear regression model. The validity and power are examined through simulation, and find that the BLUP based permutation test is the most powerful when compared with the permutation test of Raz and the asymptotic likelihood ratio test.

CHAPTER I

Introduction

The simplest form of regression in statistics is linear regression, which is usually expressed as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y}$ is a vector of a continuous outcome of interest, $\boldsymbol{X}$ is a matrix of known fixed effects or explanatory variables that are multiplied by coefficients from the vector $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ is a vector of independent and identically distributed random errors that are typically assumed to follow a normal distribution. One of the key assumptions in linear regression is that the data are independent. Often this assumption is violated when data come from clusters or groups. Examples of sources of correlation include shared genetic traits of family members, plots of land in agricultural studies, and hospitals in multi-center clinical trials.

Mixed models are an extension of regression models that incorporate random effects into typical fixed effects regression models to account for correlation in the data. Adding $\boldsymbol{Z}\boldsymbol{b}$ to the linear regression model, where $\boldsymbol{Z}$ is a matrix of random effect covariates and $\boldsymbol{b}$ is the vector of random effects, results in the Laird and Ware [1982] formulation of the linear mixed model (LMM):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}.$$

The random effects are assumed to come from a mean zero random distribution, usually a normal distribution.

Random effects can also be added to generalized linear models (GLMs) for non-normal data. GLMs were first proposed by McCullagh and Nelder [1989] and are an extension of linear models where the mean of the outcome is associated with a linear function of the independent variables through a link function and the variance of the outcomes can be a function of the mean. Just like with linear models the observations are assumed to be independent. A standard GLM is written as

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$$

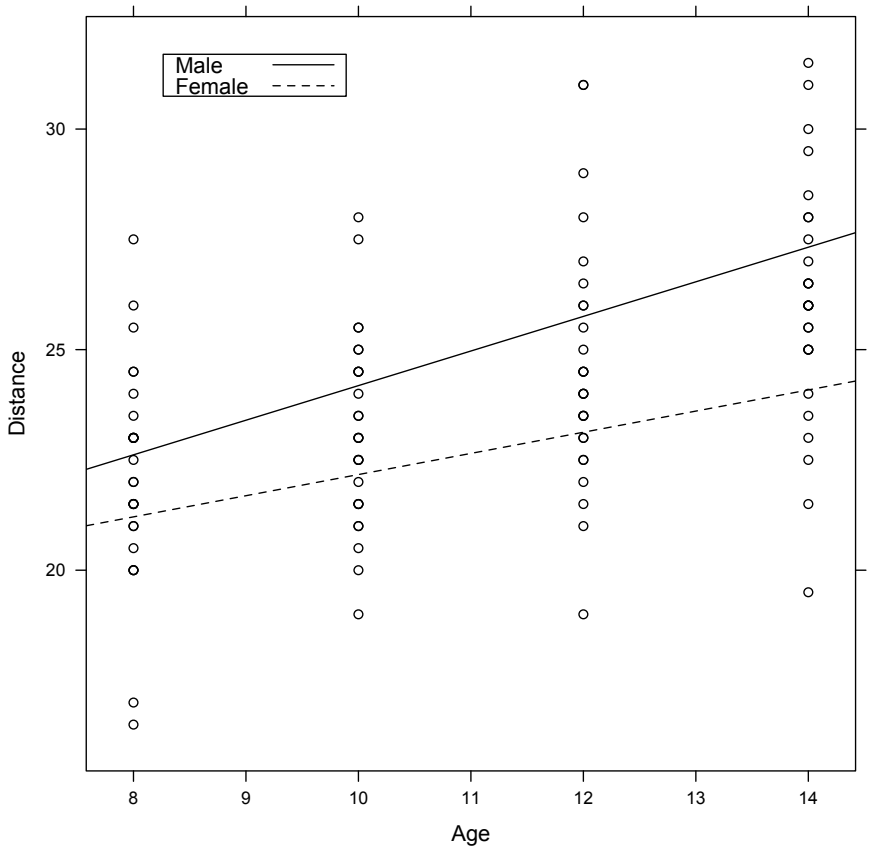
where our data, $\boldsymbol{Y}$, come from an exponential family distribution, and $\boldsymbol{\mu}$ is the mean of the distribution that is related to a linear function of fixed effects and coefficients through a link function $g(\cdot)$. Examples of GLMs include logistic regression and Poisson regression. When the data are correlated, random effects can be added extending the GLM to a generalized linear mixed model (GLMM). After adding random effects to the GLM, the subsequent GLMM can be written as

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}.$$

Additional notation for both LMMs and GLMMs will be provided in the subsequent chapters.

Our motivating application for developing methods for mixed models is in the analysis of longitudinal data [Laird and Ware, 1982, Diggle et al., 2002]. These types of data are characterized by repeated measurements of an outcome from a set of

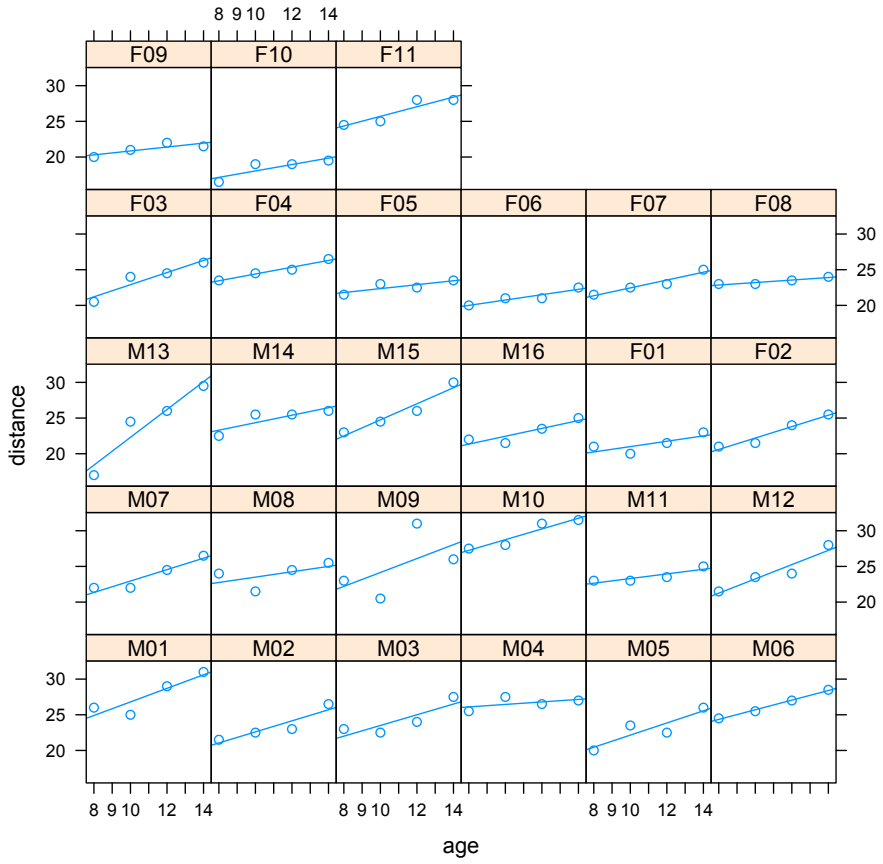
Figure 1.1: Scatterplot of the Pothoff Roy dental data.



subjects over a period of time. Mixed models are needed in order to ensure that inference for the fixed effects is valid, because measurements that originate from the same individual are likely to be correlated and assumption of independence will lead to invalid standard error estimates.

One example of longitudinal data is the Potthoff and Roy [1964] dental data where dental growth measurements of the distance (in millimeters) from the center of the pituitary gland to the pteryomaxillary fissure were recorded on 16 males and 11 females at ages 8, 10, 12, and 14. To illustrate the longitudinal trend in the data, Figure 1.1 is a plot of measurements by age overlaid with overall linear regression fits for the males and the females.

Figure 1.2: Subject specific scatterplots of the Pothoff Roy dental data.



Although the data appear to be quite variable, the regression lines have a distinct increasing trend. This is further enhanced in Figure 1.2 where each subject is individually plotted along with a regression fit that shows that each subject has an increasing slope. In addition, with a couple of exceptions such as M13, most of the children had similar growth rates and children who started with a high measurement also ended with a high measurement, providing some visual evidence of correlation among measurements from the same child.

Inference for the fixed effects in mixed models is straightforward and usually done by comparing the standard likelihood ratio, Wald, or score statistic with a χ^2 distribution with degrees of freedom equal to the number of parameters being tested.

One drawback to these tests is that they are based on asymptotic distributions and may not be valid for small sample sizes. Methods to address small sample situations have been explored and include corrections based on the Wald statistic [Kenward and Roger, 1997], corrections based on the likelihood ratio [Zucker et al., 2000], and permutation tests [Gail et al., 1992, Braun and Feng, 2001].

Apart from inference on the fixed effects, testing for the inclusion or exclusion of random effects may be of interest because estimating unnecessary random effects leads to a loss of power. Testing for random effects is the same as testing to see if the variance of the random effect distribution is equal to 0 under the null hypothesis. This is a difficult problem because 0 is on the boundary of the parameter space for variance components. As a result, the standard asymptotic hypothesis tests do not hold. In this dissertation we will address this problem through the use of permutation methods. As an added benefit, our methods will also be suitable for small data sets as opposed to asymptotic methods. This problem is further compounded when dealing with GLMMs because for many distributions the likelihood has no closed form and instead must be approximated using various methods.

We continue this chapter with a review of the asymptotic likelihood ratio test for variance components and an overview of permutation tests. In Chapter 2 we develop permutation tests for random effects in LMMs based on permuting weighted residuals. Chapter 3 extends the permutation tests developed in Chapter 2 in order to apply them to GLMs. In Chapter 4 we demonstrate the flexibility of mixed models as well as our permutation tests in an unique application by testing a linear penalized spline alternative against linear regression. We conclude this dissertation with a discussion of our work and some ideas for future research.

1.1 The Asymptotic Likelihood Ratio Test for Variance Components

Hypothesis testing for random effects in linear mixed models began with the works of Self and Liang [1987] and Stram and Lee [1994]. Although separated by seven years, these two works are almost universally cited together in the literature for testing random effects in linear mixed models. The work of Self and Liang focused on deriving the properties of the likelihood when the true parameter value may be on the boundary of its parameter space. This includes showing the existence of a maximum likelihood estimator, its large sample properties, and the asymptotic distribution of likelihood ratio statistics under these conditions.

Stram and Lee applied the results of Self and Liang [1987] specifically to likelihood ratio tests for nonzero variance components in linear mixed models. Stram and Lee showed that likelihood ratio statistics for nonzero variance components follow mixture χ^2 distributions. In their paper they illustrated this through a number of specific hypotheses of the random effect covariance matrix that they have labeled as cases. For case 1, likelihood ratio statistics for tests of one random effect against no random effects have an asymptotic distribution following a 50:50 mixture of χ^2_0 and χ^2_1. In case 2 they test a model with 2 random effects that are potentially correlated against a model with just 1 random effect. The likelihood ratio test statistic for case 2 follows a 50:50 mixture of χ^2_1 and χ^2_2. Case 3 is a generalization of case 2, when testing $q+1$ random effects against q random effects. Here, the likelihood ratio statistic follows a 50:50 mixture of χ^2_q and χ^2_{q+1}. Lastly case 4 addresses simultaneously testing multiple variance components. For the special case when the information matrix is equal to the identity under the null hypothesis then the asymptotic distribution follows a mixture

χ^2 with binomial mixing probabilities. For all other situations simulation methods are recommended. As is the case with standard asymptotic test statistics, the Wald and score statistics for hypothesis tests of nonzero variance components also follow the same mixture χ^2 distributions as the likelihood ratio test statistic [Silvapulle and Silvapulle, 1995, Verbeke and Molenberghs, 2003, Silvapulle, 1992].

An intuitive argument as to why the asymptotic distribution is a mixture of χ^2 distributions was provided in Molenberghs and Verbeke [2007]. In general, variances are constrained to be greater than or equal to 0, and this constraint has an impact on tests for random effects because the estimates of the variances are also constrained to be non-negative. When the constrained estimator, $\hat{\sigma^2_{b_l}}$, for $\sigma^2_{b_l}$ equals 0 there is no evidence against H_0, and the likelihood ratio, Wald, and score test statistics are equal to 0. However, for positive values of $\hat{\sigma^2_{b_l}}$ the likelihood ratio, Wald, and score test statistics follow a χ^2 distribution of the appropriate degrees of freedom. For example when estimating a single random effect under the null hypothesis the proportion of times that the variance estimate is equal to 0 is 50%.

1.2 Permutation Tests

The origins of permutation tests date all the way back to 1935 and *The Design of Experiements* where R. A. Fisher observed that hypotheses can be examined through randomization without the assumption of normality. Additional early work on permutation tests was conducted by Pitman [1937], Hoeffding [1952], and Kempthorne [1955]. Sometimes referred to as randomization tests, permutation tests operate under the belief that if the null hypothesis is true, then the arrangement of data that is observed is purely due to chance, and therefore, all of the possible rearrangements or

permutations of the data are equally likely.

Permutation tests are considered to be nonparametric in the sense that no assumption of a specific probability distribution is made for the underlying population. However, permutation tests are not free of assumptions, as the data must be assumed to be exchangeable for a permutation test to be valid. A vector, $\boldsymbol{Y}$, is exchangeable if, for any permutation of $\boldsymbol{Y}$ denoted as $\boldsymbol{Y^*}$, $\boldsymbol{Y^*}$ has the same distribution as $\boldsymbol{Y}$ [Commenges, 2003]. Independent and identically distributed (iid) data are exchangeable, and, in fact, iid is a stronger condition than exchangeability. Samples without replacement from a finite population and multivariate normal data where the covariance matrix has a constant variance along the diagonal and identical covariance for all of the off-diagonal elements are both exchangeable [Good, 2005]. In the latter case, this type of exchangeable covariance structure is often used in modeling correlated data.

Permutation tests proceed with the following steps. First, the observed data are used to calculate a test statistic. Then all possible permutations of the observed data are enumerated, and for each permutation of the data, a new test statistic is calculated. The collection of all of the permuted test statistics comprises the null distribution to which the test statistic from the observed data is then compared in order to obtain a p-value. The permutation p-value is the proportion of the permutation distribution with values as extreme or more extreme than the observed test statistic. While the steps are straightforward to implement, the challenge lies in selecting an appropriate test statistic and determining how to permute the data correctly.

We present a simple example of a permutation test. Suppose that investigators wish to compare two different treatments in their ability to shrink solid tumors. The null hypothesis is that there is no difference in tumor shrinkage between the two

treatments. Five patients are recruited and randomized to one of the two treatments with split of three to treatment A and two to treatment B. Table 1.1 contains the observed tumor shrinkage percentage for these five patients.

Table 1.1: Observed outcomes for a hypothetical clinical trial

Patient	Treatment	Total Percent Shrinkage
1	A	10
2	A	12
3	A	14
4	B	4
5	B	5

The test statistic will be the difference in the mean percent shrinkage between treatment A and treatment B. For the data that we collected, this is the difference between 12% and 4.5%, which are the means of treatment A and B, respectively, and equals 7.5%. Under the null hypothesis of no difference between treatments A and B each patient's measured shrinkage would have occurred regardless of the treatment to which they were assigned. Therefore, this observed set of outcomes is due to the random assignment of the patients to each treatment, and any of the ten different permutations of treatment assignments are equally likely to have occurred. Table 1.2 contains all of the ten possible permutations sorted by the absolute value of the test statistic. Only one of these ten permutations has a test statistic that is as extreme or more extreme than the observed result leaving us with an exact p-value of 0.10.

Permutation tests are appealing because they can often be used when parametric tests fail. For example, when the assumptions for parametric tests cannot be met, permutation tests can be implemented as long as the exchangeability assumption holds. Permutation tests can also be used when the asymptotic distribution of a test

Table 1.2: All possible permutations of patients to the two treatments

Treatment A			Treatment B		Difference in Means
10	12	5	4	14	0.00
10	4	14	12	5	0.83
10	12	4	14	5	-0.83
10	5	14	4	12	1.67
4	12	14	10	5	2.50
5	12	14	4	10	3.33
4	5	14	10	12	-3.33
4	12	5	10	14	-5.00
10	4	5	12	14	-6.67
10	12	14	4	5	7.50

Patients are represented by their observed measurement.

statistic is unknown or intractable. This arises when testing for multiple variance components simultaneously and will be elaborated upon in the subsequent chapters. When all of the possible permutations can be enumerated permutation tests are exact. Often enumerating all of the permutations is computationally unfeasible, but a small representative random sample [Dwass, 1957] can be used. This process is called Monte Carlo permutations, and typically between 100 and 1600 random permutations are necessary [Good, 2005]. Asymptotically permutation tests have nominal size and are nearly as powerful as parametric tests [Hoeffding, 1952]. Finally, permutation tests are often better suited for small samples than asymptotic parametric tests.

CHAPTER II

Permutation Tests for Random Effects in Linear Mixed Models

Linear mixed models (LMMs) are a rich class of models containing both fixed and random effects. LMMs are often used to fit longitudinal or repeated measures data [Laird and Ware, 1982] where outcomes for a limited number of subjects are collected repeatedly over time, or with multilevel or clustered data where random effects are used to account for the within-level or within-cluster correlations. Often, inference focuses upon the need for the inclusion of random effects. For example, subjects in a clinical trial may be recruited from a set of hospitals that are participating in the study. Homogeneity among patients from the same hospital is likely and can be accounted for through a random hospital effect in the model. However, if there is no correlation among patients from the same hospital then there would be a loss of power by estimating an unnecessary random effect variance.

The difficulty in testing for random effects lies in the fact that the variance component of the random effect is equal to 0 under the null hypothesis, a value that is on the boundary of the parameter space. As a result, the usual χ^2 asymptotic distributions of the Wald, score, and likelihood ratio test statistics do not hold. In-

stead, the correct null distribution for the likelihood ratio statistic has been shown to be a mixture of χ^2 distributions [Self and Liang, 1987, Stram and Lee, 1994]. For example, when testing for one random effect, the null distribution becomes a 50:50 mixture of χ^2_q and χ^2_{q-1} distributions where q is the total number of random effects in the alternative model. The score [Silvapulle and Silvapulle, 1995, Verbeke and Molenberghs, 2003] and Wald [Silvapulle, 1992] tests for variance components have been proven to have equivalent mixture χ^2 distributions. These modified tests also rely on asymptotic approximations and are not guaranteed to have nominal size with small sample sizes.

Other methods for variance component inference have been published. Öfversten [1993] developed an exact test for uncorrelated random effects in unbalanced linear mixed models through orthogonal transformations of the model matrix. Crainiceanu and Ruppert [2004] derived the finite sample null distribution for the likelihood ratio and restricted likelihood ratio test statistics when testing for a single variance component with no other nuisance variance components. They derived the spectral decomposition of each test statistic and they also developed a simulation algorithm that generates the approximate finite sample null distribution via the spectral decomposition. Greven et al. [2008] extended the methods of Crainiceanu and Ruppert to test for a single variance component in the presence of multiple independent nuisance random effects and also developed an approximation to the parametric bootstrap. Kinney and Dunson [2008] used a Bayesian stochastic search variable selection (SSVS) method to identify nonzero random effect variances in LMMs using a modified Cholesky decomposition of the random effect covariance matrix. By reparameterizing the LMM, the SSVS method can perform variable selection with the random effects. An alternative Bayesian method was developed by Saville and Herring [2009] in which

null and alternative models are compared via Bayes factors.

Permutation tests are a viable alternative to the methods that were covered in chapter 1, as permutation tests are known to have nominal size in finite samples while requiring only a few weak assumptions. Nonetheless, the only existing permutation approach for testing for random effects was presented by Fitzmaurice and Ibrahim [2007]. The test was specifically designed for multi-level studies where inclusion of a single random effect to quantify the heterogeneity among the different levels may be required. They compared the likelihood ratio test statistic to an empirical null distribution generated by randomly permuting the observed level assignments among the subjects. However, their test is limited to the setting at hand and cannot be generalized to longitudinal studies and other correlated data sources if there are multiple random effects or a single continuous random effect, such as time.

Our work is a generalization to the approach of Fitzmaurice and Ibrahim and leads to a pair of permutation tests that allow for inference with any number and type of random effects in a LMM. Both test statistics are a sum of weighted squared residuals with the weights determined by the among- and within- subject variance components, and the empirical null distributions generated via permutations of the residuals. The first test statistic is based on the Best Linear Unbiased Predictions (BLUP) [Robinson, 1991] and the second statistic is the restricted likelihood ratio test statistic assuming normality of the data. We will show that our tests have valid size and their powers are comparable to existing methods. We will also demonstrate that our likelihood ratio based permutation test can address simultaneous inference on multiple random effects. We begin with LMM notation and some background on permutation methods in Section 2. Section 3 follows with a presentation of our proposed methods. We present the results of simulations in Section 4 that demonstrate the validity and

power of our methods as we vary both the numbers of subjects and the numbers of observations per subject. In this section we also examine the robustness of our permutation tests to violations to the assumption of normality in the data. In Section 5 we apply our methods to data from a longitudinal study investigating the levels of adenosine deaminase in chronic myelogenous leukemia patients. We close with a discussion of our work in Section 6.

2.1 Methods

2.1.1 Linear Mixed Models

Let Y_{ij} be observation j of subject or cluster i for $i = 1, 2, ..., N$ and $j = 1, 2, ..., n_i$. Following the Laird and Ware [1982] formulation of the linear mixed model, we have

$$Y_{ij} = \beta_1 x_{1ij} + ... + \beta_p x_{pij} + b_{i1} z_{1ij} + ... + b_{iq} z_{qij} + \epsilon_{ij},$$

where $\beta_1, ..., \beta_p$ are the population level fixed-effect coefficients and $b_{i1}, ..., b_{iq}$ are the random effects for the i-th subject or cluster. The $x_{1ij}, ..., x_{pij}$ and $z_{1ij}, ..., z_{qij}$ are the observed fixed effect covariates and random effect covariates respectively for observation j of subject i. Generally, x_{1ij} and z_{1ij}, are constant and equal to 1 to represent the fixed and random intercepts, respectively. The random effects, $\boldsymbol{b}_i = \{b_{i1}, b_{i2}, \dots b_{iq}\}$ are assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, in which the respective variances for $b_{i1}, b_{i2}, \dots b_{iq}$ are denoted as $\sigma^2_{b_1}, \sigma^2_{b_2}, ..., \sigma^2_{b_q}$. The random errors, ϵ_{ij}, are independent, identically distributed normal random variables with mean 0 and variance σ^2_ϵ. For each j, $\boldsymbol{b}_i$ and ϵ_{ij} are assumed to be independent, although the elements of $\boldsymbol{b}_i$ are not necessarily independent of each

other.

Equivalently, we can write the linear mixed model for subject i using matrix notation, $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_p\}$, $\boldsymbol{\epsilon}_i = \{\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{in_i}\}$, and $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are subject-specific design matrices for the p fixed effect covariates and q random effect covariates, respectively. We then combine data from all subjects so that $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ is the $\sum_i n_i$ vector of outcomes, $\boldsymbol{\epsilon} = \{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_N\}$ is the $\sum_i n_i$ vector of errors, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the respective design matrices for the p fixed effect covariates and q random effect covariates formed by successively placing each subject's design matrices under each other. Furthermore, if we denote $\boldsymbol{b} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N\}$, we have

$$Var\begin{bmatrix} \boldsymbol{b} \\ \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} & 0 \\ \\ 0 & \boldsymbol{R} \end{bmatrix}$$

where $\boldsymbol{G} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_G$ and $\boldsymbol{R} = \sigma^2_\epsilon \boldsymbol{I}_R$, in which $\otimes$ denotes the Kroenecker product, and $\boldsymbol{I}_G$ and $\boldsymbol{I}_R$ are $N \times N$ and $\sum_i n_i \times \sum_i n_i$ identity matrices, respectively.

Estimation of the elements of $\boldsymbol{\beta}$, $\boldsymbol{G}$, and $\boldsymbol{R}$ is typically done through maximum likelihood (ML) or restricted maximum likelihood (REML). Asymptotically, the maximum likelihood and REML estimators are equivalent, but for small sample sizes, the REML estimator is expected to be less biased than the maximum likelihood estimator [Ruppert et al., 2003]. In addition, a comprehensive simulation study performed by Morrell [1998] found that the asymptotic likelihood ratio test based on the REML estimates are closer to nominal than test statistics utilizing the ML estimates. Therefore, in our proposed methods we used the REML estimators. Subject specific random effects, $b_{i1}, ..., b_{iq}$, can be predicted using best linear unbiased prediction (BLUP), the

results from which we denote $\tilde{\boldsymbol{b}} = \{\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \ldots, \tilde{\boldsymbol{b}}_N\}$, where $\tilde{\boldsymbol{b}}_i = \{\tilde{b}_{i1}, \tilde{b}_{i2}, \ldots, \tilde{b}_{iq}\}$. The estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\}$, and $\tilde{\boldsymbol{b}}$ are solutions to the following mixed model equations given by Henderson [1950]

$$\boldsymbol{X^T R^{-1} X}\hat{\boldsymbol{\beta}} + \boldsymbol{X^T R^{-1} Z}\tilde{\boldsymbol{b}} = \boldsymbol{X^T R^{-1} Y}$$
$$\boldsymbol{Z^T R^{-1} X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Z^T R^{-1} Z} + \boldsymbol{G^{-1}})\tilde{\boldsymbol{b}} = \boldsymbol{Z^T R^{-1} Y},$$

and lead to the solutions

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y} \\ \tilde{\boldsymbol{b}} &= \hat{\boldsymbol{G}}\boldsymbol{Z}\hat{\boldsymbol{V}}^{-1}\hat{e} \end{aligned} \tag{2.1}$$

where $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are the residuals and $\hat{\boldsymbol{V}} = \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}^T + \hat{\boldsymbol{R}}$ is the estimated covariance matrix for $\boldsymbol{Y}$. In general, $\tilde{\boldsymbol{b}}$ can be interpreted as realized values of the random vector $\boldsymbol{b}$ [Robinson, 1991].

Our objective in this paper is to compare a linear mixed model containing p fixed effects and q random effects to a model with the same p fixed effects but only $q - r$ random effects where $0 < r \leq q$. Performing this inference is equivalent to testing if the variances of the r random effects are all equal to 0. As stated before classical tests in this situation do not follow their typical χ^2_r distributions. Intuitive arguments as to why this is the case are presented by Molenberghs and Verbeke [2007].

2.1.2 Permutation Tests

A permutation test is one in which the null distribution of the test statistic is determined through permutations of the data; the test will have nominal size when the permutations are performed correctly. As an example, consider a study investigating the efficacy of a new treatment by comparing it to a placebo. The investigators wish to see if the treatment has an effect on some measured outcome of interest and randomize subjects equally to the treatment and placebo groups. Let X_i be the measured outcome for subject i in the treatment group, $i = 1, 2, ...n_x$, and Y_j be the outcome for subject j in the placebo group, $j = 1, 2, ..., n_y$. The X_i are assumed to have distribution $\mathcal{F}$ with mean μ_x and variance σ^2, and the Y_i are assumed to have distribution $\mathcal{F}$ with mean μ_y and variance σ^2. Under the null hypothesis of no treatment effect, $\mu_x = \mu_y$, the two groups will have the same mean response, and more importantly, the same distribution.

Therefore, we can test our null hypothesis using the mean difference in observed response between treatment and placebo groups or $T = \bar{X} - \bar{Y}$, in which $\bar{X}$ is the observed mean response in the treatment group and $\bar{Y}$ is the observed mean in the placebo group. If $\mathcal{F}$ were a normal distribution, then T, appropriately standardized by its standard error, would have a t-distribution and the appropriate critical value would be determined from this distribution. If $\mathcal{F}$ were not a normal distribution, we could still appeal to the Central Limit Theorem and use the same t-distribution as an asymptotic approximation to the exact null distribution.

However, under the null hypothesis of no treatment effect, and conditioning on the observed outcomes of the $n_x + n_y$ subjects, the observed response of each patient would have occurred independent of group assignment. Thus, we can generate the null

distribution for T by recomputing T under all $P = \binom{n_x+n_y}{n_x}$ possible permutations of group assignments. The p-value is obtained by computing the percentage of values in the permutation distribution whose magnitudes are at least as large as the magnitude of T. This permutation test is guaranteed to be nominal, meaning its size is no larger than desired [Hoeffding, 1952]. More specifically, permutation tests assume that the values being permuted are exchangeable under the null hypothesis [Good, 2005]. A vector, $\boldsymbol{Y}$, is exchangeable if, for any permutation of $\boldsymbol{Y}$ denoted as $\boldsymbol{Y^*}$, $\boldsymbol{Y^*}$ has the same distribution as $\boldsymbol{Y}$ [Commenges, 2003]. It should be noted that exchangeability is a weaker condition than independent and identically distributed.

As the amount of data increases, so does the number of possible permutations, eventually making exact enumeration of all P permutations computationally unfeasible. Instead of calculating all possible permutations, an approximate permutation distribution can be generated through Monte Carlo sampling [Dwass, 1957]. By randomly permuting the data between 100 and 1600 times [Good, 2005], an approximate permutation distribution can be generated, assuming the randomly selected permutations are drawn to sufficiently represent the tails of the exact permutation distribution.

2.2 Proposed Methods

2.2.1 Best Linear Unbiased Predictors Based Permutation Test

We begin by considering the hypothesis test for the inclusion or exclusion of a single random effect, $\boldsymbol{b_i} \sim N(0, \sigma^2_{b_i})$, in a linear mixed model with no other random effects present. This is equivalent to testing if $\sigma^2_{b_i} = 0$. Thus, we are comparing the

following models:

$$H_0 : Y_{ij} = \beta_1 x_{1ij} + ... + \beta_p x_{pij} + \epsilon_{ij} \tag{2.2}$$

$$H_1 : Y_{ij} = \beta_1 x_{1ij} + ... + \beta_p x_{pij} + b_{i1} z_{1ij} + \epsilon_{ij}. \tag{2.3}$$

We use

$$T_1 = \sum_{i=1}^{N} \tilde{b}_{i1}^2 / N \tag{2.4}$$

as our test statistic, which is the sample variance of the BLUPs for the random effect, b_i. This statistic involves the sum of the squared BLUPs where the BLUPs are treated as a random sample of $b_i \sim N(0, \sigma_{b_i}^2)$. Note that the denominator of the test statistic is constant for all of the permutations and does not affect the validity or power of our test.

To construct the permutation distribution with which to compare the observed test statistic, we permute the marginal errors, $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$. Under the null hypothesis of no random effects, the $\boldsymbol{\epsilon}$ are exchangeable, and more specifically, independent and identically normally distributed with mean 0 and variance σ_ϵ^2. By subtracting the fixed effects, $\boldsymbol{X}\boldsymbol{\beta}$ from the response $\boldsymbol{Y}$, the errors have the benefit of not requiring the continuous $\boldsymbol{X}$'s to be identical among all subjects nor do the number of observations for each subject need to be the same. Therefore, we can permute the errors both within and between subjects. In practice, the errors are estimated by the residuals, $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, calculated from estimates fit from the alternative model, and Schmoyer [1994] showed that the alternative model residuals are also asymptotically exchangeable both within- and among- subjects under the null hypothesis.

The marginal residuals from the full model are part of the calculation for the

BLUPs and lead to a straightforward permutation distribution for T_1. For each permutation $k = 1, 2, ...1,000$, we randomly permute the marginal full model residuals. Using these permuted residuals, we generate a permuted estimate $\hat{\sigma}^2_{b_i,k}$ for $\sigma^2_{b_i}$, from which we compute BLUPs for the k-th permutation which are used to compute T^*_{1k}, the test statistic T_1 for permutation k. These 1,000 permuted values of T_1 result in an approximate empirical null distribution of T_1. The re-estimation of $\sigma^2_{b_i}$ is performed because some permutations of the residuals will result in $\hat{\sigma}^2_{b_i} = 0$ and lead to the empirical null distribution having positive mass at zero. Note that the fixed effects, β, are not re-estimated. We then generate a p-value by calculating the percentage of permutations with T^*_1 greater than T_1.

Next, we extend the permutation test to test for the presence of a single random effect in a model that contains other random effects such as:

$$H_0 : Y_{ij} = \beta_1 x_{1ij} + ... + \beta_p x_{pij} + b_{i1} z_{1ij} + \epsilon_{ij} \tag{2.5}$$

$$H_1 : Y_{ij} = \beta_1 x_{1ij} + ... + \beta_p x_{pij} + b_{i1} z_{1ij} + b_{i2} z_{2ij} + \epsilon_{ij}. \tag{2.6}$$

In this setting, the null model now contains other random effects so that all $\sum_{i=1}^{N} n_i$ errors under the null hypothesis are no longer exchangeable under the null hypothesis. Instead, the errors are normally distributed with mean $\mathbf{0}$ and covariance matrix, $\boldsymbol{V_0} = \sigma^2_{b_{1_0}} \boldsymbol{Z}^T \boldsymbol{Z} + \boldsymbol{R_0}$ with $\boldsymbol{R_0} = \sigma^2_{\epsilon_0} \boldsymbol{I}$. We resolve this issue by weighting the errors by the matrix $(\boldsymbol{U}_0^T)^{-1}$, where $\boldsymbol{U}_0$ is the Cholesky decomposition of $\boldsymbol{V}_0$, i.e. $\boldsymbol{V_0} = \boldsymbol{U}_0^T \boldsymbol{U}_0$. As a result, the set of weighted errors, $(\boldsymbol{U}_0^T)^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})$, are normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$, and are thereby exchangeable, allowing once again for permutations both within and between subjects. We re-express the test

statistic T_1 in equation (2.4), to incorporate the Cholesky decomposition as:

$$T_2 = \sum_{i=1}^{N} \tilde{b}_{i2}^2 / N = \tilde{\boldsymbol{b}}^{*T} \tilde{\boldsymbol{b}}^* / N, \tag{2.7}$$

where $\boldsymbol{b}^* = \hat{\boldsymbol{G}}_1 \boldsymbol{Z} \hat{\boldsymbol{V}}_1^{-1} \boldsymbol{U}_0^T (\boldsymbol{U}_0^T)^{-1} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. Note that T_2 is only calculated for the single random effect being tested. For the observed data the statistic remains the sample variance of $\tilde{b}_{i2}$ because $\boldsymbol{U}_0^T (\boldsymbol{U}_0^T)^{-1}$ equals the identity for the unpermuted weighted residuals. Also, the earlier random intercept hypothesis test is a special case of this test, because the Cholesky decomposition in that scenario is equal to the identity, and (2.7) reduces to (2.4). With the appropriate weights, this BLUP-based permutation test can be used to perform inference on any single random effect of interest.

In simulation studies which are presented in Section 2.3, this permutation test is shown to be valid and displays power comparable to the asymptotic mixture χ^2 likelihood ratio tests. The test is very intuitive and easy to perform. However, since the test is based on the BLUPs, it does have one limitation: it can only test for one random effect at a time. In the next section, we present a likelihood ratio based permutation test that allows for testing of multiple random effects and of which the BLUP permutation test is a special case.

2.2.2 Likelihood Ratio Based Permutation Test

This permutation test is based on the likelihood ratio test statistic, $\lambda = -2\log(L_{H_0} - L_{H_1})$, where L_{H_0} and L_{H_1} are the likelihoods under the null and alternative hypotheses, respectively. Using the same linear mixed model notation as described previously where $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$ and $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, we have $\lambda = \log\left[|\boldsymbol{V_0}|/|\boldsymbol{V_1}|\right] + \boldsymbol{\epsilon}^T(\boldsymbol{V_0}^{-1} -$

$\boldsymbol{V_1}^{-1})\boldsymbol{\epsilon} + \log[|\boldsymbol{X}^T\boldsymbol{V_0}^{-1}\boldsymbol{X}|/|\boldsymbol{X}^T\boldsymbol{V_1}^{-1}\boldsymbol{X}|]$.

Let us test for a random intercept using the null and alternative hypotheses stated in (2.2) and (2.3). Similar to the BLUP based permutation test, the likelihood ratio test statistic involves the marginal residuals, $\boldsymbol{\epsilon}$, and we can permute $\boldsymbol{\epsilon}$ within and between the subjects under the null hypothesis. Therefore, the test statistic becomes

$$T_3 = \log[|\hat{\boldsymbol{V}}_{\boldsymbol{0}}|/|\hat{\boldsymbol{V}}_{\boldsymbol{1}}|] + \hat{\boldsymbol{e}}_{\boldsymbol{1}}^T(\hat{\boldsymbol{V}}_{\boldsymbol{0}}^{-1} - \hat{\boldsymbol{V}}_{\boldsymbol{1}}^{-1})\hat{\boldsymbol{e}}_{\boldsymbol{1}} + \log[|\boldsymbol{X}^T\hat{\boldsymbol{V}}_{\boldsymbol{0}}^{-1}\boldsymbol{X}|/|\boldsymbol{X}^T\hat{\boldsymbol{V}}_{\boldsymbol{1}}^{-1}\boldsymbol{X}|], \quad (2.8)$$

which is λ with all parameters replaced by their estimates under the null and alternative hypotheses as denoted by their subscripts.

Similar to the BLUP based permutation test, a new $\hat{\boldsymbol{V}}_{\boldsymbol{0}}$ and $\hat{\boldsymbol{V}}_{\boldsymbol{1}}$ is estimated for each permutation of $\hat{\boldsymbol{e}}_{\boldsymbol{1}}$ and denoted as $\hat{\boldsymbol{V}}_{\boldsymbol{0}}^*$ and $\hat{\boldsymbol{V}}_{\boldsymbol{1}}^*$. The permuted residuals are treated as an outcome, and $\hat{\boldsymbol{V}}_{\boldsymbol{0}}^*$ is estimated from a mixed model with a fixed intercept and random effects from the null hypothesis. We estimate $\hat{\boldsymbol{V}}_{\boldsymbol{1}}^*$ from a mixed model with a fixed intercept and random effects from the alternative hypothesis.

Re-estimation of $\hat{\boldsymbol{V}}_{\boldsymbol{0}}^*$ and $\hat{\boldsymbol{V}}_{\boldsymbol{1}}^*$ is necessary due to the changes that occur in the rank of $\hat{\boldsymbol{\Sigma}}$ when random effect variances are estimated to be equal to 0. If we do not re-estimate $\boldsymbol{V_0}$ and $\boldsymbol{V_1}$ (including $\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$), the permutation distribution will be completely based on estimates from the observed data. By estimating $\boldsymbol{V_0}$ and $\boldsymbol{V_1}$ for each permutation, we allow the empirical distribution to "mix" as the rank of $\hat{\boldsymbol{\Sigma}}$ varies, thereby generating a distribution similar to the mixture χ^2 asymptotic distribution of Stram and Lee [1994]. We create the permutation distribution by calculating T_3^* for each of the random permutations and determine a p-value through the location of T_3 in the permutation distribution.

When testing the presence of one random effect with one or more additional ran-

dom effects in the null hypothesis, things proceed similarly to that of the BLUP permutation test. In order to be able to permute the errors, they must first be weighted by $(\boldsymbol{U}_0^T)^{-1}$. Once weighted, the errors are exchangeable and can be permuted. The permuted weighted errors are then multiplied (unweighted) by $(\boldsymbol{U}_0^T)$ to get them back on the original scale of the residuals, and for each permutation, $\hat{\boldsymbol{V}}_{\mathbf{0}}^{\ *}$ and $\hat{\boldsymbol{V}}_{\mathbf{1}}^{\ *}$ are re-estimated using the unweighted permuted errors as described earlier. Then T_3^* is calculated, and the permutation distribution is generated for the likelihood ratio test statistic to which the observed test statistic will be compared and a p-value calculated.

If we wish to test for the inclusion of $0 < r \leq q$ random effects, we have the models:

$$H_0 : Y_{ij} = \beta_1 x_{1ij} + ... + \beta_p x_{pij} + b_{i1} z_{1ij} + ... + b_{i(q-r)} z_{(q-r)ij} + \epsilon_{ij}$$

$$H_1 : Y_{ij} = \beta_1 x_{1ij} + ... + \beta_p x_{pij} + b_{i1} z_{1ij} + ... + b_{iq} z_{qij} + \epsilon_{ij}$$

The steps for this scenario are identical to those from the previous scenario where testing for one random effect in the presence of additional random effects in the null hypothesis. Nonetheless, we emphasize the importance of re-estimating $\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$ after each permutation when testing for multiple random effects. Herein lies the largest contribution of our methods: for a general value of r, simulation is the only existing approach for calculating the correct mixing probabilities for the χ^2 distributions. In contrast, our permutation test based on the likelihood ratio statistic will automatically generate the correct mixing probabilities as the rank of $\hat{\boldsymbol{\Sigma}}^*$ changes from permutation to permutation.

2.3 Simulation Studies

2.3.1 Validity

We performed a series of simulation studies to examine the performance of our permutation tests under a number of different settings. The first study was used to evaluate the validity of our two tests under four different scenarios: (1) testing for a random intercept, (2) testing for a random slope given an independent random intercept is present in the null hypothesis, (3) testing for a random slope given a potentially correlated random intercept, and (4) simultaneously testing for both random intercept and random slope. Five hundred data sets were generated for each of the simulation scenarios using the following random intercept model:

$$Y_{ij} = \beta_1 + \beta_2 x_{2ij} + b_{i1} + \epsilon_{ij} \tag{2.9}$$

with $\beta_1 = 3$, $\beta_2 = 2.75$, $\sigma^2_\epsilon = 1$, $b_{i1} \sim N(0, \sigma^2_{i1})$, and our fixed effect, x_{2ij}, was randomly drawn from the standard normal distribution. Then, similar to Saville and Herring [2009], x_{2ij} was centered at 0 and scaled by twice its standard deviation. For scenarios 1 and 4, $\sigma^2_{b_{i1}}$ was set equal to 0, while for scenarios 2 and 3, $\sigma^2_{b_{i1}}$ was set to 1. We varied both the number of subjects, $N \in \{50, 10\}$, as well as the number of observations per subject, $n \in \{10, 5\}$, and compared the size of our permutation tests to that of the asymptotic restricted likelihood ratio test with a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom, 1 and 2 degrees of freedom, 1 and 2 degrees of freedom, and 0, 1, and 2 degrees of freedom in a 25:50:25 ratio, for scenarios 1, 2, 3, and 4 respectively. The mixing probabilities for scenario 4 were derived from Case 4 of Stram and Lee [1994] who state that when the information

matrix is equal to the identity under the null hypothesis, the likelihood ratio test has an asymptotic null distribution that is a mixture of χ^2 distributions with binomial mixing probabilities. For all other situations they recommend finding the critical value through simulations.

All estimates were performed in the statistical package R using the `lmer()` function from the R-package `lme4` Bates et al. [2011]. Unlike other linear mixed model fitting algorithms that can only estimate extremely small values for variances, `lmer()` is able to estimate 0 for the variance components. The simulations were performed using 20 cores of an Intel Xeon X5660 2.80 GHz server with 32 gigabytes of memory.

The simulation results for validity are presented in Table 1. In all settings, both permutation tests have valid size, defined as a size contained in the interval $(0.031, 0.061)$, the approximate 95% confidence interval for Type I error rate with 500 simulations. In contrast, the asymptotic test for one random effect (scenarios 1, 2, and 3) becomes more conservative as the number of subjects or the number of observations decreases. In addition, it appears that under scenario 4, the asymptotic likelihood ratio test is liberal when $N = 10$ and $n = 5$.

2.3.2 Power

The simulations to examine the power of the tests were performed for the same four scenarios in the validity study. We generated 500 data sets using the random intercept and slope model:

$$Y_{ij} = \beta_1 + \beta_2 x_{2ij} + b_{i1} + b_{i2} z_{2ij} + \epsilon_{ij} \tag{2.10}$$

Table 2.1: Size and power for the permutation tests compared to the asymptotic likelihood ratio test

			Testing Scenarios											
			(1)			(2)			(3)			(4)		
N	n	σ_i^2	B	L	A	B	L	A	B	L	A	L	A
50	10	0	5.8	5.8	5.0	5.8	5.8	5.6	5.0	4.8	4.2	4.2	4.4
		0.15	99.2	99.2	98.8	40.2	41.2	38.4	45.2	42.6	38.9	98.0	98.0
		0.20	100.0	100.0	100.0	57.4	58.4	55.4	62.2	58.4	56.4	98.8	98.8
		0.30	100.0	100.0	100.0	82.4	81.0	78.8	78.8	75.4	73.6	99.8	99.8
	5	0	3.8	3.6	2.6	6.2	5.2	5.0	5.0	5.2	4.4	4.8	4.0
		0.15	80.0	80.0	77.6	16.2	18.2	16.0	21.4	22.0	18.6	56.4	55.4
		0.20	91.2	91.2	90.2	27.8	27.4	24.4	27.0	25.4	22.0	70.1	69.3
		0.30	97.6	97.6	97.6	38.4	39.0	36.8	41.6	39.6	36.6	92.6	92.6
10	10	0	5.4	5.4	4.0	5.2	4.4	3.0	6.2	4.8	3.4	4.2	5.0
		0.15	63.4	63.2	58.8	16.2	15.2	12.6	17.3	16.3	11.0	55.6	58.3
		0.20	75.2	74.6	69.6	23.6	23.6	19.8	23.1	21.7	17.3	68.1	70.1
		0.30	89.0	89.0	87.6	34.4	34.8	29.8	30.7	27.3	22.5	88.2	89.0
	5	0	4.6	4.4	3.6	5.2	3.8	2.6	5.6	5.2	3.8	5.6	7.0
		0.15	31.6	29.4	27.0	10.0	8.6	7.0	9.8	10.0	7.2	24.8	29.1
		0.20	44.6	43.4	37.6	12.6	11.4	9.2	12.6	13.0	8.8	37.5	42.1
		0.30	63.6	62.0	58.6	12.8	13.8	11.2	15.7	15.7	10.6	47.9	53.3

Results are reported in percentages.
σ_i^2 refers to the variance component(s) being tested.
(1): Random intercept test,
(2): Random slope test with an independent random intercept present,
(3): Random slope test with a correlated random intercept present,
(4): Simultaneous test for the random intercept and random slope.
B: BLUP based permutation test,
L: Likelihood Ratio based permutation test,
A: Asymptotic likelihood ratio test

with the same fixed effects from the validity simulations and with $b_{i1} \sim N(0, \sigma^2_{i1})$, $b_{i2} \sim N(0, \sigma^2_{i2})$, and $x_{2ij} = z_{2ij}$. We varied the variance of the random effect (or random effects under scenario 4) of interest, $k \in \{1, 2\}$, $\sigma^2_{ik} \in \{0.15, 0.2, 0.3\}$ as well as both the number of subjects, $N \in \{50, 10\}$, and the number of observations per subject, $n \in \{10, 5\}$. For scenarios 3 and 4 the correlation of the random effects, ρ, was set equal to -0.3.

The results of the power simulations are shown in Table 1. With the exception of scenario 4, both permutation tests displayed strictly better power than the asymptotic test, even when the asymptotic test had nominal size. For scenario 4 the asymptotic likelihood ratio test using the 25:50:25 ratio of χ^2 distributions and the likelihood ratio based permutation test performed very similarly when $N = 50$. However, the number of rejections of the asymptotic test is higher than the permutation test for $N = 10$, and this can be explained by its inflated Type I error rate. In fact, when critical values found through simulation were used instead of the 25:50:25 mixture χ^2 null distribution, the power results for the asymptotic test were almost identical to those from the permutation test for all combinations of N and n.

Given that the residuals follow known normal distributions, it is possible that residuals could be drawn directly from those distributions (bootstrapped), rather than permuting the actual residuals, to generate the empirical null distributions of T_1, T_2, and T_3. To examine this idea, we performed simulations in which we replaced permuting the residuals with simulating new values from normal distributions with mean zero and variance equal to the error variance estimate from the null model. All other steps in the permutation tests were identical to those presented in Section 3. Both the BLUP and the restricted likelihood ratio versions were examined. N and n were set at 10, and we varied the variance of the random slope, $\sigma^2_{i2} \in \{0, 0.15, 0.2, 0.3\}$.

We tested for the presence of a random slope given a potentially correlated random intercept. The results from these simulations closely mirrored the results of the permutation tests in Table 1. Both test statistics using bootstrap residuals led to valid inference. When $\sigma^2_{i2} = 0.15$ the powers were 16.1% and 16.6% for the BLUP test and the restricted likelihood ratio tests respectively compared with the 17.3% and 16.3% from the permutation tests. For $\sigma^2_{i2} = 0.2$ the powers for the BLUP and the restricted likelihood ratio tests were 23.1% and 20.7%, respectively, and for $\sigma^2_{i2} = 0.3$, the powers were 29.1% and 27.9%, respectively.

2.3.3 Sensitivity to Non-Normality

We also investigated the sensitivity of the permutation tests to non-normality of the random effects and/or residuals when testing for a random slope given an independent random intercept in the model with $N = 10$ and $n = 10$. Both the null model with $\sigma^2_{i2} = 0$ and the alternative with $\sigma^2_{i2} = 0.3$ were run. Four different settings were studied: (1a) normal errors and normal random effects, (1b) logistically distributed errors and normal random effects, (1c) normal errors and logistically distributed random effects, and (1d) logistically distributed errors and logistically distributed random effects. Size and power estimates are given in Table 2. We see that under the null hypothesis, both permutation tests appear have size closer to nominal than the asymptotic test, with the asymptotic test being conservative in settings 1a, 1b, and 1c. Under the alternative hypothesis, we see that as expected, the permutation test is most powerful when the data truly are normally distributed (setting 1a), with slight losses in power when extra variation due to non-normality exists in the data. Nonetheless, the power losses of the permutation tests are slight, and in all settings, the permutation tests display greater power than the asymptotic test.

Table 2.2: Size and power of proposed permutation tests when random effects and/or errors are non-normally distributed.

		Method		
Model	Setting	B	L	A
$\sigma^2_{i2} = 0.0$	1a	5.2	4.4	3.0
	1b	5.4	4.4	3.6
	1c	4.4	4.4	3.2
	1d	5.0	5.0	5.6
$\sigma^2_{i2} = 0.3$	1a	34.4	34.8	29.8
	1b	29.2	29.4	26.8
	1c	29.4	30.4	25.2
	1d	29.4	30.0	27.2

Results are reported in percentages.
Settings: (1a): Normal errors and normal random effects
(1b): Logistic errors and normal random effects
(1c): Normal errors and logistic random effects
(1d): Logistic errors and logistic random effects
B: BLUP based permutation test,
L: Likelihood Ratio based permutation test,
A: Asymptotic likelihood ratio test

2.3.4 The Effect of Unbalanced Data

We also investigated the performance of the permutation tests when the data are unbalanced. We tested for a random slope given an independent random intercept in the model with $N = 10$ and $n = 10$. For the random effect of interest we used the null model with $\sigma^2_{i2} = 0$ and the alternatives as in the original power simulations with $\sigma^2_{i2} \in \{0.15, 0.2, 0.3\}$. In order to create unbalanced data the number of observations for each patient was a uniform random integer from [2, 10]. Therefore, our total number of observations can vary, but within a simulation it was fixed for all of the runs. Size and power estimates are given in Table 2.3 along with the asymptotic likelihood ratio test. We see that unbalanced data does not appear to affect the

hypothesis test. The permutation tests still appear to be slightly closer to the nominal level under the null scenario. Under the alternative the permutation tests are very slightly more powerful than the asymptotic likelihood ratio test. Since the total number of observations are not fixed to be equal to one of the above simulations where the data was balanced we cannot directly compare these results to those from the balanced data scenario.

Table 2.3: Size and power of proposed permutation tests when the number of observations for each subject is not constant.

		Method		
N	σ^2_{i2}	B	L	A
10	0.00	4.8	4.0	3.8
	0.15	11.4	10.2	9.0
	0.20	14.2	15.0	13.6
	0.30	50.6	49.2	46.4
50	0.00	5.2	5.2	4.4
	0.15	25.0	25.2	24.0
	0.20	35.8	37.8	35.0
	0.30	50.8	51.2	50.6

Results are reported in percentages.
B: BLUP based permutation test,
L: Likelihood Ratio based permutation test,
A: Asymptotic likelihood ratio test

2.3.5 Comparison to Existing Methods

In our final simulation study, we compared the permutation tests to a portion of the results published by Saville and Herring [2009] when testing for the presence of a random slope. Following their simulation settings, we generated 250 data sets from (2.10) with $\beta_0 = 2.75$, $\beta_1 = 3$, $n_i = n = 10$, $\sigma^2_{i1} = 1$, and $\rho = -0.3$. The standard deviation for the random slope, $\sigma_{i2} \in \{0, 0.15, 0.30, 0.45, 0.60\}$. Table 3 presents the

Table 2.4: Comparison of power of permutation tests to results reported by Saville and Herring when testing for the inclusion of a random slope.

N	σ_{i2}	SH1	SH2	BLUP	LRT
50	0.00	8	3	3	4
	0.15	14	7	8	9
	0.30	30	17	28	22
	0.45	56	57	66	60
	0.60	75	90	94	91
100	0.00	4	4	5	4
	0.15	12	8	14	12
	0.30	38	38	44	43
	0.45	69	87	91	90
	0.60	72	99	100	100

Results are reported in percentages.
SH1: Bayes factor as described in Saville and Herring [2009]; page 370.
SH2: Bayes factor as described in Saville and Herring [2009]; page 371.
BLUP: BLUP based permutation test.
LRT: Likelihood ratio based permutation test.

BLUP and likelihood ratio based permutation results for $N \in \{100, 50\}$ next to the published results from Saville and Herring resulting from Bayes' factors based on two different parameterizations of the model.

We see that the power for the likelihood ratio based permutation test is comparable with the approximate Bayes factors method employed by Saville and Herring. Despite some difference in results due to simulation variability, for all settings, our permutation test is as powerful or even more powerful than one or both of the tests of Saville and Herring.

2.4 Application

We applied our permutation test to a set of data presented in Klein et al. [1984] that was collected on patients with chronic myelogenous leukemia (CML). CML is characterized by a lengthy chronic phase with little to no symptoms that eventually transitions into an accelerated phase which behaves similarly to acute leukemia. The length of time until the transition from a chronic to an accelerated phase can vary greatly among patients, motivating the discovery of markers that can indicate when CML is about to change from a chronic to an accelerated stage. One potential marker is adenosine deaminase (ADA). This particular data set contains the ADA levels of 55 patients that were measured at various time points during their follow-up. Time is quantified as days following the initial observation date, and at each time point, investigators also recorded the phase of each patient's disease as chronic or accelerated. The frequency of the repeated measurements as well as the times of the measurements were not fixed and fluctuated greatly. Patients had anywhere from 2 to 59 measurements, and the repeated measurements took place from the initial observation date up to 1073 days following the diagnosis date.

We modeled the ADA measurements as patients progress from chronic to accelerated phases, and we were primarily interested in evaluating the level of heterogeneity among the patients to see if random effects are necessary in our model. Figure 1 contains boxplots stratified by stage of disease of the slopes and intercepts from individual linear regressions of each patient's ADA measurements on time. The figure indicates significant variation between the two stages, both in terms of mean ADA levels as well as changes over time, necessitating the inclusion of random effects.

We are also interested in investigating how the rate-of-change in ADA differs

Figure 2.1: Boxplots stratified by stage of the patient-specific intercepts and slopes produced from a linear regression model of ADA levels over time.

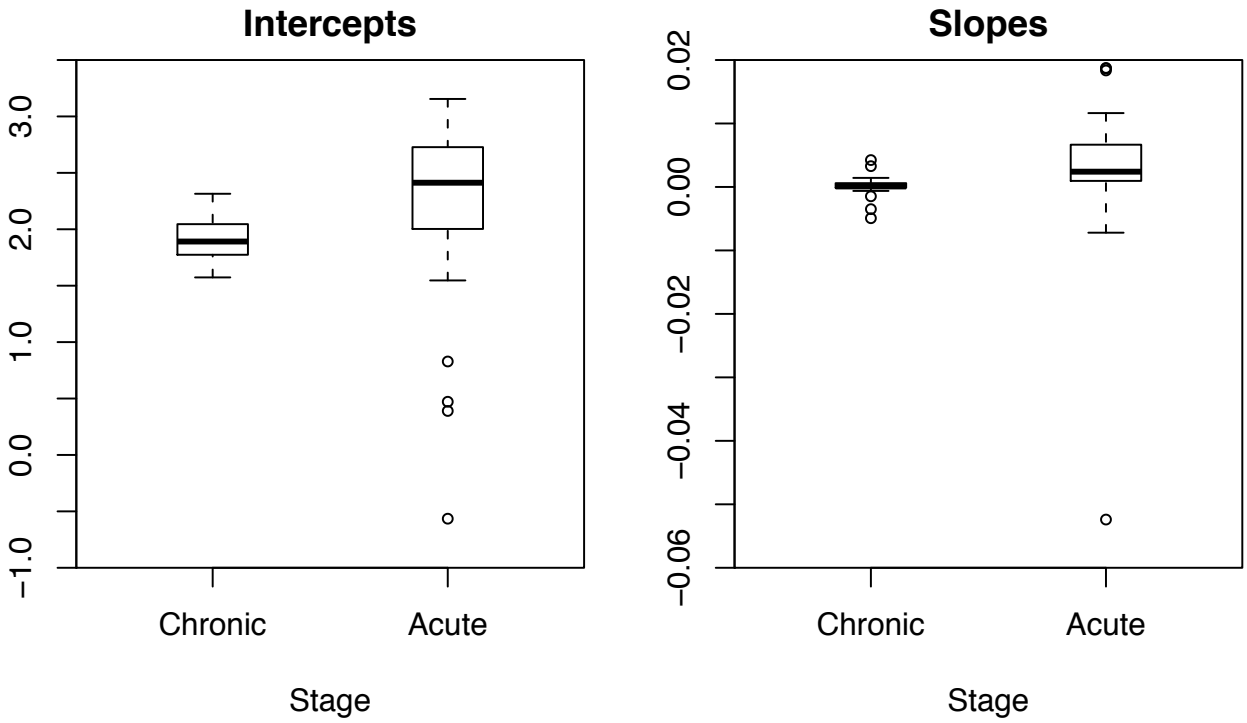


Table 2.5: Permutation and asymptotic likelihood ratio test results for inclusion of specific random effects when modeling ADA levels in patients with chronic myelogenous leukemia.

Test	Observed LRTS	Permutation p-value	Asymptotic p-value
(5) vs (4)	3.00	0.226	0.474
(4) vs (1)	234.59	<0.001	<0.001
(4) vs (2)	146.99	<0.001	<0.001
(4) vs (3)	12.88	0.019	0.004

(1): No random effects
(2): Random intercept only model
(3): Random intercepts for both stages model
(4): Random intercepts for both stages and a random slope for acute stage only
(5): Random intercepts and slopes for both stages

between chronic and accelerated phases. We applied a cubed root transformation to the ADA values so that they were approximately normally distributed, and fit a linear mixed model with the cubed root ADA assay values regressed on disease phase, with chronic as the baseline category, number of days from the initial observation date, and interaction terms between the two to allow the time effect to differ between the two disease states. Our initial model is $ADA_{ij}^{1/3} = \beta_1 + b_{i1} + (\beta_2 + b_{i2})State_{ij} + (\beta_3 + b_{i3})Days_{ij} + (\beta_4 + b_{i4})State_{ij} * Days_{ij} + \epsilon_{ij}$. The full random effects model includes four random effects, b_{i1}, b_{i2}, b_{i3}, and b_{i4}, to allow for at most a random intercept and time effect for each of the two disease stages. We wish test if any or all of these random effects should be included.

Table 4 shows the results of our permutation tests, based on 1,000 permutations, for the inclusion or exclusion of the random effects, along with results from the asymptotic likelihood ratio test. Both tests support what is seen in Figure 1: the random day effect for the chronic stage is not significant, while the other three random effects

appear to be significant. As a gauge of the computation time necessary, each of these tests takes around 4 minutes to perform when using 20 cores of an Intel Xeon X5660 2.80 GHz server with 32 gigabytes of memory.

2.5 Discussion

In this chapter, we have proposed two methods for performing inference on random effects by permuting the weighted residuals both within- and among- subjects. In some simulations, we have found that the convergence of the solutions derived from the `lmer()` function in the statistical package R appears to suffer as the number of random effects increases. Our current solution is to generate more permutations to ensure that there are enough permutations to create the null distribution.

Specifying a covariance structure to the errors should not affect the validity of the permutation tests. Examples include the autoregressive and the Toeplitz covariance structures and means that the unweighted marginal residuals are not exchangeable. When this is the case then the covariance structure and its parameter estimates are incorporated into the residual covariance matrix $\hat{\boldsymbol{R}}$. The Cholesky decomposition of $\hat{\boldsymbol{V}}_0$ will account for the covariance structure, and the weighted residuals will be asymptotically exchangeable.

As demonstrated, the proposed permutation tests perform well even when the number of patients and the number of observations per patient is small. The tests also do not require balanced data nor do the measurements need to occur at the same points in time. As a result, our methods can be applied to the use of a LMM representation of penalized spline models [Ruppert et al., 2003] in which the smoothing parameter is a random effect. Finally, implementing these permutation tests is

straightforward and can be incorporated into standard practice for analysis of linear mixed models using existing software; example computer code can be found at www.sph.umich.edu/∼/tombraun/software.html. While the methods are computationally intensive, the recent rise in parallel computing through clusters and multicore processors has made it possible to greatly reduce the amount of time necessary to implement these tests.

In the following chapter we generalize the methods presented in this chapter to allow for permutation-based inference in generalized linear mixed models (GLMMs). Our approach is based upon a first-order approximation of the GLMM to make it resemble the form of a LMM, an approach that is the foundation of penalized quasi-likelihood (PQL) [Breslow and Clayton, 1993] for estimation in GLMMs.

CHAPTER III

Permutation Tests for Random Effects in Generalized Linear Mixed Models

3.1 Introduction

Often in studies non-normal data are observed and can be modeled using generalized linear models (GLM) such as logistic regression or Poisson regression [McCullagh and Nelder, 1989]. When the data arise from common origins or clusters it is likely that the assumption of independence among all of the data is violated because observations from the same cluster can be correlated. Examples of this include clinical trials where patients who share a common hospital are potentially correlated due to similar treatment standards and longitudinal data where a set of subjects are followed and repeatedly measured over time. When the independence assumption is presumed to be violated random effects can be added to the GLM resulting in the generalized linear mixed model (GLMM) [Breslow and Clayton, 1993]. Similar to adding random effects to a linear regression model to form a linear mixed model (LMM) the random effects account for the subject or cluster specific variation. Typically, the random effects are assumed to be normally distributed with mean zero and variance equal to σ_b^2.

Maximum likelihood estimation of the mean parameters (fixed effects) of a GLMM is a challenging task because the likelihood for the mean parameters does not have a closed form when the random effects have a normal distribution. However, a closed form does exist for a GLMM with a single random effect when a so-called bridge distribution is assumed for the random effect [Wang and Louis, 2003]. With normally distributed random effects, the likelihood for the mean parameters is a possibly multi-dimensional integral, with one integral for each of the random effects, making numerical computation problematic. Therefore, many methods to circumvent this hurdle have been developed. These include a Monte Carlo EM algorithm [McCulloch, 1997, Booth and Hobert, 1999], a method of estimation by parts [Song et al., 2005], penalized quasi-likelihood [Breslow and Clayton, 1993], and adaptive Gaussian quadrature (AGQ)[Pinheiro and Bates, 1995]. Bayesian methods based on the Gibbs sampler have also been explored [Zeger and Karim, 1991].

While the majority of the literature has been focused on inference of the population or fixed effects, there is also considerable interest in testing for the inclusion or exclusion of random effects. These are tests for overdispersion, heteroscedasticity, and correlation among the outcomes of a GLMM. For the simplest case when one random effect is present, this test is equivalent to testing for the GLMM alternative against a null GLM with no random effects. Since the random effect is defined as a normally distributed random variable with mean zero, comparing the GLMM to the GLM is equivalent to testing if the variance of the random effect is equal to zero. This presents another set of problems because under the null hypothesis, the variance component is located on the boundary of its parameter space, and the typical likelihood ratio, score and Wald test statistics do not have their usual χ^2 asymptotic null distributions. Instead it was shown by Self and Liang [1987] that the likelihood

ratio test (LRT) statistic follows a mixture χ^2 distribution. Stram and Lee [1994] demonstrated these results specifically for the LMM; these results also hold for the GLMM [Wolak, 1989].

Apart from the likelihood ratio test, there has also been work on score tests for variance components in GLMMs [Lin, 1997, Hall and Præstgaard, 2001]. Lin [1997] builds upon the the quasilikelihood methods of Breslow and Clayton and utilizes a Laplace expansion of the integrated log-quasilikelihood to derive a global or omnibus score test for the null hypothesis that all random effects are unnecessary. The advantage of the score test over the LRT is that only the simpler GLM needs to be fit to obtain the null parameter estimates. Surprisingly, Lin [1997] shows that the score test follows an asymptotic χ^2_m distribution, where m is the number of independent random effects. She notes that the asymptotic tests could be less accurate when the number of levels of each random effect is small. In addition to the global score test, Lin [1997] also developed an individual variance component test, however, the performance of this test is unsatisfactory when the data are binary. Hall and Præstgard modify Lin's omnibus score test by constraining the alternative covariance matrix to be positive-semidefinite. The modified score test statistic follows a mixture of χ^2 distributions and is shown to have increased power in simulations when compared with the score test of Lin [1997].

Bootstrap tests for variance components of GLMMs have also been proposed [Sinha, 2009]. Using the score as the test statistic the parametric bootstrap proceeds by generating samples of the data under the null GLM which are then used to calculate a bootstrap specific score statistic. The bootstrap score statistics form an approximate null distribution for the score statistic with which to calculate a p-value. Bayesian alternatives have also been proposed; the work of Sinharay and Stern [2005]

provides an overview of different methods of estimating Bayes factors for GLMMs.

Permutation tests provide an alternative method of testing for random effects in mixed models. The work of Fitzmaurice and Ibrahim [2007] is applicable to GLMMs, but is limited to multi-level studies where the inclusion of a single random effect to quantify the heterogeneity among the different levels may be tested. They compared the likelihood ratio test statistic to an empirical null distribution generated by randomly permuting the level assignments among the subjects. The methods of Lee and Braun [2012] are based on permutations of weighted residuals and are applicable to any type of LMM and any number of random effects. However, these methods were developed solely for LMMs.

In this article we present a modification of the work of Lee and Braun to apply the permutation tests to the random effects of GLMMs based on a penalized quasi-likelihood (PQL) approximation developed by Breslow and Clayton [1993]. Two permutation test statistics are proposed, and both statistics are a sum of squared residuals, with the empirical null distributions generated via permutations of the residuals. The first test statistic is based on the Best Linear Unbiased Predictions (BLUPs) [Robinson, 1991] and the second statistic is the restricted likelihood ratio test statistic assuming normality of the data. We will show via simulation that our tests appear to have valid size, and that their powers are comparable to existing methods. In Section 2, we begin with notation for generalized linear mixed models. Section 3 follows with a presentation of our proposed methods. We present the results of simulations in Section 4 that demonstrate the validity and power of our methods, and compare our permutation tests to existing methods. In Section 5 we apply our methods to two data sets: a longitudinal clinical trial investigating the incidence of amenorrhea in women using contraception, and a clinical trial studying the effects of

progabide on reducing the number of seizures. We close with a discussion of our work in Section 6.

3.2 Methods

3.2.1 Generalized Linear Mixed Models

Let Y_i be an observation of subject i for $i = 1, 2, \ldots, N$, where the density of Y_i belongs to the exponential family of distributions, written as

$$f(y;\eta) = exp\left(\frac{y\eta - b(\eta)}{\phi} + c(y,\phi)\right).$$

For exponential family distributions ϕ is the dispersion parameter, η is the natural parameter, $b(\eta)$ is a known function specific to the distribution, and $c(y,\phi)$ is a function of the data that does not depend on η. The natural parameter, η is connected to the mean of the distribution, μ, through a strictly increasing link function $g(\cdot)$. A generalized linear model for Y_i relates the mean, μ_i, to a linear function of the covariates through the natural parameter η_i

$$g(\mu_i) = \eta_i = \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

where $\beta_1, \ldots, \beta_p$ are the population level fixed-effect coefficients. The $x_{1i}, \ldots, x_{pi}$ are the observed fixed effect covariates for subject i. Generally, x_{1i} is constant and equal to 1 to represent the fixed intercept.

Next, we will extend the GLM by accounting for correlation among the observations. The subscript j will denote a repeatedly measured cluster or subject. There-

fore, let Y_{ij} be the jth observation of subject or cluster i for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n_i$, where the density of Y_{ij} belongs to the exponential family of distributions. A generalized linear mixed model for Y_{ij} relates the mean, μ_{ij} conditional on the subject specific random effects, to a linear function of the covariates

$$g(\mu_{ij}|\boldsymbol{b}_i) = \eta_{ij} = \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + b_{i1} z_{1ij} + \ldots + b_{iq} z_{qij}$$

where $\beta_1, \ldots, \beta_p$ are still the population level fixed-effect coefficients and $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})$ are the subject-specific random effects for subject or cluster i. The $z_{1ij}, \ldots, z_{qij}$ are the observed random effect covariates for observation j of subject i, and usually z_{1ij}, is constant and equal to 1 to represent a random intercept. The random effects, $\boldsymbol{b}_i = \{b_{i1}, b_{i2}, \ldots, b_{iq}\}$ are assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, in which the respective variances for $b_{i1}, b_{i2}, \ldots, b_{iq}$ are denoted as $\sigma^2_{b_1}, \sigma^2_{b_2}, \ldots, \sigma^2_{b_q}$. Typically, the inverse link function is denoted as $h(\cdot)$. Therefore for a given $\boldsymbol{b}_i$, $\mu_{ij} = h(\beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + b_{i1} z_{1ij} + \ldots + b_{iq} z_{qij})$. The linear mixed model is a special case of the GLMM with y_{ij} normally distributed given $\boldsymbol{b}_i$ and the identity link $g(\mu) = \mu$. Other examples of GLMMs include logistic regression and Poisson regression with random effects.

The generalized linear mixed model for subject i can be written using matrix notation as $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$, where $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_p\}$ and $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are subject-specific design matrices for the p fixed effect covariates and q random effect covariates, respectively. We then combine data from all subjects so that $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ is the $\sum_i n_i$ vector of outcomes and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the respective design matrices for the p fixed effect covariates and q random effect covariates formed by successively placing the design matrices of each subject under each other. Furthermore, we denote

$\boldsymbol{b} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N\}$ where $Var(\boldsymbol{b}) = \boldsymbol{G} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_G$, in which $\otimes$ denotes the Kroenecker product, and $\boldsymbol{I}_G$ is a $N \times N$ identity matrix to reflect independence among subjects or clusters.

Estimation of the fixed effect parameters of the GLMM involves maximizing the marginal log-likelihood

$$\int f(\boldsymbol{Y}|\beta, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{G}) d\boldsymbol{b}$$

and requires that all of the q random effects are integrated out of the log-likelihood. However, the q-dimensional integral often does not have a closed-form solution. Therefore, methods for approximating the integral are required and have seen a lot of development.

3.3 Proposed Methods

3.3.1 Linear Mixed Model Approximation of the Generalized Linear Mixed Model

In order to apply the permutation tests from Chapter 2, an approximation for the GLMM is necessary to "linearize" the data so that an LMM can be used to fit the data to get working random errors. Let us begin with a standard GLMM,

$$g(\boldsymbol{\mu}|\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}. \tag{3.1}$$

Using an extension of working variates in generalized linear models [Nelder and Wedderburn, 1972], a Taylor expansion of the link function, g, around the conditional mean of $\boldsymbol{Y}$ given $\boldsymbol{b}$ [McCulloch et al., 2008] results in a vector of working variates

that will be denoted as $\boldsymbol{Y}^*$:

$$\boldsymbol{Y}^* = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + g'(\boldsymbol{\mu}|\boldsymbol{b})(\boldsymbol{Y} - \boldsymbol{\mu}|\boldsymbol{b}). \tag{3.2}$$

The third term of this equation, $g'(\boldsymbol{\mu}|\boldsymbol{b})(\boldsymbol{Y} - \boldsymbol{\mu}|\boldsymbol{b})$, has mean equal to $\mathbf{0}$ so it can be considered to be an error term which we will denote as $\boldsymbol{\epsilon}^*$. Therefore, $\boldsymbol{Y}^*$ can be fit using a linear mixed model with non-constant variance, as each element of $\boldsymbol{\epsilon}^*$ has variance that is a function of its mean. In fact, a method for estimation in GLMMs through iterations of LMM fits was proposed by Schall [1991]. Breslow and Clayton arrived at the same conclusion through their work on the penalized quasi-likelihood.

Because the elements of $\boldsymbol{\epsilon}^*$ do not have constant variance, they are not exchangeable and cannot be permuted. The variance of $\boldsymbol{\epsilon}^*$ conditional on $\boldsymbol{b}_i$, denoted as $\boldsymbol{W}^{-1}$, is a matrix with the diagonal elements equal to $w_{ij} = \{Var(\mu_{ij}|\boldsymbol{b}_i)(g'(\mu_{ij}|\boldsymbol{b}_i))^2\}^{-1}$. When g is the canonical link, $g'(\mu_{ij}|\boldsymbol{b}_i) = Var(\mu_{ij}|\boldsymbol{b}_i)^{-1}$, so that $w_{ij} = g'(\mu_{ij}|\boldsymbol{b}_i)$. Thus, we can weight $\boldsymbol{Y}^*$ by $\boldsymbol{W}$ so that $\boldsymbol{W}\boldsymbol{\epsilon}^*$ has variance equal to $\sigma^2_\epsilon \boldsymbol{I}$, resulting in the following equation:

$$\boldsymbol{W}\boldsymbol{Y}^* = \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{W}\boldsymbol{\epsilon}^*. \tag{3.3}$$

At this point, (3.3) is a linear mixed model of $\boldsymbol{W}\boldsymbol{Y}^*$ with parameters equal to those in (3.1). Estimates from the GLMM are needed in order to calculate $\boldsymbol{W}\boldsymbol{Y}^*$, and inserting the GLMM estimates into (3.3) results in

$$\hat{\boldsymbol{W}}\boldsymbol{Y}^* = \hat{\boldsymbol{W}}\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{W}}\boldsymbol{Z}\tilde{\boldsymbol{b}} + \hat{\boldsymbol{W}}\hat{\boldsymbol{\epsilon}}^*, \tag{3.4}$$

where $\hat{\boldsymbol{W}}^{-1}$ has diagonal elements equal to $\hat{w}_{ij} = \{Var(\hat{\mu}_{ij}|\tilde{\boldsymbol{b}}_i)(g'(\hat{\mu}_{ij}|\tilde{\boldsymbol{b}}_i))^2\}^{-1}$ and

$\hat{\boldsymbol{\epsilon}} = g'(\hat{\boldsymbol{\mu}}|\tilde{\boldsymbol{b}})(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}|\tilde{\boldsymbol{b}})$.

Now that we have rewritten (3.1) as a linear mixed model with estimates obtained from a GLMM fit we can proceed with the permutation tests.

In our simulations, we have found that performing the permutation test after transforming the data under the alternative GLMM resulted in inflated estimates of the random effect variance components, and caused the permutation tests to be liberal. As a remedy, we instead transform the data based on estimates from the null hypothesis GLM. This method estimates only the fixed effects, and all of the random effects and random error are contained in the working error term. We then use a linear mixed model to partition the model variance into random effects and random noise.

3.3.2 Permutation Tests for Random Effects

We begin by considering the hypothesis test for the inclusion or exclusion of a single random effect, $\boldsymbol{b_i} \sim N(0, \sigma^2_{b_i})$, in a GLMM with no other random effects present. Thus, we are comparing the following models:

$$H_0 : g(\mu_{ij}) = \beta_1 x_{1ij} + \ldots + \beta_p x_{pij}$$

$$H_1 : g(\mu_{ij}|\boldsymbol{b}_i) = \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + b_{i1} z_{1ij}.$$

We first fit the null GLM and calculate

$$\hat{w}_{ij} {Y_{ij}}^* = \hat{\beta}_1 \hat{w}_{ij} x_{1ij} + \ldots + \hat{\beta}_p \hat{w}_{ij} x_{pij} + \hat{w}_{ij} g'(\hat{\mu}_{ij})(Y_{ij} - \hat{\mu}_{ij}), \tag{3.5}$$

where $\hat{w}_{ij} = \{Var(\hat{\mu}_{ij})(g'(\hat{\mu}_{ij}))^2\}^{-1}$. Once we have obtained $\hat{w}_{ij}Y_{ij}{}^*$ if the alternative hypothesis is true then $e_{ij} = \hat{\epsilon}_{ij} = \hat{w}_{ij}g'(\hat{\mu}_{ij})(Y_{ij} - \hat{\mu}_{ij})$ contains both random effects as well as the random error. We propose using either T_1 or T_3 as the permutation test statistic, and to obtain observed values for T_1 or T_3, we fit the alternative linear mixed model to $\hat{w}_{ij}Y_{ij}{}^*$ computed in Equation (3.5).

The permutation distribution is achieved by permuting the marginal residuals from the linear mixed model. Under the null hypothesis, the errors, $e_{ij} = w_{ij}g'(\mu_{ij})(Y_{ij} - \mu_{ij})$, are exchangeable because they are independent and have constant variance. Asymptotically, the residuals, e_{ij} are also exchangeable. Permuting the residuals has the benefit of not requiring the continuous $\boldsymbol{X}$'s to be identical among all subjects nor do the number of observations for each subject need to be the same.

For each permutation of the residuals we must re-estimate the variance component of the random effect of interest. This allows our permutation null distribution to "mix" because a proportion of the permutations will result in a random effect variance estimate equal to zero. When testing for a single random effect, this allows the permutation test to generate the point mass at zero that is seen in the asymptotic 50:50 mixture of χ^2_0 and χ^2_1 distributions. For more complex hypotheses, this will automatically generate the appropriate mixture χ^2 distribution. The linear mixed model estimates are then used to calculate the permutation test statistics. In our simulations, 1000 Monte Carlo permutations are generated for the permutation null distribution of the test statistic. The same method can be used for hypothesis tests that simultaneously test multiple random effects at the same time.

3.3.2.1 Best Linear Unbiased Predictors Based Permutation Test Statistic

The first proposed permutation test statistic is the sample variance of the BLUPs for the random effect of interest:

$$T_1 = \sum_{i=1}^{N} \tilde{b}_{i1}^2 / N$$

where the denominator N is constant for all of the permutations and does not affect the validity or power of the test. T_1 involves the sum of the squared BLUPs where the BLUPs are treated as a random sample from the random effect distribution.

The BLUPs are predicted using solutions to the mixed model equations given by Henderson [1950].

$$\tilde{\boldsymbol{b}} = \hat{\boldsymbol{G}}\hat{\boldsymbol{W}}\boldsymbol{Z}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{e}}$$

where $\hat{\boldsymbol{e}} = \hat{\boldsymbol{W}}\boldsymbol{Y}^* - \hat{\boldsymbol{W}}\boldsymbol{X}\hat{\boldsymbol{\beta}}$. After calculating T_1 for each permutation, we generate a p-value by calculating the percentage of permutations with T_1 greater than the observed T_1. This test statistic is intuitive and the permutation test is easy to perform, but T_1 can only be used to test for one random effect at a time. The next test statistic that we propose will be able to test for multiple random effects, and, of which, T_1 is a special case.

3.3.2.2 Linear Mixed Model Restricted Likelihood Ratio Based Permutation Test Statistic

The second proposed test statistic is the restricted likelihood ratio test statistic for a linear mixed model, $\lambda = -2\log(L_{H_0} - L_{H_1})$, where L_{H_0} and L_{H_1} are the restricted likelihoods under the null and alternative hypotheses, respectively. For Equation

(3.3) we assume that $\boldsymbol{WY}^* \sim N(\boldsymbol{WX\beta}, \boldsymbol{V})$ and $\boldsymbol{\epsilon} = \boldsymbol{WY} - \boldsymbol{WX\beta}$, then we have $\lambda = \log[|\boldsymbol{V_0}|/|\boldsymbol{V_1}|] + \boldsymbol{\epsilon}^T(\boldsymbol{V_0}^{-1} - \boldsymbol{V_1}^{-1})\boldsymbol{\epsilon} + \log[|\boldsymbol{WX}^T\boldsymbol{V_0}^{-1}\boldsymbol{WX}|/|\boldsymbol{WX}^T\boldsymbol{V_1}^{-1}\boldsymbol{WX}|]$.

Inserting estimates into λ gives us the statistic

$$T_3 = \log[|\hat{\boldsymbol{V}}_{\boldsymbol{0}}|/|\hat{\boldsymbol{V}}_{\boldsymbol{1}}|] + \hat{\boldsymbol{e}}_{\boldsymbol{1}}^T(\hat{\boldsymbol{V}}_{\boldsymbol{0}}{}^{-1} - \hat{\boldsymbol{V}}_{\boldsymbol{1}}{}^{-1})\hat{\boldsymbol{e}}_{\boldsymbol{1}} + \log[|\hat{\boldsymbol{W}}\boldsymbol{X}^T\hat{\boldsymbol{V}}_{\boldsymbol{0}}{}^{-1}\hat{\boldsymbol{W}}\boldsymbol{X}|/|\hat{\boldsymbol{W}}\boldsymbol{X}^T\hat{\boldsymbol{V}}_{\boldsymbol{1}}{}^{-1}\hat{\boldsymbol{W}}\boldsymbol{X}|],$$

in which the subscripts 0 and 1 correspond to the null and alternative hypotheses, respectively. The observed T_3 is compared to the permutation distribution created by calculating T_3 for each permutation to calculate the p-value.

The main advantage of T_3 is that it can simultaneously test for multiple random effects. Coupled with re-estimation of the parameters for each permutation this means that for any number of random effects in our test the permutation test based on the likelihood ratio statistic will automatically generate the correct mixing probabilities as the rank of $\hat{\boldsymbol{\Sigma}}^*$ changes from permutation to permutation. The only other method of finding the correct mixing probabilities for the χ^2 distributions is through simulation.

3.4 Simulation Studies

3.4.1 Validity

We performed a series of simulations to examine the performance of our permutation tests under a number of different settings. The first study was used to evaluate the validity of the two tests under two different scenarios: (1) testing for a random intercept and (2) simultaneously testing for both random intercept and slope. This was done for both logistic and Poisson regression models. Five hundred data sets

were generated for each of the simulation scenarios using the following GLM

$$g(\boldsymbol{\mu}_{ij}) = \beta_1 + \beta_2 x_{ij}$$

with the appropriate link function for binary and Poisson data. The fixed effects, β_1 and β_2 were set equal to 0.25 and 0.5, respectively. Our fixed effect covariate, x_{2ij}, was randomly drawn from the standard normal distribution. We varied the number of subjects, $N \in \{50, 10\}$, and set the number of observations per subject, $n = 5$. For $N = 10$ we ran additional simulations with $n = 10$.

We compare the size of our permutation tests to that of the asymptotic restricted likelihood ratio test with a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom in scenario (1), and 0, 1, and 2 degrees of freedom in a 25:50:25 ratio in scenario (2). The mixing probabilities for scenario 2 were derived from Case 4 of Stram and Lee [1994] who state that when the information matrix is equal to the identity under the null hypothesis, the likelihood ratio test has an asymptotic null distribution that is a mixture of χ^2 distributions with binomial mixing probabilities. For all other situations, they recommend finding the critical value through simulations.

All estimates were performed in the statistical package R with the GLM estimates obtained using the `glm()` function and estimates for the LMM obtained through the `lmer()` function from the R-package `lme4` Bates et al. [2011]. Unlike other linear mixed model fitting algorithms that can only estimate extremely small values for variances, `lmer()` is able to estimate 0 for the variance components.

The simulation results for validity are presented in the first, sixth, and eleventh rows of Table 3.1. With the exception of two simulations with very small samples, $N = 10$ and $n = 5$, for the Poisson data both permutation tests have nominal

size, defined as a size contained in the interval $(0.031, 0.069)$, the approximate 95% confidence interval for Type I error rate with 500 simulations. The hypothesis tests for the binomial data appears to be more stable and closer to 0.05 than those for the Poisson data. For the Poisson data with smaller sample sizes, the asymptotic test appears to be slightly more conservative than the permutation tests when testing for one random effect and more liberal when testing for two random effects.

3.4.2 Power

The simulations to examine the power of the tests were performed for both binary and Poisson data under the same two scenarios as in the validity study. We generated 500 data sets using the random intercept and slope model model:

$$g(\mu_{ij}|\boldsymbol{b}_i) = \beta_1 + \beta_2 x_{2ij} + b_{i1} + b_{i2} z_{2ij}$$

with the same fixed effects from the validity simulations and with $b_{i1} \sim N(0, \sigma^2_{i1})$, $b_{i2} \sim N(0, \sigma^2_{i2})$, and $x_{2ij} = z_{2ij}$. We varied the variance of the random effect (or random effects under scenario 2) of interest, $k \in \{1, 2\}$, $\sigma^2_{ik} \in \{0.25, 0.5.0, 0.75, 1.00\}$ as well as both the number of subjects, $N \in \{50, 10\}$. Again, for $N = 10$ we performed simulations for $n \in \{10, 5\}$. For scenario 2 the random effects are assumed to be independent.

The results of the power simulations are shown in Table 3.1. It appears that the BLUP based and restricted likelihood ratio permutation tests have extremely similar power. When testing for a single random effect the permutation tests are more powerful than the asymptotic likelihood ratio test. The difference in power increases as the sample size gets smaller. For the Poisson data the difference in power

Table 3.1: Size and power for the permutation tests compared to the asymptotic likelihood ratio test

			Testing Scenarios									
			Binomial					Poisson				
			(1)			(2)		(1)			(2)	
N	n	σ_i^2	B	L	A	L	A	B	L	A	L	A
50	5	0	4.4	4.4	5.8	5.0	5.8	3.2	3.2	4.2	4.0	6.6
		0.25	28.8	28.8	27.2	34.6	33.2	99.6	99.6	99.6	98.4	100.0
		0.50	62.4	62.0	61.4	68.8	68.4	100.0	100.0	100.0	98.8	100.0
		0.75	81.4	81.4	81.6	85.6	85.2	100.0	100.0	100.0	98.8	100.0
		1.00	92.6	92.4	92.0	94.8	94.4	100.0	100.0	100.0	99.0	100.0
10	10	0	6.6	6.0	5.4	4.4	5.4	6.0	5.6	3.4	4.4	5.8
		0.25	26.0	26.4	21.6	31.0	30.8	92.8	92.2	91.4	98.0	99.0
		0.50	49.4	49.0	44.8	57.2	57.0	98.6	98.6	98.4	99.2	99.8
		0.75	64.4	63.6	60.6	72.6	71.8	99.0	99.0	99.0	99.2	100.0
		1.00	73.2	73.0	70.6	82.4	82.4	99.8	99.8	99.8	99.6	100.0
10	5	0	6.6	6.6	4.8	4.8	4.8	5.6	7.0	3.8	7.2	9.6
		0.25	14.4	13.8	12.0	13.6	13.8	69.0	69.0	66.8	75.6	83.0
		0.50	18.4	18.6	17.0	26.2	26.6	85.8	85.4	84.8	91.0	94.4
		0.75	32.6	32.0	29.6	34.6	34.4	96.4	96.4	95.8	94.6	98.0
		1.00	40.0	39.0	35.6	37.4	38.4	96.0	96.4	96.0	97.2	99.4

Results are reported in percentages.
σ_i^2 refers to the variance component(s) being tested.
(1): Random intercept test,
(2): Simultaneous test for the random intercept and random slope.
B: BLUP based permutation test,
L: Likelihood ratio based permutation test,
A: Asymptotic likelihood ratio test

between the permutation tests and the asymptotic test is small, but this could be attributed to how powerful all three tests are for our simulation settings. When we test for two random effects the asymptotic likelihood ratio test using the 25:50:25 ratio of χ^2 distributions displays similar power to the restricted likelihood ratio based permutation test when the data are binary. For Poisson data the asymptotic test has higher power but this could be due to the test being liberal which can be seen from the null test.

For the simulations in Table 3.1 we only performed omnibus tests where we compare the alternative GLMM to the null GLM. When we compare two GLMMs and include a nuisance random effect we have observed that the permutation tests are liberal. It has been shown that the estimators obtained through the linear mixed model approximation of non-normal data are biased [**?**]. Our hypothesis is that this bias on the nuisance random negatively impacts the permutation tests for the random effect of interest.

3.4.3 Comparison Simulations

In this section we compare our restricted likelihood ratio permutation test to the score tests of Lin [1997] and Hall and Præstgaard [2001]. We compare the size and power of our test to published results from simulations performed by Hall and Præstgaard [2001] which have identical settings to those that were initially performed by Lin [1997]. These simulations are based upon the salamander mating dataset found in McCullagh and Nelder [1989] where two populations of salamanders, rough butt (RB) and whiteside (WS), are mated together in a crossed design. The salamander experiment involves mating ten males and ten females from each of the two populations six times for each salamander resulting in 120 correlated binary observations

of whether or not the mating took place. The experiment was performed a total of three times for a grand total of 360 observations. The question of interest is whether or not there is heterogeneity across males and females. To answer this the following model is used

$$logit\{E(y_{ij}|b_i^f, b_j^m)\} = x_{ij}^T\alpha + b_i^f + b_j^m \quad (i = 1, \ldots, n_f, j = 1, \ldots, n_m),$$

where n_f and n_m are the numbers of female and male salamanders, respectively. The outcome y_{ij} is the binary mating outcome of female i with male j and $x_{ij} = (1, WS_i^f, WS_j^m, WS_{ij}^{fm})^T$ contains the four covariates of interest. The first is the intercept, WS_i^f is an indicator variable for whiteside female (0 = RB , 1 = WS), WS_j^m is an indicator variable for whiteside male (0 = RB , 1 = WS), and WS_{ij}^{fm} is their interaction. The random effects, b_i^f and b_j^m, are assumed to be independent and normally distributed with mean 0 and variances σ_f^2 and σ_m^2, respectively.

For each simulation we generated 3,000 data sets under the following settings. The number of females and males, n_f and n_m, are both set equal to 60. Next, the fixed effects were set equal to their estimates obtained from the original data through restricted maximum likelihood, $\alpha = (1.18, -0.32, -2.84, 3.35)^T$. We test the following null hypothesis: $H_0 \colon \theta = (\sigma_f^2, \sigma_m^2)^T = 0$. We assess the performance of the permutation test as we vary the random effect variances $\sigma_f^2 = \sigma_m^2 \in \{0, 0.25, 0.50, 0.75, 1.00\}$. We compare our restricted likelihood ratio based permutation test to four score tests: T, the score test of Lin [1997], T^* the bias corrected score test of Lin [1997], $\tilde{T}$ the order restricted score test of Hall and Præstgaard [2001], and $\tilde{T}^*$ the bias corrected order restricted score test of Hall and Præstgaard [2001].

From Table 3.2 we see that the restricted likelihood ratio based permutation test

Table 3.2: Estimated test size and power based on simulated data from the salamander dataset

θ	T_3	T	T^*	$\tilde{T}$	$\tilde{T}^*$
0.00	5.4	5.1	4.9	5.0	5.8
0.25	42.1	28.9	31.0	40.1	43.5
0.50	79.5	69.2	71.7	78.8	81.3
0.75	94.8	89.9	91.0	94.6	95.5
1.00	99.0	97.0	97.5	98.4	98.7

Results are reported in percentages.

displays power on par with the most powerful score test which is the small-sample bias-corrected order restricted score test of Hall and Præstgaard. Slight differences can be attributed to simulation variability.

3.5 Examples

3.5.1 Amenorrhea Events from a Clinical Trial of Contracepting Women

In our first example, we analyze data arising from a longitudinal clinical trial of contracepting women. Studies of contraceptive methods are time-to-event studies primarily interested in the time from first use of the contraceptive until discontinuation for any reason. In this particular study women were randomized to receive either a 100 mg or a 150 mg dosage of depot-medroxyprogesterone acetate (DMPA). Three additional identical dosages were given at 90-day intervals and a final follow-up visit occurred 90 days following the fourth dose. One of the major reasons for discontinuations are disturbances in menstrual bleeding. One third of the women dropped out before the completion of the trial, and of those 58.8% reported bleeding pattern disturbances as the reason. To further investigate the impact of the drug dose level

on disrupting menstrual bleeding, each woman completed a menstrual diary that recorded any bleeding pattern disturbances starting from the day of the first dosage. It was concluded that the major difference between the two dose levels was that less amenorrhea, the absence of menstrual bleeding for a specified number of days, was observed in the 100 mg group [World Health Organization, 1987].

The article of Machin et al. [1988] was interested in analyzing the occurrence of amenorrhea in these women. Of the women who participated in the clinical trial, 1,151 had a sufficiently completed menstrual diary. Each woman accounts for a sequence of up to four binary records indicating whether she experienced amenorrhea in each of the the 90 days following a contraception dose with 0 indicating no amenorrhea and 1 indicating that it occurred. In our analysis we investigate if there was indeed a higher incidence of amenorrhea in the 150 mg dose group and if there are any changes in amenorrhea incidence over time. We fit the data using a GLMM with random intercept to account for the repeated subject measurements. The suitability of this model will be checked using a permutation test for the subject level random effects.

We fit the following logistic regression model with a random intercept

$$\text{logit}\{E(Y_{ij}|b_i)\} = \beta_0 + b_{0i} + \beta_1 \text{time}_{ij} + \beta_2 \text{time}_{ij}^2 + \beta_3 \text{dose}_i \times \text{time}_{ij} + \beta_4 \text{dose}_i \times \text{time}_{ij}^2$$

where time $= 1, 2, 3, 4$ indicating one of the four 90-day intervals following a dose of DMPA, and dose equals 1 if assigned to 150 mg DMPA and 0 otherwise. Due to this parameterizing of time there is no information when time $= 0$, and therefore, no main effect of dose is included in this model. We assume that individual women may be heterogeneous in their propensity of developing amenorrhea and is the reason for

the inclusion of the random intercept. The estimate for the variance component for the random intercept is 4.3366 which strongly suggests the necessity of the random intercept. The results of the permutation tests confirm this with p-values of < 0.001 for both the likelihood ratio based and BLUP based permutation tests.

3.5.2 Comparing the Number of Epileptic Seizures between Progabide and Placebo

The second example focuses on a clinical trial of counts of epileptic seizures that took place in 1987 [Leppik et. al.]. This was a placebo-controlled trial of an anti-epileptic drug, progabide, with 59 patients suffering from epilepsy. Progabide is a drug that operates by enhancing the amount of gamma-aminobutyric acid which is the primary inhibitory neurotransmitter in the brain. Each patient was randomly assigned to either progabide or placebo as an adjuvant to the standard anti-epileptic treatment.

The researchers collected baseline data on the number of seizures that occurred for each patient in the eight weeks leading up to the treatment. The outcome of interest was the count of seizures during four two-week intervals following the treatment start date. A plot of the mean number of seizures is provided in Figure 3.1.

The goal of the analysis is to determine whether or not the addition of progabide to standard treatment reduces the rate of seizures and determine the necessity of accounting for within-subject correlation. We fit the following Poisson regression model with a random intercept and slope

$$\log\{E(Y_{ij}|b_i)\} = \log(T_{ij}) + \beta_0 + b_{0i} + (\beta_1 + b_{1i})\text{time}_{ij} + \beta_2\text{trt}_i + \beta_3\text{trt}_i \times \text{time}_{ij},$$

Figure 3.1: Plot of the mean number of seizures per week by treatment group

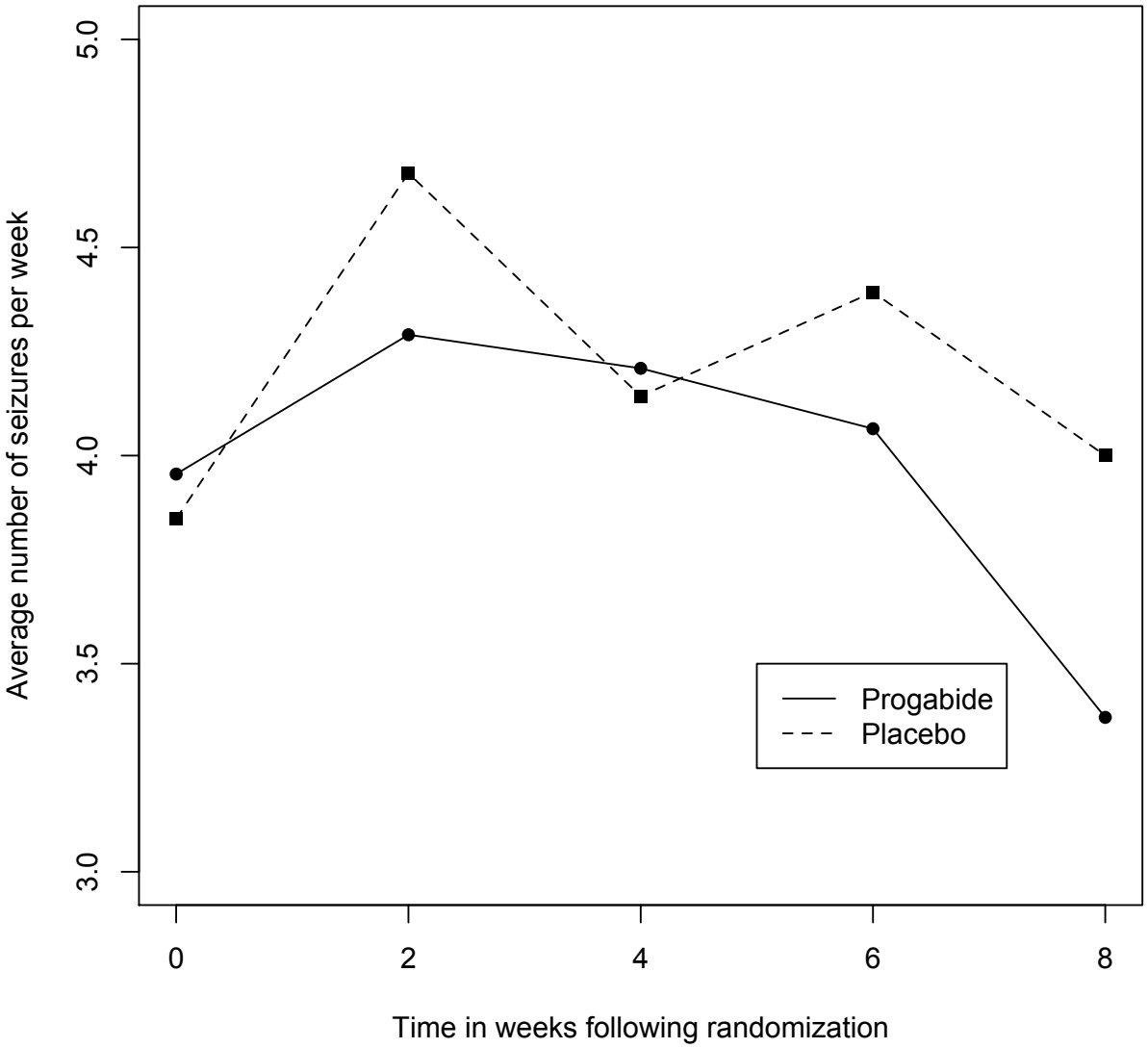
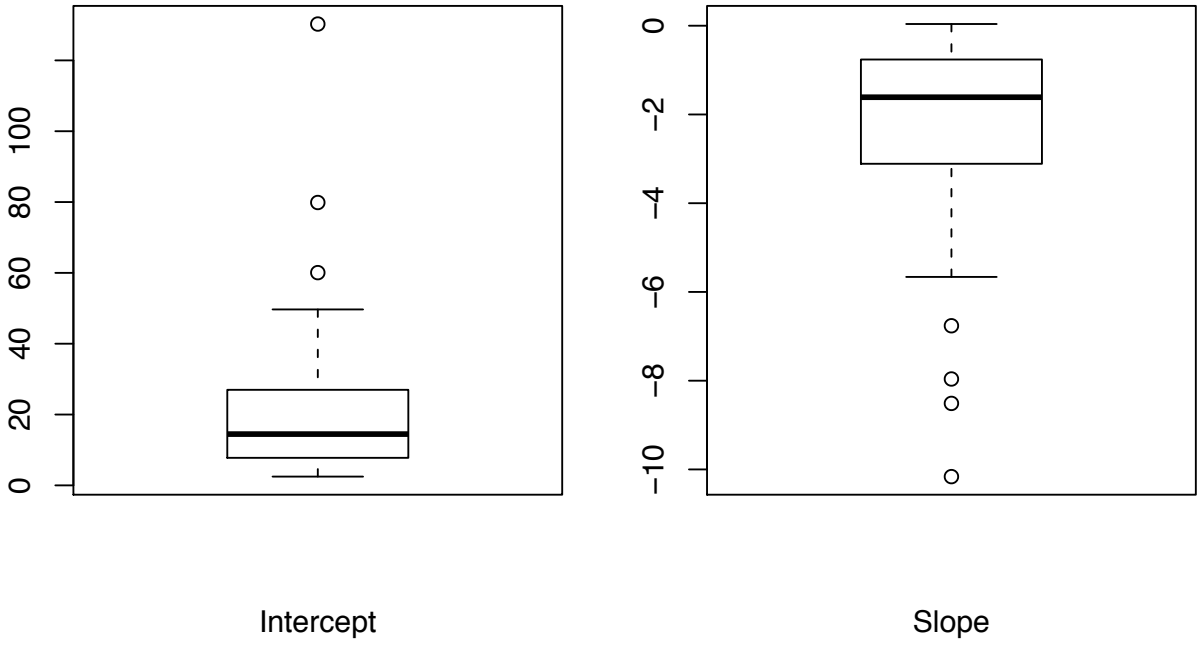


Figure 3.2: Boxplots of the subject specific intercepts and slopes



where Y_{ij} is the number of epileptic seizures for the i-th patient in the j-th 2 week period and $j = 0, 1, 2, 3, 4$. T_{ij} is the length of period j and is equal to 8 when $j = 0$ and equal to 2 when $j \in (1, 2, 3, 4)$ we include an offset $\log(T_{ij})$ to adjust for the difference in the observation time for the baseline seizure count and the post randomization follow-up periods. The treatment variable, trt, is an indicator variable equaling 1 for the progabide group and 0 for the placebo group. Finally, time $\in (0, 2, 4, 6, 8)$ tracks the number of weeks following randomization. The random intercept and slope account for heterogeneity among the patients both at the baseline level and in their response to treatment over time. The two random effects are allowed to be potentially correlated with each other.

Boxplots for subject specific intercepts and slopes were created and can be seen in Figure 3.2. These plots appear to display a significant amount of variation in the subject specific intercepts and slopes and support the decision to include random

effects for the intercept and slope. An earlier boxplot of the subject specific intercepts revealed a potential outlier. This outlier is patient number 49 who had a baseline record of 151 seizures in the 8 weeks leading up to the randomization date and 302 total seizures over the next 8 week which is more than twice as many as the next highest number of total seizures following randomization. As a result we decided to remove patient 49. The boxplots in Figure 3.2 have this observation removed.

Table 3.3: Permutation test results for inclusion of specific random effects when modeling seizure counts.

Test	Observed LRTS	LRT Permutation p-value	BLUP Permutation p-value
(4) vs (1)	151.19	<0.001	-
(3) vs (1)	140.36	<0.001	<0.001
(2) vs (1)	39.68	<0.001	<0.001

(1): No random effects
(2): Random intercept only model
(3): Random slope only model
(4): Random intercept and random slope model

Initial estimates for the variance components were 0.461 and 0.005 for the intercept and slope respectively. While the random slope for time is very small the slope estimate is equal to 0.009 which means that the random slope effect is quite large in comparison. Permutation tests were applied to this model for the global test and each of the individual variance components by themselves. From the results in Table 3.3 we reject all of the null hypotheses and conclude that the random effects in this model are necessary.

3.6 Discussion

In this chapter, we have proposed two methods for performing inference on random effects in GLMMs by linearizing the data in order to fit linear mixed model. After this step, permuting the weighted residuals both within- and among- subjects allows us to test for the inclusion or exclusion of the random effects in an omnibus test. Unfortunately, the permutation tests for individual random effects in the presence of nuisance random effects displayed Type I error rates nearly twice the nominal size. Our hypothesis as to a potential cause of this is the bias in the estimate for the nuisance random effect after linearization. A correction for the bias may be one way to resolve this problem [Lin and Breslow, 1996].

As demonstrated through our simulations and example, when we are examining omnibus hypotheses the proposed permutation tests are valid when the number of patients and the number of observations per patient is small. The performance of the tests is similar to that of the small sample bias-corrected order restricted score test of Hall and Præstgaard. The tests also do not require balanced data nor do the measurements need to occur at the same points in time. As a result, our methods can be applied to omnibus tests of any type of GLMM. Finally, these methods can be applied through standard software and can be incorporated into standard practice for analysis of generalized linear mixed models. While the methods are computationally intensive, the recent rise in parallel computing through clusters and multi-core processors has made it possible to greatly reduce the amount of time necessary to implement these tests.

In the following chapter we generalize the methods presented in Chapter 2 to demonstrate how the permutation tests can be applied to the roughness penalty of a

linear penalized spline model through a mixed model representation of the penalized spline model.

CHAPTER IV

Permutation Tests for Linear Penalized Spline Models

4.1 Introduction

Standard statistical regression methods model an outcome or dependent variable as a function of one or more independent predictor variables. The effect of each predictor variable on the dependent variable is often assumed to be constant or linear. That is, each one unit increase in the predictor variable, X, results in a mean change in the dependent variable, Y, equal to the estimated coefficient for X. However, for some data this assumption of linearity may not hold. When the predictor variable does not affect Y in a constant manner polynomials of X could be used to obtain a better fit of the data. For extremely nonlinear data higher degree polynomials could be utilized but leads to additional curvature that is often not representative of the data as a result of the degrees of the polynomial. Another alternative is to apply transformations to the data. However, the amount of correction feasible through transformations is limited. As an alternative, statisticians often turn to nonparametric smoothing methods broadly termed scatterplot smoothing.

The focus of this article is on one particular type of scatterplot smoothing known

as penalized splines which were first proposed by Eilers and Marx [1996] based on the ideas of O'Sullivan [1986]. A basic linear spline model is a linear combination of spline basis functions that forms a piecewise linear function joined at K pre-specified points known as knots. Examples of spline basis functions include truncated power bases and B-splines. The number and location of the knots have a profound impact on the fit of the linear spline model. When K is small the model may not be flexible enough to properly fit the data, but large values of K can lead to overfitting. Instead of using automated methods of selecting the optimal number and location of the knots, a penalty term is incorporated into the objective function in order to constrain the influence of the K knots. This results in the linear penalized spline model. The effect of the tuning parameter penalty term is that it controls the degree of smoothness. Furthermore, K only needs to be sufficiently large to ensure the desired flexibility in the linear penalized spline model [Ruppert, 2002]. Examples of nonlinear data that can be modeled using penalized splines include pharmacokinetic and pharmacodynamic data [Bonate, 2005]. Pharmacokinetics is the study of the time course of a drug in the human body which is affected by processes such as the metabolism and excretion. The goal of analyzing pharmacokinetic data is to devise a treatment regimen that will maintain a consistent concentration of drug in the body of a patient. Pharmacodynamic analysis deals with modeling the effect of a drug on the body over various concentrations of the drug within the body.

Extensive work has been done to devise methods of estimating the tuning parameter penalty term which include cross validation, Mallows C_p, and Akaikes information criterion. Following Brumback et al. [1999] a penalized spline model can be represented as a linear mixed model (LMM) with the penalized spline coefficients treated as random effects and the penalty parameter parametrized as a function of the vari-

ances of the random error and the random spline coefficients. As a result of this maximum likelihood or restricted maximum likelihood can also be used to estimate the penalty term.

Hypothesis testing has received less attention in the literature. For a linear penalized spline model that is represented as a linear mixed model a hypothesis test on the penalty term is equivalent to testing a standard linear regression fit versus a linear penalized spline alternative. Since the penalty term is a function of the variance of the random penalized spline coefficients the hypothesis test is equivalent to testing if the variance of the random spline coefficients is equal to 0. This hypothesis test is a variance component test.

The difficulty in testing for variance components lies in the fact that the variance component is equal to 0 under the null hypothesis, a value that is on the boundary of the parameter space. As a result, the usual χ^2 asymptotic distributions of the Wald, score, and likelihood ratio test statistics do not hold. Instead, the correct null distribution for the likelihood ratio statistic has been shown to be a mixture of χ^2 distributions [Self and Liang, 1987; Stram and Lee, 1994]. For example, when testing for one random effect, the null distribution becomes a 50:50 mixture of χ^2_q and χ^2_{q-1} distributions where q is the total number of random effects in the alternative model. The score [Silvapulle and Silvapulle, 1995; Verbeke and Molenberghs, 2003] and Wald [Silvapulle, 1992] tests for variance components have been proven to have equivalent mixture χ^2 distributions.

The asymptotic likelihood ratio test for the variance component of the random penalized spline coefficients was investigated by Crainiceanu and Ruppert [2004], and they found that the 50:50 mixture of χ^2_0 and χ^2_1 distributions is a conservative approximation and can lead to severe loss of power. For example, through simulation they

find that when comparing a constant mean versus an alternative piecewise constant spline model with $K = 20$ the χ^2_0 mixing proportion is equal to 0.65 for the restricted likelihood ratio test while for the likelihood ratio test it is essentially a degenerate distribution at 0. Therefore, alternative methods for testing variance components are needed. One such method is presented by Crainiceanu and Ruppert [2004]. The authors apply the spectral decomposition to the likelihood ratio and restricted likelihood tests. A simulation algorithm is then applied to obtain the finite sample null distribution of the likelihood ratio and restricted likelihood ratio tests. Other methods include bootstrap [Kauermann et al., 2009] and a general permutation test [Raz, 1990] designed to test any nonparametric fit of a variable versus no effect at all.

As demonstrated by Raz permutation tests are a viable method for addressing this problem, as permutation tests are known to have nominal size in finite samples while requiring only a few weak assumptions. Our work here is an application of two permutation tests that we have developed specifically for variance component tests in linear mixed models [Lee and Braun, 2012] in order to test a linear regression model against the alternative linear penalized spline model. The two permutation test statistics are a sum of weighted squared residuals, and the empirical null distributions are generated via permutations of the residuals. The first test statistic is based on the Best Linear Unbiased Predictions (BLUPs) [Robinson, 1991] and the second statistic is the restricted likelihood ratio test statistic assuming normality of the data. We will show that our tests have valid size and their powers are comparable to existing methods. In Section 4.2, we begin with notation for linear penalized spline models and show how they can be parametrized as linear mixed models. Section 4.3 follows with a presentation of our proposed methods. We present the results of simulations in Section 4.4 that demonstrate the validity and power of our methods as we vary the

number of observations. In Section 4.5 we apply our methods to data from a study investigating the concentration of ragweed pollen over the course of a season. We close with a discussion of our work in Section 4.6.

4.2 Methods

4.2.1 Linear Penalized Spline Models

Let Y_i be the i-th observation for $i = 1, 2, ..., N$, and x_i is the independent variable measured for observation i. A standard linear spline model can be written as

$$Y_i = \beta_0 + \beta_1 x_i + \sum_{k=2}^{K+1} \beta_k (x_i - \kappa_k)_+ + \epsilon_i, \tag{4.1}$$

where there are K linear spline basis functions written as $(x_i - \kappa_k)_+$. The subscript $+$ indicates that negative values of the spline basis function are set equal to 0, where each of the κ terms is fixed and known as a knot. Other spline bases such as truncated power functions, B-splines, and cubic splines can be used as well. In this model there are a total of $K + 2$ coefficients where β_0 and β_1 are the standard fixed effect coefficients for the intercept and slope, and $\beta_2, \beta_3, ...\beta_{K+1}$ are the spline coefficients. The random errors, ϵ_i, are independent and identically distributed normal with mean 0 and variance, σ^2_ϵ. Equivalently, we can write the linear spline model using matrix notation, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}$ is a $(K + 2)$-dimensional vector of the fixed effect coefficients, and $\boldsymbol{X}$ is the design matrix containing a column of ones, x_i, and all of the linear spline basis functions. $\boldsymbol{\epsilon}$ is the vector of the random errors, ϵ_i.

Estimations of the parameters of a linear spline model can be obtained through

ordinary least squares by minimizing the objective function,

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2. \tag{4.2}$$

This fit results in a piecewise linear model joined at each knot. The adequacy of the fit is sensitive to the location and number of specified knots. When the number of knots is low the fit can be poor, while too many knots can lead to overfitting where some of the effects in the model are due to random noise in the data.

To control overfitting, a roughness penalty is applied. The penalty constrains or shrinks each of the spline coefficients, $\beta_2, \beta_3, ...\beta_{K+1}$ towards 0. The linear penalized spline model does this by adding the roughness penalty to the objective function. Instead of minimizing (4.2) the penalized spline model minimizes

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda^2 \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{D} \boldsymbol{\beta}, \tag{4.3}$$

where $\lambda \geq 0$. The $\lambda^2 \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{D} \boldsymbol{\beta}$ term is the roughness penalty, where $\boldsymbol{D}$ is a $(K+2) \times (K+2)$ matrix with $D_{1,1} = 0$, $D_{2,2} = 0$, and each of the remaining diagonal terms equal to 1. The off-diagonal elements are all equal to 0. This forces the penalty to only be applied to the spline coefficients, and in turn the estimates for the spline coefficients are ridge regression estimators. The amount of smoothing is controlled by λ. When λ is equal to 0 then the penalty is not applied and the model reduces to the piecewise linear model. As λ goes to infinity, which is the test of interest, the spline coefficients are shrunk all the way to 0 which results in a standard linear regression model. Therefore the choice of λ is crucial, while the number of knots only needs to be large enough to provide the desired level of flexibility in the model [Ruppert, 2002].

Once the number of knots is selected they are located at evenly spaced quantiles of x_i [Ruppert, 2002]. Choosing λ can be done through different selection methods such as cross-validation, Mallows' C_p, and the Akaike information criterion. In the next subsection we show how maximum likelihood or restricted maximum likelihood can be used to estimate λ for linear penalized spline models.

4.2.2 Representing a Linear Penalized Spline Model as a Linear Mixed Model

As demonstrated by Brumback et al. [1999] and Ruppert et al. [2003] a linear penalized spline model can be expressed as a linear mixed model. Starting with the linear penalized spline model written as in (4.1), two design matrices are be defined. $\boldsymbol{X}$ is a $N \times 2$ matrix containing a column of ones and x_i. $\boldsymbol{Z}$ is a $N \times K$ matrix containing the K spline basis functions. Furthermore, the coefficients are also partitioned into two vectors–$\boldsymbol{\beta}^* = \{\beta_0, \beta_1\}$ and $\boldsymbol{b} = \{\beta_2, ..., \beta_{K+1}\}$. Subsequently, the objective function now becomes

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{Z}\boldsymbol{b}\|^2 + \lambda^2 \boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{b}. \tag{4.4}$$

Next, divide (4.4) by the error variance, σ_ϵ^2.

$$\frac{1}{\sigma_\epsilon^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{Z}\boldsymbol{b}\|^2 + \frac{\lambda^2}{\sigma_\epsilon^2}\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{b} \tag{4.5}$$

If $\boldsymbol{b}$ is treated as a random effect that follows a normal distribution with mean 0 and variance $\sigma_b^2 = \sigma_\epsilon^2/\lambda^2$, and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent, then (4.5) is the objective function of a linear mixed model derived from the log likelihood conditional on $\boldsymbol{b}$. As a result, the linear penalized spline model can now be written as the following linear mixed

model, $\boldsymbol{Y} = \boldsymbol{X\beta}^* + \boldsymbol{Zb} + \boldsymbol{\epsilon}$, where

$$Var \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \boldsymbol{R} \end{bmatrix}$$

with $\boldsymbol{G} = \sigma_b^2 \boldsymbol{I}$ and $\boldsymbol{R} = \sigma_\epsilon^2 \boldsymbol{I}$.

Estimation of $\boldsymbol{\beta}$, σ_ϵ^2, and σ_b^2 can be done through maximum likelihood or restricted maximum likelihood. Predictions for the the penalized spline coefficients, $b_1, ..., b_K$ are denoted by $\tilde{b}_1, ..., \tilde{b}_K$ or $\tilde{\boldsymbol{b}}$ and are obtained using best linear unbiased prediction or BLUP

$$\tilde{\boldsymbol{b}} = \boldsymbol{GZ^T V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{4.6}$$

where $\boldsymbol{V} \equiv \text{cov}(\boldsymbol{Y}) = \boldsymbol{ZGZ}^T + \boldsymbol{R}$.

Once the linear penalized spline model has been re-expressed as a linear mixed model, we can utilize LMM methods to test for the necessity of the penalized spline model opposed to a standard linear regression model. For example, we would be testing

$$H_0 : Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.7}$$

$$H_1 : Y_i = \beta_0 + \beta_1 x_i + \sum_{k=2}^{K+1} \beta_k (x_i - \kappa_k)_+ + \epsilon_i. \tag{4.8}$$

Given that the spline coefficients are random effects from a normal distribution with mean 0 and variance, σ_b^2, this hypothesis test is equivalent to testing that $\sigma_b^2 = 0$. As noted previously, estimation of the parameters of a LMM is typically done through maximum likelihood or restricted maximum likelihood. Asymptotically, the maximum likelihood and REML estimators are equivalent, but for small sample sizes, the

REML estimator is expected to be less biased than the maximum likelihood estimator [Ruppert et al., 2003]. In addition, a comprehensive simulation study performed by Morrell [1998] found that the asymptotic likelihood ratio test based on the REML estimates are closer to nominal than test statistics utilizing the ML estimates. Therefore, in our proposed methods we used the REML estimators.

4.3 Proposed Methods

4.3.1 Permutation Tests

As alternatives to the asymptotic likelihood ratio test for variance components we propose two permutation tests [Lee and Braun, 2012]. Permutation tests are nonparametric tests that have nominal size when performed correctly. They operate by generating an empirical null distribution for an observed test statistic through permutations of the data. Permutation tests assume that the values being permuted are exchangeable under the null hypothesis [Good, 2005]. A vector, $\boldsymbol{Y}$, is exchangeable if, for any permutation of $\boldsymbol{Y}$ denoted as $\boldsymbol{Y}^*$, $\boldsymbol{Y}^*$ has the same distribution as $\boldsymbol{Y}$ [Commenges, 2003]. It should be noted that exchangeability is a weaker condition than independent and identically distributed. When it is unfeasible to enumerate all possible permutations, an approximate permutation distribution can be generated through Monte Carlo sampling [Dwass, 1957].

Both proposed permutation tests are based upon permuting the marginal errors, $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$. Under the null hypothesis of penalized spline functions, the $\boldsymbol{\epsilon}$ are exchangeable, and more specifically, independent and identically normally distributed with mean 0 and variance σ^2_ϵ. In practice, the errors are estimated by the residuals, $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, calculated from estimates fit from the alternative model, and Schmoyer

[1994] showed that the residuals are also asymptotically exchangeable.

When the null hypothesis includes nuisance variance components such as random effects or penalized spline functions for a different independent variable the marginal errors are no longer identically distributed under the null hypothesis. Instead, the errors follow a normal distribution with mean $\mathbf{0}$ and covariance matrix, $\boldsymbol{V_0} = \boldsymbol{Z}\boldsymbol{G}_0\boldsymbol{Z^T} + \boldsymbol{R_0}$ with $\boldsymbol{R_0} = \sigma^2_{\epsilon_0}\boldsymbol{I}$ where $\sigma^2_{\epsilon_0}$ is the variance of the random errors under the null model, and $\boldsymbol{G}_0$ is the covariance matrix of the nuisance random effects. This problem is resolved by weighting the errors by $(\boldsymbol{U}_0^T)^{-1}$ where $\boldsymbol{U}_0$ is the Cholesky decomposition of $\boldsymbol{V}_0$. The set of weighted errors, $(\boldsymbol{U}_0^T)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, are distributed $MVN(\mathbf{0}, \boldsymbol{I})$ under the null hypothesis and can be permuted which is extended to the residuals.

The first permutation test is based on the sum of the squared BLUPs or the penalized spline coefficients and utilizes the following test statistic:

$$T_2 = \sum_{i=1}^{K} \tilde{b}^2_{i2}/K = \tilde{\boldsymbol{b}}^{*T}\tilde{\boldsymbol{b}}^*/K,, \tag{4.9}$$

where $\boldsymbol{b}^* = \hat{\boldsymbol{G}}_1\boldsymbol{Z}\hat{\boldsymbol{V}}_1^{-1}\boldsymbol{U}_0^T(\boldsymbol{U}_0^T)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. This is the BLUP formula with the weighted residuals and an unweighting matrix. For this test statistic we sum over the K squared spline coefficients because these are the random coefficients that are normally distributed with mean zero and variance equal to $\hat{\sigma}^2_b$. For the observed data this quantity is the sample variance of the BLUPs for the random spline coefficients. Note that the denominator of the test statistic is constant for all of the permutations and does not affect the validity or power of our test.

To obtain the permutation null distribution we use Monte Carlo sampling to randomly permute the weighted marginal residuals. Next, we re-estimate $\hat{\sigma}^2_b$ for each

permutation before computing T^*_{2m}, the value of T_2 computed from permutation m of the data, $m = 1, 2, ..., 1000$, to generate the approximate empirical null distribution of T_2. The re-estimation is performed because certain permutations of the residuals will result in estimates of $\hat{\sigma}^2_b$ equal to zero. We then calculate the percentage of permutations with T^*_2 greater than T_2 to generate a p-value. T_2 is only calculated for the single variance component being tested and cannot be used for tests of multiple variance components at the same time.

Our second permutation test is able to overcome this limitation. It is based on the restricted likelihood ratio test statistic, $\phi = -2\log(L_{H_0} - L_{H_1})$, where L_{H_0} and L_{H_1} are the restricted likelihoods under the null and alternative hypotheses, respectively. Our test statistic is

$$T_3 = \log\left[|\hat{\boldsymbol{V}}_{\boldsymbol{0}}|/|\hat{\boldsymbol{V}}_{\boldsymbol{1}}|\right] + \hat{\boldsymbol{e}}^T(\hat{\boldsymbol{V}}_{\boldsymbol{0}}{}^{-1} - \hat{\boldsymbol{V}}_{\boldsymbol{1}}{}^{-1})\hat{\boldsymbol{e}} - \log\left[|\boldsymbol{X}^T\hat{\boldsymbol{V}}_{\boldsymbol{0}}{}^{-1}\boldsymbol{X}|/|\boldsymbol{X}^T\hat{\boldsymbol{V}}_{\boldsymbol{1}}{}^{-1}\boldsymbol{X}|\right]. \quad (4.10)$$

Again, the marginal residuals are weighted prior to permutation. Following permutation, the unweighted permuted residuals are used to obtain permutation specific estimates of $\hat{\boldsymbol{V}}_{\boldsymbol{0}}{}^*$ and $\hat{\boldsymbol{V}}_{\boldsymbol{1}}{}^*$. When simultaneously testing for multiple variance components, re-estimation of $\hat{\boldsymbol{V}}_{\boldsymbol{0}}{}^*$ and $\hat{\boldsymbol{V}}_{\boldsymbol{1}}{}^*$ is necessary due to the changes that occur in the rank of $\hat{\boldsymbol{\Sigma}}$ when some number of the variance components of interest are estimated to be equal to 0. By taking this into account, the permutation distribution is allowed to "mix" as the rank of $\hat{\boldsymbol{G}}_1$ varies, thereby generating a distribution similar to the mixture χ^2 asymptotic distribution of Stram and Lee [1994]. Simulation is the only other method of obtaining the correct mixture distribution. We create the permutation distribution by calculating T^*_3 for each of the random permutations and determine a p-value by comparing the observed log restricted likelihood ratio statistic

to the permutation distribution.

4.4 Simulation Studies

4.4.1 Validity

We examine the validity of our permutation tests on linear penalized spline models through a series of simulations while varying the number of observations $N \in 30, 50, 100$. A total of one thousand data sets were generated from the the following linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

with $\beta_0 = 3$, $\beta_1 = 2.75$, $\sigma^2_\epsilon = 1$, and our fixed effect, x_i, are equally spaced points on the interval (0,1]: $x_i = i/N$ where $i = 1, ..., N$. The alternative model that we are testing is a penalized linear spline model with linear basis splines at 10 equally spaced knots over the range $\min(x_i)$ and $\max(x_i)$.

All estimates were performed in the statistical package R using the `amer()` function from the R-package `amer` [Scheilpl, 2011]. The `amer()` function leverages the mixed model fitting function `lmer()` from the R-package `lme4` [Bates et al., 2011] in order to estimate the spline coefficients and fit a linear penalized spline model.

To contrast against the two proposed permutation tests we also incorporate the mixture χ^2 asymptotic likelihood ratio test as well as the alternative permutation method developed by Raz [1990]. The inclusion of the asymptotic likelihood ratio test serves as an example of how poorly this approximation performs when testing for a penalized spline alternative. Raz's permutation test was developed to assess any effect of a variable when it is used in a nonparametric procedure to estimate a

smooth function. This method can be applied to several nonparameteric regression procedures such as kernel smoothing, local regression, and smoothing splines. The test statistic, R, is a ratio of sums of squares and is similar to an F-type of statistic.

$$R = (N-1)Q_1/Q_3$$

For this test statistic $Q_1 = \sum_i [\hat{f}(x_i) - \bar{Y}]^2$ and $Q_3 = \sum_i [Y_i - \bar{Y}]^2$, where $\hat{f}(x_i)$ is the nonparametric estimate for Y_i. The permutation distribution for this test statistic is obtained by permuting x_i among all of the N responses. Raz also provides a method of approximating the permutation distribution using a gamma distribution, but through parallel computing we will avoid using the approximation and instead generate 1000 Monte Carlo permutations for each simulation. One consequence of applying Raz's permutation test is that it tests for any effect of x_i on the independent variable Y_i which includes the fixed slope effect in our linear penalized spline model. Therefore, in order to make use of Raz's method for just the penalized spline portion of the model we first subtract the estimated intercept and slope of x_i from Y_i before permuting x_i for the penalized spline terms.

The simulation results for validity are presented in Table 4.1. Both the likelihood ratio based permutation test and Raz's permutation test appear to be valid with Type I error rates that lie within the 95% confidence bounds for 0.05 which are $(0.036, 0.064)$. The BLUP based permutation test is conservative when $N = 30$ but is valid as N increases. The asymptotic likelihood ratio test is quite conservative for all three settings.

4.4.2 Power

Simulations to examine the power of the permutation tests were also performed and again are compared to the asymptotic likelihood ratio test and Raz's permutation method. We generated 500 data sets from a sine curve with some random noise

$$Y_i = C\sin(x_i) + \epsilon_i,$$

where x_i are equally spaced points on the interval $(0, 4\pi] : x_i = 4i\pi/N$ where $i = 1, ..., N$ and $\sigma^2_\epsilon = 1$. C is a scalar that controls the amplitude of the sine curve and was varied from 0.5, 1, and 2. Small values of C lead to a reduction in the amount of curvature in the data relative to the random noise making a penalized spline model less necessary. Once again, we varied the sample size $N \in \{30, 50, 100\}$. The alternative model that we are testing is a penalized linear spline model with linear basis splines at 10 knots that are equally spaced over the range $\min(x_i)$ and $\max(x_i)$.

Table 4.1 contains the results of the power simulations. Overall the performance of the BLUP based permutation test and Raz's permutation test are very similar with slight advantages for the BLUP based permutation test in three of the settings. The likelihood ratio based permutation test is not as powerful as either of the other two permutation tests. Again, the 50:50 mixture of χ^2_0 and χ^2_1 asymptotic null distribution for the asymptotic likelihood ratio tests is conservative in all but the most extreme settings.

Table 4.1: Size and power for the permutation tests compared to the asymptotic likelihood ratio test and Raz's method

		Method			
N	C	B	L	R	A
30	0.0	2.8	4.5	4.2	2.3
	0.5	11.7	9.7	12.3	5.8
	1.0	34.2	26.7	34.0	21.6
	2.0	81.4	81.2	78.4	80.8
50	0.0	4.2	5.4	5.0	2.7
	0.5	20.1	16.3	20.0	10.5
	1.0	63.6	57.6	61.9	55.2
	2.0	99.8	99.8	99.6	99.8
100	0.0	5.2	6.1	5.7	2.7
	0.5	43.3	35.4	39.7	29.8
	1.0	95.7	95.6	95.7	95.5
	2.0	100	100	100	100

Results are reported in percentages.
B: BLUP based permutation test,
L: Likelihood ratio based permutation test,
R: Raz's permutation test,
A: Asymptotic likelihood ratio test.

4.5 Application

We apply our permutation tests to a set of data presented by Stark et al. [1997] who modeled ragweed pollen levels collected in Kalamazoo, MI from 1991 to 1994 as a function of meteorological data. Pollen measurements were obtained 7 days a week during this four year period on the roof of a local television station. Ragweed pollen can cause hay fever and asthma in sensitive people. One method of controlling pollen-induced allergies is through avoidance. Accurate predictions of pollen levels would benefit people seeking to avoid being exposed to pollen. The authors found that the most significant variables for predicting ragweed pollen counts were an indicator of whether or not there was significant rainfall in the late morning with 3 hours of

steady rain or brief but intense rain defined as significant rainfall, wind speed in knots, average daily temperature in degrees Fahrenheit, and the day number in the ragweed pollen season. Given that there are seasonal effects the day number in the ragweed pollen season may not affect the ragweed pollen count linearly. Wind speed has been previously shown to influence pollen dispersal in a linear fashion. However, the effects of temperature and the day number of the pollen season may not be linear. It is reasonable to believe that as the pollen season progresses in days the pollen levels increases to a peak and then declines as the days in season continues to increase. Regarding the effect of temperature, the same temperature during the peak pollen season will probably have a different effect than towards the end of the season.

We utilized penalized splines to fit the three continuous variables and used the proposed permutation tests to determine if the penalized splines are necessary. First, in order to make the pollen counts more normally distributed we transformed the ragweed pollen counts by taking square root of those values. Figure 1 contains preliminary plots of the three continuous variables with a loess curve fit in order to show the potential deviation from linearity. Neither wind speed nor temperature appear to deviate too much from linearity while the day in season variable appears to be very nonlinear.

Each continuous variable was modeled using penalized splines with 20 evenly spaced knots. Our initial full model is

$$\begin{aligned}\sqrt{Y_i} &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \sum_{k=1}^{20} \beta_{2k}(X_{2i} - \kappa_{2k})_+ + \beta_3 X_{3i} + \sum_{k=1}^{20} \beta_{3k}(X_{3i} - \kappa_{3k})_+ \\ &\quad + \beta_4 X_{4i} + \sum_{k=1}^{20} \beta_{4k}(X_{4i} - \kappa_{4k})_+ + \epsilon_{ij}.\end{aligned}$$

Figure 4.1: Plot of square root pollen counts against wind speed.

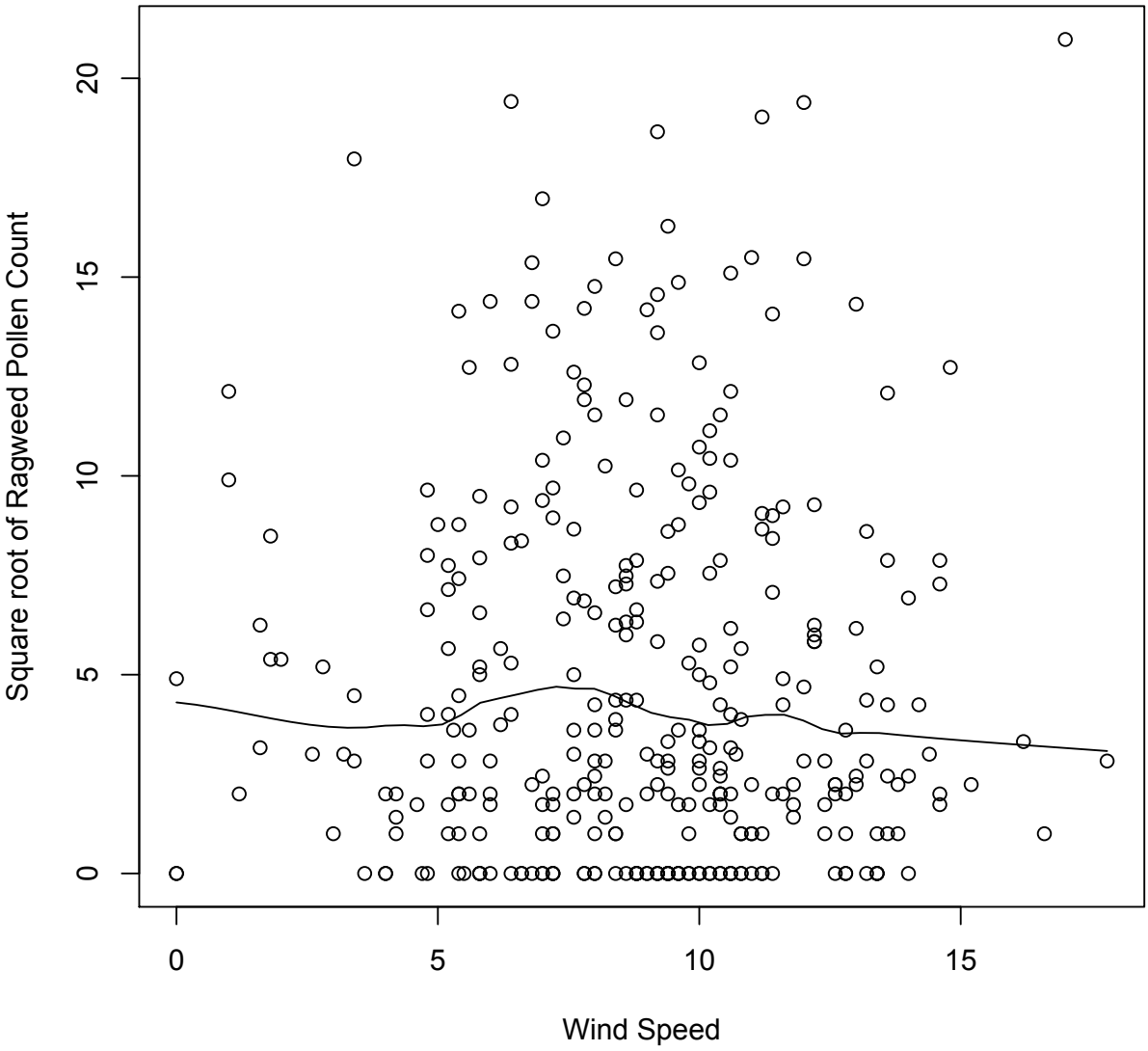


Figure 4.2: Plot of square root pollen counts against temperature.

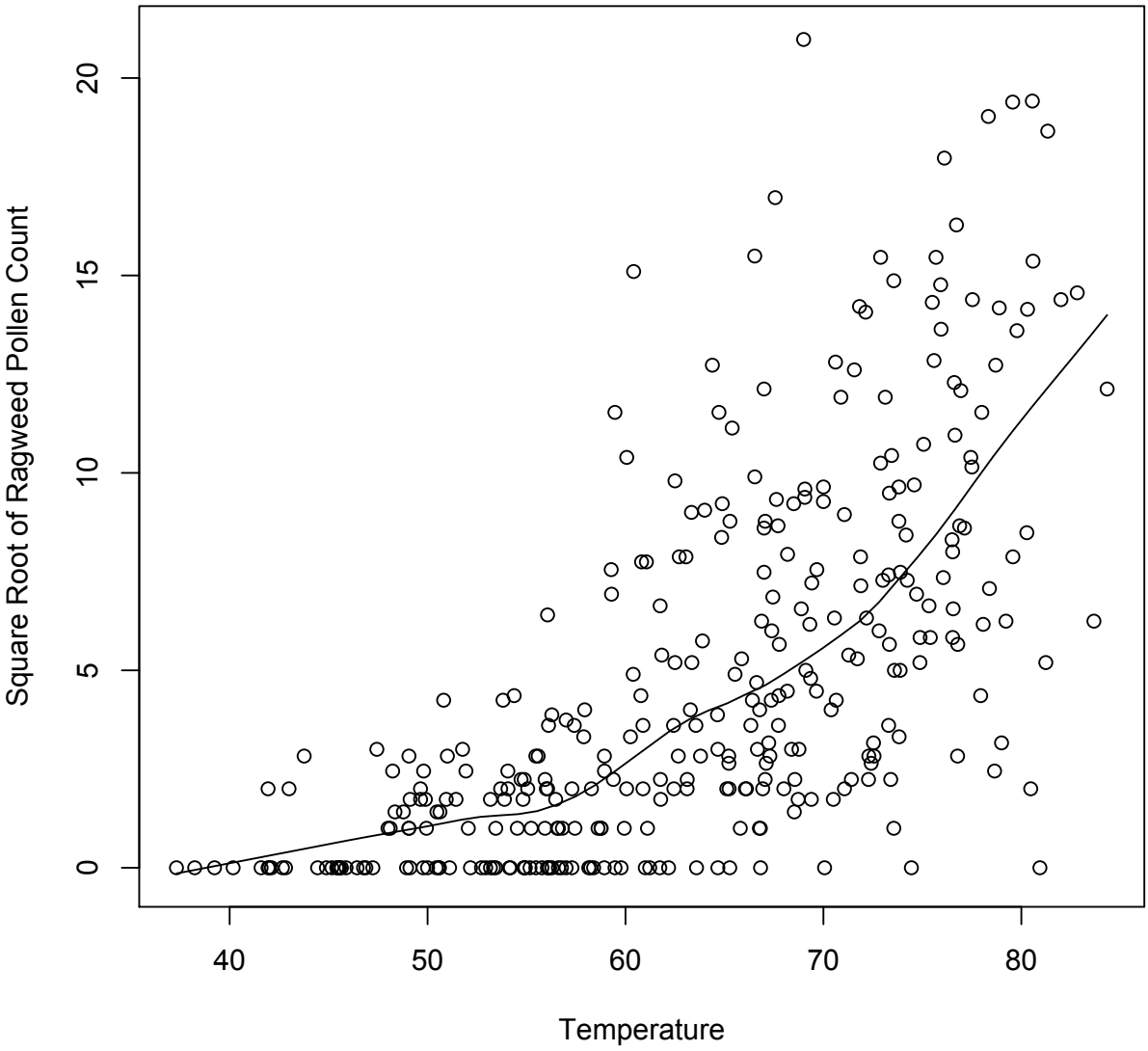


Figure 4.3: Plot of square root pollen counts against day in season.

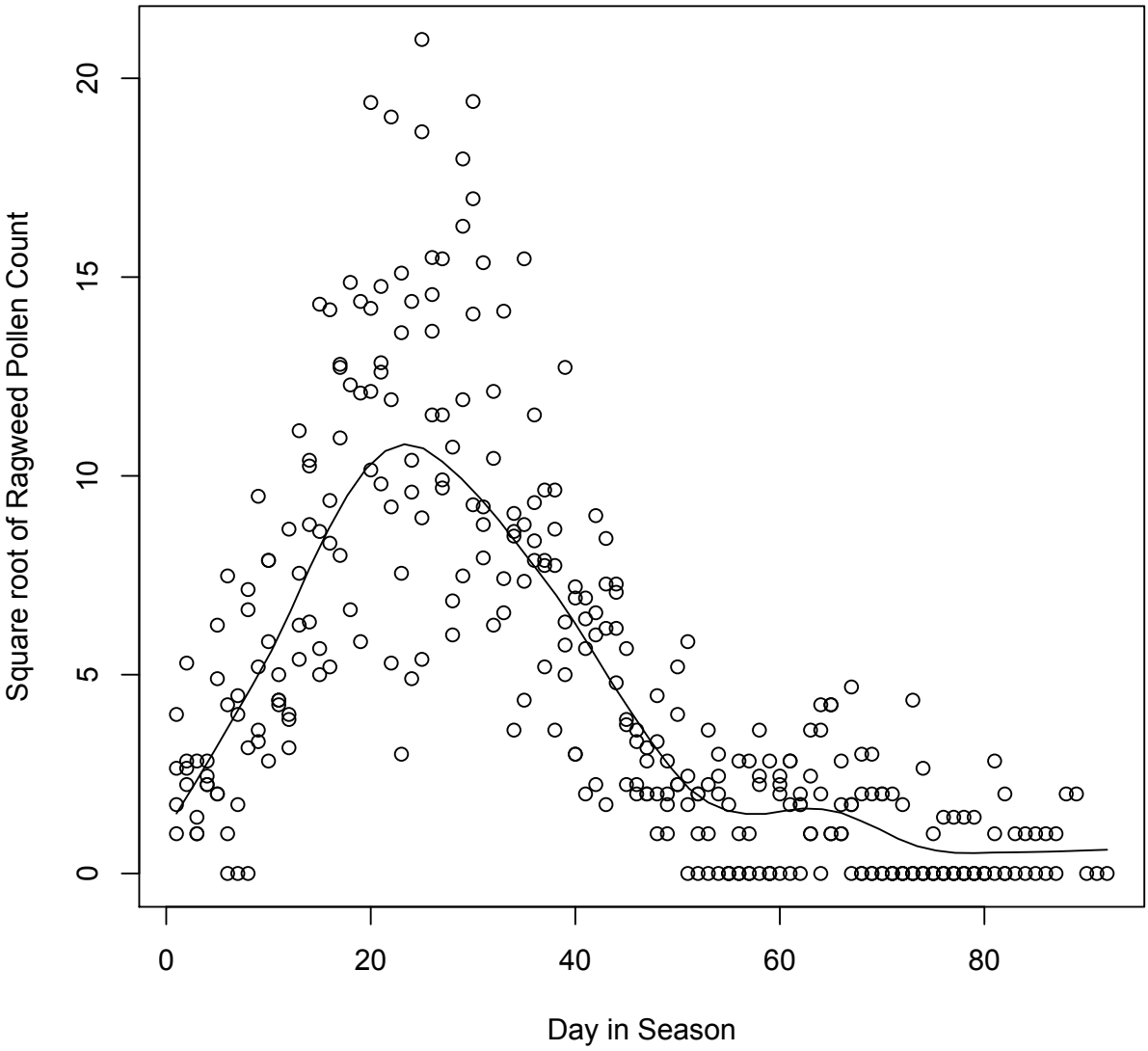


Table 4.2: Permutation test results the penalized spline terms

Test	Observed LRTS	Permutation P-value
(3) vs (1)	296.16	<0.001
(3) vs (2)	11.64	<0.001

(1): No penalized splines
(2): Penalized splines for day in season
(3): Penalized splines for temperature and day in season

Y_i is the pollen level for the i-th measurement, and X_{1i}, X_{2i}, X_{3i}, and X_{4i} are the rain, wind speed, temperature, and day in season variables, respectively, for measurement i. This model is written as a linear mixed model with $\boldsymbol{b}_2 \sim N(0, \sigma^2_{b_2})$, $\boldsymbol{b}_3 \sim N(0, \sigma^2_{b_3})$, and $\boldsymbol{b}_4 \sim N(0, \sigma^2_{b_4})$. We wish test if any or all of these variance components is equal to 0. After fitting the initial model the estimated variance for the random spline coefficients for wind speed is equal to 0. Therefore, the penalized spline terms for wind speed were removed resulting in the following model

$$\sqrt{Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \sum_{k=1}^{20} \beta_{3k}(X_{3i} - \kappa_{3k})_+ + \beta_4 X_{4i} + \sum_{k=1}^{20} \beta_{4k}(X_{4i} - \kappa_{4k})_+ + \epsilon_{ij}.$$

Table 3 shows the results of our hypothesis tests based on 1000 permutations for the variance components of the penalized spline terms using the likelihood ratio based permutation test. First, we performed the omnibus hypothesis test to simultaneously test both temperature and day in season. This test was highly significant so at least one of the two penalized spline terms is necessary. The day in season variable shows a very strong nonlinear marginal relationship with the square root of the pollen counts. Therefore, we then test for the necessity of nonparametrically modeling temperature

given that day in season is already modeled with penalized splines. The hypothesis test result is again signifiant, and we must reject the null hypothesis and include both penalized spline terms. For completeness, we also calculated the BLUP based permutation test for the second hypothesis test. This test results in $\sum \tilde{b}_0^2 = 1.43$ and a p-value equal to 0.021 supporting the conclusion that we must reject the null hypothesis.

4.6 Discussion

In this paper, we have proposed two methods based on permutations which can be used to test a linear regression model against the alternative of a linear penalized spline model. The permutation tests were originally designed for variance component tests but can be applied to the variance component of the random penalized spline coefficients when the linear penalized spline model is represented as a linear mixed model and estimated using restricted maximum likelihood. The permutation tests are based on permuting the weighted residuals and both proposed tests can be applied to hypothesis tests for a single variance component. For hypothesis tests that include multiple variance components simultaneous the likelihood ratio based permutation test is required, and the only other alternative is through simulation. As was previously reported and replicated in our simulations the asymptotic mixture of χ^2 null distribution for the likelihood ratio test statistic is very conservative under this setting.

The proposed permutation tests perform well even when the number of observations is small. We have shown that these tests have nominal Type I error rates, and that their power are equivalent to the alternative permutation test of Raz. We have

shown that the tests can be used for continuous data modeled with a linear penalized spline models, but the methods will also work for generalized linear mixed models (GLMMs) as well. The approach for GLMMs is based upon a first-order approximation of the GLMM to make it resemble the form of a LMM, an approach that is the foundation of penalized quasi-likelihood (PQL) [Breslow and Clayton, 1993] for estimation in GLMMs.

Implementing these permutation tests is straightforward and can be incorporated into standard practice for fitting linear penalized spline models using maximum likelihood. While the methods are computationally intensive, the recent rise in parallel computing through clusters and multi-core processors has made it possible to greatly reduce the amount of time necessary to implement these tests.

The permutation tests that we have presented in this article are not limited only to hypothesis tests for linear penalized spline models. The tests can be applied to any type of penalized regression method that is equivalent to a mixed model such as smoothing splines [Liu and Wang, 2002].

Throughout this paper, we have assumed that the errors are normally distributed. We have previously examined the performance of the permutation tests when the assumption of normality of the errors is violated and found that the permutation tests appear to be robust enough to continue to work well, but a more detailed examination of the behavior of the tests is necessary. In particular violations that cause poor estimates of the variance components could potentially affect the permutation test and warrants further investigation.

CHAPTER V

Discussion

5.1 Closing

In this dissertation we have developed new methods of performing inference on the random effects of mixed effects models. Through permutations of the weighted marginal residuals, our tests can be applied to any type of mixed model, and the residuals can be permuted both between and among the subjects. The tests can handle single random effect tests and simultaneous tests of multiple random effects. Through simulations we have shown that the tests are valid, more powerful, and more robust to violations in the distribution assumptions of the random effects and random error than the asymptotic likelihood ratio test.

We have demonstrated how the LMM random effect tests can be applied to GLMMs using a Taylor expansion to approximate the GLMMs. We find in this setting that the permutation tests perform well when comparing a GLMM to a GLM, i.e. no random effects, but the tests do not have the correct size when comparing two nested GLMMs. However, existing methods also cannot deal with nuisance random effects.

Finally, we demonstrate how our random effect test can be applied to nonlinear data fit using penalized linear splines where we test the spline alternative against the linear regression model. The asymptotic likelihood ratio test is conservative in this setting, but the permutation tests are valid. The power of our tests are comparable to the permutation test of Raz [1990].

5.2 Commonly Asked Questions

After several presentations of this work to others, a number of questions have been repeatedly asked. First, it has been asked how the performance of our permutation test compares to that of the bootstrap. Good [2005] has shown that the permutation test is generally more powerful than the nonparametric bootstrap in many settings, but empirical support for this conclusion specifically in variance component tests does not yet exist. For the parametric bootstrap, the linear mixed model is first fit under the null hypothesis, and then new data is simulated from the estimated linear model and these data are then used to calculate test statistics that form an empirical null distribution. Simulations that were performed under the linear mixed model setting found that the results from the parametric bootstrap were nearly identical to those of the permutation test.

However, because the parametric bootstrap generates data in conjunction with an assumed parametric model, the validity of the parametric bootstrap is questionable if the assumptions in that model are violated. Then the distribution from which the bootstrap data are generated will be incorrect. The permutation test is only affected if the residuals are not asymptotically exchangeable and therefore more robust to model misspecification than the parametric bootstrap.

Finally, the assumptions required for a general permutation test is that the errors are exchangeable under the null hypothesis. This means that there is a possibility that we can relax the assumptions of normality on the random effects and the random errors. This would lead into the realm of nonparametric regression methods for fitting mixed models and will be further elaborated upon in a subsequent sub-section.

Second, others have suggested alternative test statistics for the permutation tests. Examples include $\hat{\sigma}_b^2/\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_b^2$, the restricted maximum likelihood estimate of the random effect variance. We examined simulations that used $\hat{\sigma}_b^2$ as a test statistic and found that its size and power were, unsurprisingly, very similar to those of T_2. Perhaps gains in power could be attained through a test statistic that we have not yet explored, although we suspect that any gain in power would be very modest at best.

Third, it has been noted by others that sometimes with correlated data when the fixed effects are of primary interest, statisticians will include random effects without performing any inference on those random effects. There is extensive literature on the dangers of ignoring correlation within data and its effects on inference for the fixed effects. By including random effects, the hope is that these pitfalls will be avoided. However, very little work exists on understanding the potential consequences of incorporating unnecessary random effects into a model beyond the loss in efficiency due to estimating additional parameters. A more comprehensive study of the impact of including unnecessary random effects would be interesting and could convince more statisticians to incorporate random effect hypothesis tests into their standard procedures when building mixed models.

5.3 Further Research

5.3.1 Simultaneously Testing for Multiple Random Effects with T_2

As in Chapter 2, the statistic derived $T_2 = \sum_{i=1}^{N} \tilde{b}_{i2}^2/N$ can be interpreted as the sample variance of the random effect of interest with weights to account for nuisance random effects in the permutations. In the preceding chapters we concluded that T_2 is limited to tests of single random effects as it is a sum of BLUPs for the variance component of interest. Subsequently, we developed the restricted likelihood ratio based permutation test to handle testing for multiple random effects. However, we believe that T_2 may still be suitable for a hypothesis test comparing multiple random effects. This can be done when the random effects of interest are assumed to be independent of one another or are transformed so that each of the transformed random effects are independent.

Consider each random effect being tested as a random variable following a normal distribution with mean 0 and variance σ_k^2 where $k = 1, 2, 3, ..., K$ is the index for the K random effects of interest and each random effect is independent of the others. When we test for a single random effect, the BLUPs for the random effect of interest are treated as a random sample from the random effect distribution. If we now scale the BLUPs for random effect k by σ_k, the distribution for each set of BLUPs is now a standard normal distribution. Scaled BLUPs from all of the random effects of interest can then be incorporated into a test statistic. $T_{2b} = \dfrac{\sum_{k=1}^{K}(\sum_{i=1}^{N}(\tilde{b}_{ik}^2/\hat{\sigma}_k))}{Nk}$, as our BLUP based test statistic where again N is the number of subjects. A large value for T_{2b} would be evidence against the null hypothesis that all of the random effects of interest are equal to zero.

If the random effects were assumed to be correlated with one another, we could use a Cholesky decomposition of the covariance matrix of the K random effects in a manner very similar to Chapter 2 to create BLUPs that are iid N(0,1). The transformed random effects and the associated BLUPs could then be used to calculate T_{2b}. However, an examination of the validity of this test still needs to be performed.

5.3.2 Permutation Based Confidence Intervals for Random Effect Variance Components

After a significant hypothesis test occurs, it is natural to compute a confidence interval to provide a range of plausible values for the parameter of interest. To generate a permutation based confidence interval for a location parameter, a series of simple-vs-simple hypotheses are tested until we find the set or range of parameter values for which the tests no longer reject the null hypothesis [Good, 2005]. For example, a simple hypothesis test for a location parameter would be H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$. To test this hypothesis, θ_0 is subtracted from all of the observations, and the test statistic is calculated. This is done for all possible values of θ_0; those that accept the hypothesis at the desired α level form the confidence interval for θ.

Now for a variance component, consider a hypothesis test for a single random intercept with variance σ^2 against no random effects for an LMM. We denote σ_0^2 as one of the possible values of σ^2 and we have a test of H_0: $\sigma^2 = 0$ vs. H_1: $\sigma^2 = \sigma_0^2$. Each of the tests in the series assumes that $\boldsymbol{V}$, the variance of $\boldsymbol{Y}$, is equal to $\sigma_0^2 \boldsymbol{Z}^T \boldsymbol{Z} + \boldsymbol{R}$. Therefore, we can weight $\boldsymbol{Y}$ by $\boldsymbol{A}$ such that $\text{var}(\boldsymbol{AY}) = \boldsymbol{I}$. $\boldsymbol{A}$ is equal to $(\boldsymbol{U}^T)^{-1}$ where $\boldsymbol{U}$ is the Cholesky decomposition of $\boldsymbol{V}$. Then we can test for the inclusion or exclusion of the random intercept for $\boldsymbol{AY}$ using a permutation test. The resulting p-value will show whether or not σ_0^2 is in the confidence interval. Similar steps can

be taken when working with GLMMs to obtain confidence intervals for the random effect variances by following the methods in Chapter 3.

While relatively straightforward, one challenge that will need to be addressed is the computational intensity of this method of generating confidence intervals. This method requires that a large number of hypothesis tests are performed in order to locate the bounds of the confidence interval. As a gauge of the potential computational time necessary, each of the the simulation results from Chapter 2 can take over 8 hours to complete while the GLMM simulations from Chapter 3 can take up to 16 hours using 20 cores of an Intel Xeon X5660 2.80 GHz server with 32 gigabytes of memory.

5.3.3 Increasing Computational Efficiency

Huge improvements to the processing time can be achieved through personal computers with multi-core processors or through computing clusters by leveraging parallelization methods. For example, an array job with up to 200 total cores available on a computing cluster can reduce a 16 hour simulation from chapter 3 to a single hour by processing each of the 500 runs within a simulation as an independent job run by a single core. However, access to a computing cluster may not always be possible or the necessary resources on a cluster may already be in use. Therefore, methods to improve computational efficiency must be explored.

In this dissertation we have extensively utilized the Monte Carlo as a method of reducing the total number of permutations that we identify. However, there is minor a drawback to this approach in that because a random sample of the permutations is drawn the estimated p-value that we obtain is actually a binomial random variable Good [2005]. This can cause a small reduction in the power of the test Dwass [1957],

but is typically not an issue. We could implement importance sampling which places weights on the permutations in order to minimize the variance of the permutation distribution and improve the power of the test [Mehta et al., 1988].

When an observed test statistic is compared to the permutation distribution only the proportion of the distribution which is greater than or equal to the observed test statistic is considered. Therefore, methods that focus on the tails of the permutation distribution can significantly reduce the computation time. An example is the branch and bound method presented by Green [1977] for Fisher's one and two-sample tests of location. In our setting this might be applied once a Monte Carlo permutation is obtained by looking at the absolute value of the sum of the permuted residuals for each subject. If many patients have large values then this could be a sign that heterogeneity between subjects is present and place this permutation towards the tail of the empirical distribution. This would reduce the number of permutations that we would need to calculate test statistics for.

5.3.4 Relaxing the Assumptions of the Proposed Permutation Tests

For a general permutation test the only assumption that is required is that the data being permuted are exchangeable under the null hypothesis. However, for the permutation tests that we have developed, a few extra assumptions are necessary. These assumptions are that the random effects and in the case of the linear mixed model, the random errors, are normally distributed.

It should be possible to relax the assumptions of normality for the random effects and/or the errors. When neither random effects nor errors are assumed to be normally distributed we only need to assume that the errors are exchangeable under the null hypothesis. To do this, we would need to estimate the parameters using semiparametric

regression methods in the case where we relax one assumption and nonparametric regression methods when neither random quantity has an assumed distribution. Many of these estimation methods have been explored in the literature. Examples include Zhang and Davidian [2001], Chen et al. [2002], and Vock et al. [2011], who developed semiparametric methods with normally distributed errors. Nonparametric methods include work by Laird [1978], Mallet et al. [1988], and Chafaï and Loubes [2006].

5.3.5 Simultaneously Testing for Both Fixed and Random Effects

One might wish to simultaneously test for both fixed and random effects using a permutation test. For this hypothesis test, the full likelihood ratio is a plausible test statistic because it incorporates information about both the fixed and random effects. We can still use permutations of the errors in order to test for the fixed effects. For example, consider a very simple hypothesis test where we test for a single fixed effect in a linear regression model.

$$H_0 : Y_i = \beta_0 + \epsilon_i \tag{5.1}$$

$$H_1 : Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i \tag{5.2}$$

The fixed effect of interest is β_1, and under the null hypothesis $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and are exchangeable. Therefore, we can utilize the likelihood ratio test statistic and obtain the p-value from the permutation null distribution to test for the fixed effect.

While our two random effect permutation tests were not originally developed using the likelihood ratio test statistic it is easy to simply use the full likelihood ratio as the test statistic. Therefore, a combined test using the full likelihood ratio and through

permuting the errors both within and among the subjects appears to be reasonable. Nuisance random effects can still be dealt with by weighting the errors. One concern is the impact of the small sample bias of the full likelihood estimates of the variance components for the random effects. As mentioned in Chapter 2, this is reason that we decided to use the restricted likelihood for our estimation as well as the restricted likelihood ratio for our test statistic. The impact of this on the validity and power of our test will need to be evaluated. For large samples there should be no difference between the restricted likelihood and the full likelihood for the variance component estimates.

The simultaneous hypothesis test for both fixed and random effects potentially provides a one-stage method of building mixed models. In contrast, the most common existing approach to model building is to first determine significant fixed effects using the full likelihood with an unstructured design for the errors. Once the mean structure has been determined, the restricted likelihood is used to estimate and select the random effects. However, extensive simulations that compare the traditional approach to model building to a permutation test approach first need to be performed to determine the utility of permutation tests with selecting a "best" LMM for a set of data.

BIBLIOGRAPHY

BIBLIOGRAPHY

D. Bates, M. Maechler, and B. Bolker. lme4: Linear mixed-effects model using S4 classes. *R package version 0.999375-39*, 2011.

P. L. Bonate. *Pharmacokinetic-Pharmacodynamic Modeling And Simulation*. Springer, New York, 2005.

J. G. Booth and J. P. Hobert. Maximum generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, 61:265–285, 1999.

T. M. Braun and Z. Feng. Permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association*, pages 1424–1432, 2001.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

B. A. Brumback, D. Ruppert, and M. P. Wand. Comment on Variable selection and function estimation in additive nonparametric regression using data-based prior by Shively, Kohn, and Wood. *Journal of the American Statistical Association*, 94: 794–797, 1999.

D. Chafaï and J. M. Loubes. On nonparametric maximum likelihood for a class of stochastic inverse problems. *Statistical & Probability Letters*, 76:1225–1237, 2006.

J. Chen, D. Zhang, and M. Davidian. A monte carlo em algorithm for generalized

linear mixed models with flexible random effects distribution. *Biostatistics*, 3:347–360, 2002.

D. Commenges. Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, 88:9–25, 2003.

C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B*, 66:165–185, 2004.

P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford niversity Press, New York, 2nd edition, 2002.

M. Dwass. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 66:165–185, 1957.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121, 1996.

R. A. Fisher. *The Design of Experiments*. Hafner, New York, 1st edition, 1935.

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

M. H. Gail, D. P. Byar, T. F. Pechacek, and D. K. Corle. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials*, 123:6–21, 1992.

P. I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer-Verlag, New York, 3rd edition, 2005.

B. F. Green. A practical interactive program for randomization tests of location. *The American Statistician*, 31:37–39, 1977.

S. Greven, C. M. Crainiceanu, H. Küchenhoff, and A. Peters. Restricted likelihood

ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17:870–891, 2008.

D. B. Hall and J. T. Præstgaard. Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, 88:739–751, 2001.

C. R. Henderson. Estimation of genetic parameters. *Annals of Mathematical Statistics*, 21:309–310, 1950.

W. Hoeffding. The large-sample power of tests based on permutation of observations. *Annals of Mathematical Statistics*, 23:169–192, 1952.

G. Kauermann, G. Claeskens, and J. D. Opsomer. Bootstrapping for penalized spline regression. *Journal of Computational and Graphical Statistics*, 18:126–146, 2009.

O. Kempthorne. The randomization theory of experimental inference. *Journal of the American Statistical Society*, 50:946–967, 1955.

M. Kenward and J. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53:983–997, 1997.

S. K. Kinney and D. B. Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63:690–698, 2008.

J. P. Klein, J. H. Klotz, and M. R. Grever. A biological marker model for predicting disease transitions. *Biometrics*, 40:927–936, 1984.

N. M. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811, 1978.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

O. E. Lee and T. M. Braun. Permutation tests for random effects in linear mixed models. *Biometrics*, 68:486–493, 2012.

X. Lin. Variance component testing in generalized linear models with random effects.

Biometrika, 84:309–326, 1997.

X. Lin and N. E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016, 1996.

A. Liu and Y. Wang. Hypothesis testing in smoothing spline models. *Available from http://www. pstat.ucsb.edu/faculty/yuedong/papers/tests.pdf.*, 2002.

D. Machin, T. Farley, B. Busca, M. Campbell, and C. d'Arcangues. Assessing changes in vaginal bleeding patterns in contracepting women. *Contraception*, 38:165–179, 1988.

A. Mallet, F. Mentré, J. L. Steimer, and F. Lokiec. Nonparametric maximum likelihood estimation for population pharmacokinetics with application to cyclosporine. *Journal of Pharmacokinetics & Pharmacodynamics*, 16:311–327, 1988.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170, 1997.

C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2008.

C. R. Mehta, N. R. Patel, and P. Senchaudhuri. Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association*, 83:999–1005, 1988.

G. Molenberghs and G. Verbeke. Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician*, 61:22–27, 2007.

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

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.

J. Öfversten. Exact tests for variance components in unbalanced mixed linear models. *Biometrics*, 49:45–57, 1993.

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:505–527, 1986.

J. C. Pinheiro and D. M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35, 1995.

E. J. G. Pitman. Significance tests which may e applied to samples from any population. *Journal of the Royal Statistical Society, Supplement 4*, pages 119–130, 1937.

R. F. Potthoff and S. N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51:313–326, 1964.

J. Raz. Testing for no effect when estimating a smooth function by nonparametric regression a randomization approach. *Journal of the American Statistical Society*, 85:132–138, 1990.

G. K. Robinson. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6:15–51, 1991.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757, 2002.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, New York, 2003.

B. R. Saville and A. H. Herring. Testing random effects in the linear mixed model using approximate Bayes factors. *Biometrics*, 65:369–376, 2009.

R. Schall. Estimation in generalised linear models with random effects. *Biometricka*, 78:719–727, 1991.

F. Scheilpl. amer: Additive models with lme4. *R package version 0.6.10*, 2011.

R. L. Schmoyer. Permutation tests for correlation in regression errors. *Journal of the American Statistical Association*, 89:1507–1516, 1994.

S. G. Self and K. Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610, 1987.

M. J. Silvapulle. Robust Wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association*, 87:156–161, 1992.

M. J. Silvapulle and P. Silvapulle. A score test against one-sided alternatives. *Journal of the American Statistical Association*, 90:342–349, 1995.

S. K. Sinha. Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, 37:219–234, 2009.

S. Sinharay and H. S. Stern. An Empirical Comparison of Methods of Computing Bayes Factors in Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 14:415–435, 2005.

P. X.-K. Song, Y. Fan, and J. D. Kalbfleisch. Maximization by parts in likelihood inference (with discussion). *Journal of the American Statistical Association*, 100: 1145–1158, 2005.

P. C. Stark, L. M. Ryan, J. L. McDonald, and H. A. Burge. Using meteorologic data to predict daily ragweed pollen levels. *Aerobiologia*, 13:177–184, 1997.

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

G. Verbeke and G. Molenberghs. The use of score tests for inference on variance

components. *Biometrics*, 59:254–262, 2003.

D. M. Vock, M. Davidian, A. A. Tsiatis, and A. J. Muir. Mixed modelanalysis of censored longitudinal data with flexible random-effects density. *Biostatistics*, 2011.

Z. Wang and T. A. Louis. Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, 90: 765–775, 2003.

F. A. Wolak. Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models. *Econometric Theory*, 5:1–35, 1989.

World Health Organization. A multicentred phase III comparative clinical trial of depot-medroxyprogesterone acetate given three-monthly at doses of 100mg or 150mg. II. The comparison of bleeding patterns. *Contraception*, 35:591–610, 1987.

S. L. Zeger and R. M. Karim. Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 90:151–156, 1991.

D. Zhang and M. Davidian. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57:795–802, 2001.

D. Zucker, O. Lieberman, and O. Manor. Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. *Journal of the Royal Statistical Society, Series B*, 62:827–838, 2000.