Bayesian Multilevel Models and Medical Applications

by
Benjamin R. Saville

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2008

Approved by:

Dr. Amy Herring, Committee Chair
Dr. Gary Koch, Committee Member
Dr. Lawrence Kupper, Committee Member
Dr. Lisa LaVange, Committee Member
Dr. Andrew Olshan, Committee Member

ABSTRACT

BENJAMIN R. SAVILLE: Bayesian Multilevel Models and Medical Applications. (Under the direction of Dr. Amy Herring.)

Deciding which predictor effects may vary across subjects is a difficult issue. Standard model selection criteria are often inappropriate for comparing models with different numbers of random effects due to constraints on the parameter space of the variance components. We propose a straightforward approach for testing random effects in the linear mixed model using Bayes factors. We scale the random effects to the residual variance and introduce parameters that control the relative contributions of the random effects. The resulting integrals needed to calculate the Bayes factor are low-dimensional integrals lacking variance components and can be efficiently approximated with Laplace's method. Our method incorporates default priors and can test multiple random effects simultaneously. We illustrate our method on data from a clinical trial of patients with bipolar disorder and on data from an environmental study of water disinfection by-products and male reproductive outcomes.

We extend our method for testing random coefficients to multilevel linear models. A major contribution of our method is the ability to test several variance components from multiple factors simultaneously, and to do so for nested, non-nested, or cross-nested multilevel designs. We illustrate our method on a study investigating significant predictors of infant birth weights in New York City.

Random effects are often used for jointly modeling distributions of correlated longitudinal and survival outcomes. These methods generally require strong parametric assumptions and can be difficult to implement. We propose a straightforward approach to evaluate the effect of a treatment or baseline predictor on both longitudinal and survival outcomes simultaneously. We define cutpoints of interest in the longitudinal outcome and time-to-event endpoints based on time to reach a given cutpoint or the survival event, whichever comes first. We use multivariate time-to-event methods on the resulting endpoints to evaluate the effect of the treatment or baseline predictor. The method is particularly attractive in clinical trial settings in which the

primary analysis must be specified *a priori*. We illustrate the method on data from a study of chronic lung disease.

To my beautiful wife, Jenny, for her love and support; to my baby boy, Tyler, for his smiles and giggles; and to my parents for always believing in me.

ACKNOWLEDGMENTS

I have been very fortunate to have great mentors during my graduate studies. First, I would like to thank my adviser, Dr. Amy Herring, for her guidance, advice, and time during the past 4 years. Her high expectations have challenged me to develop a thorough knowledge of the subject matter and a clear understanding of the research process. I would also like to thank Dr. Gary Koch for guiding me through my graduate studies. His generosity with time, advice, and financial aid has been an immense aid to me in completing my dissertation. I am also appreciative of Dr. Larry Kupper for his role as my initial academic adviser and for offering me a training grant from the National Institute of Environmental Health Sciences (NIEHS-T32ES007018), which has provided financial assistance during 5 years of my graduate studies. In addition, I would like to give special thanks to Dr. Lisa LaVange and Dr. Andrew Olshan for participating on my committee.

I am appreciative of GlaxoSmithKline for generously providing data for two separate studies and of Dr. Andrew Olshan for providing data from the Healthy Men Study, which was funded by the U.S. Environmental Protection Agency (R-82932701) and the American Water Works Association Research Foundation (CR825625-01, CR827268-01, CR828216-01). This research was also supported by the U.S. Environmental Protection Agency (R-83184301-0) and the NIEHS (P30ES10126)

I am grateful for all of my classmates in the Biometric Consulting Laboratory (BCL) for their friendship, conversation, and support during the past 5 years. In particular, I am in debt to Lauren Paynter for computing help on my M.S. paper that was extremely helpful in writing my Ph.D. dissertation.

Author's note:

The views and opinion contained in this Dissertation shall not be construed or interpreted whether directly or indirectly to be the views or opinions of any of the officers or employees of GlaxoSmithKline Research and Development Limited or any of its affiliated companies forming part of the GlaxoSmithKline group of companies. Further, reliance on the information contained in this Dissertation is at sole risk of the user. The information is provided "as is" without any warranty or implied term of any kind, either express or implied, including but not limited to any implied warranties or implied terms as to quality, fitness for a particular purpose or non-infringement. All such implied terms and warranties are hereby excluded.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Literature Review

1.1 Introduction

1.1.1 Model uncertainty

Many researchers are interested in finding the "best" statistical model to make scientific inferences. This usually involves determining the model class as well as which predictors to incorporate in the model. The model class is the general model structure which determines the relationship between the predictors and the outcome. In generalized linear models, the model class is determined by the link function, such as the identity link in linear regression, or a logit link in logistic regression. Although determining the model class may be straightforward for some applications, there may be other situations in which there is potentially more than one reasonable model class. For example, two possibilities for modeling a dichotomous outcome include logistic regression and probit regression. Both model classes may reasonably explain the data, yet in practice we often choose one model class (e.g. logistic) and ignore the other (e.g. probit).

After identifying a model class, one must also decide which predictors and interactions to include in the model. There has been substantial research on this topic. The emphasis on variable selection in the literature has made the term "model selection" synonymous with the term "variable selection." The ideal study would begin with a small number of predictors specified *a priori*, or before the data are collected. In practice, however, researchers often collect

data on as many variables as they can afford or manage and use the data to determine which variables to include in the model used for inference. This may reflect their uncertainty in the relationship between the outcome and the potential predictors. Due to sample size constraints, it may not be possible to fit one model that incorporates all variables of interest. Hence one is forced to determine which of all possible combinations of variables best explains the data.

An important element of choosing a good statistical model is the selection of an appropriate covariance model. This structure can be implicitly defined by the choice of model class and predictors or it can be manipulated within the context of a chosen model. In many model classes there are a large variety of options in choosing a covariance structure. This leads to the task of formally determining whether a chosen covariance structure is appropriate for the data.

1.1.2 The longitudinal linear mixed model

Covariance model selection can be especially difficult in the context of random coefficient models. These models incorporate random coefficients that vary by group, introducing intraclass (i.e. within-group) correlation in the covariance structure. Consider a linear mixed model for longitudinal data,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{1.1}$$

in which $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ is a $n_i \times 1$ vector of responses, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip})$ is a $n_i \times p$ design matrix, $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iq})$ is a $n_i \times q$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of parameters, and $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})'$ is a $q \times 1$ vector of random coefficients (Laird and Ware, 1982). The matrix $\boldsymbol{Z}_i$ is usually considered to be a subset of $\boldsymbol{X}_i$. It is assumed that $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\psi})$, in which $\boldsymbol{\psi}$ is the $q \times q$ covariance matrix of the random effects. A popular choice for $\boldsymbol{R}$ is $\boldsymbol{\sigma}^2\boldsymbol{I}$, which assumes the observations are independent within a subject given the random coefficients.

The random coefficients $\boldsymbol{b}_i$, often referred to as random effects, allow the estimated parameters to vary by individual. This introduces intraclass correlation for observations within a given individual. In the context of model selection, different combinations of random effects lead to different covariance structures. For situations in which the covariance model is of primary inter-

est, or the covariance model has a large impact on inference, one must carefully choose random coefficients for inclusion in a model. This leads to the problem of formally testing whether a random coefficient should be included in a model.

1.1.3 Frequentist methods for testing variance components

Testing whether a random effect should be included in a model involves the test of whether the variance of that random effect, say ψ, is equal to 0. This can be written as $H_0 : \psi = 0$ versus $H_1 : \psi > 0$. Because the constrained variance component test lies on the boundary of the parameter space, classical procedures such as the likelihood ratio test can break down asymptotically (Pauler et al., 1999; Lin, 1997; Self and Liang, 1987; Stram and Lee, 1994). It has been shown that tests for a single variance component can be carried out using mixtures of chi-square distributions (Self and Liang, 1987). For a linear mixed model, Stram and Lee (1994, 1995) show that a likelihood ratio test of q versus $q+1$ correlated random effects has a null distribution of $0.5(\chi^2_q + \chi^2_{q+1})$. For example, consider a linear mixed model for a response y_{ij} at time t_{ij}, with a random intercept and a random slope for the effect of time,

$$y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t_{ij} + \varepsilon_{ij}, \tag{1.2}$$

with $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{\psi})$ independent of $\varepsilon_{ij} \sim N(0, \sigma^2)$. Suppose we wish to test for the presence of the random slope b_{i1}, or $H_0 : \psi_{22} = 0$ versus $H_1 : \psi_{22} > 0$. A model constraint placed by this hypothesis is that the covariance of the random effects (ψ_{12}) also equals 0. The likelihood ratio test statistic is equal to twice the difference of the log likelihoods, or

$$T_{LR} = -2\{l(\boldsymbol{y}|H_0, \hat{\boldsymbol{\theta}}_0) - l(\boldsymbol{y}|H_1, \hat{\boldsymbol{\theta}}_1)\}, \tag{1.3}$$

in which $l(\boldsymbol{y}|H_0, \hat{\boldsymbol{\theta}}_0)$ and $l(\boldsymbol{y}|H_1, \hat{\boldsymbol{\theta}}_1)$ are the log likelihoods under the null and alternative hypotheses evaluated at their maximum likelihood estimates, respectively. It follows that T_{LR} asymptotically follows a 50:50 mixture of chi-square distributions with 1 and 2 degrees of freedom (Stram and Lee, 1994, 1995). The critical value for an $\alpha = 0.05$ test using this mixture distribution is 5.14, indicating one would reject H_0 for $T_{LR} > 5.14$.

Crainiceanu and Ruppert (2004) show that approximations to the null distribution using the 50:50 mixture of chi-square distributions can perform poorly in simulations. They argue that the theory of Self and Liang (1987) only applies to linear mixed models in which the data vector can be partitioned into a large number of independent and identically distributed subvectors (e.g. subjects). This may be violated when the number of subjects is not sufficient to ensure an accurate asymptotic distribution. For example, consider a cluster-randomized study with 50 patients randomized within each of 5 different hospitals (250 total patients). In this case there are only 5 independent and identically distributed clusters (hospitals). Crainiceanu and Ruppert (2004) and Crainiceanu (2005) derive the finite sample and asymptotic distribution of the likelihood ratio test, and show that under general conditions the null distribution for testing a single variance component is different from a 50:50 mixture of chi-square distributions. The present a restricted likelihood ratio test based on the restricted maximum likelihood (REML), and derive its exact null distribution. Using eigenvalues based on design matrices, they use a simulation algorithm to derive the distributions of interest.

Note that these approaches using the likelihood ratio test are only applicable for testing a single variance component. In more complex model comparisons (i.e. testing more than one random effect), distributions of test statistics become more complex and are not easily applied (Pauler et al., 1999; Feng and McCulloch, 1992). For example, a test of k *uncorrelated* variance components $\boldsymbol{\psi}_m = 0$ versus $\boldsymbol{\psi}_m > 0$ $(m = 1, \ldots, k)$ has a null distribution that is a mixture of the form (Molenberghs and Verbeke, 2007; Shapiro, 1988)

$$\sum_{m=0}^{k} 2^{-k} \begin{pmatrix} k \\ m \end{pmatrix} \chi^2_m. \tag{1.4}$$

Such mixtures can be calculated from a weighted average of p values corresponding to each of the χ^2 distributions. For a broad number of cases, determining the mixture's weights is a complex and possibly numeric task (Verbeke and Molenberghs, 2003). For addressing multiple variance components, Crainiceanu (2007) suggests using the parametric bootstrap to approximate the null distribution of the restricted likelihood ratio test. In cases that are computationally demanding, the author proposes obtaining finite sample approximations according to Greven et al. (2008).

Some alternative frequentist methods for testing a single variance component include score tests (Lin, 1997; Commenges and Jacqmin-Gadda, 1997; Verbeke and Molenberghs, 2003; Molenberghs and Verbeke, 2007; Zhang and Lin, 2008), Wald tests (Molenberghs and Verbeke, 2007; Silvapulle, 1992), and generalized likelihood ratio tests (Crainiceanu and Ruppert, 2004). Similar to the likelihood ratio test, these alternative tests also have modified null distributions due to the boundary constraint (i.e. these are modified forms of the usual asymptotic tests). For one sided variance component tests, Molenberghs and Verbeke (2007) show that the null distribution of the test statistics for the likelihood ratio, score and Wald tests are asymptotically equivalent. Zhang and Lin (2008) conduct a simulation in the setting of generalized linear mixed models, and show that the one-sided score test is slightly more powerful than the likelihood ratio test for testing a single variance component. Additionally, their simulation showed that the likelihood ratio test may suffer from numerical instability when the variance component is small and numerical integration is high dimensional. Generally, the score and Wald tests are more difficult to implement than the likelihood ratio test and require substantial programming in standard statistical software packages (Molenberghs and Verbeke, 2007). As with the likelihood ratio test, these alternative approaches are not easily extended for testing multiple random effects simultaneously.

In more simple settings such as random effects ANOVA with balanced and complete data, one sided tests of the variance components can be carried out using F tests (Neter et al., 1996, pg. 959). However, in most applications the assumption of balanced and complete data is not realistic.

A common frequentist method for choosing between competing random effects models is the Akaike information criterion (AIC) (Akaike, 1973). Akaike suggested that one should choose the model that minimizes the quantity

$$\text{AIC} = -2\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})\} + 2d, \tag{1.5}$$

in which d is the number of parameters and $\hat{\boldsymbol{\theta}}$ is the MLE. The AIC is popular because the models being compared need not be nested (although the test was originally developed for nested models). Shibata (1976) and Katz (1981) (Kass and Raftery, 1995) show that the AIC tends to

overestimate the number of parameters needed. In random effects models, the AIC suffers from ambiguity in the model dimension d.

Hence there are a lack of simple and efficient frequentist-based methods for testing variance components, especially for testing multiple variance components simultaneously. As an alternative to these frequentist-based methods, we consider Bayesian methods for testing random coefficients. Before we discuss specific challenges associated with such tests, we first introduce Bayesian methodology in the context of model selection.

1.2 Bayesian methods for model selection

1.2.1 Introduction to Bayesian inference

We introduce Bayesian methods by considering a density function $p(\boldsymbol{y}|\boldsymbol{\theta})$ for observed data $\boldsymbol{y}$ and a parameter vector $\boldsymbol{\theta}$. The likelihood function in Bayesian inference is any function proportional to $p(\boldsymbol{y}|\boldsymbol{\theta})$, i.e.

$$L(\boldsymbol{\theta}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}). \tag{1.6}$$

In contrast to frequentist methods in which $\boldsymbol{\theta}$ are fixed and unknown, Bayesian methods assume that the parameter vector $\boldsymbol{\theta}$ has a prior probability distribution $\pi(\boldsymbol{\theta})$, reflecting uncertainty in the parameters $\boldsymbol{\theta}$. The word "prior" is used to denote that $\pi(\boldsymbol{\theta})$ is the density before the data $\boldsymbol{y}$ are observed. The prior distribution allows the researcher to incorporate prior knowledge about the behavior of $\boldsymbol{\theta}$ before data are collected. Bayesian inference is primarily based on the posterior distribution of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{y}$. Using Bayes' Theorem, the posterior distribution can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{1.7}$$

in which

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1.8}$$

is the marginal distribution of $\boldsymbol{y}$, also known as the normalizing constant. In most inference problems this quantity is not available in closed form. A common technique is to identify the kernel density of $p(\boldsymbol{\theta}|\boldsymbol{y})$ by recognizing that the posterior distribution of $\boldsymbol{\theta}$ is proportional to $p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, i.e.

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.9}$$

In cases in which the kernel is not identifiable, $p(\boldsymbol{y})$ must be computed directly, unless more advanced techniques are used (e.g. Markov chain Monte Carlo methods). If we can calculate (or estimate) the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$, we can use various posterior summaries for inference on the unknown parameters $\boldsymbol{\theta}$ (e.g. the mean, mode, variance, and quantiles).

1.2.2 The Bayes factor

The Bayes factor is the posterior odds of one hypothesis versus another when the prior probabilities of the two hypotheses are equal (posterior odds = Bayes factor * prior odds). From Bayes' theorem, we have

$$p(H_k|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|H_k)p(H_k)}{p(\boldsymbol{D}|H_0)p(H_0) + p(\boldsymbol{D}|H_1)p(H_1)}, k = 0, 1, \tag{1.10}$$

in which $p(\boldsymbol{D}|H_k)$ is the marginal likelihood of the data given hypothesis H_k and $p(H_k)$ is the prior probability that H_k is true. It follows that

$$\frac{p(H_1|\boldsymbol{D})}{p(H_0|\boldsymbol{D})} = \frac{p(\boldsymbol{D}|H_1)p(H_1)}{p(\boldsymbol{D}|H_0)p(H_0)}, \tag{1.11}$$

in which

$$B_{10} = \frac{p(\boldsymbol{D}|H_1)}{p(\boldsymbol{D}|H_0)} \tag{1.12}$$

is the Bayes factor, or the ratio of the posterior odds of H_1 to its prior odds divided by ratio of the posterior odds of H_0 to its prior odds (Kass and Raftery, 1995; Good, 1958). When the two hypotheses H_1 and H_0 are equally probable *a priori*, (i.e. $p(H_1) = p(H_0)$), the Bayes factor

is equal to the posterior odds in favor of H_1 versus H_0. The numerator and denominator of equation (1.12) can be written as

$$p(\boldsymbol{D}|H_k) = \int p(\boldsymbol{D}|\boldsymbol{\theta}_k, H_k)\pi(\boldsymbol{\theta}_k|H_k)d\boldsymbol{\theta}_k, \tag{1.13}$$

in which $\boldsymbol{\theta}_k$ is a vector of parameters and $\pi(\boldsymbol{\theta}_k|H_k)$ is the prior distribution of $\boldsymbol{\theta}_k$. The quantity $p(\boldsymbol{D}|H_k)$ is known as the marginal likelihood, integrated likelihood, or predictive probability of the data. One limitation of Bayes factors lies in the influence of the prior distribution. It can be seen from equation (1.13) that the Bayes factor is a function of the prior distribution imposed by the investigator. Hence, there are an infinite number of Bayes factors that arise from different priors. Because only some priors may be appropriate for the data, not all Bayes factors are scientifically meaningful. In parameter estimation, in which inference is based on the posterior distribution $p(\boldsymbol{\theta}_k|\boldsymbol{D}, H_k)$, priors are often picked for convenience under the knowledge that if the sample is large, the effect of the prior on the estimates is small. The same rationale cannot be applied to hypothesis testing, because Bayes factors tend to be more sensitive to the priors than estimates based on a posterior distribution (Kass and Raftery, 1995; Kass, 1993). For an example, see Kass (1993), in which a sensitivity analysis reveals that the Bayes factor varies more than the posterior mean estimate across a range of possible priors. Kass and Raftery (1995) point out that choosing "non-informative" priors (as often done in Bayesian inference) can force the Bayes factor to favor H_0. As a result, in practice it is important to implement a sensitivity analysis to determine the influence of a chosen prior.

The Bayes factor is a summary of evidence provided by the data of one hypothesis versus another. Jeffreys (1961) suggests interpreting the Bayes factor according to the scale in Table 1.1 (Wasserman, 2000). For example, if $B_{10} = 12$, then H_1 is 12 times more likely than H_0 (given the data), indicating strong evidence for H_1 relative to H_0. If $B_{10} = .08$, then H_0 is 12.5 (1/.08) times more likely than H_1 (given the data), indicating strong evidence for H_0 relative to H_1.

1.2.3 Bayes factors versus Frequentist hypothesis tests

The common frequentist approach to model selection is to begin with a certain model class and a potentially large number of predictor variables. Model selection methods (e.g. stepwise selection) are used to determine which variables are most likely to be associated with the outcome. A final model is chosen by including only the significant or important predictors resulting from the selection method. Estimates and inference are based on this "best" model.

One problem with this approach is that the investigator chooses one model out of many possibilities, and proceeds with inference as if it were the only model ever considered. By choosing among several models, one is increasing the probability of finding "significant" variables by chance alone. This can cause the p-values in the final model to be very misleading (Raftery, 1995; Miller, 1984; Freedman, 1983). For example, suppose a researcher collects data on one outcome and 100 predictors, and the predictors are independent of each other. Suppose also that the researcher uses a variable selection method to arrive at a final model that includes 5 "significant" predictors, all with p-values less than 0.05. One may want to interpret these p-values as the probability of observing data as extreme or more extreme than the observed data, given the null hypothesis of no association between the predictors and outcome. At the alpha = 0.05 level, this would imply a statistically significant association between each of the five predictors and the response. However, this interpretation of the p-values is no longer valid. Basic laws of probability state that even if there is no association between any one of the predictors and the response, on average 5 out of 100 variables will have p-values less than 0.05 by chance alone. This suggests one can expect about 5 out of 100 variables to be statistically significant, even when there is truly no relationship between any one of the predictors and the outcome. This means that claims of an association between these 5 carefully chosen predictors and the response *may* be completely false.

Another difficulty associated with frequentist approaches to model selection involves the hypothesis test. In frequentist settings, it is assumed that a null hypothesis (H_0) is true, and a p-value indicates the degree of evidence against H_0. If one were comparing two nested models, an appropriate null hypothesis is that the effect of interest in the larger model is equal to 0. It is generally understood that the effect cannot exactly equal 0, but can be close enough to 0 to

be clinically meaningless. Raftery (1993, 1986a) argues that p-values ask the wrong question. Instead of asking "Is the null model true?", a better question is "Which model predicts the data better?" A Bayesian approach using Bayes factors is designed to answer the latter question. The Bayes factor is the ratio of posterior to prior odds, and measures how well one model predicts the data compared to another model. This can be advantageous in the above example because one does not need to assume that an effect is arbitrarily close to 0. In addition, if a frequentist approach yields a non-significant p-value, it can be unclear whether there is evidence for the null hypothesis or whether there are not enough data. In contrast, Bayes factors allow one to assess the evidence for a null hypothesis versus the alternative hypothesis. In cases in which there are not enough data, the evidence using Bayes factors is unlikely to be strong in either direction, reflecting the uncertainty present in the data.

The differences between p-values and Bayes factors become more apparent in large samples. Frequentist methods tend to reject H_0 almost systematically in large samples while Bayes factors do not. Frequentist approaches to this problem include adjusting the level of significance (e.g. 0.01 instead of 0.05) or simply ignoring the p-values and basing inference on other criteria that appeal to common sense (Kass and Raftery, 1995). Because Bayes factors measure how well one model predicts the data versus another, no adjustments are needed for large samples. In the case of small samples, frequentist methods based on asymptotic theory may not be valid. Of course exact methods, if they can be used, do not rely on asymptotics. Bayes factors, however, do not require asymptotics for valid inference.

Frequentist approaches often assume there are only two possible hypotheses to entertain, even though there may be additional hypotheses of interest. When multiple hypotheses are compared, frequentist methods must make adjustments to the significance level in order to correctly interpret the p-values, due to the independence assumed between each of the tests. Additional complications arise if the hypotheses are non-nested. Bayes factors are well suited for comparing many models, nested or non-nested. The interpretation of a Bayes factor does not change in the presence of multiple testing, although one still may need to control for an increased rate of false positives. To limit false positives that may arise in Bayesian multiple testing, one can build information about correlated hypotheses into the prior distribution (e.g. setting the probability that $\beta = 0$ in the prior).

Bayes factors follow the likelihood principle, which says that if two distinct sampling designs yield proportional likelihood functions, then inference about the parameters of interest will be the same between the two designs. In other words, all of the information in a sample is contained in the likelihood function. This provides flexibility in studies in which data are accrued sequentially (e.g. clinical trials), in which certain aspects of the study can be modified without changing the likelihood function. For example, one can conduct an unscheduled analysis of the data without affecting the interpretation of the final analysis. This can allow one to modify the sample size or even stop a clinical trial according to pre-specified criteria. In contrast, frequentist methods generally do not follow the likelihood principle. As an example (taken from Ibrahim, 2005), consider 12 independent coin tosses, in which one observes 9 heads and 3 tails. We are interested in testing the hypothesis $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$, in which θ is the true probability of heads. Depending on the experiment, one could base the likelihood either on the binomial distribution or the negative binomial distribution. If we let $n = 12$ be fixed beforehand, and define x as the number of heads in 12 tosses, then x follows a binomial distribution. The likelihood function in this experiment is

$$L_1(\theta) = \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3. \tag{1.14}$$

As an alternative experiment, suppose one continues flipping the coin until the third tail appears, and x equals the number of heads required to complete the experiment, then x follows a negative binomial distribution. In this case, the likelihood function is

$$L_2(\theta) = \left(\begin{array}{c} 11 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3. \tag{1.15}$$

The two likelihoods differ by a constant, meaning they are proportional to each other. From a Bayesian perspective, the posterior distribution of θ is the same for both experimental designs, i.e.

$$p(\theta|x) \propto \theta^x (1-\theta)^{12-x} \pi(\theta) \tag{1.16}$$

for a chosen prior $\pi(\theta)$. Hence Bayesian inference (including those based on Bayes factors) will be the same regardless of the experimental design. From a frequentist perspective, inferences about θ are different under each design. The p-value for the binomial design is

$$p_1 = \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = 0.075, \tag{1.17}$$

while the p-value for the negative binomial design is

$$p_2 = 1 - \sum_{j=1}^{8} \binom{2+j}{j} \theta^j (1-\theta)^3 = 0.0325. \tag{1.18}$$

At a significance level of 0.05, we make two different conclusions depending on the distribution of x. We reject H_0 in the negative binomial design, yet fail to reject H_0 in the binomial design.

Now suppose that interim analyses (both Bayesian and frequentist) were conducted before all the coin flips were completed under the first experimental design. The complete likelihood based on all the observed data would not be altered by this interim analysis (as long as the experimental design is not changed). Hence conclusions based on Bayesian inference would not be affected by the interim analysis. However, the frequentist p-value based on the complete data would now have a different interpretation, because the type I error rate has been inflated by performing two tests on the same data.

Many critics of Bayesian methods argue that Bayesian inference (and hence model selection) is subjective due to the elicitation of prior distributions on the unknown parameters. These subjective prior distributions can have a large influence on inference through the posterior distribution. When there is prior knowledge about a parameter, a Bayesian approach is advantageous because it can incorporate the prior knowledge. In cases in which there is no prior knowledge, it is desirable to formulate "non-informative" (i.e. well-spread out) priors that allow the observed data likelihood to dominate the posterior distribution. There is a large literature on challenges with prior selection as the prior distributions must be carefully defined.

Another major criticism of Bayesian methodology lies in the computational challenges associated with estimating posterior distributions and Bayes factors. There are only a few statistical

software packages that offer Bayesian inference (e.g. WinBUGS, limited capabilities in SAS PROCs BGENMOD, BLIFEREG, and BPHREG), meaning many Bayesian applications must be programmed by the user. This requires a higher level of knowledge and more user time than standard frequentist approaches. Frequentist methods do not incorporate prior distributions and hence do not rely on subjective information. They also are straightforward to implement in standard statistical software packages. Hence, despite some advantages of Bayesian approaches to model selection, frequentist methods are most commonly implemented in practice.

1.2.4 Approximating the Bayes factor

The main limitation of Bayes factors is that the marginal likelihood (1.13) can be difficult to calculate, especially when $\boldsymbol{\theta}_k$ has many dimensions. For example, consider a simple logistic regression model. Suppose $y_i, \ldots, y_n$ are independent Binomial$(1, p_i)$ random variables

$$p_i = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}} \tag{1.19}$$

in which $\boldsymbol{x}_i$ is a vector of predictors with corresponding parameters $\boldsymbol{\beta}$. Suppose we specify an non-informative improper prior $\pi(\boldsymbol{\beta}) \propto 1$. This suggests that $\boldsymbol{\beta}$ has a uniform prior distribution on the entire real line, and that the odds ratio $\phi = e^{\boldsymbol{\beta}}$ has a prior distribution $p(\phi) \propto 1/\phi$ on the positive real line (which is not necessarily non-informative). Then the marginal likelihood is

$$\begin{aligned} p(\boldsymbol{y}) &= \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta}) d\boldsymbol{\beta} && (1.20) \\ &= \int \exp\left[\sum_{i=1}^{n} \{y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log\left(1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}\right)\}\right] d\boldsymbol{\beta}, && (1.21) \end{aligned}$$

which does not have a closed form solution. Hence an alternative strategy is needed to compute the marginal likelihood. Numerical methods of computing complex marginal likelihoods are not efficient nor useful in most cases. However, there are some useful approximations that perform well in certain settings.

Laplace's approximation to the marginal likelihood (Tierney and Kadane, 1986) is derived by assuming the posterior density, which is proportional to $p(\boldsymbol{D}|\boldsymbol{\theta}, H)\pi(\boldsymbol{\theta}, H)$, is highly peaked

about its maximum $\tilde{\boldsymbol{\theta}}$ (the posterior mode). This is usually the case when the likelihood function $p(\boldsymbol{D}|\boldsymbol{\theta}, H)$ is peaked near its maximum, as in large samples. Using the notation of Kass and Raftery (1995), let $\tilde{l}(\boldsymbol{\theta}) = \log\{p(\boldsymbol{D}|\boldsymbol{\theta}, H)\pi(\boldsymbol{\theta}|H)\}$. Expanding $\tilde{l}(\boldsymbol{\theta})$ as a quadratic about $\tilde{\boldsymbol{\theta}}$ and exponentiating gives the approximation

$$p(\boldsymbol{D}|H_k) \approx \exp[\tilde{l}(\boldsymbol{\theta})] \int \exp\{[1/2(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T[-\boldsymbol{D}^2\tilde{l}(\tilde{\boldsymbol{\theta}})](\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})]\}d\boldsymbol{\theta}. \tag{1.22}$$

The integral in (1.22) takes the form of a normal density with mean $\tilde{\boldsymbol{\theta}}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}} = [-\boldsymbol{D}^2\tilde{l}(\tilde{\boldsymbol{\theta}})]^{-1}$, in which $\boldsymbol{D}^2\tilde{l}(\tilde{\boldsymbol{\theta}})$ is the Hessian matrix of second derivatives. After integrating with respect to $\boldsymbol{\theta}$, the Laplace approximation is given by

$$\widehat{p}(\boldsymbol{D}|H_k) = (2\pi)^{d/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2}p(\boldsymbol{D}|\tilde{\boldsymbol{\theta}}, H_k)\pi(\tilde{\boldsymbol{\theta}}|H_k). \tag{1.23}$$

The relative error of the Laplace approximation is of the order $O(n^{-1})$. For adequate accuracy using the Laplace approximation, Kass and Raftery (1995) recommend a sample size greater than $20d$, in which d is the dimension of $\boldsymbol{\theta}$. This will be sufficient in most "reasonable" problems, in which the likelihoods are well-behaved and a good parameterization is used. A modification of (1.23) is

$$\widehat{p}(\boldsymbol{D}|H_k) = (2\pi)^{d/2}|\hat{\boldsymbol{\Sigma}}|^{1/2}p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}, H_k)\pi(\hat{\boldsymbol{\theta}}|H_k), \tag{1.24}$$

in which $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$ is the observed information matrix (Kass and Raftery, 1995; Kass and Vaidyanathan, 1992). This approximation also has a relative error of order $O(n^{-1})$, but may be less accurate than (1.23) when the prior is somewhat informative relative to the likelihood. The advantage of (1.24) is that it is easily computed from standard statistical output, because it only requires the MLE, the observed information matrix, and the maximized likelihood.

A popular approximation to the log Bayes factor is the Schwarz Criterion (Schwarz, 1978; Kass and Raftery, 1995), given by

$$S = \log\{p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_0, H_0)\} - \log\{p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_1, H_1)\} - \frac{1}{2}(d_0 - d_1)\log(n), \tag{1.25}$$

in which $\hat{\boldsymbol{\theta}}_k$ is the MLE under hypothesis H_k, d_k is the dimension of $\boldsymbol{\theta}_k$, and n is the sample size. This approximation is also known as the Bayesian information criterion (BIC) (Weakliem, 1999; Raftery, 1986a; Raftery, 1986b), formally defined as

$$\text{BIC}_{01} = -2S. \tag{1.26}$$

The BIC approximation assumes an implied non-informative prior $\pi(\boldsymbol{\theta})$, and suggests that the log marginal likelihood of a model can be approximated by

$$\log\{\widehat{p}(\boldsymbol{D}|H_k)\} = \log\{p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}, H_k)\} - \frac{d}{2}\log(n). \tag{1.27}$$

It follows that the Bayes factor B_{10} can be approximated as

$$B_{10} \approx \exp\left\{\frac{1}{2}\text{BIC}_{01}\right\}. \tag{1.28}$$

The Bayesian information criterion approximates the log Bayes factor with a relative error of order $O(1)$. Although the error of $O(1)$ implies a crude approximation, empirical experience has found the BIC to be more accurate in practice than the error term $O(1)$ suggests (Raftery, 1995, 1996) . In fact, it has been shown that under certain conditions the BIC approximation has a relative error of $O(n^{-1/2})$ (Kass and Wasserman, 1995).

One argument against the BIC is that the Bayes factor derived from the BIC may not be close to the Bayes factor derived from an appropriate prior set by the investigator (Weakliem, 1999). The Bayes factor from the BIC corresponds closely to that derived from the unit information prior, which is a prior with the amount of information equal to the amount of information contained in one observation. More specifically, the unit information prior is a multivariate normal prior with mean at the maximum likelihood estimate and variance equal to the inverse of the expected information matrix for one observation. A simple example, taken from Raftery (1999), illustrates the idea. Let $Y_i \sim N(\mu, \sigma^2)$, iid for $i = 1, \ldots, n$, with σ known, and consider the test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Then a unit information prior is $\mu \sim N(\bar{y}, \sigma^2)$, in which $\bar{y}$ is the mean of the data. Raftery (1999) points out that the unit information prior is usually a well spread out prior as it covers the range of the data. It seems unlikely that an investigator

would specify a prior outside the range of the data; hence any alternative prior imposed by an investigator is likely to be less spread out than the unit information prior. The more spread out a prior, the more it favors the null hypothesis of no effect. If the BIC favors an effect, then it is likely that any alternative prior will also find evidence for the effect. If the BIC does not favor an effect, then there still may be a justifiable prior that would show evidence for an effect. One can still use the BIC as a baseline reference even when the unit information prior is not the same as the prior chosen by the investigator (Raftery, 1999).

The BIC and Laplace approximations are based on the assumption that the dimension d is fixed as the sample size n goes to infinity. However, in random coefficient models the dimension increases as the sample size increases. For example, consider a linear mixed model with a subject-specific random intercept. For every additional subject added to the data, an additional parameter is added for the subject-specific random coefficient. Stone (1979) observed that the BIC can be inconsistent when the dimension of the parameter vector goes to infinity. Berger et al. (2003) used Stone's example to show that the BIC and Laplace approximation may not be good approximations to the Bayes factor as the dimension and sample size both tend to infinity. The authors propose a generalized Bayesian information criterion (GBIC) and a Laplace approximation to the log Bayes factor as alternatives to approximating the Bayes factor. Chakrabarti and Ghosh (2006) generalize the methods of Berger et al. (2003) to allow distributions from the exponential family and show derivations that clarify the structure of the GBIC.

For an excellent summary of other approximations available, see Kass and Raftery (1995). These include a simple Monte Carlo method, in which the marginal likelihood in (1.13) is estimated by averaging $p(\boldsymbol{D}|\boldsymbol{\theta})$ over sampled values of $\boldsymbol{\theta}$, in which the samples are taken from the prior distribution of $\boldsymbol{\theta}$. This method has been shown to be inefficient when the posterior is concentrated relative to the prior (McCulloch and Rossi, 1991). The precision of the simple Monte Carlo estimate can be improved by using importance sampling, which generates samples of $\boldsymbol{\theta}$ from a more complex density (Geweke, 1989). Another option is Gaussian quadrature, which uses numerical analysis to evaluate integrals that are peaked around a dominant mode (Genz and Kass, 1993). Other approaches discussed by Kass and Raftery (1995) involve simulating from the posterior distribution. Such methods include direct simulation, rejection sampling, a weighted

likelihood bootstrap (Newton and Raftery, 1994), and Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler or Metropolis-Hastings algorithm. Several variations of the above methods have been proposed. See Han and Carlin (2001) for a more thorough review of MCMC methods to approximate the Bayes factor. More recently, Raftery et al. (2007) proposed a modified BIC approximation to the marginal likelihood based on MCMC output. Using the harmonic mean identity and the fact that the posterior distribution of the log likelihood is approximately a shifted gamma distribution, they introduce BICM (where M stands for Monte Carlo), a posterior-based version of the BIC. One major disadvantage of this approach is the need to fit each model with MCMC methods, which can be computationally demanding.

Several methods of approximating the Bayes factor are compared by Raftery (1996) in generalized linear models. Raftery concludes that exact analytic evaluation is the most accurate approach, but is only useful for a limited class of models. The Laplace method gives accurate approximations and is usually computationally efficient. In cases of modest dimensionality, the adaptive quadrature method of Genz and Kass (Kass and Raftery, 1995; Genz and Kass, 1993) is effective. Monte Carlo integration and importance sampling are less accurate and more computationally intensive, but there may be few other options in complex models. MCMC methods seem promising, but may be difficult to use because they can require large numbers of likelihood evaluations. The Schwarz criterion (BIC) is the easiest approximation to use, and has the advantage of not depending on a prior distribution imposed by the investigator. However, it can perform poorly when the number of degrees of freedom is large (Kass and Raftery, 1995; McCulloch and Rossi, 1991).

1.2.5 Bayes factors and prior distributions

An additional challenge to model selection via Bayes factors lies in the choice of prior distributions. It is well known that Bayes factors are sensitive to the choice of priors (Kass and Raftery, 1995; Kass, 1993). This is problematic in situations in which one has no prior information on the parameters, and the goal is to choose the best model based on the data. In these situations, it is common to use default "noninformative" priors, or prior distributions that accommodate a wide range of choices for the prior mean. However, one must choose these default prior vari-

ances with care, because as the prior variance increases toward infinity the Bayes factor will increasingly favor the null model (Bartlett, 1957). It has been documented that normal priors lead to aberrant behavior in model selection problems, leading Jeffreys (1961) to suggest the Cauchy prior as a heavy-tailed and more robust alternative. This early work by Jeffreys was extended by Zellner and Siow (1980) to develop a robust class of multivariate Cauchy priors for variable selection problems. Zellner's g-prior (Zellner, 1986) has been widely adopted in linear models, and only requires the specification of one hyperparameter. Liang et al. (2008) generalized Zellner's g-prior by implementing a fully Bayes approach using mixtures of g-priors.

1.2.6 Latent variable methods

Random effects can be viewed as special cases of latent variables, generally defined as variables not directly observed. Latent variables are commonly used in the social sciences to model underlying characteristics such as self-esteem. In latent variable models, the BIC and Laplace approximations to the Bayes factor can suffer in performance due to ambiguity of the model dimension d. This can be especially problematic in Bayesian analyses and hierarchical models. For example, in a Bayesian analysis one can increase the number of parameters of a given model by incorporating hyperpriors, even though the marginal distribution of interest may be unchanged. As a result, the model dimension d is not clearly defined for computing the BIC and Laplace approximations. As a result, researchers have suggested using the "effective model dimension" in place of the standard model dimension (Berger et al., 2003). The effective model dimension is a measure of the complexity of the model that takes into account the latent variables and unknown parameters. Defining the "effective model dimension" is a non-trivial task, as the relationship between the latent variables influences the effective dimension of the model. Additionally, the sample size n used to compute the BIC must be defined carefully. In hierarchical models, this may depend on which parameters are being tested (Kass and Raftery, 1995).

Some alternative methods of Bayesian model selection have been developed for latent variable models and show promise for random effects models. One such method is the deviance information criterion (DIC), which is based on the posterior distribution of the deviance statis-

tic,

$$D(\boldsymbol{\theta}) = -2\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\}, \tag{1.29}$$

in which $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood function for the observed data vector $\boldsymbol{y}$ given the parameter vector $\boldsymbol{\theta}$ (Spiegelhalter et al., 2002). The effective number of parameters is defined as

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\tilde{\boldsymbol{\theta}}), \tag{1.30}$$

in which $\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})]$ is the posterior mean deviance, and $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\boldsymbol{y})$ is an estimate of $\boldsymbol{\theta}$ based on the data (generally taken to be $E(\boldsymbol{\theta}|\boldsymbol{y})$). The measure p_D is the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters. The deviance information criterion (DIC) is

$$\begin{aligned} \text{DIC} &= \overline{D(\boldsymbol{\theta})} + p_D \\ &= 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}), \end{aligned} \tag{1.31}$$

which is a measure of model fit penalized by the complexity of the model. The DIC can be thought of as a Bayesian analogue to the AIC. Assuming that $D(\boldsymbol{\theta})$ is available in closed form, the DIC is calculated after an MCMC run by taking twice the sample mean of the simulated values of $D(\boldsymbol{\theta})$, minus the plug-in estimate of the deviance using the sample means of the simulated values of $\boldsymbol{\theta}$. Celeux et al. (2006) introduce variations of the DIC that allow flexibility in whether the latent variables are regarding as missing data or as parameters in the model.

One poor property of the DIC discussed by Spiegelhalter et al. (2002) is that p_D can take on negative values. Additionally, p_D is not invariant to a model's parameterization as it involves the posterior mean $\bar{\theta}$. As a result, restructuring of the data can lead to different values of the DIC. Another drawback of the DIC is the need to fit each model with MCMC methods, which can be difficult when the number of models being compared is large.

1.3 Multilevel linear models

1.3.1 Introduction

Many studies collect data that have hierarchical or clustered structures. Examples include randomized studies in which patients are clustered within practices, educational studies in which students are clustered in schools, or environmental studies in which individuals are clustered in homes which are clustered in counties. An analysis that ignores the clustering in these examples regards all observations as independent, resulting in incorrect model-based standard errors that can lead to misleading scientific inferences. Multilevel models are used to account for the correlation of observations within a given group by incorporating group-specific random effects. These random effects can be be nested (e.g. repeated observations of students nested in schools, with random effects at the student and school levels), cross-nested (e.g. repeated observations of students nested in schools and neighborhoods, with random effects at the school and neighborhood levels), or even non-nested (e.g. individuals clustered within job categories and states, with random effects at the job and state level). For an introduction to multilevel models, see Gelman and Hill (2007) or Fitzmaurice et al. (2004).

1.3.2 Nested models

There can be many levels to a data hierarchy in nested multilevel modeling. A longitudinal linear mixed model is an example of a two-level model, in which the level 1 units are the repeated observations and the level 2 units are the subjects. A two-level model can be expressed as

$$Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_j + \varepsilon_{ij}, \tag{1.32}$$

in which i indexes the the first level (e.g. repeated observations) and j indexes the second level (e.g. individuals), $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of predictors with corresponding parameters $\boldsymbol{\beta}$, and $\boldsymbol{z}_{ij}$ is $q \times 1$ vector of predictors with corresponding random effects $\boldsymbol{b}_j$. The vector $\boldsymbol{z}_{ij}$ is formed as a subset of the vector $\boldsymbol{x}_{ij}$. It is assumed that the $\boldsymbol{b}_j \sim N(\boldsymbol{0}, \boldsymbol{\psi})$ are independent of $\varepsilon_{ij} \sim N(0, \sigma^2)$. This notation can be generalized to accommodate a three-level model (e.g. repeated measurements clustered within patients which are clustered within practices). Let k

index the third level, and let

$$Y_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{z}'^{(3)}_{ijk}\boldsymbol{b}^{(3)}_k + \boldsymbol{z}'^{(2)}_{ijk}\boldsymbol{b}^{(2)}_{jk} + \varepsilon_{ijk}, \tag{1.33}$$

in which $\boldsymbol{x}_{ijk}$ is a $p \times 1$ vector of predictors with corresponding fixed effects $\boldsymbol{\beta}$, $\boldsymbol{z}^{(3)}_{ijk}$ is $q_3 \times 1$ vector of predictors with corresponding random effects $\boldsymbol{b}^{(3)}_k$, and $\boldsymbol{z}^{(2)}_{ijk}$ is $q_2 \times 1$ vector of predictors with corresponding random effects $\boldsymbol{b}^{(2)}_{jk}$. Independence is assumed between $\boldsymbol{b}^{(3)}_k$, $\boldsymbol{b}^{(2)}_{jk}$, and ε_{ijk} with distributions $\boldsymbol{b}^{(3)}_k \sim N(\boldsymbol{0}, \boldsymbol{\psi}^{(3)})$, $\boldsymbol{b}^{(2)}_{jk} \sim N(\boldsymbol{0}, \boldsymbol{\psi}^{(2)})$, and $\varepsilon_{ijk} \sim N(0, \sigma^2)$. The predictors $\boldsymbol{z}^{(3)}_{ijk}$ vary across level 2 and level 1 units, while the $\boldsymbol{z}^{(2)}_{ijk}$ vary across level 1 units. The superscripts attached to $\boldsymbol{b}^{(3)}_k$ and $\boldsymbol{b}^{(2)}_{jk}$ denote the levels at which the random effects vary (levels 3 and 2 in this case, respectively). Models with more than 3 levels can be written using similar notation.

A key feature of multilevel modeling is the incorporation of covariates $\boldsymbol{x}_{ijk}$ that can be measured at any level of the hierarchy. This allows one to address the effect of a given covariate, say at the individual level, while controlling for the effect of a higher level covariate, say at the school level. However, greater care is required in the interpretation of regression parameters, because some covariates can operate at many different levels.

For example, consider a multi-center study of 229 male patients from 3 sites (Raleigh, NC; Memphis, TN; and Galveston, TX), in which investigator's are interested in evaluating the effect of disinfection by-products (DBP's) in drinking water on male reproductive outcomes in presumed fertile men. DBP exposure was measured using water system samples and data collected on individual water usage. Three exposure variables of interest for the outcome percent normal sperm are brominated haloacetic acids (HAA-Br), brominated trihalomethanes (THM-Br), and total organic halides (TOX). Our focus is to evaluate the DBP exposure effects on sperm quality (% normal sperm) using a multilvel model. In assessing the impact of DBPs on sperm quality, it is of interest to assess the heterogeneity among study sites with respect to the overall mean of percent normal sperm (i.e. intercept) and each DBP effect (i.e. slope). It may be the case that study site is a surrogate for unmeasured aspects of water quality or other unmeasured factors of interest.

We can analyze these data using a two-level model, in which the level-1 units are the male subjects (indexed by i), and the level-2 units are the study sites (indexed by j), which are

Raleigh, Memphis, and Galveston. Let Y_{ij} denote the response for subject i in study site j. For a given water exposure x_{ij}, a random intercepts model can be written as

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_j^{(2)} + \varepsilon_{ij}, \tag{1.34}$$

in which $b_j^{(2)}$ represents the random intercept of study site j. It is assumed that $b_j^{(2)} \sim N(0, \sigma_2^2)$ is independent of $\epsilon_{ij} \sim N(0, \sigma^2)$, such that the correlation of two observations within study site j is given by $\rho = \sigma_2^2/(\sigma_2^2 + \sigma^2)$. Note that in the longitudinal linear mixed model, we used i to index the subjects, or the level-2 units. In multilevel models, we change this notation and let i index the level-1 units, j index the level-2 units, k index the level-3 units, etc., even when individuals are at higher levels in the hierarchy.

Suppose that we also collect information on the county that each subject lives in, and we think that the sperm morphology may vary by county within a given study site. Reasons for this may be different environmental risk factors or demographics in each of the different counties for a given site. We extend our notation to a 3-level model by incorporating random intercepts at the county and site level, where counties are nested within sites. In this setup, i indexes the subject, j indexes the county, and k indexes the study site. Let n_3 equal the number of level-3 units (i.e. $n_3 = 3$ study sites). Each of the sites (for $k = 1, ..., n_3$) is composed of n_{2k} level-2 clusters (i.e. counties), and each of the level-2 clusters is composed of n_{1jk} level-1 units (i.e. subjects). Let Y_{ijk} denote the response for subject i in county j in study site k. The model can be written as

$$Y_{ijk} = \beta_0 + \beta_1 x_{ijk} + b_k^{(3)} + b_{jk}^{(2)} + \varepsilon_{ijk}, \tag{1.35}$$

in which $b_{jk}^{(2)}$ is the random intercept for county j (nested within the kth study site), $b_k^{(3)}$ is the random intercept for site k, and x_{ijk} is the water exposure predictor for subject i in county j in site k. It is assumed that $b_k^{(3)} \sim N(0, \sigma_3^2)$, $b_{jk}^{(2)} \sim N(0, \sigma_2^2)$, and $\epsilon_{ijk} \sim N(0, \sigma^2)$, such that the correlation of two observations within county j in site k is given by $\rho = (\sigma_3^2 + \sigma_2^2)/(\sigma_3^2 + \sigma_2^2 + \sigma^2)$.

We can extend this model to allow the water exposure effect (i.e. slope) to vary by county

and by study-site. Let

$$Y_{ijk} = \beta_0 + b^{(3)}_{k0} + b^{(2)}_{jk0} + (\beta_1 + b^{(3)}_{k1} + b^{(2)}_{jk1})x_{ijk} + \varepsilon_{ijk}, \tag{1.36}$$

in which $b^{(2)}_{jk0}$ and $b^{(2)}_{jk1}$ are the random intercept and slope for county j (nested within study site), and $b^{(3)}_{k0}$ and $b^{(3)}_{k1}$ are the random intercept and slope for study site k, respectively. It is assumed that $\boldsymbol{b}^{(3)}_k = (b^{(3)}_{k0}, b^{(3)}_{k1})' \sim N(\mathbf{0}, \boldsymbol{\psi}^{(3)})$, $\boldsymbol{b}^{(2)}_{jk} = (b^{(2)}_{jk0}, b^{(2)}_{jk1})' \sim N(\mathbf{0}, \boldsymbol{\psi}^{(2)})$, and $\epsilon_{ijk} \sim N(0, \sigma^2)$. This model allows the intercept and water exposure effect to vary at both the county and site levels.

1.3.3 Non-nested models

In multilevel linear models, it is also possible to have non-nested random coefficients. For example, suppose investigators are interested in modeling grade point average (GPA), Y_i, for subjects (i) nested in types of extra-curricular activities (j) and schools (k). Possible categories for extra-curricular activities are sports, band, drama, debate, student council, etc. In this case the subjects are nested within activities and schools, but neither activities nor schools are nested within each other.

To illustrate, we consider a simple model with no predictors, and the focus is to determine whether there is variability in GPA across schools and across extra-curricular activities. We modify our notation somewhat to account for the non-nested structure. The model can be written in terms of Y_i as

$$Y_i = \beta_0 + \alpha_{j[i]} + b_{k[i]} + \varepsilon_i, \tag{1.37}$$

with $\alpha_{j[i]} \sim N(0, \sigma^2_\alpha)$, $b_{k[i]} \sim N(0, \sigma^2_b)$, and $\varepsilon_i \sim N(0, \sigma^2)$. The random effects in this example ($\alpha_{j[i]}$ and $b_{k[i]}$) are non-nested because neither activities (j) nor schools (k) are subsets of the other. We use the notation $j[i]$ and $k[i]$ to denote that the level-1 units (i) are nested in the level-2 units (j and k).

Now suppose that investigators conduct an assessment test prior to the school year, and are interested in assessing the effect of the test score on the students' GPA, and whether heterogeneity exists across schools for this effect. Let x_i be the test score for subject i. We can fit the

model

$$Y_i = \beta_0 + \alpha_{j[i]} + b_{0k[i]} + (\beta + b_{1k[i]})x_i + \varepsilon_i, \tag{1.38}$$

in which $\boldsymbol{b}_{k[i]} = (b_{0k[i]}, b_{1k[i]})'$ denotes the random intercept and random slope for the kth school corresponding to subject i (respectively), and $\boldsymbol{b}_{k[i]} \sim N_2(\boldsymbol{0}, \boldsymbol{\psi}_b)$. This model has a random intercept and slope for school as well as a random intercept for activity, allowing heterogeneity in the mean GPA among activities and schools and heterogeneity in the effect of the test score among schools.

1.3.4 Notation

Because multilevel linear models can have nested, cross-nested, and non-nested random coefficients, we need notation that encompasses these various types of data structures. We define the general multilevel linear model with q factors as

$$\begin{aligned} Y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_{[i]} + \varepsilon_i, \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{h=1}^{q} \boldsymbol{z}_{ih}'\boldsymbol{b}_{h[i]} + \varepsilon_i, \end{aligned} \tag{1.39}$$

in which Y_i is the response for observation i, $i = 1, \ldots, m$, $\boldsymbol{x}_i$ is a $p \times 1$ vector of predictors with corresponding fixed effects $\boldsymbol{\beta}$, $\boldsymbol{b}_{[i]} = (\boldsymbol{b}_{1[i]}', \ldots, \boldsymbol{b}_{q[i]}')'$, $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}', \ldots, \boldsymbol{z}_{iq}')'$, $\boldsymbol{z}_{ih}$ is a $d_h \times 1$ vector of predictors with corresponding random effects $\boldsymbol{b}_{h[i]}$ in which $[i]$ indexes the group in factor h pertaining to the ith observation, and $\boldsymbol{b}_{h[i]} \sim N(\boldsymbol{0}, \boldsymbol{\psi}_h)$ independent of $\varepsilon \sim N(0, \sigma^2)$, with $\boldsymbol{b}_{h[i]}$ independent of $\boldsymbol{b}_{h'[i]}$ for $h \neq h'$.

To illustrate, consider the water exposure study and the nested model given in (1.36). This model can be written as

$$Y_i = \beta_0 + b_{10[i]} + b_{20[i]} + (\beta_1 + b_{11[i]} + b_{21[i]})x_i + \varepsilon_i, \tag{1.40}$$

in which $b_{10[i]}$ and $b_{11[i]}$ are the random intercept and slope for the study site corresponding to subject i, and $b_{20[i]}$ and $b_{21[i]}$ are the random intercept and slope for the county corresponding

to subject i. It is assumed that $\boldsymbol{b}_{1[i]} = (b_{10[i]}, b_{11[i]})' \sim N(\mathbf{0}, \boldsymbol{\psi}_1)$, $\boldsymbol{b}_{2[i]} = (b_{20[i]}, b_{21[i]})' \sim N(\mathbf{0}, \boldsymbol{\psi}_2)$, and $\varepsilon_i \sim N(0, \sigma^2)$.

For a non-nested illustration, consider the GPA example with non-nested random coefficients. We can express model (1.38) as

$$Y_i = \beta_0 + b_{1[i]} + b_{20[i]} + (\beta_1 + b_{21[i]})x_i + \varepsilon_i, \tag{1.41}$$

in which $b_{1[i]}$ is the random intercept for the activity corresponding to subject i, $\boldsymbol{b}_{2[i]} = (b_{20[i]}, b_{21[i]})'$ denotes the random intercept and random slope for the school corresponding to subject i. We assume $\boldsymbol{b}_{2[i]} \sim N_2(\mathbf{0}, \boldsymbol{\psi}_2)$, $b_{1[i]} \sim N(0, \psi_1)$, and $\varepsilon_i \sim N(0, \sigma^2)$.

1.3.5 Model selection in multilevel models

Testing whether a random coefficient should be included in a multilevel model involves the test of whether the variance of that random coefficient is equal to 0. This is problematic because the null hypothesis lies on the boundary of the parameter space. Such issues are addressed in the literature in the context of linear mixed models (e.g. Stram and Lee, 1994), but there is very little research specifically for testing variance components in the broader class of multilevel models. Berkhof and Snijders (2001) proposed three score tests for variance components in multilevel models and compare their method via simulation to the likelihood ratio test, fixed F test, and Wald test. However, their simulations only consider two level models and it is not clear whether generalizations to a larger number of levels are possible. Fitzmaurice et al. (2007) proposed a permutation test for variance components in multilevel generalized linear mixed models. They apply their method to two-level generalized mixed models and suggest strategies for multilevel models with greater than two levels. However, their strategy cannot be directly applied to multilevel models with crossed random effects and can only test one variance component at a time. Frequentist methods for testing variance components in the linear mixed model are useful to some extent in nested multilevel models for testing single variance components (e.g. Crainiceanu and Ruppert, 2004; Verbeke and Molenberghs, 2003), but the null distributions are not easily obtained for testing multiple variance components, and it is not clear whether these methods can be applied to non-nested variance components. Also, Bayesian MCMC methods for

testing variance components in the linear mixed model (e.g. Cai and Dunson, 2006; Kinney and Dunson, 2008) may be generalizable to multilevel models, but these methods generally suffer from computational constraints and rely on subjective choice of hyperparameters.

1.4 Joint modeling of longitudinal and time-to-event outcomes

1.4.1 Introduction

Many clinical trials evaluate the efficacy of a treatment on correlated longitudinal and time-to-event outcomes. For example, consider a randomized clinical trial evaluating the effectiveness of a treatment drug versus a control in 2,000 patients with a chronic respiratory disorder. The investigators recorded the time to death within 3 years of randomization, as well as repeated measurements at 6 month intervals of respiratory lung function FEV, or postbronchodilator forced expiratory volume at 1 second. Because these patients suffer from a chronic condition, lung function is expected to deteriorate over time and ultimately result in death. Clearly, lung function and survival are expected to be highly correlated. There are well-established methods for analyzing these longitudinal and survival outcomes separately, including the linear mixed model for longitudinal data (Laird and Ware, 1982) and the Cox proportional hazards model for survival data (Cox, 1972). However, the analysis of these longitudinal and survival outcomes separately may be inefficient or even inappropriate when the longitudinal variable is correlated with the survival endpoint (Guo and Carlin, 2004). Such approaches ignore important information in the other outcome as well as potentially informative dropout in the longitudinal process. This has led to a growing literature on jointly modeling distributions of correlated longitudinal and survival endpoints. For additional reviews of joint modeling methods, see Hogan and Laird (1997b), Tsiatis and Davidian (2004), and Yu et al. (2004).

There are many reasons to consider a joint model of longitudinal and event outcomes. Such reasons include describing the trajectory of the longitudinal process over time subject to informative censoring and how this is affected by baseline covariates; determining how the probability of an event outcome is influenced by the longitudinal process; evaluating whether the longitudi-

nal process can be used as a surrogate endpoint for the event outcome; or making predictions of future event times for subjects who are censored. Whatever the purpose is, a general strategy of joint models is to base inference on the joint distribution of the longitudinal and survival outcomes.

Hogan and Laird (1997b) discuss joint models from the perspective of repeated measures with missing, possibly non-ignorable, observations. They broadly classify the joint models as either *selection models* or *mixture models* (see also Little, 1993). A selection model is obtained by specifying the joint density function $f_{y,d}$ as a product of the conditional distribution of the failure time d_i given the longitudinal measure $\boldsymbol{y}_i$, and the unconditional distribution of $\boldsymbol{y}_i$ (i.e. $f_{y,d} = f_{d|y} f_y$). A mixture model is given by by first conditioning $\boldsymbol{y}_i$ on d_i, such that $f_{y,d} = f_{y|d} f_d$. Hogan and Laird (1997b) point out that modeling the joint distribution of longitudinal and survival outcomes is a global strategy that does not depend on which outcome is the primary endpoint. Another broad view classifies joint model approaches as either a *two-stage approach* or a *likelihood-based approach* (Yu et al., 2004). In a two-stage approach, estimates are imputed for the longitudinal process at all time points, and the estimates are treated as true values of the longitudinal process for the event outcome model. As an alternative, a likelihood-based approach bases estimation and inference on the likelihood from a joint model of both the longitudinal and event outcomes. The likelihood approaches simultaneously estimate parameters from both outcome models, and are generally more accurate and efficient at estimating the relationship between the longitudinal and event outcomes compared to two-stage approaches (Yu et al., 2004).

We adopt the notation of Hogan and Laird (1997b) such that $\boldsymbol{y}_i^o$ and $\boldsymbol{y}_i^m$ denote the vectors of observed and missing outcomes in $\boldsymbol{y}_i$, respectively. Let d_i be the time at which subject i experiences an event, which may be right censored by C_i (independent of d_i). The time-to-event data for subject i is given by $(\tilde{d}_i, \delta_i)$, in which $\tilde{d}_i = \min(d_i, C_i)$, and $\delta_i = I(d_i \leq C_i)$. Let $\boldsymbol{X}_i$ be additional (and complete) covariate information for subject i. The observed data are then given by the set $\{(\boldsymbol{y}_i^o, \tilde{d}_i, \delta_i, \boldsymbol{X}_i) : i = 1, \ldots, N\}$.

From a repeated measures perspective, a joint model should take into account the nature of the missing data on the longitudinal process. Little and Rubin (1987) discuss various definitions of missing data mechanisms. Missing data in a response $\boldsymbol{y}_i$ is missing completely at random

(MCAR) when the probability of missing does not depend on $\boldsymbol{y}_i$. It is missing at random (MAR) when the probability of missing depends on the observed data $\boldsymbol{y}_i^o$ but not on the missing data $\boldsymbol{y}_i^m$. Under MCAR and MAR, one can obtain unbiased parameter estimates based on the likelihood of the observed data. Hence the missing data mechanisms are said to be ignorable. In contrast, missing not at random (MNAR) mechanisms are non-ignorable, because the probability of missing depends on the unobserved data $\boldsymbol{y}_i^m$. Diggle and Kenward (1994) define *informative dropout* as that which induces MNAR. In the context of joint models for longitudinal and event outcomes, we focus on models that account for informative dropout in the longitudinal process but not in the survival endpoints. In other words, it is assumed that the censoring C_i is independent of the event time d_i given the covariates in the model.

1.4.2 Mixture models

A mixture model is given by by first conditioning $\boldsymbol{y}_i$ on d_i, such that $f_{y,d} = f_{y|d} f_d$. Two types of mixture models are pattern mixture models and random effects mixture models. In pattern mixture models, each possible outcome of d_i corresponds to a different model for the longitudinal process. Little (1993) proposes modeling the longitudinal process as a multivariate normal distribution conditional on dropout time. The marginal distribution $f(\boldsymbol{y}_i)$ is straightforward to estimate, but informative dropout can be difficult to detect using this formulation. As an alternative, random effects mixture models assume the conditional distribution of $\boldsymbol{y}_i$ given d_i can be modeled using a linear mixed model with d_i as a (possibly censored) covariate. This is done by specifying distributions for $(\boldsymbol{y}_i|\boldsymbol{b}_i, d_i)$, $(\boldsymbol{b}_i|d_i)$, and $(d_i|\boldsymbol{\theta}_d)$, such that the event outcome is related to the longitudinal outcome through the random effects $\boldsymbol{b}_i$. The joint distribution $p(\boldsymbol{y}_i, d_i)$ is obtained by integrating over the random effects $\boldsymbol{b}_i$.

Wu and Bailey (1988, 1989) propose a random effects mixture model by specifying the random slope as a linear function of dropout time, and calculate ordinary least squares estimates using weighted least squares methodology. Wu et al. (1994) extend this approach to provide robust variance estimation using bootstrap methods. Mori et al. (1992) modify the approach of Wu and Bailey (1989) to estimate the adjusted slope using empirical Bayes methodology. These mixture model approaches can be implemented in software with linear mixed model capabilities.

However, a drawback of these approaches is that the dropout times must be fully observed (i.e. no censored events). Hogan and Laird (1997a) propose maximum likelihood estimation (MLE) that allows censoring in d_i. Their model assumes $p(\boldsymbol{y}_i|d_i)$ is multivariate normal and can accommodate non-parametric and semi-parametric forms for the cumulative distribution function F_d of d. The authors regard the distinct outcomes of d_i as categories in a multinomial distribution and use the EM algorithm to obtain maximum likelihood estimates.

Mixture models are useful when the primary goal is inference about the unconditional survivor rates or the association between the survival time and the longitudinal process. A limitation of these models is there must be enough observed events (i.e. high levels of dropout) to reliably estimate the parameters in the longitudinal model. It can also be difficult to account for non-ignorable missing data in the longitudinal process.

1.4.3 Selection models

The majority of the literature on joint models for longitudinal and event outcomes can be classified as selection models, in which one first conditions d_i on $\boldsymbol{y}_i$, such that $f_{y,d} = f_{d|y}f_y$. One type of selection model is an outcome-dependent selection model (Hogan and Laird, 1997b), in which d_i depends on both the observed data $\boldsymbol{y}_i^o$ and the missing data $\boldsymbol{y}_i^m$. For example, an AIDS clinical trial may define death as the event outcome and CD4 counts as the longitudinal measure. A person experiencing death at time t_k will not have observed values for CD4 counts at t_k, although the probability of death may depend on the unobserved CD4 count at t_k. Diggle and Kenward (1994) define the probability of dropout at time t_k as a function of both outcome history prior to t_k, $\boldsymbol{H}_{ik} = (y_{i1}, \ldots, y_{i,k-1})$, and the unobserved y_{ik}. The longitudinal outcomes are modeled with a linear model and the probability of dropout is modeled via a logistic regression model. Diggle and Kenward (1994) suggest that one can formally test for MAR, MCAR, and MNAR structures by testing the coefficients in the dropout model (although this is controversial).

In some cases the missing data may be more directly related to a trend over time as opposed to the actual longitudinal data. In these situations it is reasonable to relate the missing data mechanism to an underlying disease or illness progression related to $\boldsymbol{y}_i$. For example, in an AIDS clinical trial we may be interested in modeling the relationship of CD4 cell count and

survival times. One can use subject-specific random effects to model this underlying process. Consider the linear mixed model of Laird and Ware (1982), with the complete data $\boldsymbol{y}$ subset into observed $\boldsymbol{y}_i^o$ and missing $\boldsymbol{y}_i^m$ data. In this setting, we can define the probability of dropout as a function of $\boldsymbol{y}_i^o$ and $\boldsymbol{y}_i^m$ through the unobserved random effects $\boldsymbol{b}_i$, i.e.

$$f(d_i|\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{\theta}_{d|y}) = f(d_i|\boldsymbol{b}_i, \boldsymbol{\theta}_{d|y}). \tag{1.42}$$

Informative dropout, or non-ignorable missing data, occurs when d_i depends on $\boldsymbol{y}_i^m$, conditional on $\boldsymbol{y}_i^o$. If the probability of missing depends on the random effects $\boldsymbol{b}_i$, then

$$\begin{aligned} f(d_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) &= \int_{\boldsymbol{b}_i} f(d_i|\boldsymbol{b}_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m) f(\boldsymbol{b}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) d\boldsymbol{b}_i & (1.43) \\ &= \int_{\boldsymbol{b}_i} f(d_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) d\boldsymbol{b}_i. & (1.44) \end{aligned}$$

Note that $f(d_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)$ only depends on $\boldsymbol{y}_i^m$ because $f(\boldsymbol{b}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)$ depends on $\boldsymbol{y}_i^m$. In other words, the probability of missing only depends on the missing data $\boldsymbol{y}_i^m$ through the random effects $\boldsymbol{b}_i$.

A more general class of selection models that incorporates random effects is a *shared parameter model* (Hogan and Laird, 1997b). These models treat the random effects $\boldsymbol{b}_i$ as parameters in the model for $\boldsymbol{y}_i$, and as predictors in the model for the event outcome. Follman and Wu (1995) discuss shared parameter models in the setting of generalized linear models and use likelihood-based estimates that can accommodate right-censored values of d_i. Wu and Carroll (1988) develop a less general shared parameter model that assumes a multivariate normal model for the longitudinal data and a probit regression model for the probability of dropout.

Other approaches to shared parameter selection models include that of Schluchter (1992) and DeGrutolla and Tu (1994), who specify a multivariate normal distribution on $(\boldsymbol{b}_i, d_i)$. Schluchter (1992) is motivated by a longitudinal study with dropouts and considers a monotone transformation $h(d_i)$ on the event outcome, such that a transformation of event time is modeled as a linear combination of the random effects. The authors assume that the subject-specific slopes are linear functions of dropout time (see also Wu and Bailey, 1988, 1989). DeGrutolla and Tu (1994) is motivated by using CD4 cell counts as a marker for survival in AIDS patients. The authors model the survival time d_i as a linear function of the random effects, in which a non-zero

coefficient of the random effects indicates a longitudinal process can serve as a biomarker of the event process.

Tsiatis et al. (1995) use a similar AIDS study motivation to propose a semiparametric proportional hazards model on the survival data as a function of individual trends in the progression of CD4 cell counts. The authors use a linear mixed model on the longitudinal data, and then use empirical Bayes (EB) estimates of the random effects as predictors in the proportional hazards model. This is referred to as a *two-stage approach* (Yu et al., 2004), in which the strategy is to impute estimates of the longitudinal process at all time points, and treat the estimates as true values of the longitudinal process for the event outcome model. In Tsiatis et al. (1995), the hazard of death at time t takes the form

$$\lambda(t|\boldsymbol{b}_i) = \lambda_0(t)\exp\{\phi(b_{0i} + b_{1i}t)\}, \tag{1.45}$$

in which $\lambda_0(t)$ is the baseline hazards at time t, and ϕ quantifies the relationship between the the event time and the CD4 counts. Bycott and Taylor (1998) propose a two-stage approach similar to Tsiatis et al. (1995), in which they incorporate a Browning motion error term in the longitudinal model. Dafni and Tsiatis (1998) investigate the two-stage approach of Tsiatis et al. (1995) by simulation and find that the use of empirical Bayes estimators in the survival model may exhibit bias due to violated normality assumptions of the $\boldsymbol{y}_i$ (see also Tsiatis and Davidian, 2001). Dafni and Tsiatis (1998) propose an alternative two-stage approach that allows a different random intercept and slope model for k different treatments.

Wulfsohn and Tsiatis (1997) use an EM algorithm for a joint model that uses a mixed model on the longitudinal data and a proportional hazards model on the survival data. Henderson et al. (2000) use a related approach, but incorporate a mean-zero stochastic process independent of the random effects and baseline covariates. Their longitudinal model takes the form

$$y_{ij} = \boldsymbol{x}'_{1i}(t)\boldsymbol{\beta}_1 + W_{1i}(t_{ij}) + \varepsilon_{ij} \tag{1.46}$$

for times $t_{i1}, \ldots, t_{in_i}$, in which $\boldsymbol{x}'_{1i}(t)\boldsymbol{\beta}_1$ is the mean response (with possibly time-varying predictors), $W_{1i}(t) = \boldsymbol{z}'_{1i}(t)\boldsymbol{b}_i$ incorporates subject-specific random effects, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The

event outcome model for a proportional hazards model takes the form

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{x}'_{2i}\boldsymbol{\eta}_2 + W_{2i}(t)\}, \tag{1.47}$$

in which $\lambda_0(t)$ is the baseline hazards at time t, $\boldsymbol{x}'_{2i}(t)$ may be a subset of $\boldsymbol{x}'_{1i}(t)$, and $W_{2i}(t)$ is specified similar to $W_{1i}(t)$. The authors assume $\boldsymbol{W}_i(t) = (W_{1i}(t), W_{2i}(t))'$ to be a non-zero Gaussian process independent across subjects with distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. For their application, they recommend a longitudinal process with a random intercept and slope,

$$W_{1i}(t) = b_{i0} + b_{i1}t, \tag{1.48}$$

and a survival model

$$W_{2i}(t) = \gamma_0 b_{i0} + \gamma_1 b_{i1} + \gamma_2(b_{i0} + b_{i1}t) + b_{i2}, \tag{1.49}$$

in which the b_{i2} are independent frailty terms, modeled as $N(0, \sigma_2^2)$ variables, independent of the $(b_{i0}, b_{i1})'$, which have variances σ_0^2 and σ_1^2 and correlation ρ. The parameters γ_0, γ_1, and γ_2 in the survival model measure the association between the two models induced by the random intercepts, slopes, and fitted longitudinal value $W_{1i}(t)$. Guo and Carlin (2004) develop a fully Bayesian version of Henderson et al. (2000) via Markov chain Monte Carlo (MCMC) methods. They implement their methods using the software WinBUGS and compare their results to a frequentist implementation of Henderson et al. (2000).

Lin et al. (2002) use a similar framework to Henderson et al. (2000), but employ shared dependency in the models through a latent class variable. The latent class membership is determined through a multinomial logistic model, and allows underlying population heterogeneity. The longitudinal and survival outcomes are modeled independently given the latent class membership. Song and Davidian (2002) consider Wulfsohn and Tsiatis (1997), but relax the assumption of normality of the random effects. They instead assume that the random effects have a distribution with a "smooth" density.

Faucett and Thomas (1996) specify a joint model similar to Wulfsohn and Tsiatis (1997) using a Bayesian approach via Monte chain Monte Carlo (MCMC) methods. They assume

the baseline hazards function is a step function, and employ Gibbs sampling to sample from the posterior distributions of the unknown parameters. Xu and Zeger (2001a) generalize this approach to allow a stochastic process in the longitudinal model. Wang and Taylor (2001) use a similar framework to implement a Bayesian MCMC approach. Brown and Ibrahim (2003b) suggest a semiparametric Bayesian joint model in which the random effects are modeled non-parametrically. Brown and Ibrahim (2003a) and Law et al. (2002) propose Bayesian joint models that account for cured fraction. For a more thorough review of Bayesian joint models, see Ibrahim et al. (2001).

Tsiatis and Davidian (2001) focus on estimation of the parameters in the hazards model. They use a set of unbiased estimating equations that yield consistent and asymptotically normal estimators without specifying distributional assumptions on the random effects. Faucett et al. (2002) and Xu and Zeger (2001b) focus on inference on the marginal event time distribution, incorporating the longitudinal data as auxiliary information. More recent developments include Tseng et al. (2005), in which the authors propose an accelerated failure time model on the event outcome. They use a Monte Carlo EM algorithm to estimate the unknown parameters and the baseline hazard function. Hsieh et al. (2006) examine the robustness of maximum likelihood estimates against departure from the normal random effects assumption. They also discuss a profile likelihood approach, and suggest using bootstrap methods to obtain reliable variance estimates. Yu et al. (2004) formulate a joint model to account for the cured fraction, and consider estimates from the Monte Carlo EM algorithm and MCMC methods.

1.4.4 Limitations of joint models

Joint modeling methods can be computationally demanding, difficult to implement, and may require specialized software (Hogan and Laird, 1997b). Many of the joint model approaches make strong parametric assumptions regarding the longitudinal and survival processes (Tsiatis and Davidian, 2004; Yu et al., 2004). These assumptions may not be obvious and can be difficult to validate.

One strong assumption commonly made in joint models, though not required, is that the missing data $\boldsymbol{y}_i^m$ are ignorable (Yu et al., 2004). However, in many instances this assumption

is not reasonable and may lead to biased inference. For example, consider the study of chronic lung disease, in which investigators are interested in assessing the effect of treatment on FEV and survival. A person with rapidly decreasing FEV values may die at time t_{ij}, but as a result of death will not have an FEV measurement at time t_{ij}. It may even be the case that the last FEV measurement prior to death showed reasonably good values and provided no indication of decreasing lung function. In this case the probability of death depends on the unobserved FEV value at time t_{ij}, resulting in non-ignorable missing data. If one is interested in evaluating baseline covariates as predictors of FEV measurements and survival, a joint model would be appropriate. However, it would need to account for the non-ignorable missing data in the FEV measurements.

In many settings such as clinical trials, the primary effect of interest may be a baseline covariate such as a treatment effect. Fitting a marginal model on either the longitudinal or survival outcomes separately ignores important information in the other outcome. Many joint models are too complex and computationally demanding to implement in practice, and make strong assumptions regarding the longitudinal and survival outcomes. An ideal joint model would incorporate information from both longitudinal and survival outcomes in a simple manner that is straightforward to implement and makes limited distributional assumptions.

TABLE 1.1: Grades of evidence of Bayes factors

Bayes factor	Interpretation
$B_{10} < 1/10$	Strong evidence for H_0
$1/10 < B_{10} < 1/3$	Moderate evidence for H_0
$1/3 < B_{10} < 1$	Weak evidence for H_0
$1 < B_{10} < 3$	Weak evidence for H_1
$3 < B_{10} < 10$	Moderate evidence for H_1
$B_{10} > 10$	Strong evidence for H_1

CHAPTER 2

Testing Random Effects in the Linear Mixed Model Using Approximate Bayes Factors

2.1 Introduction

The linear mixed model with random effects (Laird and Ware, 1982) is a popular method for fitting longitudinal data. In such models it is often of interest to test whether certain random effects should be included the model. Testing whether a random effect should be included in the model involves the test of whether the variance of that random effect is equal to 0. Because this test lies on the boundary of the parameter space, classical procedures such as the likelihood ratio test can break down asymptotically (Pauler et al., 1999; Lin, 1997; Self and Liang, 1987; Stram and Lee, 1994). It has been shown that tests for a single variance component can be carried out using mixtures of chi-square distributions (Self and Liang, 1987; Stram and Lee, 1994). In more complex model comparisons (i.e. testing more than one random effect), distributions of test statistics are more complex and are not easily applied (Pauler et al., 1999; Feng and McCulloch, 1992; Shapiro, 1988). Some alternative frequentist methods include score tests (Lin, 1997; Commenges and Jacqmin-Gadda, 1997; Verbeke and Molenberghs, 2003; Molenberghs and Verbeke, 2007; Zhang and Lin, 2008), Wald tests (Molenberghs and Verbeke, 2007; Silvapulle, 1992), and generalized likelihood ratio tests (Crainiceanu and Ruppert, 2004),

but these methods also require modified asymptotic null distributions for tests on the boundary of the parameter space.

Bayesian sampling-based estimation approaches for calculating Bayes factors can also encounter numerical problems on the boundary of the parameter space. Markov chain Monte Carlo (MCMC) methods such as the Gibbs sampler or data augmentation can fail for certain choices of default priors on the random effects (Gilks and Roberts, 1996). Some MCMC methods have been suggested to test variance components (Sinharay and Stern, 2001; Chen and Dunson, 2003; Cai and Dunson, 2006; Kinney and Dunson, 2008), but these methods are generally time consuming to implement, require special software, and rely on subjective choice of hyperparameters which are difficult to elicit. The most widely used approximation to the Bayes factor is based on the Laplace approximation (Tierney and Kadane, 1986), resulting in the Bayesian information criterion (BIC) (Schwarz, 1978) under certain assumptions. However, the required regularity conditions of the Laplace approximation fail when the parameter lies on the boundary (Pauler et al., 1999; Hsiao, 1997; Erkanli, 1994). Pauler et al. (1999) proposed estimating Bayes factors for model comparison using an importance sampling approach and a boundary Laplace approximation. Their methods are relatively complex and are only applied in the context of simple variance component models.

Because random effects involve a distinct parameter for every individual, linear mixed models can have a very large number of dimensions. This is problematic in calculating Bayes factors, because high dimensional integrals are needed to calculate marginal likelihoods. Generally these integrals are not available in closed form, and one must consider approximations. Numerical integration methods are not efficient nor useful in such high-dimensional integrals (Kuonen, 2003). Monte Carlo integration and importance sampling methods are generally recommended for approximating high-dimensional integrals, but these methods lack accuracy and are computationally demanding. The Laplace and BIC approximations also suffer in performance from high-dimensionality (Kass and Raftery, 1995). In addition, it is not entirely clear how to define the dimensional penalty, or "effective dimension", in the BIC approximation (Spiegelhalter et al., 2002).

An additional challenge to model selection via Bayes factors lies in the choice of prior distributions. It is well known that Bayes factors are sensitive to the choice of priors (Kass and

Raftery, 1995). This is problematic in situations in which one has no prior information on the parameters, and the goal is to choose the best model based on the data. In these situations, it is common to use default priors, which can be chosen based on the data without subjective inputs and that result in good frequentist and Bayesian operating characteristics. However, one must choose these default prior variances with care, because as the prior variance increases toward infinity the Bayes factor will increasingly favor the null model (Bartlett, 1957). It has been documented that normal priors lead to aberrant behavior in model selection problems, leading Jeffreys (1961) to suggest the Cauchy prior as a heavy-tailed and more robust alternative. This early work by Jeffreys was extended by Zellner and Siow (1980) to develop a robust class of multivariate Cauchy priors for variable selection problems. Zellner's g-prior (Zellner, 1986) has been widely adopted in linear models, and only requires the specification of one hyperparameter. Liang et al. (2005) generalized Zellner's g-prior by implementing a fully Bayes approach using mixtures of g-priors.

We propose a simple approach for conducting approximate Bayesian inferences on testing whether to include random effects in the linear mixed model using Bayes factors. Our approach involves a re-parameterization of the linear mixed model, and allows for accurate approximations to the Bayes factor via Laplace's approximation. In Section 2 we introduce our method in the context of a repeated measures ANOVA model, and conduct a simulation to evaluate its performance in testing a subject-specific intercept. In Section 3 we generalize our approach to the linear mixed model, and in Section 4 we illustrate our method using two data examples. We conclude with a discussion in Section 5.

2.2 Testing a random intercept

2.2.1 ANOVA model

We start by considering a simple ANOVA model with a random subject effect

$$M_1^{(1)}: \; Y_{ij} = \mu + \lambda b_i + \varepsilon_{ij}, \tag{2.1}$$

in which Y_{ij} is the jth response for subject i, μ is an intercept, $b_i \sim N(0, \sigma^2)$ is a scaled random effect multiplied by a parameter $\lambda > 0$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the disturbance term for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. This is an ANOVA model with a random effect variance equal to $\lambda^2\sigma^2$, in which λ is a parameter controlling the level of within subject correlation. The utility of this variance component decomposition will later become clear. The notation $M_k^{(a)}$ represents parameterization (a) for model k. We distinguish models parameterized in different ways in order to consider the impact of parameterization on the accuracy of the Laplace approximation to the marginal likelihood. The implied covariance matrix of $\boldsymbol{y}_i = (Y_{i1}, \ldots Y_{in_i})'$ is $\sigma^2(\boldsymbol{I}_{n_i} + \lambda^2 \mathbf{1}_{n_i}\mathbf{1}_{n_i}')$, in which $\boldsymbol{I}_{n_i}$ is the $n_i \times n_i$ identity matrix, and $\mathbf{1}_{n_i}$ is a $n_i \times 1$ vector of 1's. It follows that the implied correlation between Y_{ij} and Y_{is} for $j \neq s$ is

$$\rho(Y_{ij}, Y_{is}) = \frac{\lambda^2}{1 + \lambda^2}. \tag{2.2}$$

Our initial focus is to compare the ANOVA model to a model with no random subject effect,

$$M_0 : \; Y_{ij} = \mu + \varepsilon_{ij}, \tag{2.3}$$

in which μ is an overall mean and $\varepsilon_{ij} \sim N(0, \sigma^2)$. We are interested in estimating Bayes factors to determine which model has the largest posterior odds given equal prior odds, given by

$$B_{10}^{(a)} = \frac{p(\boldsymbol{Y}|M_1^{(a)})}{p(\boldsymbol{Y}|M_0)}, \tag{2.4}$$

in which $\boldsymbol{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')'$. Estimating the Bayes factor implies deriving estimates of $p(\boldsymbol{Y}|M_k^{(a)})$, or

$$p(\boldsymbol{Y}|M_k^{(a)}) = \int p(\boldsymbol{Y}|\boldsymbol{\theta}_k^{(a)}, M_k^{(a)}) \pi(\boldsymbol{\theta}_k^{(a)}|M_k^{(a)}) d\boldsymbol{\theta}_k^{(a)}, \tag{2.5}$$

in which $p(\boldsymbol{Y}|\boldsymbol{\theta}_k^{(a)}, M_k^{(a)})$ is the data likelihood, $\boldsymbol{\theta}_k^{(a)}$ is the vector of model parameters, and $\pi(\boldsymbol{\theta}_k^{(a)}|M_k^{(a)})$ is the prior distribution of $\boldsymbol{\theta}_k^{(a)}$. For clarity, let $M_0^{(a)} = M_0$, i.e. only one parameterization of M_0 will be considered. For $M_1^{(a)}$ and M_0, the marginal likelihoods are not generally not available in closed form for common choices of prior distributions. Let $\boldsymbol{\theta}_1^{(a)} = (\boldsymbol{\zeta}_1^{\prime(a)}, \boldsymbol{b}', \sigma^2)'$

and $\boldsymbol{\theta}_0^{(a)} = (\boldsymbol{\zeta}_0^{\prime(a)}, \sigma^2)'$, such that the vector $\boldsymbol{\zeta}_k^{(a)}$ denotes all parameters other than the random effects $\boldsymbol{b}$ and residual variance σ^2. We specify an inverse gamma prior on σ^2 with parameters v, w, in which the mean of σ^2 is $w/(v-1)$ for $v > 1$. By marginalizing out $\boldsymbol{b}$ and σ^2 in $M_1^{(1)}$ and σ^2 in M_0, it can be shown that $(\boldsymbol{Y}|\mu, \lambda, M_1^{(1)})$ and $(\boldsymbol{Y}|\mu, M_0)$ follow multivariate t-distributions with the general form

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\zeta}_k^{(a)}, M_k^{(a)}) &= \frac{\Gamma\left(\frac{2v+m}{2}\right)\prod_{i=1}^n |\frac{w}{v}\boldsymbol{\Sigma}_i|^{-1/2}}{(\pi 2v)^{m/2}\Gamma(2v/2)} \\ &\times \left\{1 + \frac{1}{2v}\sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\left(\frac{w}{v}\boldsymbol{\Sigma}_i\right)^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right\}^{-\frac{2v+m}{2}}, \end{aligned} \tag{2.6}$$

in which $m = \sum_{i=1}^n n_i$ is the total number of observations. In our ANOVA setup, $\boldsymbol{\mu}_i = \mu\mathbf{1}_{n_i}$ in both M_0 and $M_1^{(1)}$, $\boldsymbol{\Sigma}_i = \boldsymbol{I}_{n_i}$ in M_0, and $\boldsymbol{\Sigma}_i = (\boldsymbol{I}_{n_i} + \lambda^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}')$ in $M_1^{(1)}$. After specifying a suitable prior on μ, the Laplace method can be used to integrate over (μ, λ) in $M_1^{(1)}$ and μ in M_0. We then use the resulting marginal likelihood estimates to estimate the Bayes factor $B_{10}^{(1)}$. Note that the vector $\boldsymbol{\mu}_i$ and the matrix $\sigma^2\boldsymbol{\Sigma}_i$ are the mean and covariance matrix of $\boldsymbol{y}_i$ marginalized over the random effect b_i. For additional details regarding these multivariate t-distributions, see the Appendices.

The Laplace approximation is based on a linear Taylor series approximation of $\tilde{l}(\boldsymbol{\zeta}_k^{(a)}) = \log\{p(\boldsymbol{Y}|\boldsymbol{\zeta}_k^{(a)}, M_k^{(a)})\pi(\boldsymbol{\zeta}_k^{(a)}|M_k^{(a)})\}$. The marginal likelihood $p(\boldsymbol{Y}|M_k^{(a)})$ for model k and parameterization (a) is estimated by

$$\widehat{p}(\boldsymbol{Y}|M_k^{(a)}) = (2\pi)^{d/2}|\tilde{\boldsymbol{\Sigma}}_k^{(a)}|^{1/2}p(\boldsymbol{Y}|\tilde{\boldsymbol{\zeta}}_k^{(a)}, M_k^{(a)})\pi(\tilde{\boldsymbol{\zeta}}_k^{(a)}|M_k^{(a)}), \tag{2.7}$$

in which $\tilde{\boldsymbol{\Sigma}}_k^{(a)}$ is the Hessian matrix of $\tilde{l}(\boldsymbol{\zeta}_k^{(a)})$ evaluated at the posterior mode $\tilde{\boldsymbol{\zeta}}_k^{(a)}$. Because the Laplace approximation is based on a linear Taylor series approximation, it requires certain regularity conditions. When the posterior mode lies on the boundary of the parameter space these regularity conditions fail. The Laplace method can perform poorly even if the mode is close to the boundary of the parameter space; hence estimating the marginal likelihood in $M_1^{(1)}$ via Laplace can be problematic because of the restricted parameter space of $\lambda > 0$. If the posterior mode $\tilde{\lambda}$ is close to 0, this can cause problems with the accuracy of the approximation.

Hence we consider an alternate parameterization of equation (2.1),

$$M_1^{(2)}: \; Y_{ij} = \mu + e^{\phi} b_i + \varepsilon_{ij}, \tag{2.8}$$

in which $\phi = \log(\lambda)$. Note the parameter space of ϕ is unrestricted, ensuring that the posterior mode falls within the boundaries of the parameter space. Because the posterior mode of ϕ will not violate the regularity conditions of the Laplace approximation, the estimated marginal likelihoods based on $M_1^{(2)}$ may be more accurate than those based on $M_1^{(1)}$. Following the steps outlined previously, it can be shown that $(\boldsymbol{Y}|\mu, \phi, M_1^{(2)})$ follows a multivariate t-distribution with density (2.6), with $\boldsymbol{\mu}_i = \mu \mathbf{1}_{n_i}$ and $\boldsymbol{\Sigma}_i = (\boldsymbol{I}_{n_i} + e^{2\phi} \mathbf{1}_{n_i} \mathbf{1}'_{n_i})$. We use the Laplace approximation to integrate over (μ, ϕ), and use the resulting estimate of the marginal likelihood to estimate the Bayes factors $B_{10}^{(2)}$.

2.2.2 Prior choice

It is well understood that a Bayes factor is sensitive to the choice of prior distributions (Kass and Raftery, 1995). As the prior variance of the random effect increases toward infinity, the Bayes factor will increasingly favor M_0 over the random effects model. It is therefore of interest to suggest default priors that yield robust tests with respect to model selection. In our model, we have introduced a parameter λ (or e^{ϕ}) that controls the contribution of the random effect, free of the scale of the data. We propose default priors of $\lambda \sim \log N(\kappa = 0, \tau = 1)$ and $\phi \sim N(0, 1)$, in which κ and τ denote the mean and variance of the log-normal distribution on the log scale. The priors on λ and ϕ are "equivalent" priors, meaning they lead to the same marginal likelihood. Any differences in the estimated marginal likelihoods between $M_1^{(1)}$ and $M_1^{(2)}$ should be a result of differences in the accuracy of the Laplace approximation under different parameterizations.

In choosing a prior distribution for $\lambda (> 0)$, we want to avoid a prior that is concentrated around the null value of 0. Given that the the random effects are scaled to the residual error, a $\log N(0, 1)$ prior on λ is a reasonable default prior for model selection. After marginalizing out σ^2, this log normal prior is heavy-tailed and covers most reasonable mean values of the parameter.

2.2.3 Simulation study

We conducted a simulation study to evaluate the performance of parameterization $M_1^{(2)}$ and $M_1^{(1)}$ in correctly identifying models with or without random effects. We simulated 100 data sets based on parameterization (2.1), with $n = 25, 50, 100, 500, 1000, 5000$, $n_i = 3$ and $\sigma^2 = 1$. The parameter λ was varied to allow different degrees of correlation in the simulated data, or correlations of 0, 0.14, 0.33, 0.5, 0.69. In order to implement the Laplace approximation, we estimated the posterior mode using an algorithm by Nelder and Mead (1965). We used prior distributions $\mu \sim N(0, 1)$ and $\sigma^2 \sim \text{InvGam}(1, 1)$. Estimates of the Bayes factors $B_{10}^{(1)}$ and $B_{10}^{(2)}$ were calculated for each data set for a given correlation, and were interpreted according to the scale given by Wasserman (2000) and Jeffreys (1961). Table 2.1 includes the percent of times that the estimated Bayes factors fell into the respective categories, indicating weak, moderate, or strong evidence in favoring a given model.

Both parameterizations performed well in favoring the correct model, but accuracy depended on both the sample size and the simulated correlation. In general, as ρ increased our method increasingly favored $M_1^{(a)}$ over the null model. As the sample size increased, our method more accurately detected the absence of a random slope for $\rho = 0$, and more accurately detected the presence of a random slope for $\rho > 0$. For small sample sizes, we observed reasonably good performance with zero or moderate correlation ($\rho = 0, \rho \geq 0.33$). However, larger sample sizes were needed in order to detect smaller correlations close to the boundary. Figure 2.1 shows box plots of $\log \hat{B}_{10}^{(1)}$ for $\rho = 0, 0.33$. The dotted black line represents a log Bayes factor of 0 (i.e. Bayes factor equal to 1). As n goes toward infinity the estimated log Bayes factor $B_{10}^{(1)}$ goes to infinity for $\rho = 0.33$, and goes to negative infinity for $\rho = 0$, showing that our method increasingly favors the correct model as n increases.

The estimated Bayes factors comparing $M_1^{(a)}$ to M_0 were very similar across parameterizations, even close to the boundary. We also considered the use of numerical integration to more effectively compare the parameterizations. For finite integrals of low dimension, adaptive numerical integration is an accurate and efficient method for calculating integrals. We employed transformations on the parameters (λ, ϕ, μ) to map the infinite integral in (2.5) to a finite integral, and implemented Genz' (1991) adaptive numerical integration routine for sample sizes

$n = 25, 50$. We did not find either parameterization to outperform the other. In fact, the Laplace approximations from the two parameterizations were so close that it was difficult to compute a numerical integration approximation with enough precision to distinguish between the two parameterizations. Hence, given the similarities between the two parameterizations, it is fairly evident that the boundary issue of λ is not a major problem with the Laplace approximation in this model.

2.3 Testing a random slope

2.3.1 Linear mixed model

We generalize our approach for testing random effects by considering a linear mixed model (Laird and Ware, 1982) of the form

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{2.9}$$

in which $\boldsymbol{y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is a $n_i \times 1$ vector of responses, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip})$ is a $n_i \times p$ design matrix, $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iq})$ is a $n_i \times q$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of parameters, and $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})'$ is a $q \times 1$ vector of random coefficients. The matrix $\boldsymbol{Z}_i$ is usually considered to be a subset of $\boldsymbol{X}_i$. It is assumed that $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\psi})$, in which $\boldsymbol{\psi}$ is the $q \times q$ covariance matrix of random effects. A popular choice for $\boldsymbol{R}$ is $\boldsymbol{\sigma}^2\boldsymbol{I}$, which assumes the observations are independent within a subject given the random coefficients.

We choose $b_{ih} \sim N(0, \sigma^2)$, and introduce a parameter λ_h that controls the relative contribution of the hth random effect of subject i. Let $M_k^{(a)}$ refer to model k and parameterization a. Similar to the approach of Chen and Dunson (2003) (but without the assumption of $\boldsymbol{b}_{0,i} \sim N(\boldsymbol{0}, \boldsymbol{I})$), our reparameterized model takes the form

$$M_0^{(1)}: \ \boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_{0,i}\boldsymbol{\Lambda}_0^{(1)}\boldsymbol{\Gamma}_0\boldsymbol{b}_{0,i} + \boldsymbol{\varepsilon}_i, \tag{2.10}$$

in which $\boldsymbol{Z}_{0,i} = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iq})$, $\boldsymbol{b}_{0,i} = (b_{i1}, \ldots, b_{iq})'$, $\boldsymbol{\Lambda}_0^{(1)}$ is a positive diagonal matrix with

diagonal elements $\boldsymbol{\lambda}_0^{(1)} = (\lambda_1, \ldots, \lambda_q)'$, and $\boldsymbol{\lambda}_0^{(1)} \sim \log N(\mathbf{0}_q, \boldsymbol{I}_q)$ as an extension of the "default" prior suggested in 2.2.2. Let $\boldsymbol{\Gamma}_0$ be a lower triangular matrix with $\mathbf{1}_q$ along the diagonal, and lower off-diagonal elements $\boldsymbol{\gamma}_0$ which induce correlation between the respective random effects.

Our focus is to test whether to include an additional random effect $b_{i(q+1)}$. Let $\boldsymbol{Z}_{1,i}$, $\boldsymbol{\Lambda}_1^{(1)}$, $\boldsymbol{\Gamma}_1$, and $\boldsymbol{b}_{1,i}$ be equal to their counterparts from (2.10), but including the elements corresponding to the additional random effect $b_{i(q+1)}$. The full model including the additional random effect takes the form

$$M_1^{(1)} : \ \boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_{1,i}\boldsymbol{\Lambda}_1^{(1)}\boldsymbol{\Gamma}_1\boldsymbol{b}_{1,i} + \boldsymbol{\varepsilon}_i, \tag{2.11}$$

in which $\boldsymbol{Z}_{1,i} = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{i(q+1)})$, $\boldsymbol{b}_{1,i} = (b_{i1}, \ldots, b_{i(q+1)})'$, $\boldsymbol{\Lambda}_1^{(1)}$ is a positive diagonal matrix with diagonal elements $\boldsymbol{\lambda}_1^{(1)} = (\lambda_1, \ldots, \lambda_{q+1})'$ and $\boldsymbol{\lambda}_1^{(1)} \sim \log N(\mathbf{0}_{q+1}, \boldsymbol{I}_{q+1})$, and $\boldsymbol{\Gamma}_1$ is a lower triangular matrix with $\mathbf{1}_{q+1}$ along the diagonal and lower off-diagonal elements $\boldsymbol{\gamma}_1$.

As demonstrated with the ANOVA model, we also consider an alternate parameterization of (2.10) and (2.11), by setting $\boldsymbol{\lambda}_k^{(2)} = \boldsymbol{\phi}_k = \log \boldsymbol{\lambda}_k^{(1)}$, with $\boldsymbol{\lambda}_0^{(2)} \sim N(\mathbf{0}_q, \boldsymbol{I}_q)$ and $\boldsymbol{\lambda}_1^{(2)} \sim N(\mathbf{0}_{q+1}, \boldsymbol{I}_{q+1})$. We define $\boldsymbol{\Lambda}_0^{(2)}$ as a diagonal matrix with diagonal elements $e^{\boldsymbol{\lambda}_0^{(2)}} = (e^{\phi_1}, \ldots, e^{\phi_q})$, and $\boldsymbol{\Lambda}_1^{(2)}$ as a diagonal matrix with diagonal elements $e^{\boldsymbol{\lambda}_1^{(2)}} = (e^{\phi_1}, \ldots, e^{\phi_{q+1}})$. Let $M_0^{(2)}$ denote the reduced model and $M_1^{(2)}$ denote the full model under this parameterization.

2.3.2 Approximating the marginal likelihoods

In order to implement the Laplace approximation, we first marginalize out $\boldsymbol{b}$ and σ^2. Let $\sigma^2 \sim \text{InvGam}(v, w)$. It can be shown that the marginal distribution $p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda}_k^{(a)}, \boldsymbol{\gamma}_k, M_k)$ follows a multivariate t-distribution with density (2.6), in which $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_i = (\boldsymbol{I}_{n_i} + \boldsymbol{Z}_{k,i}\boldsymbol{\Lambda}_k^{(a)}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'\boldsymbol{\Lambda}_k'^{(a)}\boldsymbol{Z}_{k,i}')$. Integrating out all random effects simultaneously dramatically decreases the dimension of the integral needed for the marginal likelihoods. After specifying suitable priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_k$, we use the Laplace method to integrate over $(\boldsymbol{\beta}, \boldsymbol{\lambda}_k^{(a)}, \boldsymbol{\gamma}_k)$ to approximate the marginal likelihoods $p(\boldsymbol{Y}|M_k^{(a)})$ used to evaluate the Bayes factor $B_{10}^{(a)}$. For additional details regarding these multivariate t-distributions, see the Appendices.

As previously discussed, many of the existing methods for testing variance components are only applicable in simple settings, e.g. testing a single variance component. One major ad-

vantage of our approach is we can test multiple random effects simultaneously by modifying equation (2.11) such that the $\boldsymbol{Z}_{1,i}$, $\boldsymbol{\Lambda}_1^{(a)}$, $\boldsymbol{\Gamma}_1$, and $\boldsymbol{b}_{1,i}$ correspond to a model with several additional random effects. Incorporating default priors on the $\boldsymbol{\beta}$ coefficients, one can simultaneously compare models with varying numbers of both fixed and random effects.

2.3.3 Simulation study

We conduct a simulation study to test for the presence of a random slope. We define one predictor based on time, such that $\boldsymbol{x}_i = (1, 2, \ldots, J)'$, $\boldsymbol{X}_i = (\boldsymbol{1}, \boldsymbol{x}_i)$, and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Consistent with our previous notation, let M_1 refer to the random intercept model and M_2 refer to the random intercept and slope model. Letting $\boldsymbol{Z}_{1,i} = \boldsymbol{1}_J$, $\boldsymbol{\lambda}_1^{(a)} = \lambda_0^{(a)}$, and $\boldsymbol{b}_{1,i} = b_{i0}$, we have

$$M_1^{(a)} : \; y_{ij} = \beta_0 + \lambda_0^{(a)} b_{i0} + \beta_1 x_{ij} + \varepsilon_{ij} \tag{2.12}$$

for the random intercepts model. Letting $\boldsymbol{Z}_{2,i} = (\boldsymbol{1}_J, \boldsymbol{x}_i)$, $\boldsymbol{\lambda}_2^{(a)} = (\lambda_0^{(a)}, \lambda_1^{(a)})'$, and $\boldsymbol{b}_{2,i} = (b_{i0}, b_{i1})'$, we have

$$M_2^{(a)} : y_{ij} = \beta_0 + (\lambda_0^{(a)} + \gamma_{12} x_{ij} \lambda_1^{(a)}) b_{i0} + (\beta_1 + \lambda_1^{(a)} b_{i1}) x_{ij} + \varepsilon_{ij} \tag{2.13}$$

for the random intercept and slope model. Our focus is to compare model $M_2^{(a)}$ to $M_1^{(a)}$. After integrating out $\boldsymbol{b}$ and σ^2 to produce marginal multivariate t-distributions, the integrals needed to calculate the marginal distributions $p(\boldsymbol{Y}|M_1^{(a)})$ and $p(\boldsymbol{Y}|M_2^{(a)})$ only have 3 or 5 dimensions, respectively. Hence the Laplace method can effectively be used to integrate over $(\beta_0, \beta_1, \lambda_0^{(a)})$ in $M_1^{(a)}$ and $(\beta_0, \beta_1, \lambda_0^{(a)}, \lambda_1^{(a)}, \gamma_{12})$ in $M_2^{(a)}$.

We simulated 100 data sets based on a random intercept and slope model under the standard notation of Laird and Ware (1982) as shown in (2.9), i.e.
$Y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1}) x_{ij} + \epsilon_{ij}$. We set $\beta_0 = 0$, $\beta_1 = 0.5$, $J = 10$, $\sigma^2 = 1$, and we generated the random effects from a multivariate normal distribution $\boldsymbol{b}_i \sim N_2(\boldsymbol{0}, \boldsymbol{\psi})$, in which $\psi_{11} = 1$ and $\psi_{12} = \rho(b_{i0}, b_{i1})\sqrt{\psi_{11}\psi_{22}} = -0.3$. We considered different combinations of the random slope variance component and sample size by varying $\sqrt{\psi_{22}} = 0, 0.04, 0.08, 0.15, 0.25$ across $n = 25, 50, 100, 500, 1000, 5000$. For implementing the Laplace approximation to the marginal

likelihoods, we used prior distributions $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{I})$, $\sigma^2 \sim \text{InvGam}(1,1)$, and $\gamma_{12} \sim N(0,1)$.

As illustrated in Table 2.2, our method performed well in favoring the correct model, but accuracy depended on both the sample size and the simulated variance of the random slope. In general, as the standard deviation of b_{i1} increased, our method increasingly favored $M_2^{(a)}$ over $M_1^{(a)}$. As the sample size increased, our method more accurately detected the absence of a random slope for $\sqrt{\psi_{22}} = 0$, and more accurately detected the presence of a random slope for $\sqrt{\psi_{22}} > 0$. For smaller sample sizes, our method generally detected the random slope for $\sqrt{\psi_{22}} \geq 0.15$, indicating our method is useful even for small sample sizes with moderate to large random effects. Figure 2.2 shows box plots of $\log \hat{B}_{21}^{(1)}$ for $\sqrt{\psi_{22}} = 0$ and $\sqrt{\psi_{22}} = 0.08$. As n goes to infinity the estimated log Bayes factor $\hat{B}_{21}^{(1)}$ goes to infinity for $\sqrt{\psi_{22}} = 0.08$, and goes to negative infinity for $\sqrt{\psi_{22}} = 0$. This indicates the estimated Bayes factor increasingly favors the correct model as n increases.

As noted previously, the approximations to the marginal likelihoods do not seem to vary a great deal across parameterizations. Similar patterns were found in this simulation, with most differences extremely small. Occasionally we did observe large differences between the marginal likelihood estimates of $M_2^{(2)}$ and $M_2^{(1)}$ for simulated variances close to the boundary for $n = 5000$. It appears that this situation was due to occasional poor convergence of the maximization routine for the $\boldsymbol{\lambda}_k^{(2)}$ parameterization, and not due to the Laplace approximation itself.

2.4 Illustrative examples

2.4.1 Hamilton Rating Scale for Depression

To illustrate our method, we consider a clinical trial of patients with bipolar I disorder (Calabrese et al., 2003), GlaxoSmithKline study SCAB2003. The investigators concluded that the treatment drug, lamotrigine, significantly delays the time to intervention for a depressive episode compared to placebo. The investigators also collected repeated measurements on the Hamilton Rating Scale for Depression (HAMD), a numerical measure of the severity of depressive symptoms. As a secondary analysis, we wish to determine if lamotrigine is effective in reducing

depressive symptoms during the first year after randomization as measured by the HAMD-17 summary score. Larger HAMD-17 scores reflect higher levels of depression.

We consider 275 patients (160 lamotrigine 200/400 mg/day, 115 placebo) with at least one outcome measurement and complete covariate data. The number of repeated measurements per subject ranges from 1 to 17, and HAMD-17 scores range from 0 to 35, with a mean value of 7. To better approximate normality, we used a square root transformation of HAMD-17 (sqrt-HAMD-17). We fit a linear mixed model with sqrt-HAMD-17 as the response, predicted by sqrt-HAMD-17 at screening and baseline, time (in years), treatment, gender, age (< 30, 30-40, 40-50, ≥ 50), and the number of depressive or mixed episodes in the last year (1-2 vs. ≥ 3). Screening refers to the time at enrollment, and baseline refers to the time of randomization (after stabilization).

In assessing the impact of lamotrigine on HAMD-17 scores, it is also interesting to assess the variability among patients with regards to the overall mean and slope. One might expect patients to have different patterns of depressive episodes across time, perhaps resulting from biological mechanisms or individual responses to drug treatment. This leads to the task of testing whether to include random effects in our model. Our focus is to compare models with varying combinations of a random intercept and slope, i.e. M_0: a model without random effects; M_1: a model with a random intercept; and M_2: a model with a random intercept and slope. Based on the scale of both the response and the explanatory variables, we use vague priors on the fixed effects $\boldsymbol{\beta}$ and residual variance σ^2 that accommodate a wide range of reasonable mean values. We define these priors as $\boldsymbol{\beta} \sim N_9(\mathbf{0}, 10\boldsymbol{I})$ and $\sigma^2 \sim$InvGamma$(0.01, 0.01)$.

The estimated log Bayes factors for the respective comparisons are $\log \hat{\beta}_{21} = 61.0, \log \hat{\beta}_{20} = 348.5, \log \hat{\beta}_{10} = 287.5$. These estimates show strong evidence for M_2 versus the other models, indicating the intercepts and slopes vary significantly by individual. We fit the preferred model, M_2, using MCMC methods based on 15,000 samples, with a burn-in of 10,000. Figure 2.3 shows the predicted overall mean for each treatment group and the predicted individual sqrt-HAMD-17 for 50 random subjects. The predicted overall mean is based on a 40-50 year old female with 1-2 depressive episodes in the past year, and average values of sqrt-HAMD-17 at screening and baseline (2.2 and 4.8, respectively). There is large variability in the subject specific intercepts and slopes. Lamotrigine use, age, and sqrt-HAMD-17 at screening and baseline all appear to

be significant predictors of the outcome. A one unit increase in sqrt-HAMD-17 at baseline is associated with a 0.63 (95% CI $= 0.51, 0.74$) increase in mean sqrt-HAMD-17, and a one unit increase in sqrt-HAMD-17 at screening is associated with a 0.22 (95% CI $= 0.02, 0.48$) increase in mean sqrt-HAMD-17. On average, patients 30-40 years old have sqrt-HAMD-17 values 0.18 (95% CI $= -0.20, 0.56$) units greater than patients < 30 years old, patients 40-50 years old have sqrt-HAMD-17 values 0.47 (95% CI $= 0.13, 0.82$) units greater than patients < 30 years old, and patients ≥ 50 years old have sqrt-HAMD-17 values 0.39 (95% CI $= 0.05, 0.73$) units greater than patients < 30 years old. As the main association of interest, sqrt-HAMD-17 values for subjects on lamotrigine are on average 0.33 units lower (95% CI $= -0.54, -0.10$) than sqrt-HAMD-17 values for subjects on placebo. The 95% credible interval does not contain 0, indicating that lamotrigine may be effective at reducing depressive symptoms. These conclusions reinforce the time-to-event analysis of Calabrese et al. (2003).

2.4.2 Exposure of disinfection by-products in drinking water and male fertility

A multi-center study of 229 male patients from 3 sites (A, B, C) was conducted to evaluate the effect of disinfection by-products (DBP's) in drinking water on male reproductive outcomes in presumed fertile men. DBP exposure was measured using water system samples and data collected on individual water usage. Three exposure variables of interest for the outcome percent normal sperm are brominated haloacetic acids (HAA-Br), brominated trihalomethanes (THM-Br), and total organic halides (TOX).

Our focus is to model the response (% normal sperm) using the three DBP exposure variables. Because we are interested in each exposure's effect independent of the other exposure variables, we fit three separate models, one for each DBP exposure. In each model we control for the following baseline covariates using indicator variables: male age (< 25, 25-30, 30-35, > 35), education (high school or less, some college, graduated college), and the abstinence interval before taking the sample (2-3 days, 4-8 days, or > 8 days). We scale each predictor by subtracting the overall mean of the predictor and dividing by a constant c ($c = 10$ for HAA-Br and THM-Br, and $c = 100$ for TOX) to allow for better computational efficiency. We use a

probit transformation of percent normal sperm and multiply the result by 5, so the tranformed response has a range of -10.5 to -1.8, a mean of -5.6, and a variance of 1.8.

In assessing the impact of DBP's on sperm quality, it is also of interest to assess the variability among study sites with regards to the overall mean of percent normal sperm (i.e. intercept) and each DBP effect (i.e. slope). It may be the case that study site is a surrogate for unmeasured aspects of water quality or other unmeasured factors of interest. For each DBP exposure, we define three models based on the inclusion of random effects, i.e. M_0 : a model without random effects; M_1 : a model with a random intercept; and M_2 : a model with a random slope and intercept. Based on the scale of both the response and the explanatory variables, we use vague priors on the fixed effects $\boldsymbol{\beta}$ and residual variance σ^2 that accommodate a wide range of reasonable mean values. We define these priors as $\boldsymbol{\beta} \sim N_9(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\sigma^2 \sim$InvGamma$(0.01, 0.01)$, with $\boldsymbol{\mu} = (-5.5, \mathbf{0}_8')'$ and $\boldsymbol{\Sigma}$ a diagonal matrix with diagonal elements $(100, 10 \times \mathbf{1}_8')$.

For HAA-Br, we observe moderate evidence for M_1 versus M_2 ($\hat{B}_{12} = 6.9$) and strong evidence for M_1 versus M_0 ($\log \hat{B}_{10} = 15.3$). For THM-Br, we observe strong evidence for M_1 versus both M_2 ($\hat{B}_{12} = 10.5$) and M_0 ($\log \hat{B}_{10} = 19.2$). For TOX, we observe weak evidence for M_1 versus both M_2 ($\hat{B}_{12} = 1.1$) and strong evidence for M_1 versus M_0 ($\log \hat{B}_{10} = 12.9$). Hence the random intercepts model M_1 is favored by the Bayes factor for all three DBP exposures. For comparison, we fit both models M_1 and M_2 using MCMC methods based on 40,000 samples, with a burn-in of 40,000 for each model. We plot the predicted mean response based on M_2, for a 30-35 year old male who has graduated college and has abstained for 2-3 days (Figure 2.4). For each of the three exposure models, one can see that there is some separation of the intercepts and varying degrees of agreement between the slopes. Although the point estimates of the slopes (based on the posterior means) appear to be quite different, the large variability associated with these estimates suggests that the slopes do not vary by study site. Hence we conclude that M_1 is the preferred model for each of the predictors. Based on M_1, both HAA-Br and THM-Br have posterior distributions centered near 0, indicating little association between these DPB's and percent normal sperm. The posterior distribution of TOX tends to be centered below 0, with a posterior mean of -1.20 (95 % CI = -3.67,0.47); however, the 95% credible interval contains 0.

2.5 Discussion

We recommend our approach as a simple and efficient method in testing random effects in the linear mixed model. Our approach avoids issues with testing on the boundary of the parameter space, uses low-dimensional approximations to the Bayes factor, and incorporates default priors on the random effects. We have shown Laplace's method to be an effective approach to estimating Bayes factors, even in cases in which the variance of the random effect lies on the boundary. By scaling the random effects to the residual variance and introducing a parameter that controls the relative contribution of the random effects, we can effectively integrate out the random effects and reduce the dimensionality of the marginal likelihood. The scaling of the random effects to the residual variance makes the $\log N(\mathbf{0}, \boldsymbol{I})$ and $N(\mathbf{0}, \boldsymbol{I})$ distributions reasonable default priors for $\boldsymbol{\lambda}_k^{(1)}$ and $\boldsymbol{\lambda}_k^{(2)}$, respectively. Simulations suggest that these priors have good small sample properties and consistency in large samples. Incorporating reasonable default priors on the fixed effects, our method can be used for comparing a large class of random effects models with varying fixed and random effects.

Alternative procedures for allowing default priors for model selection via Bayes factors are discussed by Berger and Pericchi (1996). These include the authors' proposed *intrinsic Bayes factors*, the Schwarz approximation (Schwarz, 1978), and the methods of Jeffreys (1961) and Smith and Spiegelhalter (1980). Gelman (2006) discusses various approaches to default priors specifically for variance components. Common approaches include the uniform prior (e.g. Gelman, 2007), the half-t family of prior distributions, and the inverse-gamma distribution (Spiegelhalter et al., 2003). These prior distributions can encounter difficulties when the variance components are close to 0. Other discussions of selecting default priors on variance components include Natarajan and Kass (2000), Browne and Draper (2006), and Kass and Natarajan (2006).

TABLE 2.1: Testing a random intercept, $\hat{B}_{10}^{(a)}$

		Parameterization (1)						Parameterization (2)					
		Favor null			Favor random int.			Favor null			Favor random int.		
n	ρ	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
25	0	6	67	22	4	0	1	7	71	18	3	0	1
	0.14	0	31	35	22	9	3	0	36	31	21	9	3
	0.33	0	1	12	20	25	42	0	1	14	18	25	42
	0.5	0	0	0	4	7	89	0	0	1	3	7	89
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
50	0	23	50	18	7	2	0	27	51	13	7	2	0
	0.14	2	22	34	16	14	12	4	25	30	18	11	12
	0.33	0	0	7	4	8	81	0	0	7	5	8	80
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
100	0	43	43	12	2	0	0	52	36	10	2	0	0
	0.14	3	12	18	21	15	31	6	13	15	22	15	29
	0.33	0	0	1	0	3	96	0	0	1	0	4	95
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
500	0	75	18	7	0	0	0	82	14	4	0	0	0
	0.14	0	0	0	1	1	98	0	0	0	1	1	98
	0.33	0	0	0	0	0	100	0	0	0	0	0	100
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
1000	0	84	14	2	0	0	0	86	12	2	0	0	0
	0.14	0	0	0	0	0	100	0	0	0	0	0	100
	0.33	0	0	0	0	0	100	0	0	0	0	0	100
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
5000	0	96	3	1	0	0	0	96	3	1	0	0	0
	0.14	0	0	0	0	0	100	0	0	0	0	0	100
	0.33	0	0	0	0	0	100	0	0	0	0	0	100
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100

* Table includes the percent of times that the estimated Bayes factors fell into the respective categories

TABLE 2.2: Testing for a random slope, $\hat{B}_{21}^{(a)}$

		Parameterization (1)						Parameterization (2)					
		Favor null			Favor random slope			Favor null			Favor random slope		
n	$\sqrt{\psi_{22}}$	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
25	0	97	2	1	0	0	0	97	2	1	0	0	0
	0.04	98	1	1	0	0	0	98	1	1	0	0	0
	0.08	85	6	5	3	1	0	85	6	5	3	1	0
	0.15	11	5	6	4	9	65	12	4	7	3	9	65
	0.25	0	0	0	0	1	99	0	0	0	0	1	99
50	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	97	2	1	0	0	0	98	1	1	0	0	0
	0.08	65	11	8	7	1	8	65	11	8	7	1	8
	0.15	0	3	0	0	4	93	0	3	0	0	5	92
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
100	0	99	1	0	0	0	0	100	0	0	0	0	0
	0.04	98	2	0	0	0	0	98	2	0	0	0	0
	0.08	37	11	14	10	9	19	37	11	15	9	10	18
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
500	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	93	2	2	0	2	1	93	2	2	0	2	1
	0.08	0	0	1	1	0	98	0	0	1	1	0	98
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
1000	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	77	11	8	2	1	1	78	11	7	2	1	1
	0.08	0	0	0	0	0	100	0	0	0	0	0	100
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
5000	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	2	4	3	3	4	84	2	5	2	3	4	84
	0.08	0	0	0	0	0	100	0	0	0	0	0	100
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100

* Table includes the percent of times that the estimated Bayes factors fell into the respective categories

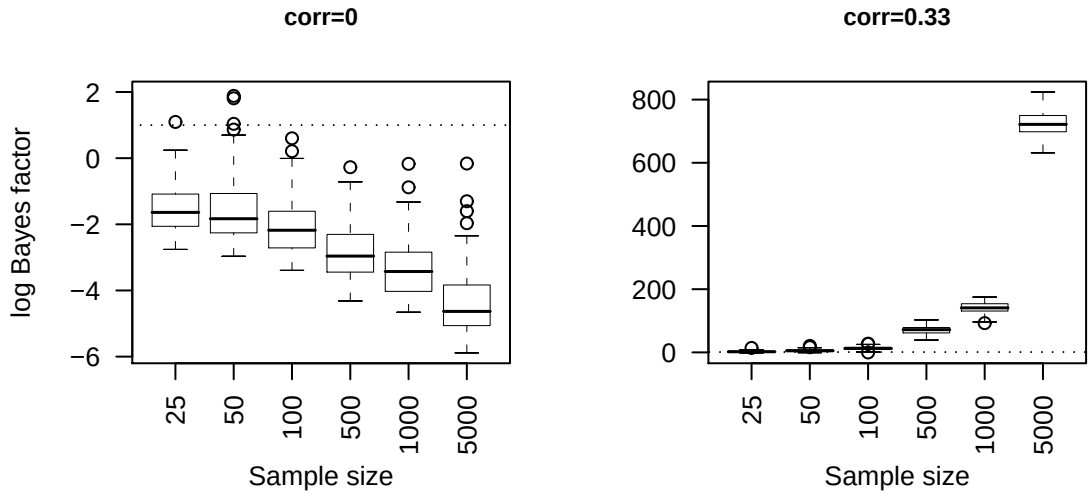


FIGURE 2.1: Box plot of $\log \hat{B}_{10}^{(1)}$, by ρ

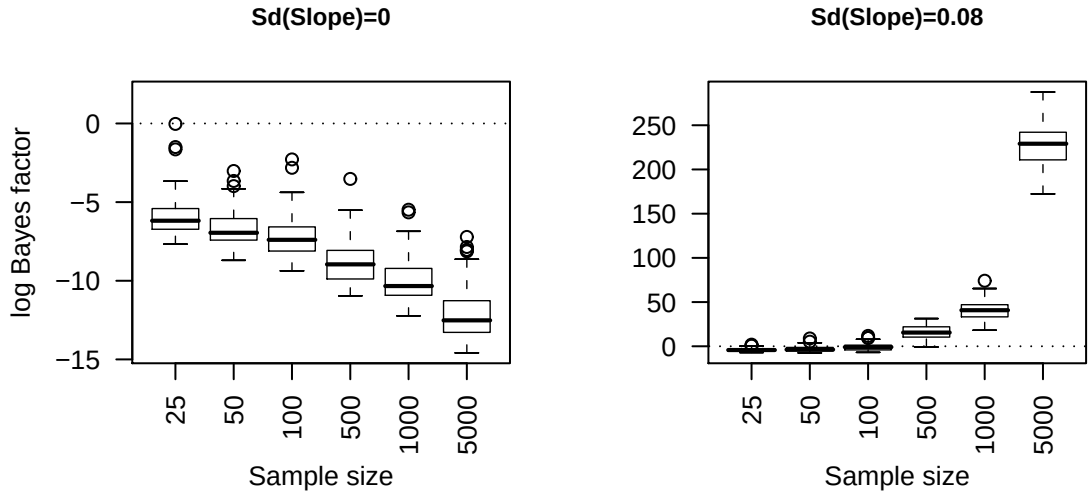


FIGURE 2.2: Box plot of $\log \hat{B}_{21}^{(1)}$, by standard deviation of random slope

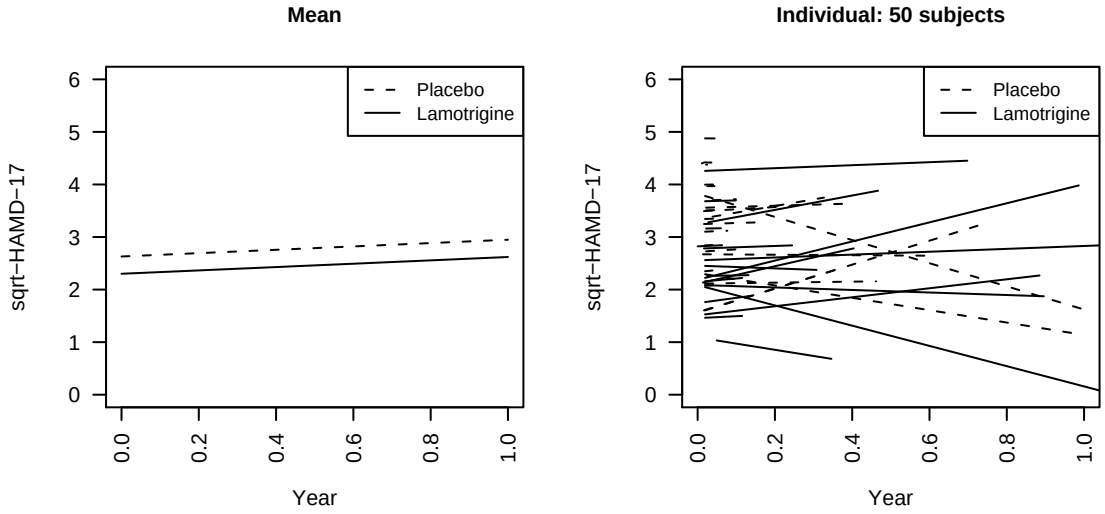
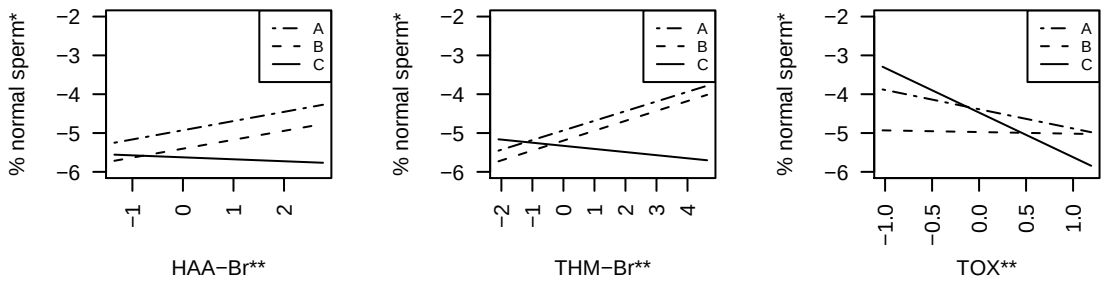


FIGURE 2.3: Predicted mean & individual HAMD-17, with random slope and intercept



* % normal sperm is on the scale of 5 times the probit
** Predictors are centered and scaled by a constant

FIGURE 2.4: Predicted mean of % normal sperm, with random slope and intercept

CHAPTER 3

Testing Variance Components in Multilevel Linear Models using Approximate Bayes Factors

3.1 Introduction

Many studies collect data that have hierarchical or clustered structures. Examples include randomized studies in which patients are clustered within practices, educational studies in which students are clustered in schools, or environmental studies in which individuals are clustered in homes clustered in counties. An analysis that ignores such clustering assumes all observations are independent, resulting in incorrect model-based standard errors that can lead to misleading scientific inferences. Multilevel models are used to account for the correlation of observations within a given group by incorporating group-specific random coefficients. These random coefficients can be be nested (e.g. repeated observations of students nested in schools, with random coefficients at the student and school levels), cross-nested (e.g. repeated observations of students nested in schools participating in different extra-curricular activities, with random coefficients at the school and activity levels), or even non-nested (e.g. individuals clustered within job categories and states, with random coefficients at the job and state level). For an introduction to multilevel models, see Gelman and Hill (2007), Fitzmaurice et al. (2004), Sullivan et al. (1999), and Bryk and Raudenbush (1992).

Birth records were obtained for all live births in New York City in 2003 and linked to the hospital discharge data from the Statewide Planning and Research Cooperative System by the New York State Department of Health. These data include information on mother's demographic characteristics, previous births, smoking, weight gain during pregnancy, maternal birth outside the U.S., and infant's gender, birth weight, and gestational age (Savitz et al., 2008), all collected from the birth certificate. These data were also linked to U.S. Census data to obtain additional demographic information at the census tract level. Investigators are interested in identifying significant predictors of birth weight among term births adjusting for gestational age. To address this, we use a multilevel linear model of infant's birth weight, predicted by infant gestational age, gender, maternal race, parity, smoking status, age, weight gain, nativity, and the neighborhood deprivation index. The neighborhood deprivation index (NDI) is a standardized score of various socioeconomic factors in which higher scores represent higher levels of deprivation, and is measured at the census tract level rather the individual level. In New York City, it is common for individuals with similar demographic characteristics to live in close proximity, resulting in social as well as biological similarities between subjects. Because of these shared characteristics, we consider random coefficients for census tracts in our model.

Research has shown a persistent racial disparity in birth outcomes in the United States (Osypuk and Acevedo-Garcia, 2008). Although individual and community-level covariates have been shown to account for some of the racial risk in low birth weight (Buka et al., 2003; Roberts, 1997; Rauh et al., 2001; OCampo et al., 1997), much of this disparity remains unexplained. Howard et al. (2006) found substantial variability in the risk of preterm birth and low birth weight among black race subgroups defined by maternal ancestry (African, American, Asian, Cuban, European, Puerto Rican, South and Central American, and West Indian and Brazilian). They also found nativity (U.S. or foreign born) to be a significant predictor that varied by ancestry. In addition to race, the NYC birth data has additional information available on maternal country of origin and nativity. We consider random coefficients in our model to allow heterogeneity in birth weights across ethnic ancestries (62 categories), and to allow the effect of race to vary by ancestry. For example, the effect of black race may depend on whether the mother has North African or Jamaican ancestry. In order to determine whether heterogeneity exists in birth weights across ancestries and census tracts, one must be able to test whether

these random coefficients should be included in the model.

Testing whether a random coefficient should be included in a multilevel model involves the test of whether the variance of that random coefficient is equal to 0. This is problematic because the null hypothesis lies on the boundary of the parameter space. Such issues are addressed in the literature in the context of linear mixed models (e.g. Stram and Lee, 1994), but there is very little research specifically for testing variance components in multilevel models. Berkhof and Snijders (2001) proposed three score tests for variance components in multilevel models and compared their method via simulation to the likelihood ratio test, fixed F test, and Wald test. However, their simulations only considered two level models, and it is not clear whether generalizations to a larger number of levels are possible. Fitzmaurice et al. (2007) proposed a permutation test for variance components in multilevel generalized linear mixed models. They applied their method to two-level generalized mixed models and suggested strategies for multilevel models with greater than two levels. However, their strategy cannot be directly applied to multilevel models with crossed random effects and can only test one variance component at a time. Frequentist methods for testing variance components in the linear mixed model are useful to some extent in nested multilevel models for testing single variance components (e.g. Crainiceanu and Ruppert, 2004; Verbeke and Molenberghs, 2003), but the null distributions are not easily obtained for testing multiple variance components, and it is not clear whether these methods can be applied to non-nested variance components. Also, Bayesian MCMC methods for testing variance components in the linear mixed model (e.g. Cai and Dunson, 2006; Kinney and Dunson, 2008) may be generalizable to multilevel models, but these methods generally suffer from computational constraints and rely on subjective choice of hyperparameters.

The potential complexity of multilevel linear models with multiple nested or non-nested random coefficients makes an approach using Bayes factors particularly challenging. In particular, one must address issues arising from testing on the boundary of the parameter space, poor performance of approximations to the Bayes factor resulting from high-dimensionality, and the specification of default non-informative priors on the random coefficients. We propose to extend the approach of Saville and Herring (2008) by scaling the random coefficients to the residual variance and introducing parameters that control the relative contribution of the random coefficients. After integrating over the random coefficients and variance components, the resulting

integrals needed to calculate the Bayes factor can be efficiently approximated with Laplace's method. The method also incorporates default prior distributions that were shown to have good frequentist properties in the linear mixed model (Saville and Herring, 2008).

We present the multilevel model and Bayesian model selection problem in Section 2. We discuss methods for approximating the marginal likelihoods in Section 3. We conduct simulation studies in Section 4 and apply our method to the NYC birth data in Section 5. We conclude with a discussion in Section 6.

3.2 Testing random coefficients in multilevel linear models

We define the general multilevel linear model with q random factors as

$$\begin{aligned} Y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_{[i]} + \varepsilon_i, \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{h=1}^{q} \boldsymbol{z}_{ih}'\boldsymbol{b}_{h[i]} + \varepsilon_i, \end{aligned} \tag{3.1}$$

in which Y_i is the response for observation i, $i = 1, \ldots, m$, $\boldsymbol{x}_i$ is a $p \times 1$ vector of predictors with corresponding fixed effects $\boldsymbol{\beta}$, $\boldsymbol{b}_{[i]} = (\boldsymbol{b}_{1[i]}', \ldots, \boldsymbol{b}_{q[i]}')'$, $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}', \ldots, \boldsymbol{z}_{iq}')'$, $\boldsymbol{z}_{ih}$ is a $d_h \times 1$ vector of predictors with corresponding random effects $\boldsymbol{b}_{h[i]}$ in which $[i]$ indexes the group in factor h pertaining to the ith observation, and $\boldsymbol{b}_{h[i]} \sim N(\boldsymbol{0}, \boldsymbol{\psi}_h)$ independent of $\varepsilon_i \sim N(0, \sigma^2)$, with $\boldsymbol{b}_{h[i]}$ independent of $\boldsymbol{b}_{h'[i]}$ for $h \neq h'$. A key feature of multilevel modeling is the incorporation of covariates $\boldsymbol{x}_i$ that can be measured at any level of the hierarchy. This allows one to address the effect of a given covariate, say at the individual level, while controlling for the effect of a higher level covariate, say at the census level. However, greater care is required in the interpretation of regression parameters, because some covariates can operate at many different levels.

To illustrate, consider the NYC birth data for 2003, in which there are 104,710 observations within 62 ethnic ancestries and 2,128 census tracts. The aims of our analysis are to identify significant predictors of infant birth weight and to determine whether there is heterogeneity across ancestry groups and census tracts. To start, we will consider the predictor maternal weight gain during pregnancy, which has been linked to infant birth weight. Because of social

and biological characteristics shared by persons of the same ancestry, the effect of maternal weight gain may vary by ancestry. This can be evaluated with a non-nested multilevel linear model, with a random intercept and slope (for weight gain) at the ancestry level and a random intercept at the census level. The model is

$$Y_i = \beta_0 + x_i\beta_1 + b_{10[i]} + b_{11[i]}x_i + b_{20[i]} + \varepsilon_i, \tag{3.2}$$

in which Y_i is the weight of infant i, x_i is the weight gain of the ith mother, β_0 is the model intercept, β_1 is the parameter corresponding to weight gain, $b_{10[i]}$ is the random intercept and $b_{11[i]}$ the random slope corresponding to the ancestry of mother i, and $b_{20[i]}$ is the random intercept corresponding to the census tract of mother i. There are a total of $2 \times 62 = 124$ random coefficients at the ancestry level and 2,128 random coefficients at the census level. In order to test whether there is heterogeneity in birth weights across ancestries ($h = 1$) or census tracts ($h = 2$), one can conduct a test of whether the variance of the respective random coefficients is equal to 0. This corresponds to a test of $H_0 : \boldsymbol{\psi}_h = \mathbf{0}$, which lies on the boundary of the parameter space.

3.2.1 Bayes factors

From a Bayesian perspective, we can test $H_0 : \boldsymbol{\psi}_h = \mathbf{0}$ by calculating the Bayes factor, or posterior odds of M_1 versus M_0 given equal prior odds, given by

$$B_{10} = \frac{p(\boldsymbol{Y}|M_1)}{p(\boldsymbol{Y}|M_0)}, \tag{3.3}$$

in which M_0 is model corresponding to the null hypothesis and M_1 is the model corresponding to the alternative hypothesis. Calculating the Bayes factor requires the marginal likelihood

$$p(\boldsymbol{Y}|M_k) = \int p(\boldsymbol{Y}|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k, \tag{3.4}$$

in which $p(\boldsymbol{Y}|\boldsymbol{\theta}_k, M_k)$ is the data likelihood for model M_k, $\boldsymbol{\theta}_k$ is the vector of model parameters, and $\pi(\boldsymbol{\theta}_k|M_k)$ is the prior distribution of $\boldsymbol{\theta}_k$. Multilevel models typically have a large number

of parameters due to the inclusion of random coefficients. This is problematic in calculating Bayes factors because high dimensional integrals are needed to calculate marginal likelihoods. Generally these integrals are not available in closed form, and one must consider approximations. Monte Carlo integration and importance sampling provide alternatives, but these methods lack accuracy and are computationally demanding. The Laplace and Bayesian Information Criterion (BIC) (Schwarz, 1978) approximations also suffer in performance from high-dimensionality (Kass and Raftery, 1995), and it is not clear how to define the penalty for dimensionality in the BIC (Spiegelhalter et al., 2002).

It is well known that Bayes factors can be sensitive to the choice of prior distributions (Kass and Raftery, 1995). This is challenging in model selection problems in which one has no prior information on the parameters. In these situations it is common to use default priors that do not require subjective inputs. However, one must choose these default priors with care, because as the prior variance increases the Bayes factor will increasingly favor the null model (Bartlett, 1957). Our goal is to propose a method that incorporates default priors on the random coefficients that result in good frequentist properties with respect to power and Type I error. Also, we aim to avoid issues with the boundary of the parameter space and high-dimensional approximations to the Bayes factor.

3.3 Approximating the marginal likelihood

3.3.1 Reparameterization

To introduce our method, we first give a modified notation for the multilevel linear model. Let

$$\begin{aligned} Y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{w}_i'\boldsymbol{b} + \varepsilon_i \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{h=1}^{q} \boldsymbol{w}_{ih}'\boldsymbol{b}_h + \varepsilon_i \end{aligned} \tag{3.5}$$

in which $\boldsymbol{w}_i = (\boldsymbol{w}_{i1}', \ldots, \boldsymbol{w}_{iq}')'$, $\boldsymbol{w}_{ih}$ is an $(r_h \times 1)$ vector of predictors with corresponding random effects $\boldsymbol{b}_h$, and $r_h = d_h c_h$ is the total number of random coefficients for factor h (d_h is the number of random coefficients for one observation for factor h, and c_h is the total number of

classifications for factor h). More specifically, $\boldsymbol{w}_{ih} = [\boldsymbol{\delta}_i \otimes \boldsymbol{z}_{ih}]$, in which $\boldsymbol{\delta}_i$ is a $(c_h \times 1)$ vector of indicator variables (equals 1 if yes, 0 if no) for group membership of observation i in each of the c_h classifications, and $\otimes$ denotes the left Kronecker product. The dimension of $\boldsymbol{w}_i$ is $(r \times 1)$, with $r = \sum_{h=1}^{q} r_h$ the total number of random coefficients in the model. Also, $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_q')'$, in which $\boldsymbol{b}_h = (\boldsymbol{b}_{h1}', \ldots, \boldsymbol{b}_{hc_h}')'$ is the vector of all random coefficients for factor h. We assume $\boldsymbol{b}_{hl} \sim N_{d_h}(\boldsymbol{0}_{d_h}, \boldsymbol{\psi}_h)$ independent of $\varepsilon_i \sim N(0, \sigma^2)$.

Extending the work of Saville and Herring (2008), we scale the random coefficients to the residual variance such that $\tilde{\boldsymbol{b}}_{hl} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. We then express the model as

$$Y_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{w}_i'\boldsymbol{\Phi}\boldsymbol{\Gamma}\tilde{\boldsymbol{b}} + \varepsilon_i, \tag{3.6}$$

in which $\tilde{\boldsymbol{b}}$ is the vector of scaled random coefficients, $\boldsymbol{\Phi} = \text{diag}(\exp(\boldsymbol{\phi}_1'^*, \ldots, \boldsymbol{\phi}_q'^*))$ with $\boldsymbol{\phi}_h^* = (\boldsymbol{1}_{c_h} \otimes \boldsymbol{\phi}_h)$, and $\boldsymbol{\phi}_h = (\phi_{h1}, \ldots, \phi_{hd_h})'$ are parameters that control the relative contribution of the random coefficients. Also, $\boldsymbol{\Gamma} = \text{blockdiag}(\boldsymbol{\Gamma}_1^*, \ldots, \boldsymbol{\Gamma}_q^*)$ with $\boldsymbol{\Gamma}_h^* = (\boldsymbol{I}_{c_h} \otimes \boldsymbol{\Gamma}_h)$, in which $\boldsymbol{\Gamma}_h$ is a lower triangular matrix with $\boldsymbol{1}_{d_h}$ along the diagonal, and lower off-diagonal elements $\boldsymbol{\gamma}_h$ that induce correlation between the random coefficients within factor h. We can also express the model in the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{\Gamma}\tilde{\boldsymbol{b}} + \boldsymbol{\varepsilon}, \tag{3.7}$$

in which $\boldsymbol{Y} = (Y_i, \ldots, Y_m)'$, $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)'$.

Let $\sigma^2 \sim \text{InvGam}(v, w)$. By integrating out $\tilde{\boldsymbol{b}}$ and σ^2 from the posterior distribution, the marginal posterior $p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$ can be shown to have the multivariate t-distribution given by

$$p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}) = \Gamma\left(\frac{2v+p}{2}\right) \frac{(\pi 2v)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}}{\Gamma(2v/2)} \left\{1 + \frac{1}{2v}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right\}^{-\frac{2v+p}{2}}, \tag{3.8}$$

in which $\Gamma()$ denotes the gamma function and $\boldsymbol{\Sigma} = (\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Phi}'\boldsymbol{W}' + \boldsymbol{I}_m)$. We assume the default prior $\phi_{hl} \sim \log N(\log(0.3), 2)$ suggested by Saville and Herring (2008), and use the Laplace method to integrate over $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$ to obtain the marginal density $p(\boldsymbol{Y})$. This default prior was shown to have good frequentist properties in simulation studies in the linear mixed model.

3.3.2 Computational considerations

Product of likelihoods

For studies with large sample size m, the covariance matrix $\boldsymbol{\Sigma}$ may be too large to handle computationally. For example, in applying model (3.2) to the complete 2003 NYC data (m =104,710), the covariance matrix $\boldsymbol{\Sigma}$ is (104,710 $\times$ 104,710). We note that this matrix has the potential to be extremely sparse, and even with very large m may be computationally feasible using sparse matrix computations. However, when the matrix is large and not sufficiently sparse, it may be advantageous to work with the product of independent likelihoods (conditional on the random coefficients) as opposed to the likelihood of the vector of response variables. To illustrate, the marginal distribution can be written as

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma}) &= \int p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma},\tilde{\boldsymbol{b}},\sigma^2)\pi(\tilde{\boldsymbol{b}})\pi(\sigma^2)d\tilde{\boldsymbol{b}}d\sigma^2 \\ &= \int \left[\prod_{i=1}^{m} p(Y_i|\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma},\tilde{\boldsymbol{b}},\sigma^2)\right]\pi(\tilde{\boldsymbol{b}})\pi(\sigma^2)d\tilde{\boldsymbol{b}}d\sigma^2 \\ &= \frac{\Gamma\left(\frac{2v+m}{2}\right)|\boldsymbol{A}|^{-1/2}}{(\pi 2v)^{m/2}\Gamma(2v/2)}\left\{1+\frac{1}{2v}\left(f(\boldsymbol{Y})-\boldsymbol{C}'\boldsymbol{A}^{-1}\boldsymbol{C}\right)\right\}^{-\frac{2v+m}{2}} \end{aligned} \tag{3.9}$$

with $\boldsymbol{A}=\{\boldsymbol{I}_r+\boldsymbol{\Gamma}'\boldsymbol{\Phi}'\left(\sum_{i=1}^{m}\boldsymbol{w}_i\boldsymbol{w}_i'\right)\boldsymbol{\Phi}\boldsymbol{\Gamma}\}$, $\boldsymbol{C}=\boldsymbol{\Gamma}'\boldsymbol{\Phi}'\left\{\sum_{i=1}^{m}\boldsymbol{w}_i(Y_i-\boldsymbol{x}_i'\boldsymbol{\beta})\right\}$, and $f(\boldsymbol{Y})=\sum_{i=1}^{m}(Y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2$, in which $\boldsymbol{I}_r$ denotes the identity matrix with dimension $(r\times r)$.

Using this approach, it should be computationally possible to approximate the marginal likelihood regardless of the size of m. The computation is limited, however, by the total number of random coefficients r. If r is very large, it may not be feasible to compute the inverse and determinant of the $(r\times r)$ matrix $\boldsymbol{A}$ (or may be very computationally expensive). For example, in applying (3.2) to the NYC data, $r=2,252$. Although it may be possible to compute the inverse and determinant of $\boldsymbol{A}$ in this example, computations are likely to be very slow. Hence, an alternative computational approach is to write the data likelihood as products of marginal likelihoods for lower-dimensional response vectors or scalars.

Alternative for non-nested models

Consider the NYC data in which there are two non-nested factors, ancestry and census tracts. We denote the factor with fewer groups as $h = 1$ (ancestry) and the factor with a larger number of groups as $h = 2$ (census tracts). We can write the marginal likelihood as

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma}) &= \int p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\phi},\tilde{\boldsymbol{b}}_2,\tilde{\boldsymbol{b}}_1,\sigma^2)\pi(\tilde{\boldsymbol{b}}_2)\pi(\tilde{\boldsymbol{b}}_1)\pi(\sigma^2)d\tilde{\boldsymbol{b}}_2 d\tilde{\boldsymbol{b}}_1 d\sigma^2, \\ &= \int \left\{ \prod_{k=1}^{c_2} p(\boldsymbol{Y}_k|\boldsymbol{\beta},\boldsymbol{\phi},\tilde{\boldsymbol{b}}_{2k},\tilde{\boldsymbol{b}}_1,\sigma^2) \right\} \pi(\tilde{\boldsymbol{b}}_2)\pi(\tilde{\boldsymbol{b}}_1)\pi(\sigma^2)d\tilde{\boldsymbol{b}}_2 d\tilde{\boldsymbol{b}}_1 d\sigma^2, \\ &= \int \left\{ \prod_{k=1}^{c_2} \int p(\boldsymbol{Y}_k|\boldsymbol{\beta},\boldsymbol{\phi},\tilde{\boldsymbol{b}}_{2k},\tilde{\boldsymbol{b}}_1,\sigma^2)\pi(\tilde{\boldsymbol{b}}_{2k})d\tilde{\boldsymbol{b}}_{2k} \right\} \pi(\tilde{\boldsymbol{b}}_1)\pi(\sigma^2)d\tilde{\boldsymbol{b}}_1 d\sigma^2 \\ &= \int \left\{ \prod_{k=1}^{c_2} \int \left[\prod_{i=1}^{m_k} p(Y_{ki}|\boldsymbol{\beta},\boldsymbol{\phi},\tilde{\boldsymbol{b}}_{2k},\tilde{\boldsymbol{b}}_1,\sigma^2) \right] \pi(\tilde{\boldsymbol{b}}_{2k})d\tilde{\boldsymbol{b}}_{2k} \right\} \pi(\tilde{\boldsymbol{b}}_1)\pi(\sigma^2)d\tilde{\boldsymbol{b}}_1 d\sigma^2, \end{aligned} \tag{3.10}$$

in which c_2 is the number of groups in factor 2, $\boldsymbol{Y}_k$ is the vector of responses for group k in factor 2, $\tilde{\boldsymbol{b}}_2$ are the random coefficients for factor 2, $\tilde{\boldsymbol{b}}_{2k}$ are the random coefficients corresponding to group k in factor 2, $\tilde{\boldsymbol{b}}_1$ are the random coefficients for factor 1, m_k is the number of subjects in group k of factor 2 and Y_{ki} is the response of the ith subject in group k of factor 2. This approach allows one to integrate out the random coefficients for factor 2 in smaller dimensions, as $\tilde{\boldsymbol{b}}_{2k}$ is only a $(d_2 \times 1)$ vector. For model (3.2) applied to the NYC data, $\tilde{\boldsymbol{b}}_{2k}$ is a scalar (representing a random intercept for census tract k) and results in matrices with smaller dimensions than those obtained from (3.9).

Alternative for nested models

Although not of particular interest in the NYC data, one could consider a 3-level nested design with subjects nested within census tracts nested within boroughs (there are 5 boroughs in NYC). In such cases one can use the nested structure for easier computation. Let $h = 1$ denote the census tract factor and $h = 2$ denote the borough factor. Then

$$\begin{aligned} & p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma}) = \int p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\phi},\tilde{\boldsymbol{b}}_2,\tilde{\boldsymbol{b}}_1,\sigma^2)\pi(\tilde{\boldsymbol{b}}_2)\pi(\tilde{\boldsymbol{b}}_1)\pi(\sigma^2)d\tilde{\boldsymbol{b}}_2 d\tilde{\boldsymbol{b}}_1 d\sigma^2 \\ = & \int \left\{ \prod_{k=1}^{c_2} p(\boldsymbol{Y}_k|\boldsymbol{\beta},\boldsymbol{\phi},\tilde{\boldsymbol{b}}_{2k},\tilde{\boldsymbol{b}}_{1k},\sigma^2) \right\} \pi(\tilde{\boldsymbol{b}}_2)\pi(\tilde{\boldsymbol{b}}_1)\pi(\sigma^2)d\tilde{\boldsymbol{b}}_2 d\tilde{\boldsymbol{b}}_1 d\sigma^2 \\ = & \int \left\{ \prod_{k=1}^{c_2} \int p(\boldsymbol{Y}_k|\boldsymbol{\beta},\boldsymbol{\phi},\tilde{\boldsymbol{b}}_{2k},\tilde{\boldsymbol{b}}_{1k},\sigma^2)\pi(\tilde{\boldsymbol{b}}_{2k})\pi(\tilde{\boldsymbol{b}}_{1k})d\tilde{\boldsymbol{b}}_{2k} d\tilde{\boldsymbol{b}}_{1k} \right\} \pi(\sigma^2)d\sigma^2 \end{aligned} \tag{3.11}$$

$$
\begin{aligned}
&= \int \left\{ \prod_{k=1}^{c_2} \int \left[\prod_{j=1}^{c_{1k}} p(\boldsymbol{Y}_{kj}|\boldsymbol{\beta}, \boldsymbol{\phi}, \tilde{\boldsymbol{b}}_{2k}, \tilde{\boldsymbol{b}}_{1kj}, \sigma^2) \right] \pi(\tilde{\boldsymbol{b}}_{2k}) \pi(\tilde{\boldsymbol{b}}_{1k}) d\tilde{\boldsymbol{b}}_{2k} d\tilde{\boldsymbol{b}}_{1k} \right\} \pi(\sigma^2) d\sigma^2 \\
&= \int \left\{ \prod_{k=1}^{c_2} \int \left[\prod_{j=1}^{c_{1k}} \int p(\boldsymbol{Y}_{kj}|\boldsymbol{\beta}, \boldsymbol{\phi}, \tilde{\boldsymbol{b}}_{2k}, \tilde{\boldsymbol{b}}_{1kj}, \sigma^2) \pi(\tilde{\boldsymbol{b}}_{1kj}) d\tilde{\boldsymbol{b}}_{1kj} \right] \pi(\tilde{\boldsymbol{b}}_{2k}) d\tilde{\boldsymbol{b}}_{2k} \right\} \pi(\sigma^2) d\sigma^2 \\
&= \int \left\{ \prod_{k=1}^{c_2} \int \left[\prod_{j=1}^{c_{1k}} \int \left(\prod_{i=1}^{m_{kj}} p(Y_{kji}|\boldsymbol{\beta}, \boldsymbol{\phi}, \tilde{\boldsymbol{b}}_{2k}, \tilde{\boldsymbol{b}}_{1kj}, \sigma^2) \right) \pi(\tilde{\boldsymbol{b}}_{1kj}) d\tilde{\boldsymbol{b}}_{1kj} \right] \pi(\tilde{\boldsymbol{b}}_{2k}) d\tilde{\boldsymbol{b}}_{2k} \right\} \pi(\sigma^2) d\sigma^2,
\end{aligned}
$$

in which c_{1k} is the number of groups for factor 1 within group k of factor 2, m_{kj} is the number of subjects in group j of factor 1 within group k of factor 2, $\boldsymbol{Y}_{kj}$ is the response vector for subjects in group j of factor 1 within group k of factor 2, Y_{kji} is the response of subject i within group j of factor 1 within group k of factor 2, $\tilde{\boldsymbol{b}}_{1k}$ are the random coefficients for factor 1 within group k of factor 2, and $\tilde{\boldsymbol{b}}_{1kj}$ are the random coefficients corresponding to group j of factor 1 within group k of factor 2. This approach allows one to integrate out the random coefficients $\tilde{\boldsymbol{b}}_{1kj}$ and $\tilde{\boldsymbol{b}}_{2k}$ which have smaller dimensions equal to $(d_1 \times 1)$ and $(d_2 \times 1)$, respectively. For the NYC data with a random intercept for census tracts and boroughs, $\tilde{\boldsymbol{b}}_{1kj}$ and $\tilde{\boldsymbol{b}}_{2k}$ are both scalars.

If there are non-nested random coefficients in addition to nested random coefficients (i.e. cross-nested) and either m or r is too large for computational feasibility, then similar strategies can be used to decrease the dimensions of the required integrals. For example, such strategies could be used on the NYC data with factors for ancestry and census tracts nested within boroughs. However, given there are only 5 boroughs in the NYC data, incorporating random coefficients at the borough level is not of particular interest for this example.

3.4 Simulation study

3.4.1 Testing random intercepts

We conducted a simulation study to evaluate the performance of our method in correctly identifying models with or without random intercepts. We consider a simple setting with two non-nested factors with 30 classifications each. We simulated $b_{10[i]} \sim N(0,1)$, $b_{20[i]} \sim N(0,1)$, $\varepsilon_i \sim N(0,1)$, and calculated

$$
Y_i = \lambda_1 b_{10[i]} + \lambda_2 b_{20[i]} + \varepsilon_i, \tag{3.12}
$$

for various combinations of $m = (100, 500, 1000)$, $\lambda_1 = (0, 0.1, 0.2, 0.3)$, and $\lambda_2 = (0, 0.1, 0.2, 0.3)$ for 1,000 datasets. Using prior distributions $\beta_0 \sim N(0, 1)$, $\sigma^2 \sim \text{InvGam}(.1, .1)$ (which are non-informative given the simulation settings), and $\phi_h \sim N(\log(0.3), 2)$, we approximated marginal likelihoods for the following models:

$$\begin{aligned} M_0 : \ Y_i &= \beta_0 + \varepsilon_i, \\ M_1 : \ Y_i &= \beta_0 + e^{\phi_1} b_{10[i]} + \varepsilon_i, \\ M_2 : \ Y_i &= \beta_0 + e^{\phi_2} b_{20[i]} + \varepsilon_i, \\ M_3 : \ Y_i &= \beta_0 + e^{\phi_1} b_{10[i]} + e^{\phi_2} b_{20[i]} + \varepsilon_i, \end{aligned} \tag{3.13}$$

in which $\phi_h = \log(\lambda_h)$ for $\lambda_h > 0$ and $h = 1, 2$. Estimates of the Bayes factors $\hat{B}_{30}$, $\hat{B}_{10}$, $\hat{B}_{20}$ were calculated for each data set and interpreted according to the scale given by Wasserman (2000) and Jeffreys (1961). For comparison with frequentist methods, we chose to reject H'_k if an estimated Bayes factor $M_{kk'}$ was greater than 1, in which model k was preferred over model k'. In this simple setting, we can use the restricted likelihood ratio test for testing M_1 and M_2 versus M_0, in which the null distributions follow a 50:50 mixture of a point mass at 0 and a chi-square distribution with 1 degree of freedom (denoted as LR_{10} and LR_{20}) (Self and Liang, 1987; Stram and Lee, 1994). We can also test M_1 and M_2 versus M_0 using the ANOVA F-test (denoted as AOV_{10} and AOV_{20}). For testing M_3 versus the other models, we implement an ad-hoc restricted likelihood ratio test, in which the standard test statistic is compared at the $\alpha = 0.10$ level to a chi-square distribution with degrees of freedom equal to the difference in the number of variance components in the models being compared (denoted as LR^*_{30}, LR^*_{31}, and LR^*_{32}). Although this approach may not be recommended from a theoretical perspective (Fitzmaurice et al., 2004), it is known to be used in practice.

In the absence of random effects, the Bayes factor approach, likelihood ratio tests, ANOVA F-tests, and ad-hoc tests all preserved the nominal Type I error rate at 0.05 for all model comparisons and all sample sizes (Table 3.1). The power for $\hat{B}_{10}$ and $\hat{B}_{20}$ in detecting a random effect was very similar to the likelihood ratio tests LR_{10} and LR_{20} and the ANOVA F-tests. For testing M_3 versus M_0, the performance of $\hat{B}_{30}$ was similar to the ad-hoc LR^*_{30}, with slighter greater power for small sample sizes and slightly less power for larger sample sizes. A similar

pattern was seen comparing $\hat{B}_{31}$ versus LR^*_{31} and B_{32} versus LR^*_{32}. These results support the claim that our method has good frequentist properties with respect to power and Type I error.

Tables 3.2-3.4 shows a more complete breakdown of the estimated Bayes factors according to the scale of Wasserman (2000) and Jeffreys (1961). As λ_1 and λ_2 increased, the estimated Bayes factor displayed greater evidence for the model with random intercepts. As the sample size increased, the estimated Bayes factors increasingly favored the null model in the absence of random intercepts, and increasingly favored the random intercept models in the presence of random intercepts. This shows large sample consistency in our method under these simulation settings.

3.4.2 Testing a random slope

We extend our simulation to test for the presence of a random slope in a two-factor non-nested multilevel model. To simulate the data, we include random intercepts for each factor as done previously, but also incorporate a random slope for one of the factors. We simulated $x_i \sim N(0, .25)$, $b_{20[i]} \sim N(0, 0.04)$, $\varepsilon_i \sim N(0, 1)$, $\boldsymbol{b}_{1[i]} \sim N_2(\mathbf{0}, \boldsymbol{\psi})$, with $\psi_{11} = 0.04$, $\psi_{12} = \rho\sqrt{\psi_{11}\psi_{22}}$, and $\rho = -0.3$, which induces a negative correlation between the random intercept and slope. The variances of the random intercepts (0.04) were chosen to match the variances from the previous simulation corresponding to $\lambda_1 = \lambda_2 = 0.2$. We calculated

$$Y_i = b_{10[i]} + b_{20[i]} + b_{11[i]}x_i + \varepsilon_i, \tag{3.14}$$

for various combinations of $m = (100, 500, 1000)$ and $\sqrt{\psi_{22}} = (0, 0.1, 0.2, 0.3, 0.6, 1.0)$ for 1,000 datasets. Using prior distributions $\beta_0 \sim N(0, 1)$, $\sigma^2 \sim \text{InvGam}(.1, .1)$ (which are non-informative given the simulation settings), and $\phi_h \sim N(\log(0.3), 2)$, we approximated marginal likelihoods for the following models:

$$\begin{aligned} M_3: \; Y_i &= \beta_0 + \beta_1 x_i + e^{\phi_1} b_{10[i]} + e^{\phi_2} b_{20[i]} + \varepsilon_i \\ M_4: \; Y_i &= \beta_0 + \beta_1 x_i + e^{\phi_{10}} b_{10[i]} + e^{\phi_{20}} b_{20[i]} + e^{\phi_{11}} b^*_{11[i]} x_i + \varepsilon_i, \end{aligned} \tag{3.15}$$

in which $b^*_{11[i]} = \gamma_1 b_{10[i]} + b_{11[i]}$. Model M_3 incorporates random intercepts for both factors with a fixed effect for the covariate, and model M_4 includes the additional random slope on the covariate for factor 1. Table 3.5 gives the power and Type I error of our approach using approximate Bayes factors and the ad-hoc restricted likelihood ratio test. Our method preserves the Type I error rate at $\alpha = 0.05$ and has similar power to the ad-hoc RLRT. Table 3.6 shows the estimated Bayes factors according to the scale of Wasserman (2000) and Jeffreys (1961). As $\sqrt{\psi_{22}}$ increased, the estimated Bayes factor displayed greater evidence for the model with the random slope. As the sample size increased, the estimated Bayes factor increasingly favored M_3 in the absence of a random slope, and increasingly favored M_4 in the presence of a random slope. These simulation results support the claim that our method has good frequentist properties and large sample consistency.

3.4.3 Choice of prior distributions

Saville and Herring (2008) considered several alternative prior distributions for this method in the context of the linear mixed model. More specifically, the authors conducted simulations with priors of the form $\phi_{hl} \sim \mathrm{N}(h, \zeta)$, with various combinations of $h = \log(1), \log(0.3), \log(0.15)$ and $\zeta = 1, 2, 3$. Additionally, they considered a t-distribution for ϕ_{hl} with 2 and 10 degrees of freedom, as well as prior distributions $\sigma^2 \propto \sigma^{-2}$, $\sigma^2 \sim$InvGamma(0.1,0.1), $\sigma^2 \sim$InvGamma(0.01,0.01), and $\boldsymbol{\beta} \propto c$ in which c is a constant. They found that alternative priors on σ^2 and $\boldsymbol{\beta}$ did not have notable influence on the estimated Bayes factors, but the priors for ϕ_{hl} did have some influence. More specifically, values of $h = \log(0.30)$ and $\zeta = 2$ resulted in power and Type I error rates that closely aligned with standard frequentist methods. Smaller values of h or ζ led to increased Type I error rates and larger values of h or ζ led to more conservative Type I error rates. Given that simulation results for the multilevel linear model using the default prior are similar to those obtained from the linear mixed model, we would expect to observe similar patterns based on alternative prior distributions in the multilevel linear model.

3.5 Application

We are interested in fitting a multilevel linear model to infant's birth weight, predicted by infant gestational age, gender, maternal race, parity, smoking status, age, weight gain, maternal nativity, and the neighborhood deprivation index, with random coefficients for census tracts and ethnic ancestries. We focus on singleton term births with a gestational age ≥ 37 weeks and a birth weight between 900 g and 5300 g. After exclusions, we have a total of 93,938 subjects with complete data available for the analysis.

The first model we investigate allows a random intercept for ancestry, defined as

$$M_1 : \; Y_i \;\; = \;\; \boldsymbol{x}_i'\boldsymbol{\beta} + b_{1[i]} + \varepsilon_i, \tag{3.16}$$

with

$$\begin{aligned} \boldsymbol{x}_i'\boldsymbol{\beta} \;\; &= \;\; \beta_0 + \beta_1 \text{Black}_i + \beta_2 \text{Hisp}_i + \beta_3 \text{Asian}_i + \beta_4 \text{Other}_i + \beta_5 \text{Gest}_i \\ &+ \;\; \beta_6 \text{Gest}_i^2 + \beta_7 \text{Pbirth}_i + \beta_8 \text{Female}_i + \beta_9 \text{Smoke}_i + \beta_{10} \text{NDI}_i \\ &+ \;\; \beta_{11} \text{Age2}_i + \beta_{12} \text{Age3}_i + \beta_{13} \text{Age4}_i + \beta_{14} \text{Age5}_i + \beta_{15} \text{Nativity}_i + \\ &+ \;\; \beta_{16} \text{Wtgain}_i + \beta_{17} \text{Wtgain}_i^2 + \beta_{18} \text{Wtgain}_i^3, \end{aligned} \tag{3.17}$$

in which $b_{1[i]}$ is the random intercept corresponding to the ancestry of subject i. The explanatory variables Black_i, Hisp_i, Asian_i, and Other_i are indicator variables for race corresponding to black, Hispanic, Asian or Pacific Islander, and other (white as the referent group). Gest_i is the gestational age of the infant for subject i and Gest_i^2 is the corresponding quadratic variable. The variables Pbirth_i, Female_i, Smoke_i, and Nativity_i are indicator variables for any previous births, female infant gender, maternal smoking, and maternal birth outside of the United States, respectively. Maternal age was categorized into the following groups: $<$ 25yrs (referent group), 26-30 yrs (Age2_i), 31-35 yrs(Age3_i), 36-40 yrs (Age4_i), and $>$ 40 yrs (Age5_i). The variable NDI_i is the neighborhood deprivation index corresponding to the census tract of subject i, and Wtgain_i is the difference in maternal pre-pregnancy weight and weight at delivery. The continuous variables NDI_i, Gest_i, and Wtgain_i are centered and standardized by 2 standard

deviations to place the regression coefficients on the same scale as the binary indicators (Gelman, 2008).

We also consider a model with a random intercept for census tracts but without random coefficients for ancestries,

$$M_2: \; Y_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + b_{2[i]} + \varepsilon_i, \tag{3.18}$$

in which $b_{2[i]}$ is the random intercept corresponding to the census tract of subject i. Incorporating random intercepts for both ancestries and census tracts, a two-factor non-nested model takes the form

$$M_3: \; Y_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + b_{1[i]} + b_{2[i]} + \varepsilon_i. \tag{3.19}$$

As discussed previously, the effect of race may depend on maternal ancestry. Hence we consider a variation of M_3 with random intercepts for both ancestry and census tract, but we allow the effect of race to vary by ancestry. This model can be written as

$$M_4: \; Y_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + b_{1p[i]} + b_{2[i]} + \varepsilon_i, \tag{3.20}$$

in which $b_{1p[i]}$ is the random intercept corresponding to the ancestry (factor 1) of subject i within race p. This model assumes that two persons of the same ancestry with different races have different random intercepts. Similarly, it may be the case that the effect of ancestry varies by nativity. Hence we consider

$$M_5: \; Y_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + b_{1s[i]} + b_{2[i]} + \varepsilon_i, \tag{3.21}$$

in which $b_{1s[i]}$ is the random intercept corresponding to the ancestry (factor 1) of subject i within nativity s. This model assumes that two persons of the same ancestry with opposite nativity have different random intercepts. Additionally, it may be the case that the effect of maternal weight gain on infant birth weight is affected by ancestry. This may result from either biological or social factors that are correlated with a given ancestry. We can model this heterogeneity by

including a random slope for weight gain for the ancestry factor. Adding this component to model M_3, we have

$$M_6: \; Y_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + b_{10[i]} + b_{2[i]} + b_{11[i]}\text{Wtgain}_i + \varepsilon_i, \tag{3.22}$$

in which $b_{11[i]}$ is the random slope for weight gain corresponding to the ancestry of subject i. Finally, we consider a model without random effects,

$$M_0: \; Y_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{3.23}$$

Our goal is to identify the preferred model using approximate Bayes factors, and to proceed with inference using this chosen model.

The mean value for infant birth weight is 3,362 grams with a standard deviation of 460 g. Converting to kilograms for computational convenience, we use prior distributions $\beta_0 \sim N(3.36, 1)$, $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{I})$, and $\sigma^2 \sim N(0.1, 0.1)$, which are non-informative priors given the scale of the response and predictors. We found very strong evidence for heterogeneity in birth weights across census tracts and across ancestries ($\log \hat{B}_{10} = 280$, $\log \hat{B}_{20} = 32$, and $\log \hat{B}_{30} = 284$), with birth weights tending to vary across maternal ancestries in greater magnitude than across census tracts. We found that the effects of race ($\log \hat{B}_{43} = -6$), nativity ($\log \hat{B}_{53} = -11$) and maternal weight gain ($\log \hat{B}_{63} = -1$) do not vary by ancestry.

We fit the preferred model, M_3, using MCMC methods and base inference on 20,000 samples after discarding 5,000. The posterior means and 95% credible intervals of the fixed effects are given in Table 3.7. Results are presented in grams for better interpretability. Predictors with 95% credible intervals greater than 0 include parity (99,111), maternal age 26-30 (45,60), maternal age 31-35 (64,80), maternal age 36-40 (75,94), maternal age >40 (60,92), and maternal nativity (3,18). Hence, previous live births, greater maternal age, and maternal birth outside the U.S. are all associated with greater infant birth weights. Predictors with 95% credible intervals that are less than 0 include maternal Asian race (-91,-24), black race $(-74, -7)$, infant female gender (-126,-115), maternal smoking (-186,-143), and higher neighborhood deprivation (95% CI=(-23,-9) for a 2 sd increase). Hence, Asian and black race (compared to white), female

infants (compared to males), smokers (compared to non-smokers), and greater NDI values are associated with lower infant birth weights. Both maternal weight gain and infant gestational age showed non-linear associations with infant birth weight. The linear effect for a 2 sd increase in maternal weight gain is significant in a positive direction (95% CI=(175,190)), the quadratic effect is significant in a positive direction (95% CI: (40,58)), and the cubic effect is significant in the negative direction (95% CI=(-41,-30,)). As shown in Figure 3.2, this implies greater maternal weight gain in the range of 8-78 lbs. is associated with greater infant birth weights, but greater maternal weight gain in the ranges of 0-8 lbs. and 78-98 lbs. is associated with smaller infant birth weights. The linear effect for a 2 sd increase in infant gestational age is highly significant in a positive direction (95% CI: (278,289)) while the quadratic effect is significant in a negative direction (95% CI: (-71,-54)). As shown in Figure 3.2, this implies greater gestational age is associated with greater infant birth weights, but this association flattens as gestational age nears the right tail of its distribution (44 weeks). The variables with the largest effects on infant birth weight are smoking ($\hat{\beta}_9 = -165$), female infant gender ($\hat{\beta}_8 = -120$), maternal weight gain (non-linear), and infant gestational age (non-linear). Variables with weaker yet "significant" associations include a 2 sd increase in NDI ($\hat{\beta}_{10} = -16$), maternal nativity ($\hat{\beta}_{15} = 11$), and black versus white race ($\hat{\beta}_1 = -40$). One must consider whether the magnitude of each of these effects is considered clinically relevant, as the statistical significance may be a result of the large sample size. The effects of Hispanic (95% CI=$(-25, 56)$) or "other" (95% CI=(-80,75)) races are not significantly associated with infant birth weight at the $\alpha = 0.05$ level. The non-significant result for Hispanic race may be due to the nature in which the variable was constructed. Data were not initially collected for Hispanic race, and investigators therefore constructed a Hispanic indicator variable using the ethnic ancestry variable. Hence this predictor may lack the precision of the other race indicator variables. The "Other" race group suffered from small sample size.

The frequency counts for ancestry by race are given in Table 3.8. We note that most ancestries correspond to predominantly one race. We give the posterior means of the random intercepts corresponding to the 62 ancestries in Table 3.9, as well as the predicted means for each of the ancestries by race for a typical subject with mean gestational age (39.3 weeks), NDI equal to 0, mean weight gain (31.2 lbs.), no previous births, male infant, non-smoker, < 25 years old, and maternal birth in the United States (with missing values for non-observed race

by ancestry classifications). 95% credible intervals for the ancestry random intercepts are given in Table 3.10 and plotted in Figure 3.1. Ancestries with the greatest estimated infant birth weights include Peru, Morocco, and Nigeria, while ancestries with the lowest estimated infant birth weights include Guyana, Bangladesh, Gambia, and Ivory Coast. There were no notable trends for certain geographical regions with respect to the ancestry effects. Also, we did not observe patterns between nativity and the ancestry random coefficients (Table 3.10), supporting the claim (based on the Bayes factors) that nativity does not modify the effect of ancestry.

In conclusion, we found heterogeneity in birth weights across maternal ancestries and census tracts. The heterogeneity in maternal ancestry exists within subgroups that Howard et al. (2006) considered homogeneous, and may be due to any of a large number of unmeasured social and biological factors. Further research is needed to determine why certain ancestries tend to have lower or higher birth weights. The effect of race was significant for Asian and black versus white race (although perhaps not clinically significant) and non-significant for Hispanic versus white race, while adjusting for the effects of ancestry and census tracts. Additionally, effects for race, nativity, and maternal weight gain did not vary by ancestry.

3.6 Discussion

We recommend our approach as a straightforward and efficient method for testing random coefficients in multilevel linear models. Our approach avoids issues with testing on the boundary of the parameter space, uses low-dimensional approximations to the Bayes factor, and incorporates a default prior on the random coefficients. The scaling of the random coefficients to the residual variance makes $\phi_{hl} \sim N(\log(0.3), 2)$ a reasonable default prior distribution. Simulations suggest that this prior has good frequentist properties and large sample consistency. A major contribution of our method is the ability to test several variance components from multiple factors simultaneously, and to do so for nested, non-nested, or cross-nested multilevel designs.

TABLE 3.1: Testing non-nested random intercepts, power and Type I error

			M_1 vs. M_0			M_2 vs. M_0			M_3 vs. M_0		M_3 vs. M_1		M_3 vs. M_2	
m	λ_1	λ_2	$\hat{B}_{10}$	LR_{10}	AOV_{10}	$\hat{B}_{20}$	LR_{20}	AOV_{20}	$\hat{B}_{30}$	LR^*_{30}	$\hat{B}_{31}$	LR^*_{31}	$\hat{B}_{32}$	LR^*_{32}
100	0	0	0.05	0.05	0.05	0.04	0.03	0.04	0.03	0.03	0.04	0.03	0.06	0.05
		0.1	0.06	0.04	0.05	0.06	0.04	0.04	0.04	0.03	0.06	0.04	0.06	0.04
		0.2	0.05	0.04	0.05	0.12	0.1	0.09	0.07	0.06	0.12	0.1	0.05	0.04
		0.3	0.06	0.04	0.05	0.26	0.22	0.2	0.16	0.14	0.27	0.21	0.05	0.04
	0.1	0	0.07	0.06	0.06	0.04	0.03	0.04	0.04	0.03	0.04	0.03	0.08	0.06
		0.1	0.07	0.06	0.06	0.06	0.04	0.04	0.05	0.04	0.06	0.04	0.07	0.06
		0.2	0.07	0.06	0.06	0.12	0.1	0.09	0.08	0.07	0.12	0.09	0.07	0.06
		0.3	0.08	0.06	0.06	0.26	0.22	0.19	0.18	0.15	0.27	0.22	0.08	0.06
	0.2	0	0.14	0.12	0.1	0.04	0.03	0.04	0.08	0.06	0.04	0.03	0.14	0.12
		0.1	0.14	0.12	0.1	0.06	0.05	0.04	0.08	0.07	0.06	0.04	0.15	0.11
		0.2	0.14	0.11	0.1	0.12	0.1	0.09	0.12	0.1	0.12	0.1	0.14	0.11
		0.3	0.13	0.11	0.1	0.26	0.22	0.19	0.22	0.19	0.26	0.22	0.14	0.11
	0.3	0	0.27	0.24	0.22	0.05	0.03	0.04	0.18	0.16	0.04	0.03	0.28	0.23
		0.1	0.26	0.23	0.22	0.06	0.05	0.04	0.2	0.17	0.06	0.04	0.27	0.23
		0.2	0.26	0.23	0.22	0.12	0.09	0.09	0.23	0.19	0.12	0.09	0.27	0.22
		0.3	0.25	0.21	0.21	0.25	0.21	0.18	0.29	0.27	0.25	0.21	0.26	0.22
500	0	0	0.04	0.04	0.05	0.05	0.06	0.07	0.02	0.04	0.05	0.05	0.04	0.04
		0.1	0.04	0.04	0.05	0.14	0.15	0.17	0.06	0.09	0.15	0.15	0.04	0.04
		0.2	0.04	0.04	0.05	0.64	0.65	0.66	0.42	0.52	0.64	0.64	0.04	0.04
		0.3	0.03	0.04	0.05	0.95	0.95	0.95	0.87	0.91	0.94	0.94	0.04	0.04
	0.1	0	0.13	0.13	0.14	0.05	0.06	0.07	0.06	0.09	0.05	0.06	0.12	0.13
		0.1	0.13	0.13	0.15	0.14	0.15	0.18	0.1	0.15	0.14	0.15	0.13	0.13
		0.2	0.13	0.13	0.15	0.62	0.63	0.64	0.48	0.57	0.62	0.63	0.13	0.13
		0.3	0.12	0.13	0.14	0.95	0.95	0.95	0.89	0.92	0.94	0.94	0.12	0.13
	0.2	0	0.6	0.61	0.63	0.05	0.06	0.07	0.41	0.49	0.06	0.06	0.61	0.61
		0.1	0.59	0.6	0.62	0.14	0.15	0.17	0.48	0.56	0.14	0.15	0.59	0.6
		0.2	0.58	0.6	0.6	0.61	0.62	0.63	0.75	0.8	0.62	0.63	0.58	0.59
		0.3	0.57	0.58	0.58	0.94	0.94	0.95	0.95	0.96	0.94	0.94	0.57	0.59
	0.3	0	0.95	0.95	0.95	0.05	0.06	0.06	0.89	0.92	0.06	0.06	0.95	0.95
		0.1	0.95	0.95	0.95	0.14	0.15	0.17	0.91	0.93	0.14	0.14	0.95	0.95
		0.2	0.95	0.95	0.95	0.56	0.57	0.58	0.95	0.96	0.58	0.59	0.94	0.95
		0.3	0.93	0.93	0.94	0.93	0.93	0.93	0.99	0.99	0.94	0.94	0.94	0.95
1000	0	0	0.04	0.06	0.06	0.03	0.04	0.05	0.01	0.04	0.03	0.04	0.04	0.06
		0.1	0.04	0.05	0.06	0.24	0.29	0.31	0.1	0.2	0.24	0.29	0.04	0.05
		0.2	0.03	0.05	0.06	0.92	0.93	0.94	0.8	0.88	0.92	0.93	0.04	0.05
		0.3	0.03	0.05	0.06	1	1	1	1	1	1	1	0.04	0.05
	0.1	0	0.24	0.29	0.32	0.03	0.04	0.05	0.12	0.21	0.03	0.04	0.24	0.28
		0.1	0.24	0.29	0.31	0.24	0.3	0.32	0.24	0.39	0.24	0.29	0.24	0.28
		0.2	0.23	0.28	0.3	0.92	0.93	0.94	0.86	0.92	0.91	0.93	0.25	0.28
		0.3	0.23	0.26	0.28	1	1	1	1	1	1	1	0.24	0.28
	0.2	0	0.92	0.94	0.94	0.03	0.05	0.05	0.81	0.89	0.03	0.04	0.92	0.93
		0.1	0.92	0.94	0.94	0.23	0.28	0.3	0.87	0.93	0.23	0.28	0.92	0.93
		0.2	0.92	0.93	0.94	0.9	0.92	0.94	0.98	1	0.91	0.93	0.92	0.93
		0.3	0.9	0.93	0.93	1	1	1	1	1	1	1	0.91	0.93
	0.3	0	1	1	1	0.03	0.04	0.05	0.99	1	0.03	0.04	1	1
		0.1	1	1	1	0.21	0.27	0.28	1	1	0.23	0.28	1	1
		0.2	1	1	1	0.88	0.91	0.92	1	1	0.9	0.93	1	1
		0.3	1	1	1	1	1	1	1	1	1	1	1	1

Table gives percent of times the null hypothesis was rejected out of 1000 simulations
Type I error is given by $\lambda_1 = 0$ or $\lambda_2 = 0$
$\hat{\beta}_{kk'}$: estimated Bayes factor for M_k vs. $M_{k'}$
LR_{k0}: restricted likelihood ratio test for M_k vs. M_0 using a mixture of chi-square distributions
AOV_{k0}: ANOVA F-test for M_k vs. M_0
$\text{LR}^*_{kk'}$: Ad-hoc restricted likelihood ratio test for M_k vs. $M_{k'}$ using $\alpha = 0.10$

TABLE 3.2: Estimated Bayes factors for comparing M_1 and M_2 versus M_0

			$\hat{B}_{10}$						$\hat{B}_{20}$					
			Favor M_0			Favor M_1			Favor M_0			Favor M_2		
m	λ_1	λ_2	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
100	0	0	0	2	92	4	1	0	0	2	94	3	0	0
		0.1	0	2	92	4	1	0	0	2	92	5	1	0
		0.2	0	2	93	4	1	0	0	1	88	9	2	1
		0.3	0	1	93	4	1	1	0	0	73	17	6	3
	0.1	0	0	1	91	6	1	0	0	2	93	4	1	0
		0.1	0	1	92	6	1	1	0	2	92	5	1	0
		0.2	0	1	91	6	1	1	0	1	87	9	2	1
		0.3	0	1	91	6	1	1	0	0	73	17	6	3
	0.2	0	0	1	85	10	3	1	0	2	93	3	1	0
		0.1	0	1	85	10	2	1	0	2	92	5	1	0
		0.2	0	1	85	11	2	1	0	1	87	10	2	1
		0.3	0	1	86	10	2	1	0	0	74	18	5	3
	0.3	0	0	0	72	17	6	4	0	2	93	4	1	0
		0.1	0	0	73	16	7	4	0	2	92	5	1	0
		0.2	0	0	74	16	7	3	0	1	87	9	2	1
		0.3	0	0	74	17	6	3	0	0	75	17	5	2
500	0	0	0	67	29	3	1	0	0	62	33	4	1	0
		0.1	0	67	29	3	0	0	0	39	47	9	4	2
		0.2	0	66	30	3	1	0	0	8	28	21	16	26
		0.3	0	66	31	3	1	0	0	0	5	7	10	78
	0.1	0	0	42	45	9	2	1	0	62	33	4	1	0
		0.1	0	43	44	9	2	1	0	40	46	9	4	2
		0.2	0	44	43	9	3	1	0	8	30	21	16	26
		0.3	0	45	43	9	2	1	0	0	5	8	9	78
	0.2	0	0	7	32	20	16	25	0	63	32	4	1	0
		0.1	0	7	33	19	15	25	0	41	45	9	4	1
		0.2	0	8	33	20	15	23	0	9	30	21	17	23
		0.3	0	9	34	21	15	21	0	0	5	9	8	77
	0.3	0	0	0	4	6	10	79	0	63	32	4	1	0
		0.1	0	0	5	6	10	78	0	42	43	10	3	1
		0.2	0	0	5	9	12	73	0	10	33	22	16	18
		0.3	0	1	6	8	13	72	0	1	7	9	9	74
1000	0	0	0	82	13	3	1	0	0	83	14	2	1	0
		0.1	0	82	13	3	1	0	0	43	31	12	7	5
		0.2	0	81	15	2	0	1	0	2	5	8	12	71
		0.3	0	81	14	2	1	1	0	0	0	0	1	99
	0.1	0	0	41	33	10	8	6	0	83	14	2	1	0
		0.1	0	42	33	10	7	7	0	43	31	12	7	5
		0.2	0	43	32	10	7	6	0	2	5	9	13	69
		0.3	0	46	29	11	7	6	0	0	0	0	1	99
	0.2	0	0	2	6	8	12	72	0	83	13	2	1	0
		0.1	0	2	6	8	11	72	0	45	30	13	6	4
		0.2	0	3	5	9	12	70	0	2	7	10	14	67
		0.3	0	3	6	10	13	66	0	0	0	0	1	99
	0.3	0	0	0	0	0	1	99	0	83	14	2	0	0
		0.1	0	0	0	0	1	99	0	47	29	12	6	3
		0.2	0	0	0	0	1	98	0	2	9	10	14	63
		0.3	0	0	0	1	1	98	0	0	0	0	1	98

Table includes the percent of times that the estimated Bayes factors fell into the respective categories

TABLE 3.3: Estimated Bayes factors for comparing M_3 versus M_1 and M_2

			$\hat{B}_{31}$						$\hat{B}_{32}$					
			Favor M_1			Favor M_3			Favor M_2			Favor M_3		
m	λ_1	λ_2	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
100	0	0	0	2	94	3	1	0	0	2	92	5	1	0
		0.1	0	1	93	5	1	0	0	2	93	4	1	0
		0.2	0	1	87	10	2	0	0	1	93	4	1	1
		0.3	0	0	73	18	7	1	0	1	93	4	1	1
	0.1	0	0	1	94	3	1	0	0	1	91	6	1	0
		0.1	0	1	93	5	1	0	0	1	92	6	1	1
		0.2	0	0	87	10	2	0	0	1	92	5	1	1
		0.3	0	0	73	18	7	1	0	1	92	6	1	1
	0.2	0	0	2	94	3	1	0	0	1	85	10	3	1
		0.1	0	1	93	4	1	0	0	1	84	11	2	1
		0.2	0	0	87	9	2	0	0	0	85	11	2	1
		0.3	0	0	73	18	7	1	0	0	85	11	2	1
	0.3	0	0	1	94	4	1	0	0	0	72	17	6	4
		0.1	0	1	93	4	1	0	0	0	72	16	7	4
		0.2	0	0	87	10	2	0	0	0	73	16	7	3
		0.3	0	0	75	17	7	1	0	0	74	16	7	3
500	0	0	0	61	33	4	1	0	0	67	29	3	1	0
		0.1	0	38	47	9	5	1	0	67	29	3	0	0
		0.2	0	8	28	21	27	16	0	66	30	3	0	0
		0.3	0	0	5	6	19	69	0	66	30	3	1	0
	0.1	0	0	63	32	4	1	0	0	42	45	9	3	1
		0.1	0	40	46	9	5	0	0	43	44	9	3	1
		0.2	0	8	29	20	26	15	0	44	43	9	3	1
		0.3	0	0	5	7	18	69	0	44	43	8	3	1
	0.2	0	0	63	31	5	1	0	0	7	32	20	16	25
		0.1	0	41	45	9	5	0	0	7	33	19	15	25
		0.2	0	9	29	20	27	15	0	8	34	19	15	25
		0.3	0	0	6	8	18	68	0	8	34	19	16	23
	0.3	0	0	63	32	5	1	0	0	0	4	7	10	79
		0.1	0	41	45	9	5	0	0	0	5	6	10	79
		0.2	0	9	33	22	25	11	0	0	5	9	12	74
		0.3	0	0	6	8	19	67	0	0	5	6	12	76
1000	0	0	0	83	14	2	1	0	0	82	13	3	1	0
		0.1	0	43	31	12	9	3	0	83	13	3	1	0
		0.2	0	2	6	9	25	58	0	83	13	3	1	0
		0.3	0	0	0	0	1	98	0	83	13	3	1	0
	0.1	0	0	84	13	2	1	0	0	41	33	10	8	6
		0.1	0	44	30	13	9	2	0	42	32	10	8	6
		0.2	0	2	6	8	25	58	0	43	31	11	7	6
		0.3	0	0	0	0	1	98	0	43	31	10	7	6
	0.2	0	0	84	12	2	1	0	0	2	6	8	12	73
		0.1	0	44	30	13	8	2	0	2	5	9	11	73
		0.2	0	2	6	8	25	57	0	2	6	9	12	71
		0.3	0	0	0	0	1	98	0	2	6	8	12	71
	0.3	0	0	84	12	2	1	0	0	0	0	0	1	99
		0.1	0	45	30	13	8	2	0	0	0	0	1	99
		0.2	0	2	7	8	26	57	0	0	0	0	1	99
		0.3	0	0	0	0	1	98	0	0	0	0	1	99

Table includes the percent of times that the estimated Bayes factors fell into the respective categories

TABLE 3.4: Estimated Bayes factors for comparing M_3 versus M_0

			Favor M_0			Favor M_3		
m	λ_1	λ_2	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
100	0	0	0	68	29	2	0	1
		0.1	0	65	31	3	1	1
		0.2	0	57	35	5	1	1
		0.3	0	41	42	11	3	2
	0.1	0	0	65	31	3	1	1
		0.1	0	62	33	4	1	1
		0.2	0	53	39	5	2	1
		0.3	0	40	42	12	4	2
	0.2	0	0	55	37	5	2	1
		0.1	0	53	38	6	2	1
		0.2	0	47	41	8	3	1
		0.3	0	34	44	15	4	3
	0.3	0	0	43	39	12	4	3
		0.1	0	40	40	13	4	3
		0.2	0	34	43	14	5	3
		0.3	0	23	47	17	8	5
500	0	0	55	38	4	2	0	0
		0.1	37	47	11	3	2	1
		0.2	10	26	22	15	11	15
		0.3	1	5	7	8	9	69
	0.1	0	38	46	11	4	1	1
		0.1	25	48	17	7	2	2
		0.2	5	24	23	17	14	18
		0.3	0	4	7	8	9	72
	0.2	0	8	31	20	14	12	15
		0.1	5	25	21	16	15	18
		0.2	1	11	13	15	17	43
		0.3	0	2	3	5	7	83
	0.3	0	1	5	6	9	12	68
		0.1	1	4	5	8	12	70
		0.2	0	2	3	7	9	79
		0.3	0	0	1	2	3	95
1000	0	0	79	16	3	1	0	0
		0.1	47	31	11	5	3	2
		0.2	3	8	8	9	15	56
		0.3	0	0	0	1	1	98
	0.1	0	48	30	10	5	3	3
		0.1	23	33	19	10	7	7
		0.2	1	6	7	7	12	66
		0.3	0	0	0	0	1	98
	0.2	0	3	8	8	9	12	59
		0.1	1	5	6	8	13	66
		0.2	0	0	1	1	4	93
		0.3	0	0	0	0	0	100
	0.3	0	0	0	0	1	1	98
		0.1	0	0	0	1	1	98
		0.2	0	0	0	0	0	100
		0.3	0	0	0	0	0	100

Table includes the percent of times that the estimated Bayes factors fell into the respective categories

TABLE 3.5: Testing a random slope, power and Type I error

m	$\sqrt{\psi_{22}}$	$\hat{B}_{43}$	LR^*_{43}
100	0	0.05	0.03
	0.1	0.05	0.04
	0.2	0.07	0.05
	0.3	0.09	0.07
	0.6	0.26	0.2
	1	0.66	0.56
500	0	0.04	0.05
	0.1	0.06	0.07
	0.2	0.13	0.14
	0.3	0.29	0.29
	0.6	0.92	0.91
	1	1	1
1000	0	0.03	0.04
	0.1	0.06	0.09
	0.2	0.24	0.28
	0.3	0.59	0.61
	0.6	1	1
	1	1	1

Rejection rate for 1000 simulations
Type I error: $\sqrt{\psi_{22}} = 0$
$\hat{\beta}_{43}$ = Bayes factor, M_4 vs. M_3
LR^*_{43} = ad-hoc RLRT, M_4 vs. M_3

TABLE 3.6: Estimated Bayes factor, $\hat{B}_{43}$, for comparing M_4 versus M_3

		Favor M_3			Favor M_4		
m	$\sqrt{\psi_{22}}$	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
100	0	0	0	95	5	0	0
	0.1	0	0	95	4	1	0
	0.2	0	0	93	6	1	0
	0.3	0	0	91	6	1	1
	0.6	0	0	73	16	6	4
	1	0	0	34	22	16	27
500	0	0	54	43	3	1	0
	0.1	0	50	43	6	1	0
	0.2	0	37	50	9	4	1
	0.3	0	17	54	14	7	8
	0.6	0	1	7	10	12	70
	1	0	0	0	0	0	100
1000	0	0	85	12	2	0	0
	0.1	0	74	19	4	2	1
	0.2	0	47	28	11	7	6
	0.3	0	16	23	15	16	28
	0.6	0	0	0	0	1	99
	1	0	0	0	0	0	100

Table includes the percent of times that the estimated Bayes factor fell into the respective categories

TABLE 3.7: Model posterior means and 95% credible interval

Parameter	Posterior Mean	2.5%	97.5 %
β_0	3329	3294	3363
β_1 (Black)	-40	-74	-7
β_2 (Hisp)	14	-25	56
β_3 (Asian)	-57	-91	-24
β_4 (Other)	-2	-80	75
β_5 (Gest)*	284	278	289
β_6 (Gest^2)	-63	-71	-54
β_7 (Previous birth)	105	99	111
β_8 (Female)	-120	-126	-115
β_9 (Smoke)	-165	-186	-143
β_{10} (Deprivation)*	-16	-23	-9
β_{11} (Age 26-30)	53	45	60
β_{12} (Age 31-35)	72	64	80
β_{13} (Age 36-40)	84	75	94
β_{14} (Age $>$ 40)	76	60	92
β_{15} (Foreign)	11	3	18
β_{16} (Wtgain)*	183	175	190
β_{17} (Wtgain^2)	49	40	58
β_{18} (Wtgain^3)	-35	-41	-30

* Estimates for a 2 sd increase
All estimates given in grams

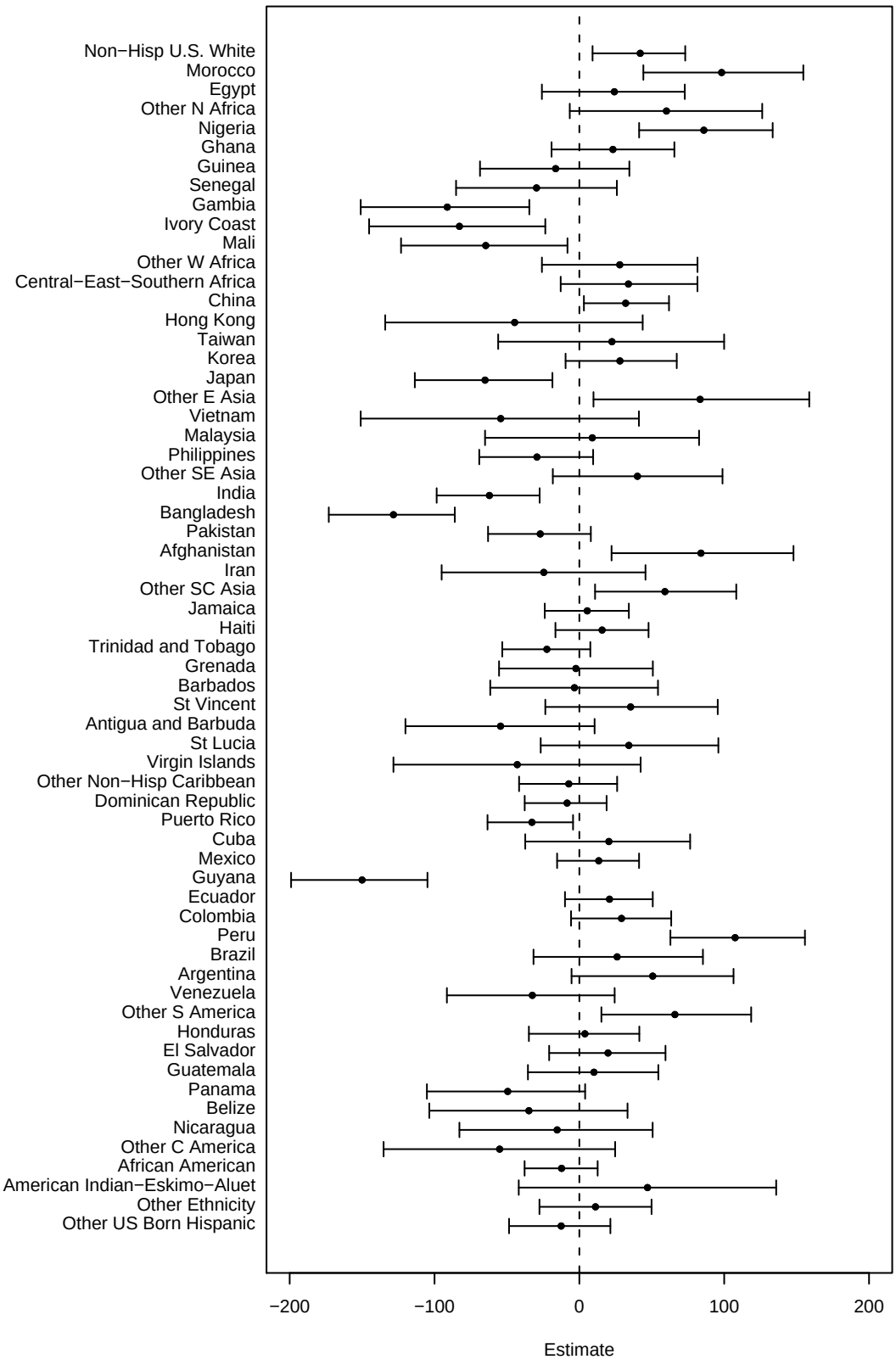


FIGURE 3.1: Posterior means and 95% credible intervals of random intercepts

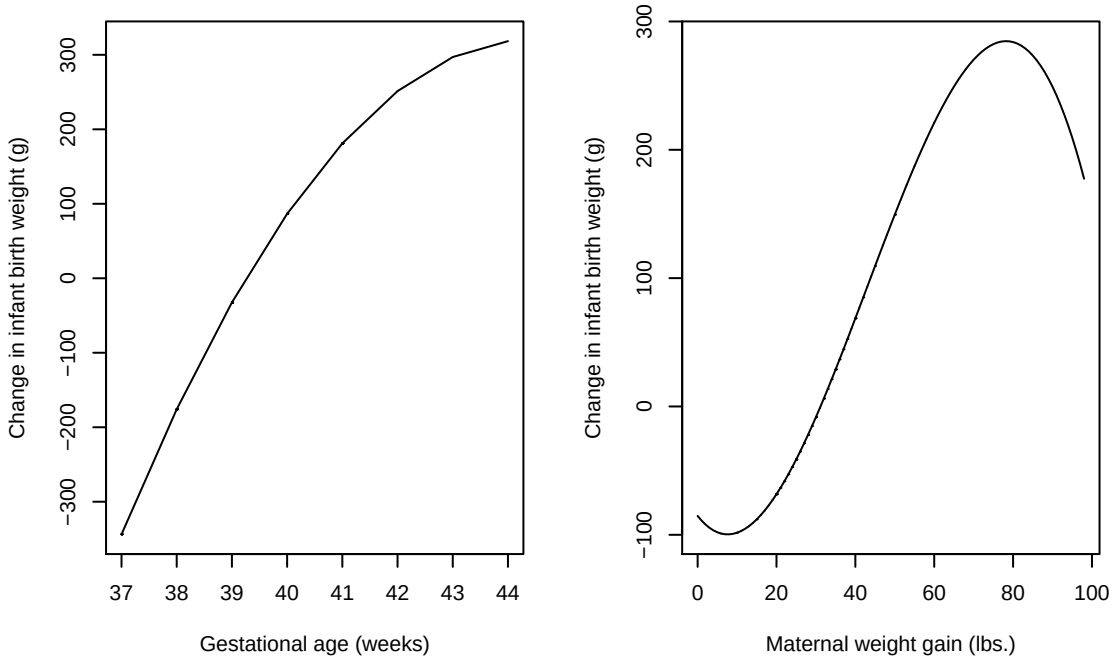


FIGURE 3.2: Estimated change in infant birth weight by gestational age and maternal weight gain

TABLE 3.8: Frequency counts for ancestry by race

Region	Ancestry	White	Black	Hispanic	Asian	Other	Total
Non-Hisp U.S. White	Non-Hisp U.S. White	24749	0	0	0	0	24749
N Africa	Morocco	203	21	0	4	0	228
	Egypt	347	0	0	7	0	354
	Other N Africa	65	44	0	4	0	113
Subsaharan Africa	Nigeria	3	410	0	3	0	416
	Ghana	2	450	0	0	0	452
	Guinea	0	256	0	0	0	256
	Senegal	1	206	0	1	0	208
	Gambia	0	177	0	0	0	177
	Ivory Coast	0	161	0	0	0	161
	Mali	2	187	0	0	0	189
	Other W Africa	5	219	0	1	0	225
	Central-East-Southern Africa	38	283	0	4	0	325
E Asia	China	25	13	0	5506	0	5544
	Hong Kong	0	0	0	36	0	36
	Taiwan	1	0	0	65	0	66
	Korea	8	2	0	784	0	794
	Japan	9	3	0	352	0	364
	Other E Asia	19	3	0	51	0	73
SE Asia-Pac Islands	Vietnam	6	4	0	13	0	23
	Malaysia	0	0	0	78	2	80
	Philippines	22	9	0	646	0	677
	Other SE Asia	12	5	0	151	0	168
SC Asia	India	8	56	0	1374	7	1445
	Bangladesh	30	20	0	1190	0	1240
	Pakistan	40	10	0	960	0	1010
	Afghanistan	65	2	0	70	0	137
	Iran	96	0	0	2	0	98
	Other SC Asia	149	3	0	148	0	300
Non-Hisp Caribbean	Jamaica	5	2076	0	14	0	2095
	Haiti	6	1269	0	0	0	1275
	Trinidad and Tobago	12	1140	0	283	0	1435
	Grenada	0	220	0	3	0	223
	Barbados	0	175	0	0	0	175
	St Vincent	0	160	0	0	0	160
	Antigua and Barbuda	0	118	0	0	0	118
	St Lucia	1	142	0	1	0	144
	Virgin Islands	2	40	0	0	0	42
	Other Non-Hisp Caribbean	16	956	0	13	0	985
Hisp Caribbean	Dominican Republic	0	0	8426	0	1	8427
	Puerto Rico	0	0	7997	0	3	8000
	Cuba	0	0	192	0	0	192
Mexico	Mexico	0	0	6585	0	0	6585
S America	Guyana	0	0	1785	0	73	1858
	Ecuador	0	0	3053	0	0	3053
	Colombia	0	0	1239	0	1	1240
	Peru	0	0	521	0	0	521
	Brazil	0	0	178	0	0	178
	Argentina	0	0	198	0	0	198
	Venezuela	0	0	181	0	0	181
	Other S America	0	0	283	0	0	283
C American	Honduras	0	0	740	0	23	763
	El Salvador	0	0	640	0	0	640
	Guatemala	0	0	397	0	13	410
	Panama	0	0	226	0	0	226
	Belize	0	0	109	0	0	109
	Nicaragua	0	0	114	0	0	114
	Other C America	0	0	59	0	0	59
African American	African American	62	12323	0	12	6	12403
American Indian	American Indian-Eskimo-Aluet	5	18	0	0	12	35
Other Ethnicity	Other Ethnicity	59	344	0	137	7	547
Other US Born Hispanic	Other US Born Hispanic	0	0	1356	0	0	1356
Total		26073	21525	34279	11913	148	93938

TABLE 3.9: Posterior means of ancestry random intercepts, and predicted means by race

Region	Ancestry	$\hat{b}_{1j}$	$\hat{y}_{\text{white}}$	$\hat{y}_{\text{black}}$	$\hat{y}_{\text{Hispanic}}$	$\hat{y}_{\text{Asian}}$	$\hat{y}_{\text{Other}}$
Non-Hisp U.S. White	Non-Hisp U.S. White	42	91	.	.	.	.
N Africa	Morocco	98	147	107	.	90	.
	Egypt	24	73	.	.	16	.
	Other N Africa	60	109	69	.	52	.
Subsaharan Africa	Nigeria	86	135	95	.	78	.
	Ghana	23	72	32	.	.	.
	Guinea	-16	.	-7	.	.	.
	Senegal	-30	19	-20	.	-38	.
	Gambia	-91	.	-82	.	.	.
	Ivory Coast	-83	.	-74	.	.	.
	Mali	-65	-16	-55	.	.	.
	Other W Africa	28	77	37	.	20	.
	Central-East-Southern Africa	34	83	43	.	26	.
E Asia	China	32	81	41	.	24	.
	Hong Kong	-45	.	.	.	-53	.
	Taiwan	22	71	.	.	14	.
	Korea	28	77	37	.	20	.
	Japan	-65	-16	-56	.	-73	.
	Other E Asia	83	132	93	.	75	.
SE Asia-Pac Islands	Vietnam	-54	-5	-45	.	-62	.
	Malaysia	9	.	.	.	1	55
	Philippines	-29	19	-20	.	-37	.
	Other SE Asia	40	89	49	.	32	.
SC Asia	India	-62	-13	-53	.	-70	-16
	Bangladesh	-128	-80	-119	.	-136	.
	Pakistan	-27	22	-18	.	-35	.
	Afghanistan	84	133	93	.	76	.
	Iran	-25	24	.	.	-33	.
	Other SC Asia	59	108	68	.	51	.
Non-Hisp Caribbean	Jamaica	5	54	15	.	-3	.
	Haiti	16	64	25	.	.	.
	Trinidad and Tobago	-22	26	-13	.	-31	.
	Grenada	-2	.	7	.	-10	.
	Barbados	-3	.	6	.	.	.
	St Vincent	35	.	45	.	.	.
	Antigua and Barbuda	-54	.	-45	.	.	.
	St Lucia	34	83	43	.	26	.
	Virgin Islands	-43	6	-34	.	.	.
	Other Non-Hisp Caribbean	-7	41	2	.	-15	.
Hisp Caribbean	Dominican Republic	-8	.	.	55	.	38
	Puerto Rico	-33	.	.	30	.	14
	Cuba	20	.	.	84	.	.
Mexico	Mexico	13	.	.	77	.	.
S America	Guyana	-150	.	.	-87	.	-104
	Ecuador	21	.	.	84	.	.
	Colombia	29	.	.	92	.	76
	Peru	108	.	.	171	.	.
	Brazil	26	.	.	89	.	.
	Argentina	51	.	.	114	.	.
	Venezuela	-32	.	.	31	.	.
	Other S America	66	.	.	129	.	.
C American	Honduras	4	.	.	67	.	50
	El Salvador	20	.	.	83	.	.
	Guatemala	10	.	.	73	.	57
	Panama	-49	.	.	14	.	.
	Belize	-35	.	.	28	.	.
	Nicaragua	-15	.	.	48	.	.
	Other C America	-55	.	.	8	.	.
African American	African American	-12	36	-3	.	-20	34
American Indian	American Indian-Eskimo-Aluet	47	96	56	.	.	93
Other Ethnicity	Other Ethnicity	11	60	20	.	3	58
Other US Born Hispanic	Other US Born Hispanic	-13	.	.	51	.	.

$\hat{b}_{1j}$ is the posterior mean of the random intercept for ancestry j
$\hat{y}$ is the predicted mean for a subject with gestational age = 39.3 weeks, NDI=0, male infant, maternal weight gain = 31.2 lbs, no previous births, non-smoker,< 25 years old, born in the U.S.
Estimates given in grams; '.' denotes no observed subjects in the category

TABLE 3.10: Posterior means of ancestry random intercepts with CI's, with nativity

Region	Ancestry	$\hat{b}_{1j}$	LL	UL	N	%Fgn
Non-Hisp U.S. White	Non-Hisp U.S. White	42	9	73	24749	29
N Africa	Morocco	98	44	155	228	97
	Egypt	24	-26	73	354	94
	Other N Africa	60	-7	126	113	99
Subsaharan Africa	Nigeria	86	41	133	416	98
	Ghana	23	-19	66	452	100
	Guinea	-16	-69	35	256	100
	Senegal	-30	-85	26	208	100
	Gambia	-91	-151	-35	177	99
	Ivory Coast	-83	-145	-24	161	100
	Mali	-65	-123	-8	189	99
	Other W Africa	28	-26	82	225	98
	Central-East-Southern Africa	34	-13	82	325	95
E Asia	China	32	3	62	5544	95
	Hong Kong	-45	-134	44	36	97
	Taiwan	22	-56	100	66	94
	Korea	28	-10	67	794	93
	Japan	-65	-114	-19	364	92
	Other E Asia	83	10	159	73	74
SE Asia-Pac Islands	Vietnam	-54	-151	41	23	100
	Malaysia	9	-65	83	80	97
	Philippines	-29	-69	10	677	90
	Other SE Asia	40	-18	99	168	95
SC Asia	India	-62	-99	-27	1445	94
	Bangladesh	-128	-173	-86	1240	100
	Pakistan	-27	-63	8	1010	98
	Afghanistan	84	22	148	137	98
	Iran	-25	-95	46	98	92
	Other SC Asia	59	11	108	300	100
Non-Hisp Caribbean	Jamaica	5	-24	34	2095	95
	Haiti	16	-17	48	1275	90
	Trinidad and Tobago	-22	-53	8	1435	96
	Grenada	-2	-55	51	223	99
	Barbados	-3	-62	54	175	95
	St Vincent	35	-24	96	160	100
	Antigua and Barbuda	-54	-120	11	118	97
	St Lucia	34	-27	96	144	100
	Virgin Islands	-43	-128	42	42	95
	Other Non-Hisp Caribbean	-7	-42	26	985	89
Hisp Caribbean	Dominican Republic	-8	-38	19	8427	81
	Puerto Rico	-33	-63	-4	8000	20
	Cuba	20	-37	76	192	26
Mexico	Mexico	13	-15	41	6585	96
S America	Guyana	-150	-199	-105	1858	96
	Ecuador	21	-10	51	3053	91
	Colombia	29	-6	63	1240	83
	Peru	108	63	156	521	92
	Brazil	26	-32	85	178	94
	Argentina	51	-5	106	198	88
	Venezuela	-32	-91	24	181	95
	Other S America	66	15	119	283	93
C American	Honduras	4	-35	41	763	90
	El Salvador	20	-21	59	640	92
	Guatemala	10	-36	55	410	92
	Panama	-49	-105	4	226	74
	Belize	-35	-104	33	109	83
	Nicaragua	-15	-83	51	114	96
	Other C America	-55	-135	25	59	80
African American	African American	-12	-38	13	12403	12
American Indian	American Indian-Eskimo-Aluet	47	-42	136	35	3
Other Ethnicity	Other Ethnicity	11	-28	50	547	43
Other US Born Hispanic	Other US Born Hispanic	-13	-49	21	1356	3

$\hat{b}_{1j}$ = posterior mean of the random intercept with 95% credible interval (LL,UL)
%Fgn = percent born outside the U.S.

CHAPTER 4

Analyzing Correlated Longitudinal and Survival Data in Clinical Trials Using Multivariate Time-to-Event Methods

4.1 Introduction

Many clinical trials evaluate the efficacy of a treatment on correlated longitudinal and time-to-event outcomes. For example, consider a randomized clinical trial evaluating the effectiveness of a treatment drug versus a control in 2,000 patients with a chronic respiratory disorder. The investigators recorded the time to death within 3 years of randomization, as well as repeated measurements at 6 month intervals of respiratory lung function FEV, or postbronchodilator forced expiratory volume at 1 second. Because these patients suffer from a chronic condition, lung function is expected to deteriorate over time and ultimately result in death. Clearly, lung function and survival are expected to be highly correlated. There are well-established methods for analyzing these longitudinal and survival outcomes separately, including the linear mixed model for longitudinal data (Laird and Ware, 1982) and the Cox proportional hazards model for survival data (Cox, 1972). However, the analysis of these longitudinal and survival outcomes separately may be inefficient or even inappropriate when the longitudinal variable is correlated with the survival endpoint (Guo and Carlin, 2004). Such approaches ignore important information in the other outcome as well as potentially informative dropout in the longitudinal

process. This has led to a growing literature on jointly modeling distributions of correlated longitudinal and survival endpoints.

There are many reasons to consider a joint model of longitudinal and event outcomes. Such reasons include describing the trajectory of the longitudinal process over time subject to informative censoring and how this is affected by baseline predictors; determining how the probability of an event outcome is influenced by the longitudinal process; evaluating whether the longitudinal process can be used as a surrogate endpoint for the event outcome; or making predictions of future event times for subjects who are censored. Joint models generally base inference on the joint distribution of the longitudinal and survival outcomes (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Tsiatis and Davidian, 2001; Lin et al., 2002; Guo and Carlin, 2004; Tseng et al., 2005, and more recently Elashoff et al., 2007; Dang et al., 2007). For more complete reviews of joint modeling methods, see Hogan and Laird (1997b), Tsiatis and Davidian (2004), Yu et al. (2004), and Ibrahim et al. (2001). Although joint models may be conceptually appealing, they can be computationally demanding, difficult to implement, and may require specialized software (Hogan and Laird, 1997b). Many of the joint model approaches make strong parametric assumptions regarding the longitudinal and survival processes (Tsiatis and Davidian, 2004; Yu et al., 2004). These assumptions may not be obvious and can be difficult to validate.

We propose a strategy that uses multivariate time-to-event methods to evaluate effects of a treatment or baseline predictor on both longitudinal and survival outcomes simultaneously. We first create multiple time-to-event endpoints based on the survival and longitudinal outcomes. These endpoints are defined as time to reach various thresholds in the longitudinal outcome or death, whichever comes first. We then use semiparametric and nonparametric methods to evaluate the treatment effect on these multivariate time-to-event outcomes. Our approach is straightforward to implement for a randomized clinical trial using standard software (SAS) and makes minimal or no assumptions regarding underlying distributions. More specifically, the multivariate time-to-event methods that we utilize include the Wei-Lin-Weissfeld method (Wei et al., 1989) and nonparametric analysis of covariance (NPANCOVA) with logrank scores as defined by Tangen and Koch (Tangen and Koch, 1999b). Although these multivariate approaches are well-established methods, they are typically applied in settings in which multivariate events are clearly defined. These events are usually distinct outcomes (e.g. time to relapse or time to

death) or repeated events of the same kind (e.g. time to hospitalization). Our contribution in this paper is to apply the multivariate time-to-event methods to longitudinal and survival data simultaneously. In Section 2 we introduce the multivariate methods used in our approach. In Section 3 we present simulation studies. In Section 4 we apply our method to a clinical trial involving chronic lung disease and conclude with a discussion in Section 5.

4.2 Application of Multivariate Time-to-Event Methods

4.2.1 Wei-Lin-Weissfeld Method

Suppose there are M time-to-event outcomes. To apply the method of Wei et al. (1989) (referred to as the WLW method), one fits a marginal Cox proportional hazards model for each of the M events

$$\lambda_{mi}(t) = \lambda_{m0}(t) \exp\{\boldsymbol{x}_i' \boldsymbol{\beta}_m\}, \tag{4.1}$$

in which $\boldsymbol{\beta}_m = (\beta_{m1}, \ldots, \beta_{mp})'$ is the vector of parameters for the mth marginal model, $\boldsymbol{x}_i'$ is a vector of baseline predictors, and $\lambda_{mi}(t)$ is the hazard for subject i proportional to the baseline hazard $\lambda_{m0}(t)$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_M')'$ be the vector of all parameters and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \ldots, \hat{\boldsymbol{\beta}}_M')'$ be the maximum partial likelihood estimates from all M models. Wei et al. (1989) showed that the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is normal with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{V}$, in which an estimator $\hat{\boldsymbol{V}}$ of the variance is a function of the score residuals and information matrix (see Appendix). Given the asymptotic normal distribution of $\hat{\boldsymbol{\beta}}$ and variance estimate $\hat{\boldsymbol{V}}$, it is straightforward to construct a model-averaged log hazards ratio to summarize the treatment effect. Let $\boldsymbol{\beta}_e = (\beta_{1e}, \ldots, \beta_{Me})'$ represent the vector of parameters for the marginal treatment effect (e indexes the experimental or treatment effect). Wei et al. (1989) suggested estimating a model-averaged log hazards ratio using the estimate

$$\hat{\theta} = \boldsymbol{C}' \hat{\boldsymbol{\beta}}_e, \tag{4.2}$$

with $\boldsymbol{C} = (\mathbf{1}'_M \hat{\boldsymbol{V}}_e^{-1} \mathbf{1}_M)^{-1} \hat{\boldsymbol{V}}_e^{-1} \mathbf{1}_M$ and $\hat{\boldsymbol{V}}_e$ equal to the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_e$ (constructed from the appropriate elements of $\hat{\boldsymbol{V}}$). This estimator was proposed as the optimal estimator because it has the smallest asymptotic variance among all linear estimators. A test statistic for testing whether the average log hazards ratio is equal to 0 can be constructed as

$$Z^2 = \frac{(\boldsymbol{C}'\hat{\boldsymbol{\beta}}_e)^2}{\boldsymbol{C}'\hat{\boldsymbol{V}}_e\boldsymbol{C}}, \tag{4.3}$$

which follows an asymptotic chi-square distribution with one degree of freedom. In SAS version 9.1 one can obtain this test directly using the procedure PROC PHREG (see SAS documentation) or by fitting the marginal models and constructing the appropriate covariance matrix using the "dfbeta" residuals. These residuals are equivalent to the product of score residuals and the information matrix (see Appendix).

4.2.2 Nonparametric ANCOVA

Logrank scores are a set of values which are used in nonparametric testing procedures for comparing the survival times of two or more groups with possible censoring (Peto and Peto, 1972; Koch et al., 1985). These scores are centered about zero starting with 1 and decreasing as endpoints lengthen (see Appendix). For M time-to-event outcomes, one can compute logrank scores for each of the M events separately to obtain M vectors of logrank scores. One can then use multivariate nonparametric ANCOVA to evaluate a treatment effect on all outcomes simultaneously adjusting for relevant covariables (Tangen and Koch, 1999b; Tangen and Koch, 1999a). This method uses weighted least squares methods to produce an estimated treatment effect $\hat{\boldsymbol{\beta}}$ and corresponding variance estimate $\hat{V}_{\hat{\boldsymbol{\beta}}}$, in which $\hat{\boldsymbol{\beta}}$ is the estimated mean difference in logrank scores between the treatment groups (see Appendix). This model restricts the vector(s) of differences between means for the covariates to zeros on the basis of randomization. One can use (4.2) to obtain an average difference in logrank scores between treatments (averaged across the M events) and its corresponding test statistic as given by (4.3). This approach is straightforward to conduct in statistical software packages. SAS macros are available to compute the logrank scores (please contact authors) and to perform multivariate nonparametric ANCOVA for comparing two treatment groups (Zink and Koch, 2002).

4.2.3 Defining the Multivariate Outcomes

We define M thresholds or cutpoints of interest in the longitudinal outcome and use these cutpoints to construct $(M+1)$ "threshold endpoints". We define the first M threshold endpoints as time to the mth cutpoint or terminating event (e.g. death), whichever comes first. The final threshold endpoint is defined as time to the terminating event. Subjects who do not experience a threshold event in the study are considered censored. For example, consider the study of chronic lung disease with FEV threshold events at $\leq 1300ml$, $\leq 1010ml$, and $\leq 740ml$. Suppose three subjects have FEV values and time of death as given in Table 4.1. For subject 1, the first threshold event is observed at 18 months, the second at 24 months, and the third and fourth at 26 months. For subject 2, all four threshold events are censored at 36 months. For subject 3, the first threshold event is observed at 6 months, and threshold events 2,3,4 are observed at 11 months.

We note that various definitions of the thresholds are possible depending on the clinical relevance. For example, one could define the first M threshold endpoints as time to reach a certain longitudinal value that is sustained for at least (say) three observations or death, whichever comes first. The definition of these thresholds should be tailored toward the clinical application such that the interpretations of the threshold endpoints are clinically relevant. Our application is most relevant to studies in which there exists non-reversible deterioration in the longitudinal process subject to censoring due to the survival endpoint.

We implicitly make assumptions in both the WLW and logrank approaches. For the WLW approach, we assume that the observed time to reach a given threshold event has an underlying continuous nature and that the hazards ratio for reaching an event is constant across time for each predictor. We also assume that there is a log-linear relationship between the independent variables and the underlying hazard function. For the logrank approach, essentially the only assumption is that the patients are randomized to their respective treatment groups. The logrank approach makes no modeling assumptions and does not require a continuous failure time. Both approaches also make the assumption that censoring is independent of treatment and is noninformative.

In a regulated clinical trial with correlated longitudinal and survival outcomes, it is often not

known *a priori* whether the primary hypothesis should be based on the longitudinal or survival outcome. The power of a survival analysis would increase with a larger number of events, which would also be associated with increasing (informative) dropout and decreasing power for a longitudinal analysis. Conversely, one has increasing power in the longitudinal process as missing data due to terminating events decreases, implying fewer events and decreasing power in the survival process. Even in cases in which the amount of missing data is predictable, it may be unknown which process is likely to have greater sensitivity to treatment differences.

Our method is attractive in such situations, as one can incorporate our multivariate approach in the study protocol with the understanding that it can lead to increased sensitivity to treatment differences compared to the standard longitudinal and survival approaches and at worst should lead to a "second best" approach. For example, consider a study in which a survival analysis may have much greater sensitivity to treatment differences than a longitudinal analysis. Because the multivariate approach incorporates information from each of these processes, it is reasonable to expect the multivariate approach to have sensitivity to treatment differences somewhere in between these two extremes. The same argument applies in the case in which the longitudinal analysis has much greater sensitivity to treatment differences than the survival analysis. By specifying the multivariate approach as the primary analysis *a priori*, one can reduce the risk of selecting the outcome with the least sensitivity to treatment differences and have a reasonably good chance of selecting the approach with the greatest sensitivity to treatment differences, as shown in simulation studies in Section 3.

4.3 Simulation Studies

We conduct two simulations to evaluate the performance of our multivariate approach (using the WLW or logrank strategy) relative to standard approaches using either of the longitudinal or survival outcomes separately and to the joint model approach of Henderson et al. (2000). Our proposed approach is most useful in settings with a small to moderate treatment effect on both the longitudinal and survival outcomes and fairly large samples sizes (e.g. ≥ 300 per group). If the treatment effects were known to be large *a priori* in both of these processes, there would be little need for our method.

We first simulate the longitudinal data with a trend over time for the mean and a random intercept inducing an exchangeable correlation structure. We then generate terminating events using a piecewise exponential model at fixed time points. The hazards function depends on treatment, baseline covariates, and the population mean of the longitudinal variable for a given interval. In the second simulation, we simulate the longitudinal data in the same format as the first simulation, but we simulate deaths based upon subjects reaching a pre-determined threshold. When subjects reach this threshold, the probability of death is set to 60% for each observed Y_{ij} below the threshold. We compare the models based on power and Type I error.

Wei et al. (1989) proposed an "optimal" estimator $\hat{\boldsymbol{\beta}}$ that weights the marginal estimates by the inverse of the covariance matrix. In our approach, one will observe a greater number of events for earlier cutpoints, causing the "optimal" estimator to place more weight on the estimates from the earlier cutpoints than on the later cutpoints. However, in many studies one may expect a greater treatment effect in the later cutpoints. Hence we consider a modified WLW approach that weights the parameter estimates of the respective events using a specified contrast matrix, potentially weighting estimates from the later events more heavily compared to the weighting of the "optimal" estimator. Let $\hat{\boldsymbol{\beta}}_e$ be the vector of treatment effects from the marginal Cox models for the respective threshold endpoints and $\hat{\boldsymbol{V}}_e$ be the corresponding covariance matrix. In the case of four events, we define a contrast matrix $\boldsymbol{C}_2 = (0.25, 0.25, 0.25, 0.25)'$ to weight the treatment effects from the various threshold endpoints (time to 1st cutpoint or death, time to 2nd cutpoint or death, time to 3rd cutpoint or death, and time to death). This contrast weights each estimate equally, placing more emphasis on estimates from the later events as compared to the optimal estimator of Wei et al. (1989). We then compute the test statistic as shown in (4.3). We denote this approach as WLW_2 and use WLW_1 to denote the WLW approach using the optimal estimator.

4.3.1 Comparing Methods

Let Y_{ij} be the longitudinal response of subject i at observation j, for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Additionally, let y_{i0} be the baseline value of the observed response (the longitudinal response at randomization) and x_i be the treatment indicator. Let T_i denote the time to death

of the ith subject, and $Z_i = \min(T_i, U_i)$, in which U_i is a censoring time for survival of patient i. In both simulation setups we compare the following methods:

- WLW$_1$: The standard WLW approach using the optimal estimator of Wei et al. (1989), i.e. $\boldsymbol{C}_1 = (\mathbf{1}_4' \hat{\boldsymbol{V}}_e^{-1} \mathbf{1}_4)^{-1} \hat{\boldsymbol{V}}_e^{-1} \mathbf{1}_4$.

- WLW$_2$: The modified WLW approach with $\boldsymbol{C}_2 = (0.25, 0.25, 0.25, 0.25)'$.

- LR: The multivariate logrank analysis using nonparametric ANCOVA based on the test statistic with equal weights, i.e. $\boldsymbol{C}_2 = (0.25, 0.25, 0.25, 0.25)$.

- Cox: A Cox proportional hazards of the form

$$\lambda_i(t) = \lambda_0(t) \exp(\gamma_1 x_i + \gamma_2 y_{i0}), \tag{4.4}$$

  in which $\lambda_i(t)$ is the hazard of subject i at time t, $\lambda_0(t)$ is an unspecified baseline hazard function at time t, and γ_1 and γ_2 are parameters indicating treatment and baseline measurement effects, respectively. To account for tied event times, we use both the approximation of Efron (1977) and the discrete logistic likelihood.

- LM$_1$: A linear mixed model (with missing data due to failure) evaluating the treatment main effect,

$$Y_{ij} = \beta_0 + b_{i0} + \beta_1 t_{ij} + \beta_2 x_i + \beta_3 y_{i0} + \varepsilon_{ij}, \tag{4.5}$$

  in which t_{ij} is the observation time for subject i and observation j, β_0 is a model intercept, b_{i0} is a random subject intercept, and ε_{ij} is the residual error. We assume $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ independent of $b_{i0} \sim \mathrm{N}(0, \psi)$.

- LM$_2$: A linear mixed model with time as a class variable (i.e. using indicator variables for each time point) and a time by treatment interaction. The treatment effect is evaluated at the last time point in which at least 50% of the subjects have an observed response. Observations are discarded for the later time points with fewer than 5% observed data,

as this would not allow for precise estimates of the time effect and treatment by time interaction at these time points.

- Hen: A joint model based on the method of Henderson et al. (2000) using SAS code from Guo and Carlin (2004). The longitudinal process takes the form of (4.5) and the time to event T_i follows an exponential distribution with hazard function

$$\lambda_i = \exp\{\gamma_0 + \gamma_1 x_i + \gamma_2 y_{i0} + \gamma_3 b_{i0}\}, \tag{4.6}$$

in which γ_0 determines the baseline hazard function and γ_1, γ_2, and γ_3 indicate the effect of the treatment, baseline measurement, and random coefficient, respectively. The longitudinal and survival processes are linked through the random coefficient b_{i0}. A joint test $H_0 : \beta_2 = \gamma_1 = 0$ will test for a treatment effect in both the longitudinal and survival processes simultaneously. This joint model assumes an exponential distribution on the hazards function, i.e. a constant hazard over time. Use of a Weibull distribution, allowing the hazard to vary over time, led to an inflated Type I error rate (data not shown).

As an alternative to the logrank approach, Tangen and Koch (1999b) discuss using nonparametric ANCOVA on the Wilcoxon scores (Gehan, 1965). However, Wilcoxon scores tend to give more weight than the logrank test to early failures and relatively less weight to later failures (Prentice and Marek, 1979), and thus this method had decreased power in our setting (results not shown).

4.3.2 Simulation One

To generate the longitudinal data, we set $n = 600$ and sample $\varepsilon \sim N(0, 1)$, $b_{i0} \sim N(0, 1)$, and calculate

$$Y_{ij} = b_{i0} + \beta_1 t_{ij} + \beta_2 t_{ij} x_i + \varepsilon_{ij} z, \tag{4.7}$$

in which $\beta_1 = (0, -0.2, -0.5)$ and $\beta_2 = (0, .01, .02, .03, .04, .05)$ in different settings, with $x_i \sim$Bernoulli(0.5) and $t_{ij} = j$ for $j = (1, \ldots, 10)$. We calculate a baseline value $y_{i0} = b_{i0} + \varepsilon_{ij}$

to be used as a predictor in model fitting. We then generate the survival data using a piecewise exponential model with hazard function

$$\lambda_{ij} = \exp(\gamma_0 + \gamma_1 E(Y_{ij}) + \gamma_2 x_i) \tag{4.8}$$

for the interval $(j-1, j]$, in which $E(Y_{ij})$ is the expected value of the longitudinal outcome, $\gamma_0 = -2$, $\gamma_1 = -0.5$, and $\gamma_2 = (0, -.05, -.10, -.15)$ over the simulations. For subject i with death in the interval $(j-1, j]$, we set Y_{ij} and all subsequent Y_{ij} to missing. We generate 5,000 datasets and calculate Type I error rates and power at the $\alpha = 0.05$ significance level. The threshold endpoints for the simulation are defined as time to the 1st, 2nd, and 3rd quartiles of the individual minimum longitudinal values or time to death, whichever comes first.

Figures 4.1-4.2 display the predicted longitudinal mean and survival probabilities for $\beta_1 = (0, -0.5)$, $\beta_2 = (0.02, 0.05)$, and $\gamma_2 = (0, -0.15)$, which represent a range of the parameter settings. Figure 4.1 shows that the treatment differences under consideration in the longitudinal measures are not very large. The survival probabilities shown in Figure 4.2 also show relatively small treatment effects. This is mainly due to the fact that we have a large sample size, and any moderate to large treatment effect would be easily detectable by a linear mixed model or a Cox model.

With the exception of Henderson's joint model, which was overly conservative, all methods consistently preserved the Type I error rate at 0.05. In general, the WLW approach had slightly greater power than the logrank approach, but the difference was very minimal. With minimal sensitivity to treatment differences in the longitudinal process ($\beta_2 = 0, 0.01$), the Cox model had the greatest power, followed by the multivariate methods and then the linear mixed models. For cases with no direct treatment effect on survival (i.e. $\gamma_2 = 0$, though treatment indirectly impacts survival through γ_1), the linear mixed model LM generally had the greatest power, followed by the multivariate methods and then the Cox model. Generally, for cases in which the longitudinal and survival processes displayed somewhat equal sensitivity to treatment differences, the multivariate methods had greater power for detecting a treatment effect than either the Cox or linear mixed models. Also, the modified (weighted) WLW approach (WLW_2) had greater power than WLW_1. The performance of Henderson's joint model varied over the simula-

tions. It generally had less power than the multivariate approaches for $\beta_1 = (-0.2, -0.5)$ (more longitudinal dropouts induced by failure), and greater power than the multivariate approaches for $\beta_1 = 0$ (fewer longitudinal dropouts).

Figure 4.3 displays the power of the various methods for detecting a treatment effect for $\beta_1 = (0, -0.5)$, $\beta_2 = (0.02, 0.05)$, and $\gamma_2 = (0, -0.15)$. For cases in which treatment does not directly impact survival ($\gamma_2 = 0$), the linear mixed models LM_1 and LM_2 have the greatest power, followed by the multivariate approaches, and then the Cox model. For datasets with greater sensitivity to treatment differences in the survival process ($\gamma_2 = -0.15$), the multivariate approach and Cox model have about equal power, while the linear mixed models LM_1 and LM_2 have the least power. Henderson's joint model is very competitive compared to the other methods in the case of little missing data ($\beta_1 = 0$) but has fairly low power with increased missing data ($\beta_1 = -0.5$) in the longitudinal process due to death.

4.3.3 Simulation Two

The second simulation generates the longitudinal data in the same manner but simulates deaths based on an increased probability of death upon reaching a pre-determined threshold rather than assuming the piecewise exponential model. We set the probability of death equal to 0.6 at all time points with $Y_{ij} < -2.5$. For subject i with death event at time j, we set Y_{ij} and all subsequent Y_{ij} to missing (and manage the patient as death at time j). Note in this setup, a subject may have technically died in the interval $(j-1, j]$ but may not have an observed death until time j. We sample 5,000 datasets and calculated Type I error rates and power at the $\alpha = 0.05$ significance level. We use the same parameter values as simulation one, except $\beta_2 = (0, .02, .03, .04, .06, .08)$ and $\beta_1 = (-0.05, -0.15, -0.20, -0.25, -0.30, -0.40, -0.50, -0.70, -0.90, -1.4)$. One could view the failure times for the terminating event as interval-censored because deaths can only occur at $j = (1, \ldots, J)$. Hence we use the discrete logistic likelihood for the Cox survival model. Figures 4.4-4.5 display the predicted longitudinal mean and survival probabilities for $\beta_2 = (0.02, 0.05, 0.08)$ and $\beta_1 = (-0.05, -0.4, -0.9)$, which represent a range of the parameter settings. As in the first simulation, the simulated treatment differences shown in Figures 4.4 and 4.5 are relatively small due to our large sample size.

With the exception of Henderson's joint model, which again was overly conservative, all methods consistently preserved the Type I error rate at 0.05. In general, the logrank approach had slightly greater power than the WLW method, but the difference was minimal. For data sets with small amounts of missing data in the longitudinal outcome due to failure, i.e. $< 20\%$ ($\beta_1 \leq -0.15$), the linear mixed models performed best, followed by the multivariate methods and then the Cox model. For data sets with 20% to 50% missing data ($-0.15 \leq \beta_1 \leq -0.5$), the multivariate approaches performed best, followed by the Cox model and then the linear mixed models. For data sets with 50% to 70% missing data ($-0.70 \leq \beta_1 \leq -0.90$), the multivariate approaches again performed best, followed by LM_1, the Cox model, and then LM_2. For data sets with greater than 70% missing data ($\beta_1 = -1.4$), LM_1 performs best followed by the multivariate methods, LM_2, and the Cox model. For data sets with $> 70\%$ missing data, the linear mixed model LM_1 had greater power than the Cox model, multivariate approaches, and the linear mixed model LM_2. One might expect the Cox model to perform best in this type of setting (i.e. extreme amounts of missing data in the longitudinal outcome). However, given the discrete sampling of our survival outcomes, most subjects experience death at time point 1 or 2, resulting in a large number of ties, and minimal sensitivity for detecting a treatment effect in the survival process. Hence, in this simulation with extreme missing data, most of the information regarding the treatment effect is contained in the longitudinal process. The modified WLW approach (WLW_2) again had greater power than WLW_1. Henderson's model had less power than the multivariate approaches for all simulations except $\beta_1 \leq -0.15$ (little missing data), with particularly low power (and a very conservative Type I error) when there was a large amount of missing data due to failure.

Figure 4.6 displays the power of the various models of detecting a treatment effect for $\beta_2 = (0.02, 0.05, 0.08)$ and $\beta_1 = (-0.05, -0.4, -0.9)$. For datasets with substantially greater sensitivity to treatment differences in the longitudinal process compared to the survival process ($\beta_2 = 0.05, \beta_1 = -0.05$), the linear mixed models perform best, followed by Henderson's joint model, the multivariate methods, and then the Cox model. In this simulation setup, there are no examples in which the sensitivity to treatment differences is substantially greater in the survival process than in the longitudinal process. For datasets with about equal sensitivity to treatment differences in the survival and longitudinal processes, the multivariate methods generally

perform better than both the Cox and linear mixed models, as well as Henderson's joint model.

4.4 Application

We illustrate our method on the previously discussed clinical trial of 2,000 patients with a chronic respiratory disorder. Due to reasons of confidentiality, these patients (1,000 treatment and 1,000 control) are a random sample from the true study population, in which patients were randomized to either treatment or control in permuted blocks with stratification by country and smoking status.

In the original analysis, time to death within 3 years was chosen *a priori* as the primary endpoint. We first evaluated the treatment effect on time to death using a Cox proportional hazards model. We regressed the survival outcome on the following predictors: treatment, baseline FEV, current smoking status (yes, no), age (< 55, 55-64, 65-74, ≥ 75), gender, body mass index (< 20, 20-25, 25-29, ≥ 29), race (white, other), and geographical region (USA, Asia-Pacific, Eastern Europe, Western Europe, other). Also, the exact days of death were available for all patients with an event, resulting in very few ties with respect to event time. There were 139 deaths (13.9%) for patients on the treatment drug and 153 deaths (15.3%) for patients on control. Survival data were available for all 2,000 subjects. Based on visual inspection of Kaplan-Meier curves, there was no evidence to contradict the proportional hazards assumption. The estimated hazard ratio for treatment versus control adjusting for covariates was 0.81 (p-value = 0.07) with a 95% confidence interval of (0.64,1.02). An unadjusted log-rank test comparing survival functions for the two treatment groups yields a test statistic of 0.84 with a p-value of 0.36, and a nonparametric ANCOVA approach adjusting for covariates yields a mean difference in log-rank scores of -.026 with a p-value of 0.12. Hence these standard analyses with survival as the primary endpoint result in non-significant results at the $\alpha = 0.05$ level.

We analyzed the longitudinal outcome FEV using a linear mixed model (LM_2) with the same predictors as the Cox model, but also including time (6, 12, 18, 24, 20, and 36 months) and a treatment by time interaction. The observation time was regarded as a class variable using indicator variables for each observation time. We included a random intercept to account for the intra-subject correlation. There were 297 subjects who did not have at least one FEV

measurement post-randomization in the 3 year period, of which 80 were dead at 3 years. A total of 8,372 observations from 1,703 subjects were available for this longitudinal analysis. 68% of these patients had observed FEV measurements at the 3 year mark, and 18% of the observations were missing across all possible (1703×6) measurements, mostly due to death. The treatment by time interaction term was not significant at the $\alpha = 0.05$ level (p-value = 0.10 with 5 df), but the estimates did show a trend for larger treatment differences as time increased. We evaluated the treatment difference at the last observation time (3 years), at which the greatest treatment difference was expected, resulting in an estimated difference of $60.5ml$ (p-value ≤ 0.0001) with a 95% confidence interval of (34.0, 87.0) for subjects on treatment versus control. This result is clearly signficant at the $\alpha = 0.05$ level, and leads to the conclusion that treatment is associated with higher FEV at 3 years.

We implemented the multivariate approaches to evaluate both outcomes simultaneously, adjusting for baseline FEV, current smoking status, age, gender, body mass index, race, and region. FEV measurements post randomization range from $210ml$ to $4{,}030ml$ with a median of $1{,}180ml$. We defined three cutpoints based on the quartiles of the individual minimum FEV measurements. This results in four threshold endpoints: time to FEV $\leq 1,300ml$ or death, time to FEV $\leq 1,010ml$ or death, time to FEV $\leq 740ml$ or death, and time to death. Although we could have required subjects to maintain FEV values below a cutpoint for two or more observations to observe a threshold event, it is clinically relevant in this example to simply define a threshold event as one observed FEV value below a given cutpoint (or death). For subjects with no FEV measurements and death (or censored) times greater than 130 days, we censored the first three threshold events at 130 days, the earliest time at which an FEV measurement was recorded.

The estimated differences in logrank scores are -0.050, -0.047, -0.035, and -0.026 for the four threshold events, respectively. The average difference in logrank scores for treatment versus control using multivariate nonparametric ANCOVA is -0.040 (p-value = 0.005) with a 95% confidence interval of (-0.068, -0.012). The goodness of fit statistic is 16.7 with 14 degrees of freedom (p-value = 0.27), showing lack of evidence for imbalance of covariates at randomization between treatments. Based on the logrank approach, we conclude that the treatment drug is associated with smaller logrank scores and extended survival times for reaching a threshold event

compared to control.

The respective marginal hazards ratios for the WLW model are 0.89, 0.84, 0.80, and 0.81. Using the optimal estimator, the WLW_1 hazards ratio is 0.87 (p-value$\leq$0.0001) with a 95% CI of (0.81,0.93). The modified estimator WLW_2 results in a hazards ratio of 0.83 (p-value=0.0004) with a 95% confidence interval of (0.75, 0.92). Based on the WLW approach, we conclude that the treatment drug is associated with lower hazard of reaching a threshold event compared to control, and hence is simultaneously associated with larger values of FEV and a decreased probability of death.

The results of our proposed approach versus the Cox and linear mixed model are not surprising based on our simulation results. In these data, we have a moderate to strong association between the treatment drug and FEV and moderate missing data (18% among those with longitudinal measurements, or 30% among all 2,000 patients). This leads to a very small p-value evaluating the longitudinal outcome. The sensitivity to treatment differences is much smaller in the survival process compared to the longitutinal process and results in a non-sigificant p-value for the survival endpoint. Because the WLW approach incorporates information from both processes, in this situation we would expect it to have sensitivity to treatment differences somewhere in between that of the longitudinal and survival approaches.

The investigators of this study had prior evidence of a strong treatment effect on FEV and conducted this study specifically to evaluate the treatment effect on mortality. However, had it been the case that the investigators did not know which outcome was more sensitive to treatment differences *a priori*, a better approach may have been to specify either the logrank or WLW approach as the primary analysis, which would hedge their planning with respect to selecting the endpoint with the greatest sensitivity to treatment differences.

4.5 Discussion

Our simulation studies show two examples in which the multivariate methods are shown to have good properties compared to standard approaches. One distinction in the first simulation setup is that the data are generated in two extreme circumstances. First, there are many examples in which the sensitivity to treatment differences is large in the longitudinal process, but very

small in the survival process. Second, there are many examples in which the sensitivity to treatment differences is large in the survival process, but very small in the longitudinal process. In both extreme cases, the multivariate methods consistently have performance in between longitudinal or survival methods alone. The multivariate methods generally do best relative to the separate methods when there is a somewhat equal sensitivity to treatment differences in both the longitudinal and survival processes. The second simulation does not have similar extreme differences in sensitivity due to the larger correlation between the longitudinal and survival processes.

In comparing the multivariate approaches to the joint model approach of Henderson et al. (2000), we observed greater power for the multivariate methods for most simulations with moderate missing data due to failure. In addition, in many simulations Henderson's approach had less power than both the linear mixed model and the Cox model. Based on simulation evidence, the multivariate methods have better overall performance, make fewer distributional assumptions, and are easier to implement than Henderson's joint model.

The modified contrast matrix in WLW_2 provided better performance than WLW_1 in both simulations. Although results are not shown here, we also implemented the weighted inverse matrix $\boldsymbol{C}_1$ on the logrank method and found that the simple contrast matrix $\boldsymbol{C}_2$ resulted in greater power. Additionally, we investigated several alternatives for the contrast matrix in both multivariate approaches, including $\boldsymbol{C}_3 = (0.30, 0.25, 0.25, 0.20)$, $\boldsymbol{C}_4 = (0.35, 0.3, 0.2, 0.15)$, $\boldsymbol{C}_5 = (0.35, 0.35, 0.30, 0)$, and $\boldsymbol{C}_6 = (0.6, 0, 0.4, 0)$, and found that the performance of the WLW and logrank approaches did not change much compared to the results based on $\boldsymbol{C}_2$. In particular, the small variations observed in power were more likely to occur in the WLW approach than in the logrank approach. Hence, in the context of our simulations, our methods are not overly sensitive to the specification of weights (excluding the weighted inverse matrix $\boldsymbol{C}_1$) and one can use a smaller number of cutpoints in the longitudinal process ($M = 1, 2$ resulting in 2 or 3 threshold endpoints) to achieve a similar result.

Both the WLW and logrank approaches had very similar performance in terms of power and Type I error. In general, the logrank method had slightly less power than the WLW model in the first simulation and slightly greater power than the WLW model in the second simulation. The logrank approach makes fewer assumptions than the WLW approach, but the analysis

cannot evaluate effects of covariables or treatment by covariables interactions. Also, the logrank estimated treatment parameter (i.e. mean difference in log ranks) only applies to populations which have the same distributions for covariables as the study population. The interpretation of the hazards ratio in the WLW approach may be more appealing than the corresponding parameter estimate from the logrank approach. Additionally, the WLW approach allows one to evaluate effects of covariables by treatment interactions, and the estimated treatment parameter is generalizable to populations which might not have similar distributions for the covariables. For a more thorough discussion of the advantages and disadvantages of nonparametric ANCOVA versus a modeling procedure, see Koch et al. (1998). Their proposed strategy is to specify the non-parametric analysis (e.g. logrank) as the primary evaluation of the treatment effect and to use the statistical model (e.g. WLW) as a supportive analysis.

The WLW approach assumes that failure time (in this case time to a threshold event) is continuous. However, because longitudinal measurements are usually taken at set time points, there is some ambiguity as to whether the continuous time assumption is satisfied for the threshold endpoints. For cases in which the exact time of the terminating event is known, most ties in the WLW model will be due to the longitudinal process (i.e. time to cutpoint). If one views time to longitudinal cutpoint as interval-censored, then time to reach a threshold event may be viewed as a combination of a continuous failure time and an interval-censored time. For cases in which the terminating event is observed only at the time the longitudinal measurement is taken, then time to a threshold event may be regarded as interval-censored. Hence the continuous failure time assumption required by the WLW may not be satisfied. However, our simulations did not show adversity for Type I error. One can still justify the WLW approach by taking the view that time to the terminating event and time to the longitudinal cutpoints are continuous outcomes. This is not unreasonable, especially when the exact time to terminating event is known and the time between longitudinal measurements is small.

Guo and Lin (1994) proposed a discrete version of the WLW approach for interval-censored data. Other discrete extensions of multivariate survival analysis include Kim and Xue (2002) and Goggins and Finkelstein (2000). However, these methods are not easily implemented in standard statistical software packages and are therefore not currently practical alternatives. One could use a discrete logistic model with generalized estimating equations to account for the

repeated events. We attempted to implement this model in our simulations but found that the Type I error rate was conservative for some settings and inflated in other settings, perhaps due to the large number of parameters in the model, as it requires an estimate for each interval and threshold as well as the interaction between interval and threshold (data not shown).

Although the WLW method is well suited for multivariate survival data with different types of events, some have criticized the WLW method in settings with recurrent events data. In such settings, an individual is not at risk for the $(k+1)$th event until the person experiences the kth event. However, the WLW method includes an individual in the risk set for each event from time zero until the individual has the event. This may cause the regression estimates to be overestimates of the regression parameters (Kelly and Lim, 2000). Others have argued against this criticism, claiming that the WLW method is appropriate for recurrent events data as long as the parameter estimates are interpreted correctly (Metcalfe and Thompson, 2007). In the context of our method, the threshold events represent progressively greater levels of deterioration or death, but are not recurrent events such as those discussed in Kelly and Lim (2000). Although an individual cannot experience the $(k+1)$th threshold event without having experienced the kth event, these events can happen simultaneously. Hence an individual is at risk for the $(k+1)$th event even without having had the kth event. This implies the criticisms of the WLW method for recurrent events data are not directly applicable to our method.

Our multivariate methods mainly address situations in which there exists attrition or non-reversible deterioration in the longitudinal process subject to censoring due to the survival endpoint. Our proposed strategy is very attractive in situations in which the best primary outcome is not known *a priori*. Choosing a multivariate approach as the primary analysis would ensure that the study does not choose a primary endpoint with the least sensitivity to treatment differences, and may result in greater sensitivity to treatment differences than either of the longitudinal or survival approaches separately.

The example in this article is based on random samples of a real clinical trial that was conducted to compare three treatment arms versus a placebo for patients with chronic obstructive pulmonary disease. The background, design, results, and interpretation of this trial are reported by Calverley et al. (2007). Our methods are applicable to other studies with similar types of data.

TABLE 4.1: Individual FEV measurements

	Observation time in months						
Subject	6	12	18	24	30	36	Time of death
1	2500	1800	1100	900	-	-	26
2	2000	1900	1800	1700	1650	1600	-
3	1200	-	-	-	-	-	11

Table gives FEV values at each observation time and time of death

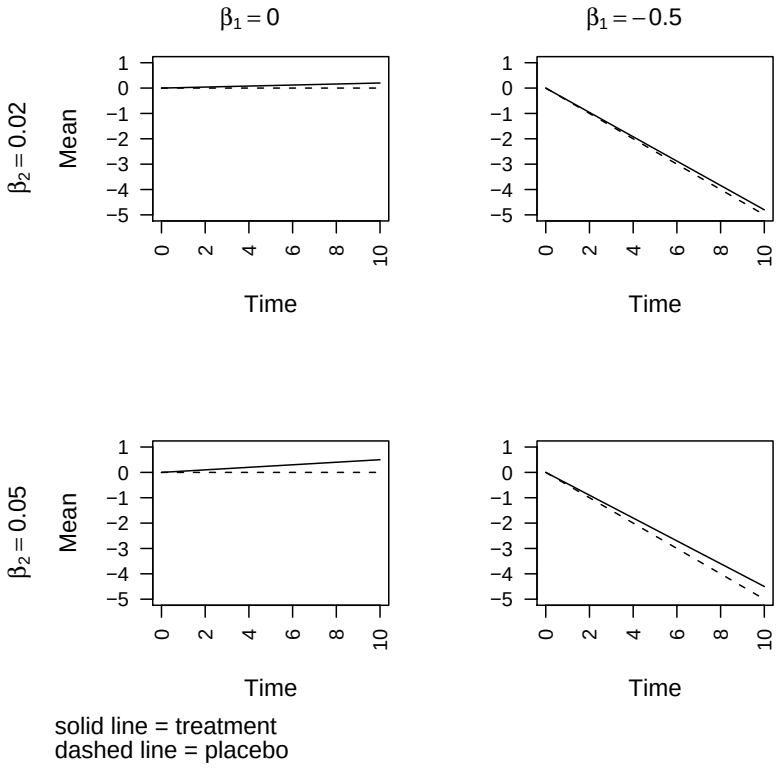


FIGURE 4.1: Simulation One: Longitudinal predicted mean

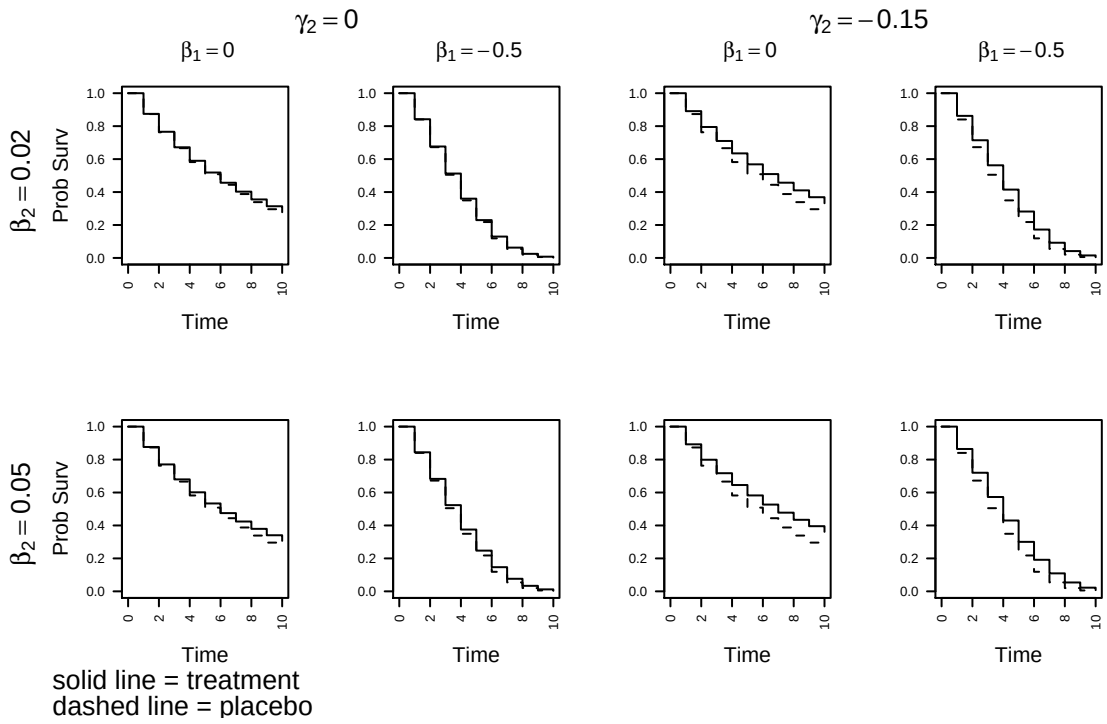


FIGURE 4.2: Simulation One: Survival probabilities

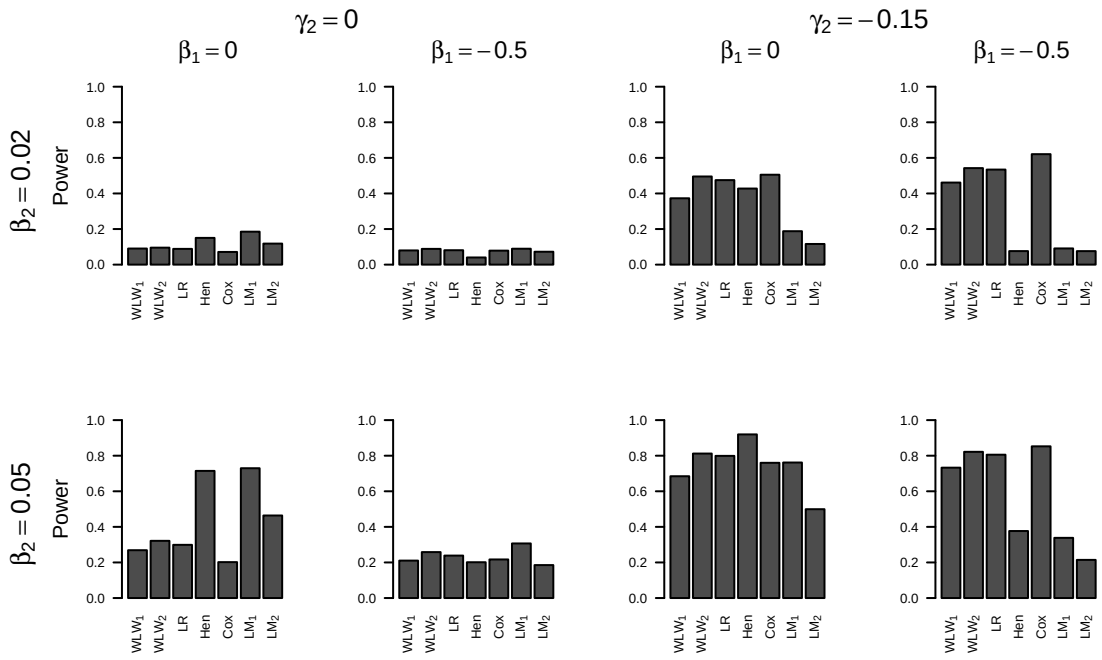


FIGURE 4.3: Simulation One: Power

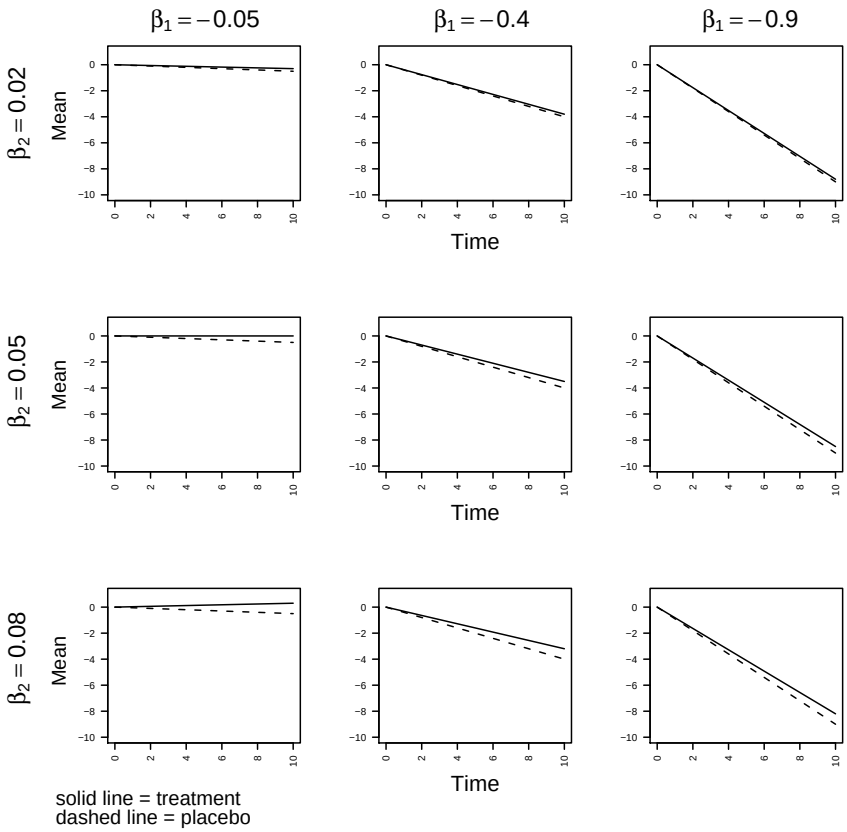


FIGURE 4.4: Simulation Two: Longitudinal predicted mean

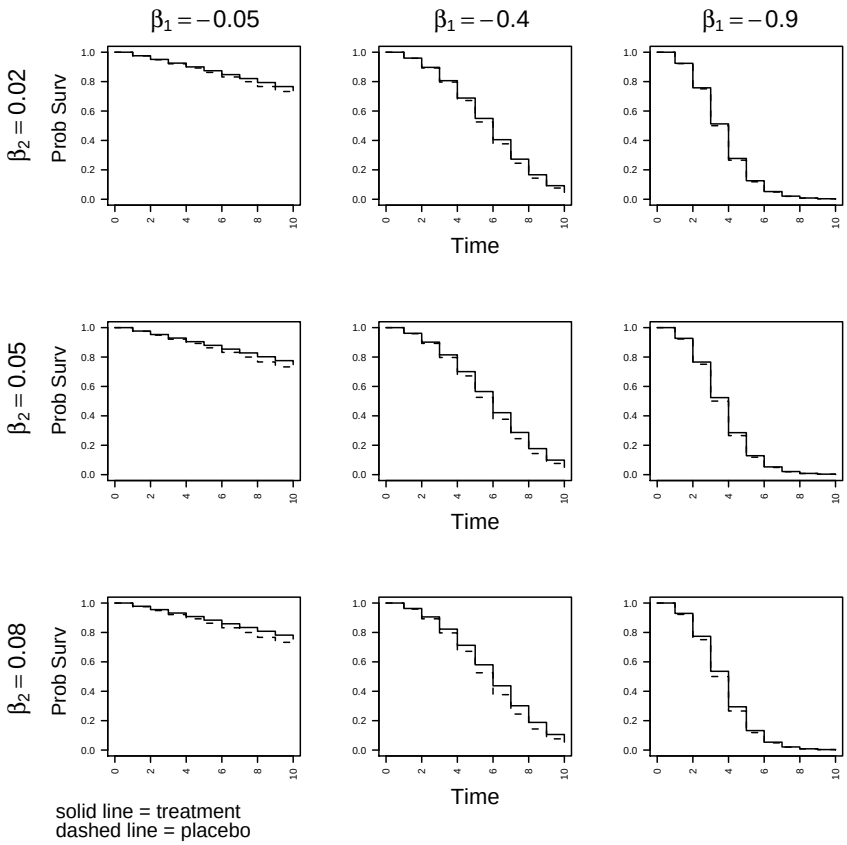


FIGURE 4.5: Simulation Two: Survival probabilities

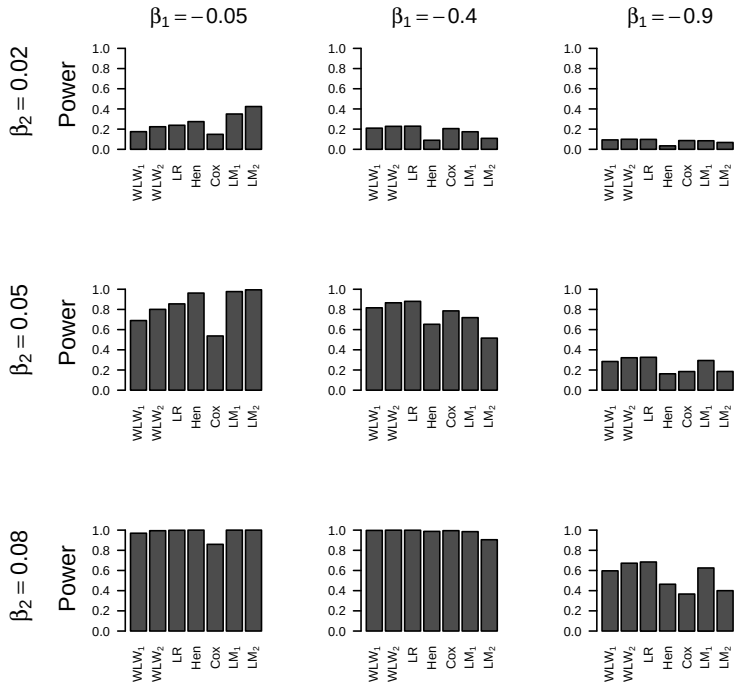


FIGURE 4.6: Simulation Two: Power

CHAPTER 5

Discussion

We have presented methods for testing random coefficients in the linear mixed model, and more generally multilevel linear models, using approximate Bayes factors. Our method incorporates default prior distributions on the random coefficients that are shown to have good frequentist properties and large-sample consistency. A major contribution of our method is the ability to test multiple random coefficients simultaneously, and to do so with relative computational efficiency. Our method does not involve computationally expensive MCMC algorithms, and only requires a maximization algorithm and numerical second derivatives. In multilevel linear models, our method is applicable to models with nested, non-nested, or cross-nested random coefficients. Hence our method is a practical and useful approach for testing random coefficients in multilevel linear models.

We also have proposed a straightforward approach for evaluating a treatment effect in correlated longitudinal and survival outcomes. Our method mainly addresses situations in which there exists attrition or non-reversible deterioration in the longitudinal process subject to censoring due to the survival endpoint. Simulations studies show that this method consistently performs either best or second best compared to standard survival or longitudinal methods alone. Our method is particularly attractive in clinical trial settings in which the primary analysis must be specified *a priori*. Our method is straightforward to implement, makes limited to no assumptions, and is a practical alternative for analyzing correlated longitudinal and survival endpoints.

APPENDIX A

Testing random effects in the linear mixed model using approximate Bayes factors

A.1 Marginal distributions for testing a random intercept

The marginal distributions in Sections 2.2.1 and 2.3.2 can be derived from the integral

$$\begin{aligned} p(\boldsymbol{Y}|M_k^{(a)}) &= \int\int \left\{ \prod_{i=1}^{n} \int p(\boldsymbol{y}_i|\boldsymbol{\zeta}_k^{(a)}, b_i, \sigma^2)\pi(b_i)db_i \right\} \pi(\sigma^2)d\sigma^2\pi(\boldsymbol{\zeta}_k^{(a)})d\boldsymbol{\zeta}_k^{(a)} \qquad \text{(A.1)} \\ &= \int\int \left\{ \prod_{i=1}^{n} p(\boldsymbol{y}_i|\boldsymbol{\zeta}_k^{(a)}, \sigma^2) \right\} \pi(\sigma^2)d\sigma^2\pi(\boldsymbol{\zeta}_k^{(a)})d\boldsymbol{\zeta}_k^{(a)} \\ &= \int\int p(\boldsymbol{Y}|\boldsymbol{\zeta}_k^{(a)}, \sigma^2)\pi(\sigma^2)d\sigma^2\pi(\boldsymbol{\zeta}_k^{(a)})d\boldsymbol{\zeta}_k^{(a)} \\ &= \int p(\boldsymbol{Y}|\boldsymbol{\zeta}_k^{(a)})\pi(\boldsymbol{\zeta}_k^{(a)})d\boldsymbol{\zeta}_k^{(a)}. \end{aligned}$$

A multivariate t-distribution for a random vector $\boldsymbol{x}$ is typically denoted as $\boldsymbol{x}_{(p\times 1)} \sim t_p(d, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, in which

$$p(\boldsymbol{x}) = \Gamma\left(\frac{d+p}{2}\right) \frac{(\pi d)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}}{\Gamma(d/2)} \left\{1 + \frac{1}{d}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}^{-\frac{d+p}{2}}. \qquad \text{(A.2)}$$

In (A.2), p is the dimension of $\boldsymbol{x}$, d is the number of degrees of freedom, $\boldsymbol{\mu}$ is the non-centrality parameter, and $\boldsymbol{\Sigma}$ is the covariance matrix. In order to express the marginal distributions $p(\boldsymbol{Y}|\boldsymbol{\zeta}_k^{(a)}, M_k^{(a)})$ in the form of (A.2), we must express the models in terms of the vector $\boldsymbol{Y} = (\boldsymbol{y}_1' \ldots, \boldsymbol{y}_n')'$.

For the ANOVA model in Section 2.2.1, we can write $M_1^{(1)}$ in terms of $\boldsymbol{Y}$ as

$$M_1^{(1)} : \boldsymbol{Y} = \mu \mathbf{1}_m + \lambda \boldsymbol{W}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{A.3}$$

in which $\mathbf{1}_m$ is a $(m \times 1)$ vector of ones, $\boldsymbol{W}$ is a $(m \times n)$ block diagonal matrix of $(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_n})$, $\boldsymbol{b} = (b_1, \ldots, b_n)'$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_n')'$. It can then be shown that $(\boldsymbol{Y}|\mu, \lambda, M_1^{(1)}) \sim t_m(2v, \mu\mathbf{1}_m, \frac{w}{v}\boldsymbol{\Sigma}^{(1)})$, in which $\boldsymbol{\Sigma}^{(1)} = (\boldsymbol{I}_m + \lambda^2 \boldsymbol{W}\boldsymbol{W}')$. Similarly, M_0 can be expressed in terms of the vector $\boldsymbol{Y}$ as

$$M_0 : \boldsymbol{Y} = \mu \mathbf{1}_m + \boldsymbol{\varepsilon}. \tag{A.4}$$

It can then be shown that $(\boldsymbol{Y}|\mu, M_0) \sim t_m(2v, \mu\mathbf{1}_m, \frac{w}{v}\boldsymbol{I}_m)$. It is also straightforward to show that $(\boldsymbol{Y}|\mu, \phi, M_1^{(2)}) \sim t_m(2v, \mu\mathbf{1}_m, \frac{w}{v}\boldsymbol{\Sigma}^{(2)})$, in which $\boldsymbol{\Sigma}^{(2)} = (\boldsymbol{I}_m + e^{2\phi}\boldsymbol{W}\boldsymbol{W}')$. For large datasets, the covariance matrix $\boldsymbol{\Sigma}$ may be too large to handle computationally in a mixed model setting. Hence it is preferable to express the multivariate t-distribution in terms of the subject-specific (independent) covariance matrices $\boldsymbol{\Sigma}_i$, as shown in equation (6)

A.2 Marginal distributions for testing a random slope

For the linear mixed model in Section 2.3.2, $M_0^{(a)}$ can be expressed in terms of the vector $\boldsymbol{Y}$ as

$$M_0^{(a)} : \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_0\boldsymbol{W}_0^{(a)}\boldsymbol{b}_0 + \boldsymbol{\varepsilon}, \tag{A.5}$$

in which $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_n')'$ is a $(m \times p)$ design matrix, $\boldsymbol{Z}_0$ is a $(m \times q)$ block diagonal matrix of $(\boldsymbol{Z}_{0,1}, \ldots, \boldsymbol{Z}_{0,n})$, $\boldsymbol{b}_0 = (\boldsymbol{b}_{0,1}', \ldots, \boldsymbol{b}_{0,n}')'$ is a $(nq \times 1)$ vector of all random effects, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_n')'$, $\boldsymbol{W}_0^{(1)} = \boldsymbol{I}_n \otimes (\boldsymbol{\Lambda}_0^{(1)}\boldsymbol{\Gamma}_0)$ and $\boldsymbol{W}_0^{(2)} = \boldsymbol{I}_n \otimes (\boldsymbol{\Lambda}_0^{(2)}\boldsymbol{\Gamma}_0)$ are $(nq \times nq)$ matrices, in which $\otimes$ denotes the right Kronecker product (whereby the matrix on the right multiplies each element of the matrix on the left). We can express $M_1^{(a)}$ in terms of the vector $\boldsymbol{Y}$ as

$$M_1^{(a)} : \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{W}_1^{(a)}\boldsymbol{b}_1 + \boldsymbol{\varepsilon}, \tag{A.6}$$

in which $\boldsymbol{Z}_1$ is a block diagonal matrix of $(\boldsymbol{Z}_{1,1}, \ldots, \boldsymbol{Z}_{1,n})$, $\boldsymbol{b}_1 = (\boldsymbol{b}'_{1,1}, \ldots, \boldsymbol{b}'_{1,n})'$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \ldots, \boldsymbol{\varepsilon}'_n)'$, $\boldsymbol{W}_1^{(1)} = \boldsymbol{I}_n \otimes (\boldsymbol{\Lambda}_1^{(1)}\boldsymbol{\Gamma}_1)$, and $\boldsymbol{W}_1^{(2)} = \boldsymbol{I}_n \otimes (\boldsymbol{\Lambda}_1^{(2)}\boldsymbol{\Gamma}_1)$. It can then be shown that $p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda}_k^{(a)}, \boldsymbol{\gamma}_k, M_k^{(a)}) \sim t_m(2v, \boldsymbol{X}\boldsymbol{\beta}, \frac{w}{v}\boldsymbol{\Sigma}_k^{(a)})$, in which $\boldsymbol{\Sigma}_k^{(a)} = (\boldsymbol{I}_m + \boldsymbol{Z}_k\boldsymbol{W}_k^{(a)}\boldsymbol{W}_k'^{(a)}\boldsymbol{Z}_k')$. As with testing a random intercept in the ANOVA setup, the covariance matrix $\boldsymbol{\Sigma}_k^{(a)}$ may be too large to handle computationally in a mixed model setting.

APPENDIX B

Analyzing Correlated Longitudinal and Survival Data in Clinical Trials Using Multivariate Time-to-Event Methods

B.1 The Wei-Lin-Weissfeld Method

Wei *et al.* Wei et al. (1989) showed that

$$\hat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \boldsymbol{V}), \tag{B.1}$$

in which $\boldsymbol{V}$ is estimated by

$$\hat{\boldsymbol{V}} = \begin{bmatrix} \hat{\boldsymbol{V}}_{11} & \hat{\boldsymbol{V}}_{12} & \dots & \hat{\boldsymbol{V}}_{1M} \\ \hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{22} & \dots & \hat{\boldsymbol{V}}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\boldsymbol{V}}_{M1} & \hat{\boldsymbol{V}}_{M2} & \dots & \hat{\boldsymbol{V}}_{MM} \end{bmatrix}. \tag{B.2}$$

The estimated covariance matrix $\hat{\boldsymbol{V}}$ is composed of the sub-matrices $\hat{\boldsymbol{V}}_{mm'} = (\boldsymbol{R}_m \hat{\boldsymbol{A}}_m)'(\boldsymbol{R}_{m'} \hat{\boldsymbol{A}}_{m'})$, in which $\hat{\boldsymbol{A}}_m$ is the inverse of the information matrix and $\boldsymbol{R}_m$ is the matrix of score residuals for event outcome m. Conveniently, the quantity $\boldsymbol{R}_m \hat{\boldsymbol{A}}_m$ is common output in most software package and is known as the matrix of "dfbeta" residuals. The "dfbeta" residuals represent the approximate change in a parameter estimate when the ith observation is omitted. It follows that the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be obtained as a function of the "dfbeta" residuals.

B.2 Nonparametric Analysis of Covariance with Logrank Scores

Let n_j be the number of observations at risk at the beginning of the jth interval,

$$n_j = \begin{cases} N, \ j = 1 \\ N - \sum_k (n_{k0} + n_{k1}), \ k = 1, \ldots, (j-1) \text{ and } j > 1, \end{cases} \tag{B.3}$$

in which N is the sample size, n_{k0} is the number of censored observations in the kth interval, and n_{k1} is the number of observed endpoints in the kth interval. Then the logrank scores for the jth interval are

$$C_{jd} = d - \sum_k (n_{k1}/n_k), \ k = 1, \ldots, j, \tag{B.4}$$

in which $d = 1$ for observed endpoints and $d = 0$ for censored endpoints.

Suppose we are interested in comparing two treatments for M logrank outcomes adjusting for p covariates. Let treatment i have sample size n_i, mean response $\bar{\boldsymbol{y}}_i$ of dimension $(M \times 1)$ and a mean of covariates $\bar{\boldsymbol{x}}_i$ of dimension $(p \times 1)$. Let $\boldsymbol{d} = (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)$ and $\boldsymbol{u} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$. We fit the model

$$E[\boldsymbol{f}] = E \begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{u} \end{bmatrix} \hat{=} \begin{bmatrix} \boldsymbol{I}_M \\ \boldsymbol{0}_{(p \times M)} \end{bmatrix} \hat{\boldsymbol{\beta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{B.5}$$

using weighted least squares with weights based on the covariance matrix $\boldsymbol{V}_0$. Under H_0,

$$\boldsymbol{V}_0 = \frac{n_1 + n_2}{n_1 n_2 (n_1 + n_2 - 1)} \left\{ \sum_{i=1}^{2} \sum_{k=1}^{n_i} \begin{bmatrix} (\boldsymbol{y}_{ik} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{ik} - \bar{\boldsymbol{y}})' & (\boldsymbol{y}_{ik} - \bar{\boldsymbol{y}})(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}})' \\ (\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}})(\boldsymbol{y}_{ik} - \bar{\boldsymbol{y}})' & (\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}})' \end{bmatrix} \right\} \tag{B.6}$$

in which $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{x}}$ are means for all patients with treatments ignored. Additional covariance estimates are possible under the alternative hypothesis of a treatment difference Tangen and Koch (1999b), Tangen and Koch (1999a). The weighted least squares estimator $\hat{\boldsymbol{\beta}}$ is obtained from

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_0^{-1}\boldsymbol{f} \tag{B.7}$$

and its estimated variance is given by

$$\hat{V}_{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}'\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}. \tag{B.8}$$

A criterion for departures from (B.5) in terms of random imbalances takes the form

$$Q = (\boldsymbol{f} - \hat{\boldsymbol{f}})'\boldsymbol{V}_0^{-1}(\boldsymbol{f} - \hat{\boldsymbol{f}}) \tag{B.9}$$

in which $\hat{\boldsymbol{f}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The statistic Q approximately has a chi-square distribution with p degrees of freedom and assesses the amount of random imbalance in the covariates at randomization.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrox, B. N. and Caski, F., editors, *In Second International Symposium on Information Theory*, page 267, Budapest. Akademiai Kiado.

Bartlett, M. S. (1957). Comment on "A Statistical Paradox" by D. V. Lindley. *Biometrika* **44**, 533–534.

Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference* **112**, 241–258.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

Berkhof, J. and Snijders, T. A. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics* **26**, 133–152.

Brown, E. R. and Ibrahim, J. G. (2003a). Bayesian approaches to join cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, 686–693.

Brown, E. R. and Ibrahim, J. G. (2003b). A Bayesian semiparametric joint hierarchical model for longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, 221–228.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models (with discussion). *Bayesian Analysis* **1**, 473–514.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA.

Buka, S., Brennan, R., Rich-Edwards, J., Raudenbush, S. and Earls, F. (2003). Neighborhood support and the birth weight of urban infants. *American Journal of Epidemiology* **157**, 1–8.

Bycott, P. and Taylor, J. (1998). A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* **17**, 2061–2077.

Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized mixed models. *Biometrics* **62**, 446–457.

Calabrese, J. R., Bowden, C. L., Sachs, G., Yatham, L. N. and Behnke, K. (2003). A placebo-controlled 18-month trial of lamotrigine and lithium maintenance treatment in recently depressed patients with bipolar I disorder. *Journal of Clinical Psychiatry* **64**, 1013–1024.

Calverley, P. M. A., Anderson, J. A., Celli, B., Ferguson, G. T., Jenkins, C., Jones, P. W., Yates, J. C. and Vestbo, J. (2007). Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *The New England Journal of Medicine* **356**, 775–789.

Celeux, G., Forbes, F., Robert, C. and Titterington, M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1**, 651–674.

Chakrabarti, A. and Ghosh, J. K. (2006). A generalization of BIC for the general exponential family. *Journal of Statistical Planning and Inference* **136**, 2847–2872.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.

Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society, Series B* **59**, 157–171.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Crainiceanu, C. M. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* **92**, 91–103.

Crainiceanu, C. M. (2007). *Likelihood ratio testing for zero variance components in linear mixed models*, . Model Uncertainty in Latent Variable and Random Effects Models (*to appear*).

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

Dafni, U. G. and Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics* **54**, 1445–1462.

Dang, Q. Y., Mazumdar, S., Anderson, S. J., Houck, P. R. and Reynolds, C. F. (2007). Using trajectories from a bivariate growth curve as predictors in a Cox regression model. *Statistics in Medicine* **26**, 800–811.

DeGrutolla, V. and Tu, X. M. (1994). Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50**, 1003–1014.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* **43**, 49–93.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–565.

Elashoff, R. M., Li, G. and Li, N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine* **26**, 2813–2835.

Erkanli, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *Journal of the American Statistical Association* **89**, 250–258.

Faucett, C. J. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1685.

Faucett, C. L., Schenker, N. and Taylor, J. M. G. (2002). Survival analysis using auxiliary

variables via multiple imputation, with application to AIDS clinical trials data. *Biometrics* **58**, 37–47.

Feng, Z. and McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter lies on the boundary of the parameter space. *Statistics & Probability Letters* **11**, 325–332.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

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

Follman, D. and Wu, M. C. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151–168.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician* **37**, 152–155.

Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrary single-censored samples. *Biometrika* **52**, 203–223.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515–533.

Gelman, A. (2007). Running WinBugs and OpenBugs from R. Available at www.stat.columbia.edu/∼gelman/bugsR/.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* **27**, 2865–2873.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

Genz, A. (1991). An adaptive numerical integration algorithm for simplices. In Sherwani, N. A., de Doncker, E. and Kapenga, J. A., editors, *Computing in the 90's, Proceedings of the First Great Lake Computer Science Conference*, pages 279–292, New York. Springer.

Genz, A. and Kass, R. E. (1993). Subregion-adaptive integration of functions having a dominant peak. *Journal of computational and graphical Statistics* **6**, 92–111.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1340.

Gilks, W. R. and Roberts, G. O. (1996). *Strategies for Improving MCMC*, page 96. Chapman and Hall, New York.

Goggins, W. B. and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* **56**, 940–943.

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association* **53**, 799–813.

Greven, S., Crainiceanu, C. M., Kuechenoff, H. and Peters, A. (2008). Restricted likelihood ratio

testing for zero variance components in the linear mixed models. *Journal of Computational and Graphical Statistics* (to appear) .

Guo, S. W. and Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics* **50**, 632–639.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician* **58**, 16–24.

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association* **96**, 1122–1133.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Hogan, J. W. and Laird, N. M. (1997a). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16**, 239–257.

Hogan, J. W. and Laird, N. M. (1997b). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* **16**, 259–272.

Howard, D. L., Marshall, S. S., Kaufman, J. S. and Savitz, D. A. (2006). Variations in low birth weight and preterm delivery among blacks in relation to ancestry and nativity: New York City, 19982002. *Pediatrics* **118**, E1399–E1405.

Hsiao, C. K. (1997). Approximate Bayes factors when a mode occurs on the boundary. *Journal of the American Statistical Association* **92**, 656–663.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics* **62**, 1037–1043.

Ibrahim, J. (2005). Course notes: Biostatistics 321, Bayesian statistics. University of North Carolina at Chapel Hill.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. Oxford University Press, Oxford, U.K.

Kass, R. E. (1993). Bayes factors in practice. *The Statistician* **42**, 551–560.

Kass, R. E. and Natarajan, R. (2006). A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 535–542.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kass, R. E. and Vaidyanathan, S. (1992). Improving the Laplace approximation using posterior simulation. *Journal of the Royal Statistical Society, Series B* **54**, 129–144.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**,

928–934.

Katz, R. W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics* **23**, 243–249.

Kelly, P. J. and Lim, L. L.-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine* **19**, 13–33.

Kim, M. Y. and Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine* **21**, 3715–3726.

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

Koch, G. G., Sen, P. K. and Amara, I. A. (1985). Log-rank scores, statistics, and tests. In Kotz, S. and Johnson, N. L. (eds), *Encyclopedia of Statistical Sciences*, Wiley, New York.

Koch, G. G., Tangen, C. M., Jung, J.-W. and Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* **17**, 1863–1892.

Kuonen, D. (2003). Numerical integration in S-plus or R: A survey. *Journal of Statistical Software* **8**, 1–14.

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

Law, N. J., Taylor, J. M. G. and Sandler, H. M. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics* **3**, 547–563.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of Statistical Software* **103**, 410–423.

Lin, H., Turnbull, B. W., McCulloch, C. E. and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *The American Statistician* **97**, 53–65.

Lin, X. (1997). Variance components testing in generalised linear models with random effects. *Biometrika* **84**, 309–326.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

McCulloch, R. E. and Rossi, P. E. (1991). A Bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics* **49**, 141–168.

Metcalfe, C. and Thompson, S. G. (2007). Wei, Lin, and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome? *Statistical methods in medical research* **16**, 103–122.

Miller, A. J. (1984). Selection of subsets of regression variables (with discussion). *Journal of the Royal Statistical Society, Series A* **147**, 389–425.

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

Mori, M., Woodworth, G. G. and Woolson, R. F. (1992). Application of empirical Bayes inference to estimation of rate of change in the presence of informative right censoring. *Statistics in Medicine* **11**, 621–631.

Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association* **95**, 227–237.

Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal* **7**, 308–313.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Statistical Models*. The McGraw-Hill Companies, Inc.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B* **56**, 3–48.

OCampo, P., Xue, X., Wang, M. and Caughy, M. (1997). Neighborhood risk factors for low birthweight in Baltimore: A multilevel analysis. *American Journal of Public Health* **87**, 1113–1118.

Osypuk, T. L. and Acevedo-Garcia, D. (2008). Are racial disparities in preterm birth larger in hypersegregated areas? *American Journal of Epidemiology* **167**, 1295–1304.

Pauler, D. K., Wakefield, J. C. and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association* **94**, 1242–1253.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 205–207.

Prentice, R. L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* **35**, 861–867.

Raftery, A. E. (1986a). Choosing models for cross-classifications. *American Sociological Review* **51**, 145–146.

Raftery, A. E. (1986b). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, Series B* **48**, 259–250.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In Bollen, K. and Long, J., editors, *Testing Structural Equation Models*, pages 163–180. Sage, Newbury Park, CA.

Raftery, A. E. (1995). Bayesian model selection in social research. *The American Sociological Association* **25**, 111–163.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251–266.

Raftery, A. E. (1999). Bayes factors and BIC: Comment on "A critique of the Bayesian information criterion for model selection". *Sociological Methods and Research* **27**, 411–427.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). *Bayesian Statistics* **8**, 1–45.

Rauh, V., Andrews, H. and Garfinkel, R. (2001). The contribution of maternal age to racial disparities in birthweight: A multilevel perspective. *American Journal of Public Health* **91**, 1815–1824.

Roberts, E. (1997). Neighborhood social environments and the distribution of low birthweights in Chicago. *American Journal of Public Health* **87**, 597–603.

Saville, B. and Herring, A. (2008). Testing random effects in the linear mixed model using approximate Bayes factors. *Biometrics* (to appear) .

Savitz, D. A., Janevic, T. M., Engel, S. M., Kaufman, J. S. and Herring, A. H. (2008). Ethnicity and gestational diabetes in New York City, 1995-2003. *British Journal of Obstetrics & Gynecology* **115**, 969–978.

Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* **11**, 1861–1870.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

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

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* **56**, 49–62.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117–126.

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

Sinharay, S. and Stern, H. S. (2001). Bayes factors for variance component models in generalized mixed models. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, pages 507–516. ISBA 2000 Proceedings.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B* **42**, 213–220.

Song, X. and Davidian, M. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58**, 742–753.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–640.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R. and Lunn, D. (2003). WinBUGS User Manual, Version 1.4. Available at www.mrc-bsu.cam.ac.uk/bugs.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series B* **41**, 276–278.

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

Stram, D. O. and Lee, J. W. (1995). Correction to "Variance components testing in the longitudinal mixed effects model". *Biometrics* **51**, 1196.

Sullivan, L. M., Dukes, K. A. and Losina, E. (1999). Tutorial in Biostatistics: An introduction to hierarchical linear modelling. *Statistics in Medicine* **18**, 855–888.

Tangen, C. M. and Koch, G. G. (1999a). Complementary nonparametric analysis of covariance of logistic regression in a randomized clinical trial setting. *Journal of Biopharmaceutical Statistics* **9**, 45–66.

Tangen, C. M. and Koch, G. G. (1999b). Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. *Journal of Biopharmaceutical Statistics* **9**, 307–338.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistician* **81**, 82–86.

Tseng, Y.-K., Hsigh, F. and Wang, J.-L. (2005). Joint modeling of accelerated failure time and longitudinal data. *Biometrika* **92**, 587–603.

Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88**, 447–458.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834.

Tsiatis, A. A., Degruttola, V. and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.

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

Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895–905.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92–107.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research* **27**, 359–397.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**, 1065–1073.

Wu, M. C. and Bailey, K. R. (1988). Analysing changes in the presence of informative right

censoring caused by death and withdrawal. *Statistics in Medicine* **7**, 337–346.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* **45**, 939–955.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics* **44**, 175–188.

Wu, M. C., Hunsberger, S. and Zucker, D. (1994). Testing for differences in changes in the presence of censoring: parametric and non-parametric methods. *Statistics in Medicine* **13**, 635–646.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Xu, J. and Zeger, S. L. (2001a). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics* **57**, 795–802.

Xu, J. and Zeger, S. L. (2001b). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society Series C - Applied Statistics* **50**, 375–387.

Yu, M., Law, N. J., Taylor, J. M. G. and Sandler, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14**, 835–862.

Zellner, A. (1986). *On assessing prior distributions and Bayesian regression analysis with g-prior distributions*, pages 233–243. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pages 585–603.

Zhang, D. and Lin, X. (2008). *Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics*, . Model Uncertainty in Latent Variable and Random Effects Models (*to appear*).

Zink, R. C. and Koch, G. G. (2002). SAS Macro NParCov Version 2, Non-Parametric Analysis of Covariance. Biometric Consulting Laboratory, Department of Biostatistics, University of North Carolina at Chapel Hill.