Marginally Interpretable Generalized Linear Mixed Models

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Jeffrey J. Gory, B.S., M.S.

Graduate Program in Statistics

The Ohio State University

2017

Dissertation Committee:

Peter Craigmile, Ph.D., Advisor

Steven MacEachern, Ph.D., Co-Advisor

Eloise Kaizar, Ph.D.

Jennifer Sinnott, Ph.D.

Abstract

A popular approach for relating correlated measurements of a non-Gaussian response variable to a set of predictors is to introduce latent random variables and fit a generalized linear mixed model. The conventional strategy for specifying such a model leads to parameter estimates that must be interpreted conditional on the latent variables. In many cases, interest lies not in these conditional parameters, but rather in marginal parameters that summarize the average effect of the predictors across the entire population. Due to the structure of the generalized linear mixed model, the average effect across all individuals in a population is generally not the same as the effect for an average individual. Further complicating matters, obtaining marginal summaries from a generalized linear mixed model often requires evaluation of an analytically intractable integral or use of an approximation. Another popular approach in this setting is to fit a marginal model using generalized estimating equations. This strategy is effective for estimating marginal parameters, but leaves one without a formal model for the data with which to assess quality of fit or make predictions for future observations. Thus, there exists a need for a better approach.

We define a class of marginally interpretable generalized linear mixed models that lead to parameter estimates with a marginal interpretation while maintaining the desirable statistical properties of a conditionally specified model. The distinguishing feature of these models is an additive adjustment that accounts for the curvature of the link function and thereby preserves a specific form for the marginal mean after integrating out the latent random variables. We discuss the form and interpretation of marginally interpretable generalized linear mixed models under various common link functions and compare inferences obtained from these models to those obtained from conventional generalized linear mixed models, highlighting the advantages of the marginally interpretable formulation over the conventional one. We also address computational issues associated with marginally interpretable generalized linear mixed models in both a classical framework and a Bayesian framework. Namely, we introduce an accurate and efficient method for evaluating the logistic-normal integral that arises in logistic mixed effects models and, for the Bayesian setting, we propose a modification of a standard Markov chain Monte Carlo algorithm that allows for more efficient posterior simulation in models with many latent random variables.

Acknowledgments

Throughout my many years as a student – from grade school and high school in the Strongsville City School District to my time as an undergraduate at NC State to graduate school at Ohio State – I have had the good fortune of being taught by many quality educators, and I am grateful to all of them. I would specifically like to thank my co-advisors, Dr. Peter Craigmile and Dr. Steven MacEachern, for their guidance over the last few years, and also Dr. Eloise Kaizar and Dr. Jennifer Sinnott for serving on my committee.

I am also grateful for the financial support I have received during my time as a graduate student. Namely, I would like to thank the Graduate School for selecting me to receive the Susan L. Huntington Dean's Distinguished University Fellowship, which allowed me to focus on coursework during my first two years and on writing this dissertation during my final year. Additionally, I am indebted to Ohio State's Statistical Consulting Service and the Nationwide Center for Advanced Customer Insights within the Fisher College of Business, which provided support during the intervening years as well as enriching experiences that will serve me well in my future career.

Vita

2012	B.S., Applied Mathematics, North Carolina State University
2012	B.S., Statistics, North Carolina State University
2014	M.S., Statistics, The Ohio State University

Fields of Study

Major Field: Statistics

Table of Contents

Page

List of Tables

List of Figures

Figure **Page**

Chapter 1: Introduction

The generalized linear model (GLM) provides a framework for modeling a non-Gaussian response variable as a function of a set of predictor variables. Given a response Y and a corresponding set of p predictors $\mathbf{x} = (x_0, \ldots, x_{p-1})^T$, such a model assumes that Y represents a draw from a distribution with mean μ and then relates μ to a linear combination of $\mathbf{x}$ through a *link function* $g(\cdot)$. Specifically, one models

$$\eta = g(\mu) = g(\mathrm{E}[Y]) = \mathbf{x}^T\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a p-vector of unknown parameters and η is standard notation for the *linear predictor* on the link scale. This mean function is also commonly written as

$$\mu = \mathrm{E}[Y] = h(\mathbf{x}^T\boldsymbol{\beta}),$$

where $h(\cdot) = g^{-1}(\cdot)$ is the inverse link function. Table 1.1 lists several common link functions and their corresponding inverse link functions. The normal-theory linear model can be viewed as a GLM with an identity link for which the data are assumed to follow a Gaussian distribution. Two popular classes of GLMs are log-linear models and logistic regression models. The former employs a log link and a Poisson distribution for Y whereas the latter uses a logit link and a binomial distribution. Additional details about GLMs can be found in Nelder and Wedderburn (1972) and McCullagh and Nelder (1989).

Table 1.1: Common link functions for GLMs and their inverses

Link Function	$g(\mu)$	$h(\eta)$
identity	μ	η
natural logarithm	$\log(\mu)$	$\exp(\eta)$
probit[a]	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
logit	$\log\big(\frac{\mu}{1-\mu}\big)$	$\frac{\exp(\eta)}{1+\exp(\eta)} = \frac{1}{1+\exp(-\eta)}$
complementary log-log	$\log\big(-\log(1-\mu)\big)$	$1 - \exp\big(-\exp(\eta)\big)$
square root[b]	$\sqrt{\mu} = \mu^{1/2}$	η^2
reciprocal[b]	$\frac{1}{\mu} = \mu^{-1}$	$\frac{1}{\eta} = \eta^{-1}$

[a] $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution

[b] Typically assume $\mu, \eta > 0$

When fitting a GLM, one typically assumes that the data represent independent observations from some underlying distribution. This assumption may not be reasonable if, for example, observations naturally fall into groups or clusters. In such a case, observations from the same group or cluster are likely to be more similar than observations from different groups or clusters. Dependence among observations can also arise from measurements taken across time or space. Here, observations taken near one another are likely more similar than observations taken farther apart. Regardless of its source, the presence of correlation in the data renders independence assumptions invalid and is an aspect of the data that must be accounted for to obtain reliable inference on the parameters in the model.

One might also find that the variation observed in the data exceeds that which is expected based on the distributional assumption for the data. Many non-Gaussian distributions have a well-defined mean-variance relationship from which the variance can be computed as a function of the mean. The data are said to be *overdispersed* if the observed variation in the data is greater than one would expect for a particular distribution with a particular mean (see McCullagh and Nelder, 1989, Section 4.5). The presence of overdispersion typically indicates the existence of some underlying structure to the data that is not captured by the standard independence assumption. For example, if the data are clustered into groups, heterogeneity among the clusters could lead the data to exhibit greater variability than would be expected if there were no clustering.

When the link function in a GLM is nonlinear, how one chooses to account for dependence or overdispersion has a meaningful impact on the interpretation of the model and, in turn, on the conclusions one can draw from the model. Two broadly defined modeling strategies are *conditional models*, which assume that dependence arises from the presence of unobserved latent random variables, and purely *marginal models*, which largely treat the correlation among the observations as a nuisance. This dissertation discusses the impact of modeling decisions related to GLMs for dependent data and proposes a modeling framework that has many advantages over existing strategies.

1.1 Conditional Models

In the spirit of linear mixed models for normally distributed data (see, for example, Henderson et al., 1959; Henderson, 1975; Laird and Ware, 1982; Verbeke and Molenberghs, 2000), one might account for dependence in a GLM by introducing latent random variables known as *random effects*. Using this strategy, the correlation or overdispersion present in the data is assumed to arise from an explainable source of variation that, unlike the predictors $x_0, \ldots, x_{p-1}$, is not of direct interest to the study. As an example, consider a longitudinal study with repeated measurements taken on the same individuals over time. Although the presence of different subjects is a known source of variation, the variability in the response due to the predictor variables $x_0, \ldots, x_{p-1}$ is of greater interest. Whereas $x_0, \ldots, x_{p-1}$ might represent specific treatment levels that one wants to compare, differences in the subjects are a consequence of the study design. Interest lies not in the specific subjects that happened to be observed, but rather in understanding the heterogeneity in the population of subjects and how it relates to the variability present in the data. The predictors $x_0, \ldots, x_{p-1}$ have specific, fixed values of interest and are therefore known as *fixed effects*. The random effects, in contrast, are viewed as a random sample from a population of such effects, and we would likely see different realizations of the random effects if we were to repeat the study. A model that incorporates both fixed effects and random effects is known as a *mixed effects model* or simply as a *mixed model*. A GLM that includes random effects is therefore known as a *generalized linear mixed model* (GLMM). For more information about the general form of a GLMM, see McCulloch et al. (2008).

Much like a GLM, a GLMM relates the mean of a response Y to a set of p predictors $\mathbf{x}$ through a link function $g(\cdot)$. In addition to the fixed predictors $\mathbf{x}$, the linear component of a GLMM also includes q random effects $\mathbf{U}$ with q-variate density $f_{\mathbf{U}}$. Conditional on $\mathbf{U}$, one typically assumes the data are independent observations from a parametric distribution with density $f_{Y|\mathbf{U}}$ and mean $\mathrm{E}[Y|\mathbf{U} = \mathbf{u}]$, and then models the conditional mean as

$$\mu = \mathrm{E}[Y|\mathbf{U} = \mathbf{u}] = h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u}),$$

where $\mathbf{d}$ is a q-vector of covariates associated with the random effects. This model is hierarchical in structure and does not directly assume a marginal distribution for Y. Rather, distributional assumptions are made for $\mathbf{U}$ and for Y given $\mathbf{U}$, and one must integrate over the random effects density $f_{\mathbf{U}}$ to obtain the marginal distribution for Y. The GLMM is a *conditional model* because the mean μ is conditioned on the random effects $\mathbf{U}$. The fixed effects parameters $\boldsymbol{\beta}$ have a *subject-specific* or *cluster-specific* interpretation, meaning each element of $\boldsymbol{\beta}$ provides information about the effect of the corresponding predictor on the response for a specific subject or cluster with a specific realization of the random effect. This interpretation does not always make sense because some predictors, such as indicators of race or sex, never change within a single subject.

Since one assumes an underlying distribution for the data, a GLMM is a fully specified, probabilistic model and can be fit using maximum likelihood estimation. However, for many choices of link function and random effects distribution, evaluation of the marginal likelihood involves an analytically intractable integral. To overcome this issue, one could use numerical integration or maximize an approximation of the marginal likelihood instead of the true likelihood and thereby avoid the intractable integral. An alternative to maximum likelihood estimation is to adopt a Bayesian framework and employ *Markov chain Monte Carlo* (MCMC) to sample from the posterior distribution of the unknown parameters. Details regarding popular strategies for fitting GLMMs are provided in Chapter 4.

1.2 Marginal Models

Interest often lies in *marginal* or *population-averaged* effects rather than subject-specific effects. That is, one often wants to know the average effect of a particular covariate on the response across the entire population instead of the corresponding effect for a specific individual in the population. Although one could obtain marginal predictions from a GLMM, it is common to directly model the marginal mean using what is known as a *marginal model*. Such a model involves specification of a mean structure, typically written as $\mu = \mathrm{E}[Y] = h(\mathbf{x}^T\boldsymbol{\beta})$, and a covariance structure, usually with no distributional form explicitly assumed for the data.

Estimation of $\boldsymbol{\beta}$ for this type of model is ordinarily accomplished using *generalized estimating equations* (GEE). Introduced by Liang and Zeger (1986) and Zeger and Liang (1986), GEE involves solving a set of score equations analogous to those used in estimation of ordinary GLMs (see Green, 1984; McCullagh and Nelder, 1989, Section 2.5). The score equations used in the GEE approach include a ”working” correlation matrix to account for the dependence among the observations. Liang and Zeger (1986) showed that GEE can yield consistent estimates of the fixed effects parameters $\boldsymbol{\beta}$ even when the assumed covariance structure is incorrect, but these estimates can be inefficient if the “working” correlation structure does not accurately represent the true correlation structure (see Fitzmaurice, 1995; Mancl and Leroux, 1996). The parameters that characterize the correlation, which we denote by $\boldsymbol{\alpha}$, are usually treated as nuisance parameters. For situations where the correlation parameters $\boldsymbol{\alpha}$ are also of interest, a related approach simultaneously estimates both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ (Prentice, 1988; Zhao and Prentice, 1990; Prentice and Zhao, 1991). This extension was termed GEE2 by Liang et al. (1992), who showed that, unlike GEE, it requires correct specification of the correlation structure to obtain consistent parameter estimates.

Purely marginal models are popular because they yield parameters with a desirable, population-averaged interpretation and are easy to fit via GEE. They are not, however, satisfactory statistical models. Marginal models specified only in terms of the mean and covariance of the data, with no assumption for the underlying distribution of the data, are not generative and cannot easily be used to make predictions at the individual level. In fact, it is possible that the score equations used in the GEE method do not correspond to the gradient of any scalar function and therefore cannot be integrated to obtain a likelihood function (McCullagh and Nelder, 1989, Section 9.3). Thus, a marginal model specified only in terms of the first two moments of the data does not necessarily correspond to any probabilistic model (see Lindsey and Lambert, 1998). The absence of a likelihood function in purely marginal models also makes it difficult to check and compare models.

1.3 Conditional Models Versus Marginal Models

The distinction between marginal and conditional models is important because the parameters in these two types of models are generally not the same when the link function $g(\cdot)$ is nonlinear (see Zeger et al., 1988; Neuhaus et al., 1991; Neuhaus and Jewell, 1993; Diggle et al., 2002; Ritz and Spiegelman, 2004). Specifically, the two modeling strategies generally lead to different parameter estimates with different interpretations. The lack of equivalence stems from the curvature of the link function and is a consequence of Jensen's inequality. Consider, for example, a GLMM with a single random effect, where the conditional mean is given by $\mathrm{E}[Y|U] = h(\mathbf{x}^T\boldsymbol{\beta} + U)$ and we assume $\mathrm{E}[U] = 0$. Using this expression for $\mathrm{E}[Y|U]$, we can express the marginal mean of Y as

$$\mathrm{E}[Y] = \mathrm{E}\big[\mathrm{E}[Y|U]\big] = \mathrm{E}[h(\mathbf{x}^T\boldsymbol{\beta} + U)].$$

In a marginal model, one directly models the marginal mean as $\mathrm{E}[Y] = h(\mathbf{x}^T\boldsymbol{\beta})$. Given a random variable U with $\mathrm{E}[U] = 0$, we can express $\mathrm{E}[Y]$ as

$$\mathrm{E}[Y] = h(\mathbf{x}^T\boldsymbol{\beta}) = h(\mathbf{x}^T\boldsymbol{\beta} + 0) = h(\mathrm{E}[\mathbf{x}^T\boldsymbol{\beta} + U]).$$

In order for these two expressions for $\mathrm{E}[Y]$ to be equivalent, the operators $\mathrm{E}[\cdot]$ and $h(\cdot)$ must commute, but they only do so when $h(\cdot)$ is a linear function. Thus, when the link function is nonlinear, the parameters $\boldsymbol{\beta}$ obtained from the two models must be different in order for the two expressions for $\mathrm{E}[Y]$ to be equal. For a linear model, which can be viewed as a GLM with an identity link, the distinction between the two modeling strategies is not important because the link function is linear. Here, the marginal and conditional model formulations are equivalent because $\mathrm{E}[h(\mathbf{x}^T\boldsymbol{\beta}+U)] = \mathrm{E}[\mathbf{x}^T\boldsymbol{\beta}+U] = \mathbf{x}^T\boldsymbol{\beta} = h(\mathbf{x}^T\boldsymbol{\beta})$ and the two expressions for the marginal mean $\mathrm{E}[Y]$ are the same.

Given their general lack of equivalence, much has been written about the relative merits of marginal models and GLMMs. Advocates of the conditional model balk at the absence of a likelihood for the marginal model and also dislike that GEE largely treats the parameters characterizing the covariance structure as a nuisance. They point to the difficulty of checking and comparing models without an underlying density for the data and feel that restricting oneself to only marginal inferences limits, and possibly distorts, the information that can be extracted from the data. Lindsey and Lambert (1998) provided examples of Simpson's paradox where conclusions about the population on average (marginal inferences) differ from what is seen for each individual. Lee and Nelder (2004) argued that the conditional model is ”fundamental” because marginal predictions can be made from it if necessary, but individual-level predictions cannot be made from a purely marginal model.

Advocates of the marginal model point to the robustness of GEE to misspecification of the covariance structure and argue against the usefulness of a subject-specific interpretation. For instance, Hubbard et al. (2010) favor marginal models because they feel mixed effects models depend too heavily on unverifiable assumptions about the random effects. Neuhaus et al. (1991) commented on the awkwardness of interpreting subject-specific parameters associated with covariates that do not vary within a subject. Heagerty (1999) pointed out that these coefficients measure contrasts that are never directly observed, while Heagerty and Zeger (2000) called such contrasts ”model-based extrapolations” and argued that they are highly sensitive to model assumptions. As an example, Swihart et al. (2014) discussed a model in which race (black or white) was used as a predictor and noted that the conditional interpretation of the race coefficient describes the difference in the response between the same person as a black person and, counterfactually, as a white person. Since nobody in the study was observed to change race, they argued that a marginal parameter comparing the black population to the white population would be more sensible.

Diggle et al. (2002), among others, advocate catering one's model to one's objective. That is, one should fit a marginal model if seeking to draw conclusions about the population and should fit a conditional model if individual-level predictions are of interest. Nonetheless, considerable effort has been made to reconcile the differences between marginal and conditional models, especially for modeling data with a binary response.

1.4 Attenuation Factors

A popular "solution" to the discrepancy between marginal and conditional models for binary response data has been to to find a proportional relationship between the two types of models. Neuhaus et al. (1991) showed that the parameters of the marginal model are generally smaller in magnitude than those of the mixed effects model. That is, the marginal model parameters are attenuated toward zero relative to the conditional model parameters. Denoting the parameters from the marginal model as $\boldsymbol{\beta}^*$, this relationship has led to efforts to find an *attenuation factor* c $(0 < c < 1)$ such that $\boldsymbol{\beta}^* = c\boldsymbol{\beta}$. More formally, attenuation factors aim to identify a relationship in which the following equation holds:

$$\int h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = h(c\, \mathbf{x}^T\boldsymbol{\beta}). \tag{1.1}$$

For most combinations of random effects distribution $f_{\mathbf{U}}$ and inverse link function $h(\cdot)$, no attenuation factor c exists such that (1.1) holds exactly for all $\mathbf{x}^T\boldsymbol{\beta}$. In such cases, one can only find an approximate attenuation factor, meaning that $\boldsymbol{\beta}^* \approx c\boldsymbol{\beta}$ or, more specifically, that (1.1) holds approximately for a range of values of $\mathbf{x}^T\boldsymbol{\beta}$. Zeger et al. (1988) and Neuhaus et al. (1991) provided two competing formulas for the approximate attenuation factor in the case of a logit link. The method of Neuhaus et al. (1991), which is based on a Taylor series expansion of an expression for the univariate marginal parameter β^*, can be generalized to other link functions. For example, Jewell and Shiboski (1990) used a similar approach to derive an approximate attenuation factor for a model with a complementary log-log link.

Rather than resorting to approximate attenuation factors, one could specify a random effects distribution such that (1.1) holds exactly for some constant c and all $\mathbf{x}^T\boldsymbol{\beta} \in \mathbb{R}$. Given an inverse link function $h(\cdot)$ and a model with a single random intercept, Wang and Louis (2003) used Fourier transforms to derive what they call a *bridge distribution*, which allows for an exact proportional relationship between the marginal and conditional parameters. Typically, when one integrates over the random effects in a GLMM, the resulting model for the marginal mean has a different functional form than the model for the conditional mean. The bridge distribution preserves the functional form after integration and therefore yields the desired proportional relationship.

Wang and Louis (2003) focused on a model with a logit link and a single random intercept. The bridge distribution in this case is symmetric and mound-shaped like a Gaussian distribution, but has heavier tails. To justify use of a bridge distribution instead of the more familiar Gaussian distribution, Wang and Louis (2003) noted that assumptions about the random effects distribution are difficult to verify and cited earlier work (Neuhaus et al., 1992; Heagerty and Kurland, 2001) claiming that coefficient estimates in GLMMs with random intercepts are not very sensitive to misspecification of the shape of the random effects distribution. They argued that replacing the standard assumption of a Gaussian distribution for the random effects with a bridge distribution should not have much impact on inference for $\boldsymbol{\beta}$, but is convenient for model interpretation.

The argument that the assumed random effects distribution has little impact on inference for the fixed effects parameters $\boldsymbol{\beta}$ is not unfounded. Neuhaus et al. (1992) argued that bias in the fixed effects parameter estimates due to misspecifying the form of the random effects distribution is generally small. Others (Agresti et al., 2004; Litière et al., 2007, 2008) have claimed that substantial bias can occur in certain cases, but these are mostly extreme cases and some of these results have been disputed (see Neuhaus et al., 2011). More recently, Neuhaus et al. (2013) argued that the estimate of a covariate effect is severely biased due to random effects misspecification only if the misspecified random effect is tied to the covariate of interest. For an overview of findings related to random effects misspecification, see McCulloch and Neuhaus (2011).

One can find a bridge distribution for almost any link function. For example, Wang and Louis (2003) derived such a distribution for a complementary log-log link in addition to a logit link. However, since a linear combination of bridge distributions is generally not a bridge distribution, this strategy for dealing with the discrepancy between marginal and conditional models cannot be applied to models with multivariate random effects or random slopes. Nonetheless, recent work has used copulas to extend the idea of a bridge distribution to models with multiple correlated random intercepts that arise in longitudinal studies (Parzen et al., 2011) and in spatial applications (Boehm et al., 2013).

An alternative to choosing a random effects distribution that allows (1.1) to hold exactly for a particular link function is to choose a link function that allows (1.1) to hold exactly for a particular random effects distribution. With Gaussian random effects, a probit link allows for the desired relationship. However, a model with a probit link is more difficult to interpret than one with a logit link because a probit model does not have the convenient odds-ratio interpretation associated with the logit. With this in mind, Caffo et al.

(2007) characterized the link function as the inverse of a cumulative distribution function and showed that the logit can be closely approximated by a mixture of five normal distributions. This approximation maintains (at least approximately) the log-odds interpretation of the logit link while also exhibiting proportionality between the marginal and conditional models. Further, Caffo et al. (2007) demonstrated that expressing both the random effects distribution and the link function as a mixture of normal distributions is convenient for developing a Gibbs sampler to fit the model and thereby simplifies computation.

Although these proportionality-seeking strategies recognize the difference between conditional and marginal models, they fail to provide a single model that both has parameters with a marginal interpretation and allows one to easily make predictions at the individual level. Instead, they provide a relationship that one could use to obtain parameters with an alternative interpretation after a conditional or marginal model has been fit.

1.5 Marginalized Multilevel Models

Heagerty (1999) and Heagerty and Zeger (2000) took a different approach. They proposed the *marginalized multilevel model*, which reparameterizes the conditional mean as $\mathrm{E}[Y|U] = h(\Delta + U)$, where Δ is a function of $\mathbf{x}^T\boldsymbol{\beta}$ and f_U that is defined implicitly by

$$\int h(\Delta + u) f_U(u) du = h(\mathbf{x}^T\boldsymbol{\beta}). \tag{1.2}$$

This model yields parameters with a marginal interpretation, but is a conditionally specified, generative model and therefore avoids the pitfalls associated with the absence of a likelihood in purely marginal models fit via GEE. The parameters $\boldsymbol{\beta}$ are marginal parameters, but one is still able to make individual-level predictions through the implicitly defined function Δ. Heagerty and Zeger (2000) argued that this parameterization allows one to separate the target of inference from the estimation procedure. Instead of relying on GEE for marginal inferences and either maximum likelihood estimation or a Bayesian approach for subject-specific inferences, the marginalized multilevel model allows one to tackle either type of question using a single model. Further, Heagerty and Kurland (2001) demonstrated that estimates of the marginal fixed effects parameters in a marginalized multilevel model are less sensitive to assumptions regarding the form of the random effects distribution than estimates of the subject-specific fixed effects parameters in a conventional GLMM.

Several extensions of the marginalized multilevel model have been introduced in recent years. Heagerty (2002) modified the dependence structure to allow for serial correlation in what he called a *marginalized transition model*. Schildcrout and Heagerty (2007) took this idea further to allow for both serial and long-range dependence. Miglioretti and Heagerty (2004) presented a Bayesian approach to fitting marginalized multilevel models. Wang and Louis (2004) proposed a marginalized multilevel model with a random intercept that is assumed to follow a bridge distribution. This leads to a form for Δ that is a simple rescaling of the marginal mean structure. Finally, Swihart et al. (2014) related marginalized multilevel models to copula models.

1.6 Marginally Interpretable Models

We focus on a class of conditionally specified models with a direct marginal interpretation for the parameters. We call these models *marginally interpretable models* and say that a GLMM is *marginally interpretable* if and only if for all $\mathbf{x}^T\boldsymbol{\beta}$

$$\int h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = h(\mathbf{x}^T\boldsymbol{\beta}), \tag{1.3}$$

where $f_{\mathbf{U}}$ is the joint density of the random effects. Such a model is defined so that the marginal mean $h(\mathbf{x}^T\boldsymbol{\beta})$ is preserved after integrating out the random effects. This property is sensible because if one had independent data and fit a fixed effects GLM, the mean structure would typically have the form $\mathrm{E}[Y] = h(\mathbf{x}^T\boldsymbol{\beta})$. When the data are not independent, random effects are introduced to account for the dependence, but these random effects should not alter the mean structure. Unfortunately, for many common choices of link function $g(\cdot)$ and random effects density $f_{\mathbf{U}}$, namely when mean-zero normal random effects are paired with canonical link functions, the marginal mean $h(\mathbf{x}^T\boldsymbol{\beta})$ is not preserved after integration over $f_{\mathbf{U}}$ and (1.3) is generally not satisfied.

Beyond ensuring that the inclusion of random effects does not distort the marginal mean structure, the motivation for marginally interpretable GLMMs is twofold. First, we view a conditionally specified model as superior to a purely marginal one. Treatments act on individuals, not on averages, and models should therefore be built on the individual level. Unlike a purely marginal model, a GLMM is a fully specified model with a well-defined density for the data that can be used for model comparison and for making individual-level predictions. Second, to understand how certain factors impact a population, it is ordinarily more informative to investigate the average effect of those factors across all units in the population than to investigate the effect for a specific unit. Thus, marginal parameters are often of greater interest than the subject-specific parameters that arise from conventional GLMMs, and it is useful to construct a model to have parameters with a population-averaged interpretation. Preserving the marginal mean results in parameters with the desired marginal interpretation.

The marginalized multilevel model of Heagerty (1999) and Heagerty and Zeger (2000) achieves these dual goals of a fully specified, generative model and parameters with a direct marginal interpretation. Consequently, the marginally interpretable GLMM is closely related to the marginalized multilevel model. The key distinction between the marginalized multilevel model and our marginally interpretable model is how we conceptualize the random effects. Our model has the same basic structure as the marginalized multilevel model, but we parameterize $\Delta = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{a}$. We call the quantity $\mathbf{d}^T\mathbf{a}$ the *adjustment* and define it such that an analogue of (1.2) holds, where we allow for multivariate random effects $\mathbf{U}$. The adjustment $\mathbf{d}^T\mathbf{a}$ potentially depends on the fixed portion of the model $\mathbf{x}^T\boldsymbol{\beta}$, the parameters characterizing the random effects distribution $f_{\mathbf{U}}$, and the random effects design $\mathbf{d}$. Despite being a deterministic piece of the model, $\mathbf{d}^T\mathbf{a}$ is viewed as a location shift of the random effects distribution. As such, we cease to conceptualize each realization of a random effect as a single value shared by all observations in the same group or cluster. Rather, observations sharing the same random effect are viewed as having a value representing the same quantile of a location family of distributions. Since the location of the random effects distribution for a particular observation depends on the covariates for that observation, the value associated with a specific realization of a random effect varies across observations in the same group or cluster. An example of when different observations sharing the same random effect could be associated with different values for the random effect is when there are repeated measurements on an individual over time and the covariates vary with time. We discuss this characterization of the random effects in greater detail in Chapter 2.

1.7 Organization of this Thesis

The remainder of this thesis is organized as follows. Chapter 2 formally introduces marginally interpretable GLMMs. Further detail is provided regarding the interpretation of the random effects in these models, and the form of the adjustment is given for several common link functions. We also address issues associated with models that include random slopes. Chapter 3 contrasts inference based on a marginally interpretable GLMM with inference based on a conventional GLMM, and argues that the properties of the marginally interpretable model are preferred over those of the conventional GLMM. Inference is discussed in both a classical framework and a Bayesian framework. Topics include hypothesis testing, reproducibility, and consistent estimation of the random effects variance. Examples using both empirical data and simulated data highlight advantages of marginally interpretable GLMMs over conventional GLMMs. Chapter 4 describes popular techniques for fitting GLMMs and explains how those methods can be adapted for fitting marginally interpretable models. A novel algorithm for efficiently computing the logistic-normal integral is introduced that is directly applicable to fitting marginally interpretable binomial GLMMs with Gaussian random effects, but could also be useful in other contexts. Both frequentist and Bayesian approaches to model fitting are discussed. Chapter 5 summarizes the findings of the preceding chapters, discusses implications of these findings, and suggests avenues for future research. Appendix A contains the data used in the examples presented in Chapters 2, 3, and 4.

Chapter 2: Model Structure and Interpretation

Suppose one has N observations, indexed by $i = 1, \ldots, N$, consisting of a response Y_i and a p-vector of predictors $\mathbf{x}_i$. Further, assume that each observation is associated with q unobserved latent random variables. These random effects are expressed as a q-vector $\mathbf{U}_i$ and have a corresponding design vector $\mathbf{d}_i$. The elements of $\mathbf{d}_i$ are often a subset of the elements of $\mathbf{x}_i$, but such a relationship is not required. We propose a marginally interpretable GLMM that expresses the conditional mean of the response as

$$\mathrm{E}[Y_i | \mathbf{U}_i = \mathbf{u}] = h(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i), \quad i = 1, \ldots, N. \tag{2.1}$$

We assume that the $\mathbf{U}_i$ have joint density given by $f_{\mathbf{U}}$ and, conditional on the $\mathbf{U}_i$, that the Y_i are mutually independent with density $f_{Y|\mathbf{U}}$ for each i. The adjustment $\mathbf{d}_i^T \mathbf{a}_i$, when it exists, is defined implicitly by the equation

$$h(\mathbf{x}_i^T \boldsymbol{\beta}) = \int h(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}, \tag{2.2}$$

and is included to ensure that the model is marginally interpretable as defined in (1.3).

Although it is written as a term in the conditional mean and potentially depends on the fixed effects parameters, the adjustment $\mathbf{d}_i^T \mathbf{a}_i$ is viewed as a location shift of the random effects distribution. The remainder of this chapter elaborates on this conceptualization of the model, provides details regarding the form of the adjustment for specific link functions, and discusses issues related to models with random slopes.

In many applications of GLMMs, $f_{\mathbf{U}}$ is assumed to be a Gaussian density. Several results discussed in this chapter and succeeding chapters rely on the following proposition, which allows one to reduce a q-dimensional integral to a univariate one when dealing with multivariate normal random effects.

Proposition 2.1. *For the case when $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{\Sigma})$, if the q-dimensional integral*

$$\int_{\mathbb{R}^q} h(\kappa + \mathbf{d}^T\mathbf{u} + a)\Big(\frac{1}{2\pi}\Big)^{\frac{q}{2}} |\mathbf{\Sigma}|^{-1/2} \exp\Big(- \frac{1}{2}\mathbf{u}^T\mathbf{\Sigma}^{-1}\mathbf{u}\Big) d\mathbf{u}$$

exists, then it can be expressed as a univariate integral of the form

$$\int_{\mathbb{R}} h(\kappa + v + a)\frac{1}{\sqrt{2\pi\tau^2}} \exp\Big(- \frac{1}{2\tau^2}v^2\Big) dv.$$

Proof of Proposition 2.1: Suppose for each $i = 1, \ldots, N$ we have $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{\Sigma})$ and write $\mathbf{U}_i = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}_i$, where $\mathbf{\Sigma}^{\frac{1}{2}}$ is a square root matrix for $\mathbf{\Sigma}$ and $\mathbf{Z}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{I}_q)$. We can define a random variable $V = \mathbf{d}^T\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}$ such that $V \sim \mathrm{N}(0, \tau^2)$, where $\tau^2 = \mathbf{d}^T\mathbf{\Sigma}\mathbf{d}$. It is also possible to define $q-1$ additional random variables $\mathbf{W} = (W_1, \ldots, W_{q-1})^T$ that span the orthogonal complement of V relative to $\mathbb{R}^q$ such that $\mathbf{W}$ follows a $(q-1)$-dimensional normal distribution with density $f_{\mathbf{W}}$. Given such a V and $\mathbf{W}$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^q} h(\kappa + \mathbf{d}^T\mathbf{u} + a)\Big(\frac{1}{2\pi}\Big)^{\frac{q}{2}} |\mathbf{\Sigma}|^{-1/2} \exp\Big(- \frac{1}{2}\mathbf{u}^T\mathbf{\Sigma}^{-1}\mathbf{u}\Big) d\mathbf{u} \\
&= \int_{\mathbb{R}^q} h(\kappa + \mathbf{d}^T\mathbf{\Sigma}^{1/2}\mathbf{z} + a)\Big(\frac{1}{2\pi}\Big)^{\frac{q}{2}} \exp\Big(- \frac{1}{2}\mathbf{z}^T\mathbf{z}\Big) d\mathbf{z} \\
&= \int_{\mathbb{R}^q} h(\kappa + v + a)\frac{1}{\sqrt{2\pi\tau^2}} \exp\Big(- \frac{1}{2\tau^2}v^2\Big) f_{\mathbf{W}}(\mathbf{w}) d\mathbf{w} dv \\
&= \int_{\mathbb{R}} h(\kappa + v + a)\frac{1}{\sqrt{2\pi\tau^2}} \exp\Big(- \frac{1}{2\tau^2}v^2\Big) dv \int_{\mathbb{R}^{q-1}} f_{\mathbf{W}}(\mathbf{w}) d\mathbf{w} \\
&= \int_{\mathbb{R}} h(\kappa + v + a)\frac{1}{\sqrt{2\pi\tau^2}} \exp\Big(- \frac{1}{2\tau^2}v^2\Big) dv.
\end{aligned}$$

□

2.1 Interpretation of Random Effects

As described in Section 1.1, random effects ordinarily represent sources of variation that are not of direct interest to a study. To understand the variability in one's data, it is important to understand the heterogeneity among the random effects, but learning, for example, that a specific subject has a greater inherent risk for a negative outcome than another subject is not usually the focus of one's analysis. Traditionally, each random effect is viewed as a random variable with some underlying distribution and each observation is associated with a specific realization of this random variable. The specific realizations of the random effect are not important because interest lies primarily in understanding its underlying distribution. Nonetheless, thinking about the random effect for a particular observation as a specific realization of a random variable can be helpful.

Consider, for example, a random intercept model for data clustered into groups. Taking the viewpoint that each group of observations shares its own unique realization of the random intercept, one can think of each realization of the random effect as a shift in the mean response that applies to all units in the group of observations sharing that random effect. Consequently, observations in one group might be systematically greater than observations in another group due to the random intercept taking a larger value in the former group relative to the latter. As such, the random intercept accounts for additional variation across groups and helps to model situations where observations within the same group are generally more similar than observations in different groups. This idea is illustrated in Figure 2.1 for a random intercept model with a single continuous predictor x. Specifically, the conditional mean has the form $\mathrm{E}[Y|U=u] = h(\beta_0 + \beta_1 x + u)$, where $\beta_0 = \beta_1 = 1$ and the random effects follow a standard normal distribution. The solid line represents $\beta_0 + \beta_1 x$, which is the linear predictor η on the link scale when $U = 0$. The dashed lines represent

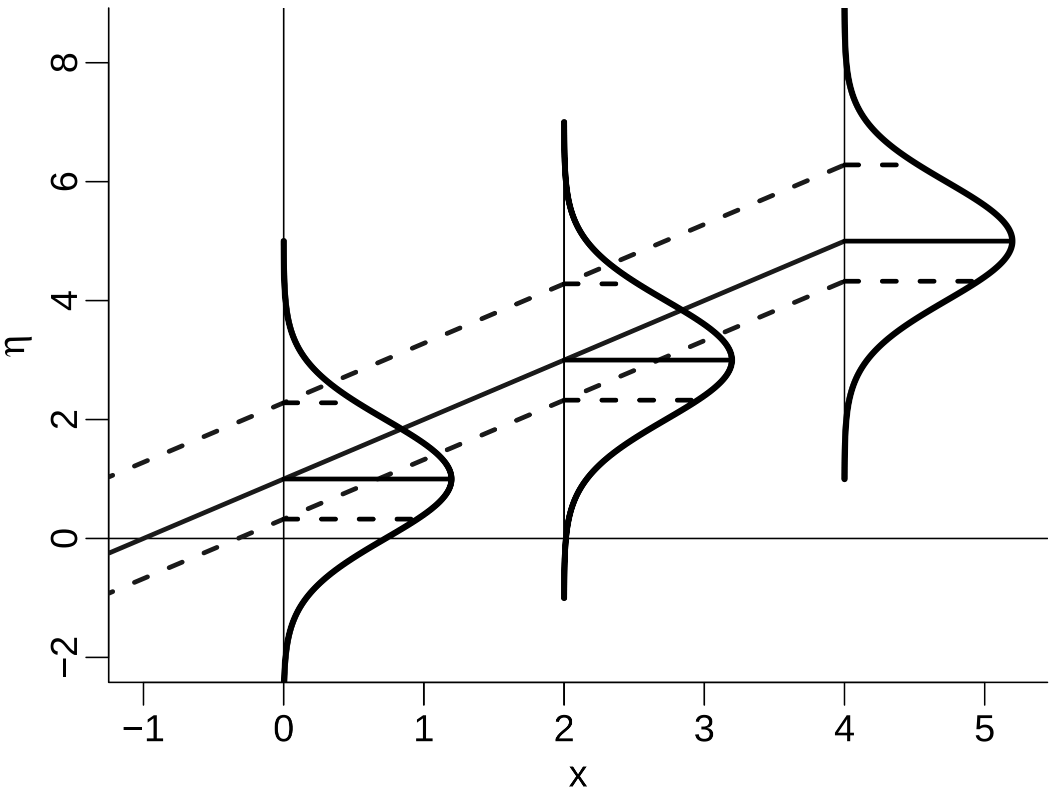

Figure 2.1: Illustration of a simple random intercept model, showing the linear predictor on the link scale along with the distribution of the random effects

$\beta_0 + \beta_1 x + U$ when $U = -0.7$ and $U = 1.3$, which are the 25^{th} percentile and the 90^{th} percentile of the standard normal distribution, respectively. Observations corresponding to $U = -0.7$ will be systematically smaller than observations corresponding to $U = 1.3$, and the difference between the two on the link scale is the same regardless of the value of x. The normal densities overlaid on the plot emphasize that the random effects are assumed to come from a normal distribution and that, in this conventional model, the random effects distribution is always centered at $\beta_0 + \beta_1 x$.

The idea of shifting all units sharing the same realization of a random effect by the same amount is not always applicable in a marginally interpretable model. Marginally interpretable GLMMs are defined such that when one integrates over the random effects distribution the marginal mean is preserved. When the link function is nonlinear, an adjustment is required to account for the curvature of the link and preserve the marginal mean. When the curvature of the link function is not uniform across the range of observed covariates, it is necessary to make different adjustments for different covariate values. That is, one might need to shift each unit within the same group of observations by a different amount in order to preserve the marginal mean.

We think of the adjustments made to preserve the marginal mean as shifts in the location of the random effects distribution. Under this conceptualization, all units in the same group are not necessarily associated with the same value for the random effect. Rather, all units in a group share the same quantile of a location family of random effects distributions. The location of the distribution for a particular unit could depend on the observed covariates, and units sharing the same quantile will not necessarily share the same value of the random effect if their covariates are not equal. Consequently, each realization of a random effect does not represent a specific value by which to shift the observations sharing that random effect, but instead represents a set of potential values with the specific value for a particular observation determined by $\mathbf{x}_i^T\boldsymbol{\beta}$. By deviating from the traditional formulation of a random effect, we are able to separate systematic variation in the population, captured by $\mathbf{x}_i^T\boldsymbol{\beta}$, from individual-level variation, captured by $\mathbf{d}_i^T\mathbf{U}_i + \mathbf{d}_i^T\mathbf{a}_i$.

A situation where this new formulation of a random effect might arise is in a multilevel model where the covariates $\mathbf{x}_i$ differ across individual units in the same group or cluster. For example, students within the same class are liable to have different characteristics, or measurements on the same subject in a longitudinal study might change over time. Depending on the choice of link function, the adjustment could be different for different units in

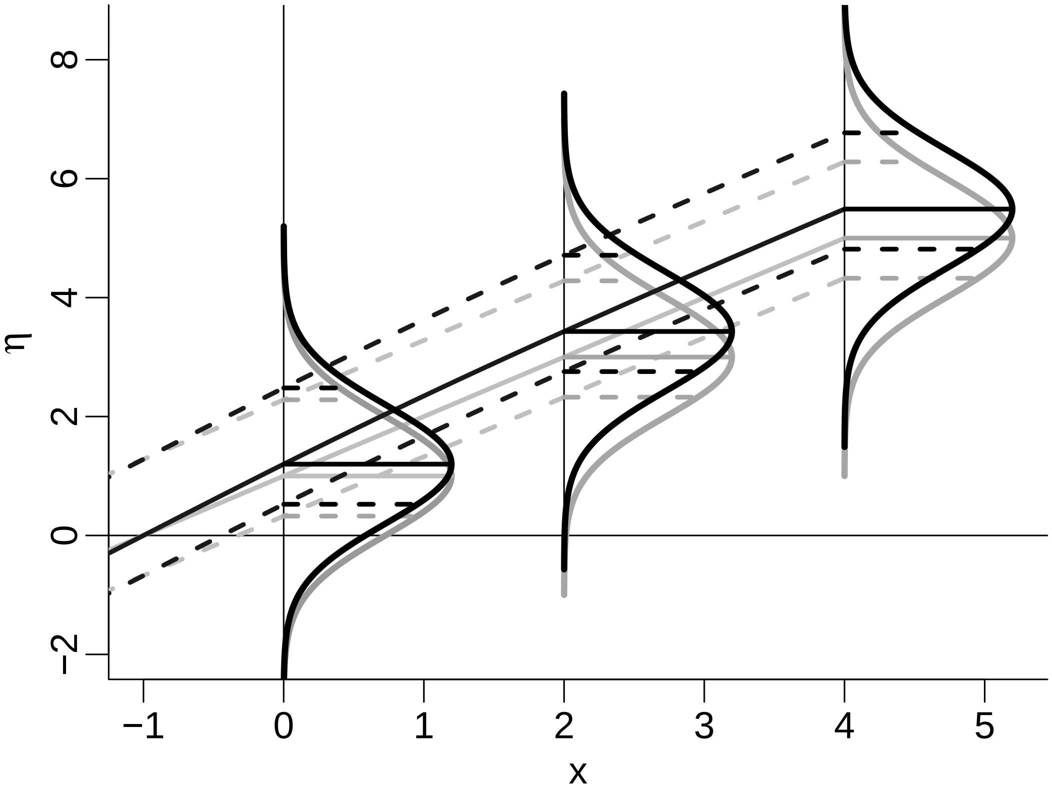

Figure 2.2: Illustration of a simple marginally interpretable random intercept model on the logit scale (in black), contrasted with a conventional random intercept model (in gray)

the same group. Thus, within a single group, the shift in the mean response associated with the random effect for that group could vary with the measured covariates for the individual units sharing that random effect.

This idea is illustrated in Figure 2.2, which depicts a model with the same basic structure as the model in Figure 2.1, but with an adjustment included. We now model the conditional mean as the $\mathrm{E}[Y|U=u] = h(\beta_0 + \beta_1 x + u + a)$, and assume a logit link function for computing a. The gray lines represent a conventional GLMM and are identical to Figure 2.1. The black curves demonstrate how the model changes when we include the adjustment. The solid black curve represents $\beta_0 + \beta_1 x + a$, and the dashed black curves represent $\beta_0 + \beta_1 x + U + a$ when U is either the 25^{th} percentile or the 90^{th} percentile of the random effects distribution. The normal densities are now centered at $\beta_0 + \beta_1 x + a$, where a depends on the value of x. The quantity $U + a$ for a particular realization of the random effect always correpsonds to the same quantile of the random effects distribution, but its magnitude varies with x because the adjustment a varies with x.

2.2 The Form of the Adjustment

The adjustment is included in the proposed model to account for the curvature of the inverse link function and ensure that the model is marginally interpretable as defined in (1.3). The form of the adjustment is determined by the choice of link function and random effects distribution, whereas its specific value typically depends on $\mathbf{x}_i^T\boldsymbol{\beta}$. Exceptions for which $\mathbf{d}_i^T\mathbf{a}_i$ does not depend on $\mathbf{x}_i^T\boldsymbol{\beta}$ are models with an identity link or a log link. For a model with an identity link, (1.3) holds as long as $\mathrm{E}[\mathbf{U}_i] = \mathbf{0}$. Thus, a standard linear mixed model is marginally interpretable without including an adjustment. Table 2.1 summarizes the form and existence of $\mathbf{d}_i^T\mathbf{a}_i$ for several common choices of link function. Specifically, this table describes the relationship between $\mathbf{x}_i^T\boldsymbol{\beta}$ and $\mathbf{d}_i^T\mathbf{a}_i$ for various link functions and random effects distributions, and also indicates whether or not there exists a closed-form solution for $\mathbf{d}_i^T\mathbf{a}_i$. The remainder of this section explores the interplay between $h(\cdot)$ and $f_\mathbf{U}$ in greater depth.

2.2.1 Log Link

Consider a GLMM with a log link. That is, let the link function be $g(\cdot) = \log(\cdot)$ with the inverse link $h(\cdot) = \exp(\cdot)$. In this case, $\mathbf{d}_i^T\mathbf{a}_i$ is defined such that

$$\exp(\mathbf{x}_i^T\boldsymbol{\beta}) = \int \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{u} + \mathbf{d}_i^T\mathbf{a}_i) f_\mathbf{U}(\mathbf{u}) d\mathbf{u}. \tag{2.3}$$

Table 2.1: Form and existence of the adjustment for common link functions

Link Function	Distribution of $\mathbf{U}_i$	Form of $\mathbf{d}_i^T\mathbf{a}_i$	Closed Form?
identity	mean exists and equals zero	zero	yes
log	exponential tails	independent of $\mathbf{x}_i^T\boldsymbol{\beta}$	yes
probit	Gaussian	linear in $\mathbf{x}_i^T\boldsymbol{\beta}$	yes
	non-Gaussian	nonlinear in $\mathbf{x}_i^T\boldsymbol{\beta}$	no
logit	bridge distribution	linear in $\mathbf{x}_i^T\boldsymbol{\beta}$	yes
	most other distributions	nonlinear in $\mathbf{x}_i^T\boldsymbol{\beta}$	no
complementary	bridge distribution	linear in $\mathbf{x}_i^T\boldsymbol{\beta}$	yes
log-log	most other distributions	nonlinear in $\mathbf{x}_i^T\boldsymbol{\beta}$	no
square root	restrictions on domain	nonlinear in $\mathbf{x}_i^T\boldsymbol{\beta}$	yes
reciprocal	$\mathrm{E}[1/(\mathbf{x}_i^T\boldsymbol{\beta}+\mathbf{d}_i^T\mathbf{U}_i)]$ exists	see Section 2.2.3	see Section 2.2.3

Solving (2.3) for $\mathbf{d}_i^T\mathbf{a}_i$ leads to the following proposition:

Proposition 2.2. *For $h(\cdot) = \exp(\cdot)$, a model of the form given by (2.1) and (2.2) is marginally interpretable if and only if $\mathbf{d}_i^T\mathbf{a}_i = -\log\big(M_{\mathbf{U}}(\mathbf{d}_i)\big)$, where $M_{\mathbf{U}}(\mathbf{d}_i)$ is the moment-generating function of $\mathbf{U}_i$ evaluated at $\mathbf{d}_i$ and is given by $M_{\mathbf{U}}(\mathbf{d}_i) = \mathrm{E}[\exp(\mathbf{d}_i^T\mathbf{U}_i)]$.*

Proof of Proposition 2.2: For each $i = 1, \ldots, N$, dividing both sides of (2.3) by the quantity $\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{a_i})$ and then taking the natural logarithm of both sides yields

$$-\mathbf{d}_i^T\mathbf{a_i} = \log\bigg(\int \exp(\mathbf{d}_i^T\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}\bigg).$$

The integral in this equation is equal to $M_{\mathbf{U}}(\mathbf{d}_i)$. Thus, multiplying both sides by -1, we obtain $\mathbf{d}_i^T\mathbf{a_i} = -\log\big(M_{\mathbf{U}}(\mathbf{d}_i)\big)$, as required. □

From Proposition 2.2 we obtain the following corollary:

Corollary 2.2.1. *For a GLMM with inverse link function $h(\cdot) = \exp(\cdot)$, an adjustment $\mathbf{d}_i^T\mathbf{a}_i$ that makes the model marginally interpretable exists if and only if $M_{\mathbf{U}}(\mathbf{d}_i)$ exists.*

Proof of Corollary 2.2.1: This result is an immediate consequence of Proposition 2.2. For each $i = 1, \ldots, N$, if $\mathbf{d}_i^T\mathbf{a_i}$ exists, then $\mathbf{d}_i^T\mathbf{a_i} = -\log\big(M_{\mathbf{U}}(\mathbf{d}_i)\big)$, and $M_{\mathbf{U}}(\mathbf{d}_i)$ must exist.

Conversely, if $M_{\mathbf{U}}(\mathbf{d}_i)$ exists, then $\mathbf{d}_i^T\mathbf{a_i} = -\log\big(M_{\mathbf{U}}(\mathbf{d}_i)\big)$ also exists because $M_{\mathbf{U}}(\mathbf{d}_i)$ is strictly positive and therefore lies in the domain of $\log(\cdot)$. $\square$

This result constrains the set of possible random effects distributions that can be used with this model to those with exponential tails. A t-distribution, for example, is not a valid random effects distribution for a marginally interpretable GLMM with a log link.

To better understand the role of the adjustment for a model with a log link, consider the case of a single random intercept $U_i \sim \mathrm{N}(0, \sigma^2)$. In this case, $d_i = 1$ for all $i = 1, \dots, N$, and we express the adjustment as a_i. From Proposition 2.2 we have the following result:

Corollary 2.2.2. *A model for which* $\mathrm{E}[Y_i|U_i = u] = \exp(\mathbf{x}_i^T\boldsymbol{\beta} + u + a_i)$ *and* $U_i \sim \mathrm{N}(0, \sigma^2)$ *is marginally interpretable if and only if* $a_i = -\sigma^2/2$ *for all* $i = 1, \dots, N$.

Proof of Corollary 2.2.2: For this model, $d_i = 1$ for all $i = 1, \dots, N$ and the moment-generating function of U_i is $M_U(t) = \exp(\sigma^2 t^2/2)$. Thus, for all $i = 1, \dots, N$, we have $a_i = -\log\big(M_U(d_i)\big) = -\log\big(M_U(1)\big) = -\log\big(\exp(\sigma^2/2)\big) = -\sigma^2/2$, as required. $\square$

As a consequence of Proposition 2.1, the same result applies for q-variate normal random effects $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, where σ^2 is replaced by $\mathbf{d}_i^T\boldsymbol{\Sigma}\mathbf{d}_i$. For simplicity, we continue to focus on a random intercept model with variance σ^2 for the random effect.

It is evident from Corollary 2.2.2 that, in this situation, the adjustment depends only on the random effects variance σ^2 and is independent of the fixed effects. It is simply an additive offset on the log scale that pulls the conditional mean $\mathrm{E}[Y_i|U_i]$ down by the same amount for every $i = 1, \dots, N$. This effectively shifts the location of the random effects distribution in a manner that makes the model marginally interpretable by ensuring that $\mathrm{E}[Y_i] = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$. A conventional GLMM would express the conditional mean as

$\mathrm{E}[Y_i|U_i] = \exp(\mathbf{x}_i^T\boldsymbol{\beta} + U_i)$, and integrating over f_U would yield

$$\begin{aligned}
\mathrm{E}[Y_i] &= \mathrm{E}\big[\mathrm{E}[Y_i|U_i]\big] = \mathrm{E}[\exp(\mathbf{x}_i^T\boldsymbol{\beta} + U_i)] = \mathrm{E}[\exp(\mathbf{x}_i^T\boldsymbol{\beta})\exp(U_i)] \\
&= \exp(\mathbf{x}_i^T\boldsymbol{\beta})\mathrm{E}[\exp(U_i)] = \exp(\mathbf{x}_i^T\boldsymbol{\beta})M_U(1) = \exp(\mathbf{x}_i^T\boldsymbol{\beta})\exp\left(\frac{\sigma^2}{2}\right),
\end{aligned}$$

where $M_U(1) = \mathrm{E}[\exp(U_i)]$ is the moment-generating function of a $\mathrm{N}(0, \sigma^2)$ random variable evaluated at $t = 1$. Thus, provided $\sigma^2 \neq 0$, this conditional model does not preserve the marginal mean $\mathrm{E}[Y_i] = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$.

The failure of the conventional GLMM to preserve the marginal mean stems from the nonlinearity of the link function. Since $\exp(\cdot)$ is a convex function, differences between positive random effects are amplified on the original scale of the data whereas differences between negative random effects are reduced. That is, the impact of the inverse link transformation is asymmetric. To illustrate this phenomenon, the black curves in Figure 2.3 represent a standard normal density and the same density when all the values on the horizontal axis are transformed via the function $\exp(\cdot)$. The transformation causes the symmetric distribution to become skewed right. As a consequence, even though the initial distribution had mean zero, the mean after the transformation is not $\exp(0) = 1$, but rather $\exp(1/2) \approx 1.65$. Since an additive random effect on the log scale has a multiplicative effect on the original scale of the data and $\mathrm{E}[\exp(U_i)] > 1$, the impact of a symmetric, mean-zero random effect is not negligible on average on the original scale of the data despite having no effect on average on the link scale.

It is to account for the asymmetry induced by the convexity of the inverse link function that we introduce the adjustment. In this instance, the adjustment is an additive offset on the log scale and is always negative to counteract the asymmetric pull of the convex inverse link function illustrated in Figure 2.3. The adjustment serves to shift the distribution of U_i so that its mean effect on the original scale, as opposed to its mean effect on the log

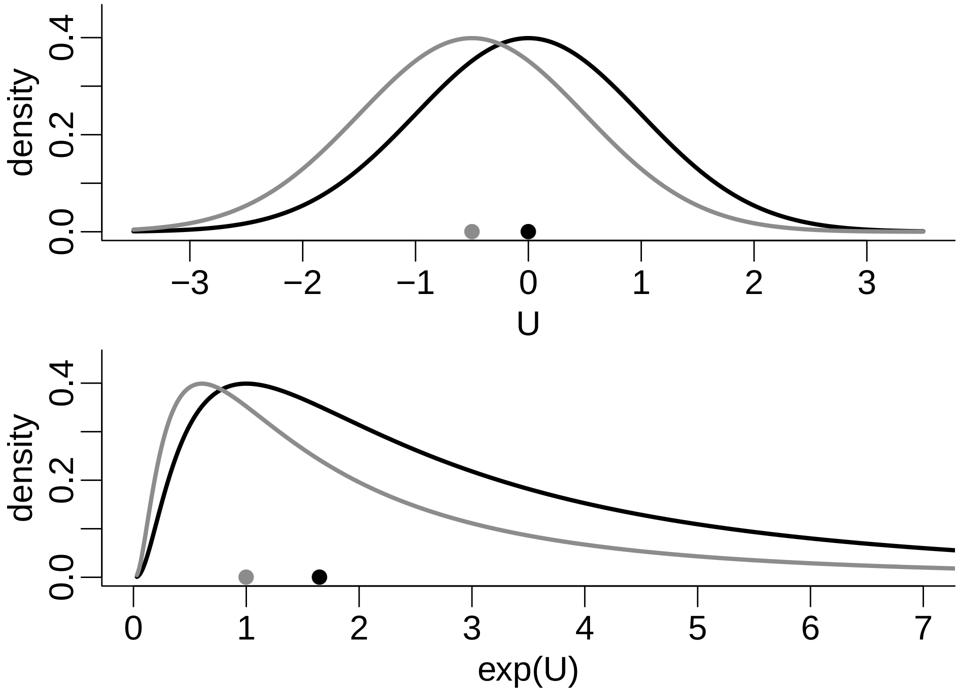

Figure 2.3: Density of $U_1 \sim \mathrm{N}(0,1)$ (in black) and $U_2 \sim \mathrm{N}(-1/2,1)$ (in gray) on their original scale and after being transformed via $\exp(\cdot)$; each dot on the horizontal axis represents the mean of the corresponding distribution

scale, is negligible. The gray curves in Figure 2.3 are analogous to the black curves, but for a $\mathrm{N}(-1/2,1)$ distribution instead of a $\mathrm{N}(0,1)$ distribution. By shifting the mean in this manner we obtain a distribution that has mean one after transforming via $\exp(\cdot)$. This allows the marginal mean to be preserved after integration, as desired.

2.2.2 Links with Bounded Domain

Several common link functions, including the probit, logit, and complementary log-log, are defined only on a bounded subset of the real line. In turn, the range of the corresponding inverse link function $h(\cdot)$ is constrained to a bounded interval. For models with such a link function, the following theorem applies:

Theorem 2.1. *Consider a model of the form given in (2.1) with $h : \mathbb{R} \to [\ell, u]$. Suppose $h(\cdot)$ is increasing and continuous, $h(\eta) \to \ell$ as $\eta \to -\infty$, and $h(\eta) \to u$ as $\eta \to \infty$. Then an adjustment $\mathbf{d}_i^T \mathbf{a}_i$ that satisfies (2.2) exists for any choice of random effects distribution.*

Proof of Theorem 2.1: To simplify notation, we suppress the subscript i. For any $f_{\mathbf{U}}$,

$$-\infty < \ell = \ell \int f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \leq \int h(\mathbf{x}^T\boldsymbol{\beta}+\mathbf{d}^T\mathbf{u}+\mathbf{d}^T\mathbf{a}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \leq u \int f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = u < \infty.$$

Thus, the integral $\int h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u} + \mathbf{d}^T\mathbf{a}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}$ exists. Further, this integral is a continuous function of $\mathbf{d}^T\mathbf{a}$ and, provided $\mathbf{d} \neq \mathbf{0}$, the following two limits hold:

$$\lim_{\mathbf{d}^T\mathbf{a} \to -\infty} \int h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u} + \mathbf{d}^T\mathbf{a}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = \ell,$$

$$\lim_{\mathbf{d}^T\mathbf{a} \to \infty} \int h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u} + \mathbf{d}^T\mathbf{a}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = u.$$

Since $\ell \leq h(\mathbf{x}^T\boldsymbol{\beta}) \leq u$, continuity implies that for any value of $h(\mathbf{x}^T\boldsymbol{\beta})$ there exists an adjustment $\mathbf{d}^T\mathbf{a}$ such that (2.2) holds. When $\mathbf{d} = \mathbf{0}$, $h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u} + \mathbf{d}^T\mathbf{a}) = h(\mathbf{x}^T\boldsymbol{\beta})$ and (2.2) trivially holds. □

Thus, one can always construct a model to be marginally interpretable when the link function is defined only on a bounded interval. We now discuss link functions with this property.

Probit Link

Consider a model with a probit link. That is, let $g(\cdot) = \Phi^{-1}(\cdot)$ and $h(\cdot) = \Phi(\cdot)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The range of the inverse link function $h(\cdot)$ is the bounded interval $(0, 1)$. Therefore, Theorem 2.1 applies and an adjustment $\mathbf{d}_i^T \mathbf{a}_i$ that makes the model marginally interpretable exists for any choice of random effects distribution.

For a model with a probit link and normal random effects, $\mathbf{d}_i^T \mathbf{a}_i$ has a closed form. Specifically, let $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a covariance matrix. One can show that with no adjustment this model satisfies a multivariate analogue to (1.1) with $c = (1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)^{-1/2}$ (see McCulloch et al., 2008, page 208). This leads to the following proposition:

Proposition 2.3. *For $h(\cdot) = \Phi(\cdot)$ and $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, a model of the form given by (2.1) and (2.2) is marginally interpretable if and only if $\mathbf{d}_i^T \mathbf{a}_i = \big((1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)^{1/2} - 1\big)\mathbf{x}_i^T \boldsymbol{\beta}$.*

Proof of Proposition 2.3: Let $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ and $\epsilon \sim \mathrm{N}(0, 1)$, and define $W = \epsilon - \mathbf{d}_i^T \mathbf{U}_i$ so that $W \sim \mathrm{N}(0, 1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)$. Then,

$$\begin{aligned}
\int \Phi(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} &= \int \mathrm{P}(\epsilon \leq \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \\
&= \mathrm{P}(\epsilon \leq \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{U}_i + \mathbf{d}_i^T \mathbf{a}_i) = \mathrm{P}(\epsilon - \mathbf{d}_i^T \mathbf{U}_i \leq \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i) \\
&= \mathrm{P}(W \leq \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i) = \Phi\left(\frac{\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i}{(1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)^{1/2}} \right).
\end{aligned}$$

Consequently,

$$\Phi(\mathbf{x}_i^T \boldsymbol{\beta}) = \int \Phi(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = \Phi\left(\frac{\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i}{(1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)^{1/2}} \right).$$

Applying $\Phi^{-1}(\cdot)$ to both sides yields $\mathbf{x}_i^T \boldsymbol{\beta} = (\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i)(1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)^{-1/2}$. Solving for $\mathbf{d}_i^T \mathbf{a}_i$ we obtain $\mathbf{d}_i^T \mathbf{a}_i = \big((1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)^{1/2} - 1\big)\mathbf{x}_i^T \boldsymbol{\beta}$, as required. □

Thus, $\mathbf{d}_i^T \mathbf{a}_i$ is a linear function of $\mathbf{x}_i^T \boldsymbol{\beta}$, which means there exists an exact proportional relationship between the marginal parameters of the marginally interpretable model and the cluster-specific parameters of the conventional model. Further, the computation required to fit the model is fairly straightforward. Nonetheless, models with a probit link are generally difficult to interpret because, in contrast to the logit link with its convenient log-odds interpretation, the probit transformation does not represent an intuitive relationship between the covariates and the response. We shall therefore focus on models with a logit link.

Logit Link

Consider a GLMM with link function $g(\mu) = \log\big(\mu/(1-\mu)\big)$ and inverse link function $h(\eta) = \exp(\eta)/\big(1+\exp(\eta)\big) = 1/\big(1+\exp(-\eta)\big)$. This function $g(\cdot)$ is known both as the *logit link* and as the *logistic link*. The adjustment $\mathbf{d}_i^T\mathbf{a}_i$ for this model is defined such that

$$\frac{1}{1+e^{-\mathbf{x}_i^T\boldsymbol{\beta}}} = \int \frac{1}{1+e^{-(\mathbf{x}_i^T\boldsymbol{\beta}+\mathbf{d}_i^T\mathbf{u}+\mathbf{d}_i^T\mathbf{a}_i)}} f_{\mathbf{U}}(\mathbf{u})d\mathbf{u}. \tag{2.4}$$

Once again, the range of the inverse link function $h(\cdot)$ is the bounded interval $(0,1)$. Thus, by Theorem 2.1, there are no restrictions on the choice of the random effects distribution. However, for most choices of random effects distribution the integral on the right-hand side of (2.4) is analytically intractable and there is no closed-form solution for $\mathbf{d}_i^T\mathbf{a}_i$. One exception is the *bridge distribution* derived by Wang and Louis (2003). Provided the model contains just a single random intercept, the bridge distribution leads to a closed-form solution for $\mathbf{d}_i^T\mathbf{a}_i$ that is linear as a function of $\mathbf{x}_i^T\boldsymbol{\beta}$.

For a model with a logit link, both the direction and magnitude of the adjustment depend on $\mathbf{x}_i^T\boldsymbol{\beta}$. The direction of the adjustment is driven entirely by the convexity of the inverse link function. The function $h(\eta)$ is convex for $\eta < 0$ and concave for $\eta > 0$. Thus, the adjustment is negative when $\mathbf{x}_i^T\boldsymbol{\beta} < 0$ and positive when $\mathbf{x}_i^T\boldsymbol{\beta} > 0$. The magnitude of $\mathbf{d}_i^T\mathbf{a}_i$ is, for most choices of $f_{\mathbf{U}}$, a nonlinear function of $\mathbf{x}_i^T\boldsymbol{\beta}$. This is illustrated in Figure 2.4 for the case of a single normal random intercept $U_i \sim \mathrm{N}(0,\sigma^2)$. In light of Proposition 2.1, the same picture would apply for q normal random effects $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0},\boldsymbol{\Sigma})$ if we were to replace σ^2 with $\mathbf{d}_i^T\boldsymbol{\Sigma}\mathbf{d}_i$. It is evident from Figure 2.4 that the magnitude of $\mathbf{d}_i^T\mathbf{a}_i$ is increasing in both σ^2 and $|\mathbf{x}_i^T\boldsymbol{\beta}|$. Further, for very large $\mathbf{x}_i^T\boldsymbol{\beta}$ we have the following result:

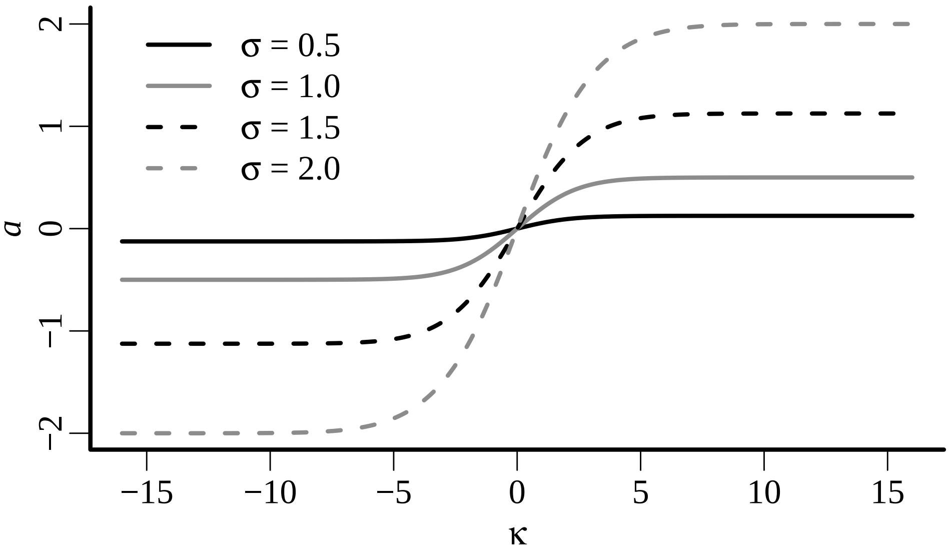

Figure 2.4: Plot of the adjustment a as a function of the fixed portion of the model κ for various values of σ in a model for which $\mathrm{E}[Y|U=u] = h(\kappa+u+a)$, $h(\cdot)$ is the inverse logit function, and $U \sim \mathrm{N}(0,\sigma^2)$

Proposition 2.4. *For $h(\cdot) = \mathrm{logit}^{-1}(\cdot)$ and $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0},\mathbf{\Sigma})$, the value of $\mathbf{d}_i^T\mathbf{a}_i$ that allows a model of the form given by (2.1) to satisfy (2.2) converges to $\frac{1}{2}\mathbf{d}_i^T\mathbf{\Sigma}\mathbf{d}_i \times \mathrm{sign}(\mathbf{x}_i^T\boldsymbol{\beta})$ as $|\mathbf{x}_i^T\boldsymbol{\beta}| \to \infty$.*

Proof of Proposition 2.4: We show the limit for $\mathbf{x}_i^T\boldsymbol{\beta} \to -\infty$; the limit for $\mathbf{x}_i^T\boldsymbol{\beta} \to \infty$ follows from symmetry. Let $\kappa = \mathbf{x}_i^T\boldsymbol{\beta}$, $a = \mathbf{d}_i^T\mathbf{a}_i$, and $\tau^2 = \mathbf{d}_i^T\mathbf{\Sigma}\mathbf{d}_i$. We want to satisfy

$$\frac{e^\kappa}{1+e^\kappa} = \int_{\mathbb{R}^q} \frac{e^{\kappa+\mathbf{u}+a}}{1+e^{\kappa+\mathbf{u}+a}} f_{\mathbf{U}}(\mathbf{u})d\mathbf{u}.$$

Using Proposition 2.1, this simplifies to

$$\frac{e^\kappa}{1+e^\kappa} = \int_{\mathbb{R}} \frac{e^{\kappa+v+a}}{1+e^{\kappa+v+a}} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}v^2\right)dv.$$

Dividing both sides by $\exp(\kappa)$ and taking the limit as $\kappa \to -\infty$ we obtain

$$1 = e^a \int_{\mathbb{R}} e^v \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}v^2\right)dv.$$

Recognizing the integral on the right-hand side of this equation as the moment generating function of a $\mathrm{N}(0, \tau^2)$ random variable evaluated with argument $t = 1$, we obtain $\exp(-a) = \exp(\tau^2/2)$, which implies that $a = -\tau^2/2$. Thus, $\mathbf{d}_i^T \mathbf{a}_i = -\frac{1}{2}\mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i$ for each $i = 1, \ldots, N$, as required. □

Figure 2.4 also shows that, for a model with a logit link, observations that have different values of the covariates $\mathbf{x}_i$ also have different adjustments. This helps illustrate the point from Section 2.1 that with a logit link, observations on units that share the same realization of a random effect but have different measured covariates do not have their mean shifted by the same amount. Rather, the magnitude of the shift associated with the random effect for each observation is determined by the value of $\mathbf{x}_i^T \boldsymbol{\beta}$ for that observation.

Complementary Log-Log Link

Consider a GLMM with a complementary log-log link. Here, $g(\mu) = \log\big(-\log(1-\mu)\big)$ and $h(\eta) = 1 - \exp\big(-\exp(\eta)\big)$. The adjustment $\mathbf{d}_i^T \mathbf{a}_i$ for this model is defined such that

$$\exp\big(-\exp(\mathbf{x}_i^T \boldsymbol{\beta})\big) = \int \exp\big(-\exp(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i)\big) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}. \tag{2.5}$$

As with a logit link, there are no restrictions on the choice of the random effects distribution, but in most cases there is no closed-form solution for $\mathbf{d}_i^T \mathbf{a}_i$. Wang and Louis (2003) also derived a bridge distribution for this link function that leads to a closed-form adjustment that is linear in $\mathbf{x}_i^T \boldsymbol{\beta}$. For more conventional choices of random effects distribution, namely Gaussian random effects, one must use an approximation or numerical integration to evaluate the integral in (2.5) when computing $\mathbf{d}_i^T \mathbf{a}_i$. See Asmussen et al. (2016) for a discussion of methods for approximating univariate integrals analogous to the integral in (2.5) for which f_U is a normal density. Since the complementary log-log link is used far less frequently than the logit link, further computational details will not be provided here.

2.2.3 Links with Range Restrictions

A number of common link functions map into a proper subset of the real line and therefore require conditions on $\mathbf{x}_i^T\boldsymbol{\beta}$ to ensure that the model is defined. For example, the square root transformation is typically defined to have nonnegative range, and no real number has a reciprocal of zero. Additive random effects with support on the entire real line could lead to problems in models with these link functions because $\mathbf{x}_i^T\boldsymbol{\beta}+\mathbf{d}_i^T\mathbf{U}_i$ could fall outside the domain of the inverse link function $h(\cdot)$. Thus, special care must be taken with these link functions, as described below.

Square Root Link

Consider a GLMM with the square root link function $g(\mu) = \mu^{1/2}$ and inverse link function $h(\eta) = \eta^2$. For such a model one typically includes the restriction that $\mathbf{x}_i^T\boldsymbol{\beta} \geq 0$. Including the adjustment, we adopt the restriction that $\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{a}_i \geq 0$. The adjustment is defined such that

$$(\mathbf{x}_i^T\boldsymbol{\beta})^2 = \int (\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{u} + \mathbf{d}_i^T\mathbf{a}_i)^2 f_{\mathbf{U}}(\mathbf{u})d\mathbf{u}. \tag{2.6}$$

If we assume $\mathrm{E}[\mathbf{U}_i] = \mathbf{0}$, then (2.6) reduces to

$$(\mathbf{x}_i^T\boldsymbol{\beta})^2 = (\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{a}_i)^2 + \mathrm{Var}(\mathbf{d}_i^T\mathbf{U}_i), \tag{2.7}$$

which is quadratic in $\mathbf{d}_i^T\mathbf{a}_i$ and leads to the following result:

Proposition 2.5. *For $h(\eta) = \eta^2$ and $\mathrm{E}[\mathbf{U}_i] = \mathbf{0}$, a model of the form given by (2.1) and (2.2) subject to the restriction that $\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{a}_i \geq 0$ is marginally interpretable if and only if $\mathbf{d}_i^T\mathbf{a}_i = -\mathbf{x}_i^T\boldsymbol{\beta} + \left((\mathbf{x}_i^T\boldsymbol{\beta})^2 - \mathrm{Var}(\mathbf{d}_i^T\mathbf{U}_i)\right)^{1/2}$.*

Proof of Proposition 2.5: Expanding the square in (2.7) and then rearranging terms yields

$$(\mathbf{d}_i^T\mathbf{a}_i)^2 + 2(\mathbf{x}_i^T\boldsymbol{\beta})(\mathbf{d}_i^T\mathbf{a}_i) + \mathrm{Var}(\mathbf{d}_i^T\mathbf{U}_i) = 0.$$

Application of the quadratic formula then leads to

$$\mathbf{d}_i^T\mathbf{a}_i = \frac{1}{2}\Big(-2\mathbf{x}_i^T\boldsymbol{\beta} \pm \big((2\mathbf{x}_i^T\boldsymbol{\beta})^2 - 4\mathrm{Var}(\mathbf{d}_i^T\mathbf{U}_i)\big)^{1/2}\Big) = -\mathbf{x}_i^T\boldsymbol{\beta} \pm \big((\mathbf{x}_i^T\boldsymbol{\beta})^2 - \mathrm{Var}(\mathbf{d}_i^T\mathbf{U}_i)\big)^{1/2}.$$

Subject to the constraint $\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{a}_i \geq 0$, we use the greater of the two roots and obtain $\mathbf{d}_i^T\mathbf{a}_i = -\mathbf{x}_i^T\boldsymbol{\beta} + \big((\mathbf{x}_i^T\boldsymbol{\beta})^2 - \mathrm{Var}(\mathbf{d}_i^T\mathbf{U}_i)\big)^{1/2}$, as required. □

Thus, for a model with a square root link, $\mathbf{d}_i^T\mathbf{a}_i$ is a nonlinear function of $\mathbf{x}_i^T\boldsymbol{\beta}$ and is only defined when $\mathbf{x}_i^T\boldsymbol{\beta} \geq \big(\mathrm{Var}(\mathbf{d}_i^T\mathbf{U}_i)\big)^{1/2}$. If the random effects variance is too large, the model cannot be fit.

Reciprocal Link

Finally, consider a model with a reciprocal link, for which $g(\mu) = h(\mu) = 1/\mu$. For a fixed effects model with this link one typically includes the restriction that $\mathbf{x}_i^T\boldsymbol{\beta} > 0$. When a random intercept U_i is included in the model, the fact that $h(\cdot)$ tends to infinity as its argument approaches zero forces us to also include restrictions on the distribution of U_i. In particular, we want a model for which

$$\frac{1}{\mathbf{x}_i^T\boldsymbol{\beta}} = \int \frac{1}{\mathbf{x}_i^T\boldsymbol{\beta} + u} f_U(u)du. \tag{2.8}$$

Therefore, f_U must be defined such that the integral in (2.8) exists. If we define U_i to have positive support, then the integral exists because

$$0 \leq \int \frac{1}{\mathbf{x}_i^T\boldsymbol{\beta} + u} f_U(u)du \leq \int \frac{1}{\mathbf{x}_i^T\boldsymbol{\beta}} f_U(u)du = \frac{1}{\mathbf{x}_i^T\boldsymbol{\beta}}.$$

Rewriting (2.8) in the form given by (2.2), we have

$$\frac{1}{\mathbf{x}_i^T\boldsymbol{\beta}} = \int \frac{1}{\mathbf{x}_i^T\boldsymbol{\beta} + u + a_i} f_U(u)du. \tag{2.9}$$

Here, the inverse link function $h(\cdot)$ is convex, which means the adjustment a_i is negative and the distribution of U_i is shifted down. For the integral in (2.9) to be defined, we want

$U_i + a_i > 0$ for all $i = 1, \ldots, N$. Thus, we require the support of U_i to be bounded below by the maximum value of $-a_i$ across all $i = 1, \ldots, N$. Although this restriction ensures the existence of the integral in (2.9), it is an awkward restriction in that the support of f_U depends on the adjustment, which itself depends on the amount of variation in f_U.

To avoid this circular argument, we move away from models of the form given by (2.1) to obtain a marginally interpretable model with a reciprocal link. Rather than adjusting the location of the random effect based on each unit's observed covariates, we instead alter the shape of the distribution of the random effect based on the observed covariates. We still must satisfy (2.8), but we no longer use an additive offset to meet this condition. The solution is to define a family of distributions for U_i such that some distribution in the family satisfies (2.8). One option is to assume that U_i follows a shifted gamma distribution. Specifically, let $\mathbf{x}_i^T\boldsymbol{\beta} + U_i$ follow a gamma distribution with shape parameter α_i and rate parameter β_i so that $\mathrm{E}[U_i] = \alpha_i\beta_i - \mathbf{x}_i^T\boldsymbol{\beta}$. Then the integral on the right-hand side of (2.8) is equal to $\big(\beta_i(\alpha_i - 1)\big)^{-1}$, and $\mathbf{x}_i^T\boldsymbol{\beta} = \beta_i(\alpha_i - 1)$. By placing additional conditions on α_i and β_i one can determine the appropriate gamma distribution for U_i for each $\mathbf{x}_i^T\boldsymbol{\beta}$. Alternatively, one could let $\mathbf{x}_i^T\boldsymbol{\beta}+U_i$ follow an inverse gamma distribution with parameters α_i and β_i, and be constrained by the relationship $\mathbf{x}_i^T\boldsymbol{\beta} = (\alpha_i\beta_i)^{-1}$. In either case it is the shape, not the location, of the random effects distribution that varies with $\mathbf{x}_i^T\boldsymbol{\beta}$ in this marginally interpretable model.

2.3 Models with Random Slopes

Up to this point, we have focused on the form of the adjustment $\mathbf{d}_i^T\mathbf{a}_i$ as a function of the fixed portion of the model $\mathbf{x}_i^T\boldsymbol{\beta}$ and the random effects density $f_{\mathbf{U}}$. How one defines $\mathbf{d}_i$ also affects the adjustment. This is particularly important in models with *random slopes*.

The notion of a random slope is most easily understood in the context of linear models. In a simple linear model with continuous predictor x and mean function given by

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x,$$

the mean $\mathrm{E}[Y]$ is represented as a linear function of x with slope β_1. As such, β_1 is known as a *slope parameter*. Extending this concept to a linear mixed model, suppose we have random effects U and V, and the conditional mean function is given by

$$\mathrm{E}[Y|U,V] = \beta_0 + \beta_1 x + U + Vx = (\beta_0 + U) + (\beta_1 + V)x. \tag{2.10}$$

Here, $\mathrm{E}[Y|U,V]$ is represented as a linear function of x with slope $\beta_1 + V$. Since V is a random variable, $\beta_1 + V$ is a random quantity and is therefore known as a random slope. In a GLMM, the notion of slope is distorted somewhat by the nonlinear link function, but we still have a linear predictor on the link scale and therefore use the same terminology whenever a continuous component of $\mathbf{x}_i$ is contained in $\mathbf{d}_i$. Grömping (1996) noted that when $\mathbf{d}_i$ is a subset of $\mathbf{x}_i$, the corresponding components of $\boldsymbol{\beta}$ can be regarded as the average of a distribution of individual effects. For instance, in the model given by (2.10) each cluster of observations has its own realization of the random slope $\beta_1 + V$. If V has mean zero, then β_1 represents the average of the distribution of $\beta_1 + V$.

In any regression model, how the predictors $\mathbf{x}_i$ are defined affects the model's interpretation. As an example, consider a model with a log link and two independent Gaussian random effects: $U_i \sim \mathrm{N}(0, \sigma^2)$ and $V_i \sim \mathrm{N}(0, \tau^2)$. Using the subscript i to index clusters of observations sharing the same realization of the random effects and the subscript j to index observations within those clusters, we model the conditional mean as

$$\mathrm{E}[Y_{ij}|U_i, V_i] = \exp(\beta_0 + \beta_1 x_{ij} + U_i + V_i x_{ij} + \mathbf{d}_{ij}^T \mathbf{a}_{ij}), \tag{2.11}$$

where $\mathbf{d}_{ij} = (1, x_{ij})^T$ for each i and j. This is a marginally interpretable model of the form given by (2.1) with $\mathbf{U}_i = (U_i, V_i)^T$ for each i. Defining the vector of predictors for each observation as $\mathbf{x}_{ij} = (1, x_{ij})^T$, we interpret $\exp(\beta_0)$ as the mean response across the entire population when $x_{ij} = 0$. This only makes sense if zero is a reasonable value for x_{ij}. One might instead choose to center the covariates x_{ij} about their grand sample mean $\bar{x}$ and define $\mathbf{x}_{ij} = (1, x_{ij} - \bar{x})^T$ for each i and j. In this case, $\exp(\beta_0)$ represents the mean response across the entire population when the predictor x_{ij} is equal to its average value in the population. Another option would be to center the covariates x_{ij} about their cluster means $\bar{x}_{i\cdot}$ and to define $\mathbf{x}_{ij} = (1, x_{ij} - \bar{x}_{i\cdot})^T$ for each i and j. For this choice, $\exp(\beta_0)$ represents the mean response across the entire population when the predictor x_{ij} is equal to its average value within cluster i. Since the value of $\bar{x}_{i\cdot}$ varies across clusters, this definition of $\mathbf{x}_{ij}$ leads one to estimate greater variability in the random intercept because $\beta_0 + U_i$ corresponds to a different value of x_{ij} for each cluster.

Due to the connection between the fixed predictors $\mathbf{x}_{ij}$ and the random effects $\mathbf{U}_i$ in a GLMM for which $\mathbf{d}_{ij} = \mathbf{x}_{ij}$, the fit of the model, in addition to its interpretaion, is affected by the definition of $\mathbf{x}_{ij}$. The fact that centering the x_{ij} about their cluster means changes one's estimate of the variance of the random intercept is one example of how the choice of covariate influences model fit. Additionally, the adjustment $\mathbf{d}_{ij}^T\mathbf{a}_{ij}$ explicitly depends on $\mathbf{d}_{ij}$ and its magnitude is therefore driven by how one defines $\mathbf{x}_{ij}$. For a model with conditional mean given by (2.11), the adjustment, as a consequence of Proposition 2.2, is $\mathbf{d}_{ij}^T\mathbf{a}_{ij} = -(\sigma^2 + \tau^2 x_{ij}^2)/2$. As such, inclusion of the adjustment takes an expression for the conditional mean that was linear in x_{ij} on the log scale and makes it quadratic in x_{ij} on the log scale. Thus, the form of a marginally interpretable GLMM as a function of x_{ij} is fundamentally different from the form of a conventional GLMM in this context.

More generally, in a marginally interpretable GLMM with log link and multivariate normal random effects, we have the following proposition:

Proposition 2.6. *Suppose* $h(\cdot) = \exp(\cdot)$ *and we have a marginally interpretable GLMM of the form given by (2.1) and (2.2) for which* $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{\Sigma})$ *and* $\mathbf{d}_i = \mathbf{x}_i$*. Then the adjustment* $\mathbf{d}_i^T \mathbf{a}_i$ *is a quadratic form in both* $\mathbf{d}_i$ *and* $\mathbf{x}_i$*.*

Proof of Proposition 2.6: By Proposition 2.2, for each $i = 1, \ldots, N$ the adjustment is given by $\mathbf{d}_i^T \mathbf{a_i} = -\log\big(M_{\mathbf{U}}(\mathbf{d}_i)\big)$. Since, $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{\Sigma})$, $M_{\mathbf{U}}(\mathbf{d}_i) = \exp(\mathbf{d}_i^T \mathbf{\Sigma} \mathbf{d}_i/2)$. Thus, $\mathbf{d}_i^T \mathbf{a}_i = -\mathbf{d}_i^T \mathbf{\Sigma} \mathbf{d}_i/2 = -\mathbf{x}_i^T \mathbf{\Sigma} \mathbf{x}_i/2$, where the final equality holds because $\mathbf{d}_i = \mathbf{x}_i$. Hence, $\mathbf{d}_i^T \mathbf{a}_i$ is a quadratic form in both $\mathbf{d}_i$ and $\mathbf{x}_i$, as required. □

In this setting, because the adjustment is a quadratic form in $\mathbf{x}_i$, if the conditional mean is written as a polynomial on the log scale, then inclusion of the adjustment doubles the degree of the polynomial. This is stated formally in the following corollary:

Corollary 2.6.1. *Suppose we have a model of the form described in Proposition 2.6 and that the conditional mean, excluding the adjustment, is expressed as a polynomial of degree* r *on the log scale. Then the expression for the conditional mean with the adjustment included is a polynomial of degree* $2r$ *on the log scale.*

Proof of Corollary 2.6.1: For each $i = 1, \ldots, N$, let $\mathbf{d}_i = \mathbf{x}_i = (x_{i,1}, \ldots, x_{i,q})^T$ contain x_i^r and a subset of $q-1$ elements of $\{1, x_i, x_i^2, \ldots, x_i^{r-1}\}$, where $x_i \in \mathbb{R}$. Additionally, let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q-1})^T$ and $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{\Sigma})$, where the j^{th} row and k^{th} column of $\mathbf{\Sigma}$ is given by σ_{jk}. Then $\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{U}_i$ is a polynomial of degree r in x_i. By Proposition 2.6, the adjustment is $\mathbf{d}_i^T \mathbf{a}_i = -\mathbf{x}_i^T \mathbf{\Sigma} \mathbf{x}_i/2 = -\sum_{j=1}^q \sum_{k=1}^q x_{i,j} x_{i,k} \sigma_{jk}/2$. Thus, $\mathbf{d}_i^T \mathbf{a}_i$ is also a polynomial in x_i. Further, since the maximum degree of $x_{i,j}$ for any $j = 1, \ldots, q$ is r, the degree of $\mathbf{d}_i^T \mathbf{a}_i$ is $2r$. Hence, $\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{U}_i + \mathbf{d}_i^T \mathbf{a}_i$ is a polynomial of degree $2r$ in x_i. □

To illustrate Corollary 2.6.1, consider a marginally interpretable model with log link and let $\mathbf{x}_i = \mathbf{d}_i = (1, x_i, x_i^2, \ldots, x_i^r)^T$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_r)^T$, and $\mathbf{U}_i = (U_{i,0}, \ldots, U_{i,r})^T$, where we assume the $U_{i,k}$ are independent for $k = 0, \ldots, r$ and that $U_{i,k} \sim \mathrm{N}(0, \sigma_k^2)$. We express the conditional mean of the data Y_i as

$$\mathrm{E}[Y_i | U_{i,0}, \ldots, U_{i,r}] = \exp\big((\beta_0 + U_{i,0}) + (\beta_1 + U_{i,1})x_i + \cdots + (\beta_r + U_{i,r})x_i^r + \mathbf{d}_i^T \mathbf{a}_i).$$

Without the adjustment, we have a polynomial of degree r on the log scale. The adjustment for this model is given by $\mathbf{d}_i^T \mathbf{a}_i = -(\sigma_0^2 + \sigma_1^2 x_i^2 + \sigma_2^2 x_i^4 + \cdots + \sigma_r^2 x_i^{2r})/2$. This is a polynomial of degree $2r$, which is double the original degree r.

For models with link functions other than the natural logarithm, the adjustment $\mathbf{d}_i^T \mathbf{a}_i$ cannot typically be written as a quadratic form in $\mathbf{d}_i$. However, if the random effects in a GLMM are multivariate normal, then the adjustment often depends on a quadratic form of $\mathbf{d}_i$. For example, from Proposition 2.3, the expression for $\mathbf{d}_i^T \mathbf{a}_i$ for a model with a probit link includes $\mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i$. Also, the proof of Proposition 2.1 suggests that the adjustment is always a function of $\mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i$ for models with multivariate normal random effects, regardless of the choice of link function.

To demonstrate the impact of different definitions of $\mathbf{x}_{ij}$ on the parameter estimates and overall fit of a GLMM with a random slope, we consider data related to roadway fatalities in Scotland. These data were obtained from Scotland's publicly available online database of official statistics (`statistics.gov.scot`) and can be found in Tables A.1 and A.2 of Appendix A. We model the number of deaths per 1,000 people as a function of the per capita fuel consumption measured in tonnes of petrol and diesel consumed per person. The response Y_{ij} is the number of deaths on roadways in each of Scotland's 29 mainland council areas during each year from 2006 through 2011, where $i = 1, \ldots, 29$ indexes the council areas and $j = 1, \ldots, 6$ indexes the years. The corresponding covariate x_{ij} is the number of tonnes of petrol and diesel consumed per capita. For each council area and year, we denote the population (in thousands of people) by z_{ij}.

We begin by fitting a model with a random intercept but no random slope. Specifically, we model the expected number of deaths per thousand people as

$$\mathrm{E}[Y_{ij}/z_{ij}|U_i] = \exp(\beta_0 + \beta_1 x_{ij} + U_i + a_{ij}),$$

where we assume $Y_{ij}|U_i \sim \mathrm{Poisson}(\mathrm{E}[Y_{ij}|U_i])$ and $U_i \sim \mathrm{N}(0, \sigma^2)$. We could equivalently express this conditional mean structure as

$$\mathrm{E}[Y_{ij}|U_i] = \exp\big(\log(z_{ij}) + \beta_0 + \beta_1 x_{ij} + U_i + a_{ij}\big).$$

From Corollary 2.2.2, we have $a_{ij} = -\sigma^2/2$ for all i and j. This model assumes that there is variability in the roadway fatality rate across council areas, but that the relationship between fuel consumption and the roadway fatality rate is constant across council areas.

We fit both this marginally interpretable model and an analogous model without the adjustment via maximum likelihood estimation three times, replacing x_{ij} with $x_{ij} - \bar{x}$ the second time and with $x_{ij} - \bar{x}_{i\cdot}$ the third time. Details of the estimation procedure are provided in Chapter 4 and the resulting parameter estimates are given in Table 2.2. Also included in Table 2.2 is the Akaike Information Criterion (AIC) for each model, which provides a measure of model fit subject to a penalty for model complexity (Akaike, 1973). Smaller values of AIC indicate better fit. Regardless of how we define the predictor in this model, the adjustment has no impact on the fit of the model. In fact, the only difference between the marginally interpretable model and the conventional model for each of the three choices of predictor is that the estimate of β_0 is shifted up in the marginally interpretable model relative to its value in the conventional model. We discuss the relationship between the marginal and cluster-specific model parameterizations in greater detail in Section 3.1.

Table 2.2: Parameter estimates for several random intercept Poisson GLMMs fit to the Scottish roadway fatalities data

Marginally Interpretable Model

	x_{ij}		$x_{ij} - \bar{x}$		$x_{ij} - \bar{x}_{i\cdot}$	
	Point	Standard	Point	Standard	Point	Standard
Parameter	Estimate	Error	Estimate	Error	Estimate	Error
intercept (β_0)	-4.60	0.27	-3.06	0.09	-3.04	0.10
fuel consumption (β_1)	2.56	0.42	2.56	0.42	0.72	0.15
intercept variance (σ^2)	0.20	—	0.20	—	0.22	—
	AIC=913.1		AIC=913.1		AIC=927.7	

Conventional Model

	x_{ij}		$x_{ij} - \bar{x}$		$x_{ij} - \bar{x}_{i\cdot}$	
	Point	Standard	Point	Standard	Point	Standard
Parameter	Estimate	Error	Estimate	Error	Estimate	Error
intercept (β_0)	-4.71	0.28	-3.16	0.09	-3.16	0.09
fuel consumption (β_1)	2.56	0.42	2.56	0.42	0.72	0.15
intercept variance (σ^2)	0.20	—	0.20	—	0.22	—
	AIC=913.1		AIC=913.1		AIC=927.7	

Comparing different definitions of the predictor (x_{ij}, $x_{ij} - \bar{x}$, and $x_{ij} - \bar{x}_{i\cdot}$), the model that centers the x_{ij} about their grand mean and the model with no centering fit equally well, while the model that centers the x_{ij} about their cluster means does not fit as well as the other two. Centering about the cluster means leads, as expected, to a larger estimate of the random intercept variance σ^2, and also has an impact on the estimate of the slope parameter β_1. Both types of centering for x_{ij} alter the estimate of the intercept parameter β_0 because β_0 provides information about the expected death rate for a different value of x_{ij} when the covariate is centered versus when it is not.

Adding a random slope to the model, we assume $Y_{ij}|U_i, V_i \sim \text{Poisson}(\text{E}[Y_{ij}|U_i, V_i])$, $U_i \sim \text{N}(0, \sigma^2)$, and $V_i \sim \text{N}(0, \tau^2)$, define $\mathbf{d}_{ij} = (1, x_{ij})^T$, and write the conditional mean as

$$\text{E}[Y_{ij}|U_i, V_i] = \exp\big(\log(z_{ij}) + \beta_0 + \beta_1 x_{ij} + U_i + V_i x_{ij} + \mathbf{d}_{ij}^T \mathbf{a}_{ij}\big).$$

Table 2.3: Parameter estimates for several random intercept and slope Poisson GLMMs fit to the Scottish roadway fatalities data

Marginally Interpretable Model

	x_{ij}		$x_{ij} - \bar{x}$		$x_{ij} - \bar{x}_{i\cdot}$	
Parameter	Point Estimate	Standard Error	Point Estimate	Standard Error	Point Estimate	Standard Error
intercept (β_0)	-5.08	0.30	-3.01	0.10	-3.06	0.10
fuel consumption(β_1)	3.42	0.57	3.70	0.68	0.83	0.20
intercept variance (σ^2)	0.001	—	0.15	—	0.21	—
slope variance (τ^2)	0.51	—	2.92	—	0.16	—
	AIC=905.8		AIC=906.4		AIC=928.7	

Conventional Model

	x_{ij}		$x_{ij} - \bar{x}$		$x_{ij} - \bar{x}_{i\cdot}$	
Parameter	Point Estimate	Standard Error	Point Estimate	Standard Error	Point Estimate	Standard Error
intercept (β_0)	-4.98	0.29	-3.11	0.09	-3.18	0.09
fuel consumption (β_1)	3.07	0.52	3.58	0.69	0.86	0.21
intercept variance (σ^2)	0.001	—	0.15	—	0.21	—
slope variance (τ^2)	0.50	—	3.39	—	0.23	—
	AIC=906.3		AIC=907.7		AIC=927.8	

In addition to assuming that there is variability across council areas in the roadway fatality rate, this model assumes that the relationship between fuel consumption and the roadway fatality rate also varies by council area.

As before, we fit both this marginally interpretable model and an analogous model without the adjustment via maximum likelihood estimation three times, replacing x_{ij} with $x_{ij} - \bar{x}$ and $x_{ij} - \bar{x}_{i\cdot}$ the second and third times. The resulting parameter estimates along with the corresponding AIC values are given in Table 2.3. Unlike the model with only a random intercept, inclusion of the adjustment changes the fit of the model. The adjustment in this case is quadratic in the predictor (be it x_{ij}, $x_{ij} - \bar{x}$, or $x_{ij} - \bar{x}_{i\cdot}$) and therefore fundamentally changes the stucture of the model. We also see more variation in the parameter estimates across the different definitions of the predictor than we did in the model with only a random intercept. This is especially true for the estimates of the random effects variances. Once again, the largest estimate for the random intercept variance σ^2 corresponds to the model that centers the x_{ij} about their cluster means, but now there are differences among the estimates of the random slope variance τ^2 and the fixed effects parameters β_0 and β_1 across all three models. For these data, not centering the x_{ij} yields the best-fitting model in terms of AIC, but this will not always be the case.

In this particular application, every model we fit suggests that there is a positive association between fuel consumption and roadway fatalities. However, the different models suggest differing amounts of variation across council areas, both in the fatality rate and in the relationship between the fatality rate and fuel consumption. Depending on the goals of one's analysis, these differences can have a meaningful impact on one's conclusions. The key takeaway from Table 2.3 is that for a model that includes a random slope it is critical that one think carefully about how to parameterize the model and define the predictors; these decisions have a bigger impact on the fit of a model with a random slope than they do on the fit of a simpler model.

Chapter 3: Inference

Whether conducting inference in a classical framework or a Bayesian framework, the choice between a marginally interpretable GLMM and a conventional GLMM that fails to preserve the marginal mean can have a sizable impact on one's conclusions. In this chapter, we compare and contrast inferences from a marginally interpretable GLMM and a conventional GLMM. We identify situations wherein likelihood-based inference is identical under the two formulations, but show through simulations and examples that common inferential procedures, such as Wald tests and confidence intervals, can lead to markedly different results for the two model parameterizations. We argue that inference based on population-averaged, marginal parameters is typically of greater interest than inference based on cluster-specific, conditional parameters, and that marginal parameter estimates are more stable across different samples from the same population than their cluster-specific counterparts. We also address the consistency of parameter estimates in GLMMs.

In order to clearly distinguish between the two sets of parameters, in this chapter we denote the marginal fixed effects parameters in the marginally interpretable model as $\boldsymbol{\beta}^*$ while continuing to denote the cluster-specific fixed effects parameters in the conventional model as $\boldsymbol{\beta}$. Specifically, we write the conditional mean of Y_i in a conventional GLMM as

$$\mu_i = \mathrm{E}[Y_i|\mathbf{U}_i = \mathbf{u}] = h(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{u}), \tag{3.1}$$

and the conditional mean of Y_i in a marginally interpretable GLMM as

$$\mu_i^* = \mathrm{E}[Y_i|\mathbf{U}_i = \mathbf{u}] = h(\mathbf{x}_i^T\boldsymbol{\beta}^* + \mathbf{d}_i^T\mathbf{u} + \mathbf{d}_i^T\mathbf{a}_i). \tag{3.2}$$

The adjustment $\mathbf{d}_i^T\mathbf{a}_i$ is defined implicitly by the equation

$$h(\mathbf{x}_i^T\boldsymbol{\beta}^*) = \int h(\mathbf{x}_i^T\boldsymbol{\beta}^* + \mathbf{d}_i^T\mathbf{u} + \mathbf{d}_i^T\mathbf{a}_i)f_{\mathbf{U}}(\mathbf{u})d\mathbf{u}. \tag{3.3}$$

Where convenient, we write the adjustment as $\mathbf{d}_i^T\mathbf{a}_i = a(\mathbf{x}_i^T\boldsymbol{\beta}^*, \boldsymbol{\alpha})$ to emphasize that it is a function of the fixed effects portion of the model $\mathbf{x}_i^T\boldsymbol{\beta}^*$ and the parameters $\boldsymbol{\alpha}$ that characterize the random effects density $f_{\mathbf{U}}$.

3.1 Equivalence Between GLMM Parameterizations

In some cases, a marginally interpretable GLMM with conditional mean (3.2) is equivalent to a conventional GLMM with conditional mean (3.1) in the sense that the two models provide equal fit to the data. We consider two models *equivalent* if the joint marginal density $f_{\mathbf{Y}}$, also called the *marginal likelihood*, is the same for both models. A GLMM with conditional density $f_{Y|\mathbf{U}}$ and random effects density $f_{\mathbf{U}}$ has joint marginal density given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{N} \int f_{\mathbf{U}}(\mathbf{u})f_{Y|\mathbf{U}}(y_i|\mathbf{U}_i = \mathbf{u})d\mathbf{u}.$$

If this quantity is the same for two models, then those models yield identical predictions and identical likelihood-based inferences. For example, likelihood ratio tests would yield the same results for two equivalent models and, in a Bayesian framework, the posterior distribution of the marginal mean should be the same for both models as long as the prior distributions are also the same. We will show in Section 3.2 that, even with identical likelihoods, two models can yield markedly different inferences if the parameters on which inference is being made do not match. In this section, we characterize situations wherein the marginally interpretable and conventional GLMMs are, and are not, equivalent.

If we assume the same $f_{\mathbf{U}}$ for both a marginally interpretable GLMM and a conventional GLMM, any differences in the marginal density $f_{\mathbf{Y}}$ must arise through differences in the conditional density $f_{Y|\mathbf{U}}$. Such differences are a consequence of the two models having two distinct expressions for the conditional mean. In order for a marginally interpretable GLMM and a conventional GLMM to be equivalent, the two expressions for the conditional mean must be equal for all $i = 1, \ldots, N$. That is, for all $\mathbf{x}_i$ and $\mathbf{d}_i$, we need

$$h(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u}_i) = h(\mathbf{x}_i^T \boldsymbol{\beta}^* + \mathbf{d}_i^T \mathbf{u}_i + \mathbf{d}_i^T \mathbf{a}_i).$$

This equation reduces to

$$\mathbf{x}_i^T \boldsymbol{\beta} = \mathbf{x}_i^T \boldsymbol{\beta}^* + \mathbf{d}_i^T \mathbf{a}_i. \tag{3.4}$$

If, given the parameters that characterize $f_{\mathbf{U}}$, which we denote by $\boldsymbol{\alpha}$, the cluster-specific parameters $\boldsymbol{\beta}$ can be written as a function of the marginal parameters $\boldsymbol{\beta}^*$ such that (3.4) is satisfied for all $i = 1, \ldots, N$, then the two models are equivalent. In such a case, there exists an isomorphism between the marginal parameters $\boldsymbol{\beta}^*$ of the marginally interpretable model and the cluster-specific parameters $\boldsymbol{\beta}$ of the conventional model. When this occurs, the difference between $\mathbf{x}_i^T \boldsymbol{\beta}$ and $\mathbf{x}_i^T \boldsymbol{\beta}^*$ is exactly equal to the adjustment $\mathbf{d}_i^T \mathbf{a}_i$, and is compensated for with an appropriate shift in the location of the random effects distribution. To determine if a marginally interpretable GLMM is equivalent to an analogous conventional GLMM, we therefore seek to find a one-to-one correspondence between the marginal parameters $\boldsymbol{\beta}^*$ and the cluster-specific parameters $\boldsymbol{\beta}$.

3.1.1 One-Way Layout

Experiments are often set up to have a *one-way layout*, meaning that the data are divided into a finite number of distinct groups, and the model for the data includes a separate fixed effects parameter for each group. Data of this form can be used to conduct a one-way analysis of variance (or analysis of deviance), and such a model is known as a *saturated model* or *cell means model*. For a one-way layout, the following proposition applies:

Proposition 3.1. *Consider two GLMMs, one with conditional mean given by*

$$\mu_k = \mathrm{E}[Y_k|\mathbf{U}_k = \mathbf{u}] = h(\kappa_k + \boldsymbol{\delta}_k^T\mathbf{u}), \tag{3.5}$$

and one with conditional mean given by

$$\mu_k^* = \mathrm{E}[Y_k|\mathbf{U}_k = \mathbf{u}] = h(\kappa_k^* + \boldsymbol{\delta}_k^T\mathbf{u} + \boldsymbol{\delta}_k^T\boldsymbol{\omega}_k), \tag{3.6}$$

where $k = 1, \ldots, p$, the κ_k and κ_k^* are fixed effects parameters, each $\boldsymbol{\delta}_k$ is a q-vector of covariates, and each $\boldsymbol{\delta}_k^T\boldsymbol{\omega}_k$ is an adjustment that makes the model marginally interpretable. Suppose $f_{\mathbf{U}}$ is the same for both models, $f_{Y|\mathbf{U}}$ is from the same family of distributions for both models, and the parameters $\boldsymbol{\alpha}$ characterizing $f_{\mathbf{U}}$ are constant across $k = 1, \ldots, p$. Then, if the adjustment $\boldsymbol{\delta}_k^T\boldsymbol{\omega}_k$ exists for all $k = 1, \ldots, p$, the two models are equivalent.

Proof of Proposition 3.1: For $k = 1, \ldots, p$, each $\boldsymbol{\delta}_k^T\boldsymbol{\omega}_k = a(\kappa_k^*, \boldsymbol{\alpha})$ is a function of κ_k^* and $\boldsymbol{\alpha}$. Since the p groups are distinct, each $a(\kappa_k^*, \boldsymbol{\alpha})$ is computed independently of the other $p - 1$ adjustments, and we can write $\kappa_k = \kappa_k^* + a(\kappa_k^*, \boldsymbol{\alpha})$ for each $k = 1, \ldots, p$ for which $a(\kappa_k^*, \boldsymbol{\alpha})$ exists. This satisfies an analogue to (3.4) for every $k = 1, \ldots, p$ and the two models are equivalent. □

More generally, if a pair of GLMMs with conditional means given by (3.1) and (3.2) can be reparameterized into a one-way layout, then Proposition 3.1 still applies. For this to occur, $\mathbf{x}_i$ must take p unique values and $\mathbf{d}_i$ must take at most p unique values, where $i = 1, \ldots, N$, each $\mathbf{x}_i$ is a p-vector, and $p \leq N$. Further, each index $i = 1, \ldots, N$ must map to one of p groups indexed by $k = 1, \ldots, p$ such that each $\mathbf{x}_i^T\boldsymbol{\beta}$ is equal to one of p values κ_k and each $\mathbf{x}_i^T\boldsymbol{\beta}^*$ is equal to one of p values κ_k^*. Under these circumstances, (3.4) holds for all $i = 1, \ldots, N$ and a marginally interpretable version of the model is equivalent to a conventional version. This is stated more formally in the following corollary:

Corollary 3.1.1. *Consider two GLMMs, one with conditional mean given by (3.1) and one with conditional mean given by (3.2), each with p fixed effects parameters, where $p \leq N$. If these two models can be reparameterized such that the all of the assumptions of Proposition 3.1 are satisfied, then the two models are equivalent.*

Proof of Corollary 3.1.1: By Proposition 3.1, given two models with conditional means (3.5) and (3.6) we can write $\kappa_k = \kappa_k^* + \boldsymbol{\delta}_k^T \boldsymbol{\omega}_k$ for each $k = 1, \ldots, p$ for which $\boldsymbol{\delta}_k^T \boldsymbol{\omega}_k$ exists. If there exists a mapping between the indices $i = 1, \ldots, N$ and $k = 1, \ldots, p$ that allows (3.1) to be written as (3.5) and (3.2) to be written as (3.6), then for every $i = 1, \ldots, N$ there exists a $k = 1, \ldots, p$ such that $\kappa_k = \kappa_k^* + \boldsymbol{\delta}_k^T \boldsymbol{\omega}_k$ translates to $\mathbf{x}_i^T \boldsymbol{\beta} = \mathbf{x}_i^T \boldsymbol{\beta}^* + \mathbf{d}_i^T \mathbf{a}_i$. Thus, (3.4) is satisfied for all $i = 1, \ldots, N$ and the two models are equivalent. □

As an example of when this result applies, consider a model with a single random intercept $U_i \sim \mathrm{N}(0, \sigma^2)$ and a single binary predictor. Let each $\mathbf{x}_i = (1, z_i)^T$, where $z_i \in \{0, 1\}$ for all $i = 1, \ldots, N$. Then either $\mathbf{x}_i = (1, 0)^T$ or $\mathbf{x}_i = (1, 1)^T$ and, in the notation of Proposition 3.1, we have $\kappa_1 = \beta_0$ and $\kappa_2 = \beta_0 + \beta_1$. Also, the only parameter needed to characterize the distribution of the random effects is σ^2. To satisfy (3.4) we require the following two conditions to be met:

1. $\beta_0 = \beta_0^* + a(\beta_0^*, \sigma^2)$;

2. $\beta_0 + \beta_1 = \beta_0^* + \beta_1^* + a(\beta_0^* + \beta_1^*, \sigma^2)$.

The first condition provides an expression for β_0 as a function of β_0^*. Given this β_0, the second condition leads to the following expression for β_1:

$$\beta_1 = \beta_1^* + a(\beta_0^* + \beta_1^*, \sigma^2) - a(\beta_0^*, \sigma^2).$$

Thus, $\boldsymbol{\beta}$ can be expressed as a function of $\boldsymbol{\beta}^*$ such that (3.4) is satisfied for all $i = 1, \ldots, N$, and the marginally interpretable GLMM is equivalent to the conventional GLMM.

If we continue to assume a single Gaussian random intercept, but add a second binary predictor and include the interaction between the two predictors in the model, we obtain a model for which $p = 4$ and each $\mathbf{x}_i = (1, z_i, w_i, z_i \times w_i)^T$, where $z_i, w_i \in \{0, 1\}$. Here, $\mathbf{x}_i \in \{(1,0,0,0)^T, (1,1,0,0)^T, (1,0,1,0)^T, (1,1,1,1)^T\}$. This set has cardinality four, and we can define $\kappa_1 = \beta_0$, $\kappa_2 = \beta_0 + \beta_1$, $\kappa_3 = \beta_0 + \beta_2$, and $\kappa_4 = \beta_0 + \beta_1 + \beta_2 + \beta_3$. For equivalence to hold, the following conditions must all be satisfied:

1. $\beta_0 = \beta_0^* + a(\beta_0^*, \sigma^2)$;
2. $\beta_0 + \beta_1 = \beta_0^* + \beta_1^* + a(\beta_0^* + \beta_1^*, \sigma^2)$;
3. $\beta_0 + \beta_2 = \beta_0^* + \beta_2^* + a(\beta_0^* + \beta_2^*, \sigma^2)$;
4. $\beta_0 + \beta_1 + \beta_2 + \beta_3 = \beta_0^* + \beta_1^* + \beta_2^* + \beta_3^* + a(\beta_0^* + \beta_1^* + \beta_2^* + \beta_3^*, \sigma^2)$.

Given the marginal parameters β_0^*, β_1^*, β_2^*, and β_3^*, we can write the cluster-specific parameters β_0, β_1, β_2, and β_3 as

1′. $\beta_0 = \beta_0^* + a(\beta_0^*, \sigma^2)$;

2′. $\beta_1 = \beta_1^* + a(\beta_0^* + \beta_1^*, \sigma^2) - a(\beta_0^*, \sigma^2)$;

3′. $\beta_2 = \beta_2^* + a(\beta_0^* + \beta_2^*, \sigma^2) - a(\beta_0^*, \sigma^2)$;

4′. $\beta_3 = \beta_3^* + a(\beta_0^* + \beta_1^* + \beta_2^* + \beta_3^*, \sigma^2) - a(\beta_0^* + \beta_1^*, \sigma^2) - a(\beta_0^* + \beta_2^*, \sigma^2) + a(\beta_0^*, \sigma^2)$.

Defining β_0, β_1, β_2, and β_3 as in 1′-4′ satisfies (3.4) for all $i = 1, \ldots, N$, and the two models are equivalent.

When there are more than p unique values for $\mathbf{x}_i$, Corollary 3.1.1 no longer applies and equivalence is not guaranteed. Take, for example, a model with a single random intercept $U_i \sim \mathrm{N}(0, \sigma^2)$ and two binary predictors that does not include the interaction between the two predictors. That is, each $\mathbf{x}_i = (1, z_i, w_i)^T$, where $z_i, w_i \in \{0, 1\}$. Thus, $p = 3$ but $\mathbf{x}_i \in \{(1,0,0)^T, (1,1,0)^T, (1,0,1)^T, (1,1,1)^T\}$, which has cardinality four. The following equations must all be satisfied for (3.4) to hold for all $i = 1, \ldots, N$:

1. $\beta_0 = \beta_0^* + a(\beta_0^*, \sigma^2)$;

2. $\beta_0 + \beta_1 = \beta_0^* + \beta_1^* + a(\beta_0^* + \beta_1^*, \sigma^2)$;

3. $\beta_0 + \beta_2 = \beta_0^* + \beta_2^* + a(\beta_0^* + \beta_2^*, \sigma^2)$;

4. $\beta_0 + \beta_1 + \beta_2 = \beta_0^* + \beta_1^* + \beta_2^* + a(\beta_0^* + \beta_1^* + \beta_2^*, \sigma^2)$.

One implication of these equations is that

$$a(\beta_0^* + \beta_1^* + \beta_2^*, \sigma^2) - a(\beta_0^* + \beta_1^*, \sigma^2) = a(\beta_0^* + \beta_2^*, \sigma^2) - a(\beta_0^*, \sigma^2). \tag{3.7}$$

If the adjustment $a(\cdot, \sigma^2)$ were linear in its first argument, then the additivity property of linear functions ($f(x+y) = f(x) + f(y)\ \forall x, y \in \mathbb{R}$) would imply that (3.7) is always true. Since $a(\cdot, \sigma^2)$ is not necessarily linear in its first argument, (3.7) does not hold in general. An example of when this condition does not hold is a model with a logit link and a normal random intercept. For such a model, $a(\cdot, \sigma^2)$ is nonlinear in its first argument, and (3.7) only holds in special cases, such as when either $\beta_1^* = 0$ or $\beta_2^* = 0$ and the model essentially reduces to the case of a single binary predictor.

Corollary 3.1.1 also does not apply when a predictor is continuous. Consider the case when $p = 2$ and each $\mathbf{x}_i = (1, z_i)^T$, where $z_i \in \mathbb{R}$. Here, $\mathbf{x}_i$ can take more than two values.

For equivalence between the conventional model and the marginally interpretable model, we require that the following two conditions be met for all $i = 1, \dots, N$:

1. $\beta_0 = \beta_0^* + a(\beta_0^*, \sigma^2)$;

2. $\beta_0 + \beta_1 z_i = \beta_0^* + \beta_1^* z_i + a(\beta_0^* + \beta_1^* z_i, \sigma^2)$.

If $a(\cdot, \sigma^2)$ is not linear in its first argument, then these conditions will generally not hold for all values of z_i, $i = 1, \dots, N$.

In the statement of Corollary 3.1.1, we assume that the parameters $\boldsymbol{\alpha}$ that characterize $f_{\mathbf{U}}$ remain constant for all $i = 1, \dots, N$. This does not necessarily need to be true for Corollary 3.1.1 to apply, as stated in the following corollary:

Corollary 3.1.2. *Suppose all of the assumptions of Corollary 3.1.1 hold, except we allow the random effects parameters $\boldsymbol{\alpha}_i$ to vary with i. If the $\boldsymbol{\alpha}_i$ map to the same p groups, indexed by $k = 1, \dots, p$, as the $\mathbf{x}_i$ and $\mathbf{d}_i$, then the marginally interpretable GLMM and conventional GLMM are still equivalent.*

Proof of Corollary 3.1.2: Let each vector of variance components $\boldsymbol{\alpha}_i$, $i = 1, \dots, N$, take one of p values $\boldsymbol{\alpha}'_k$, where $k = 1, \dots, p$. Since calculation of the adjustment $a(\kappa_k^*, \boldsymbol{\alpha}'_k)$ for each group $k = 1, \dots, p$ depends only on the variance components $\boldsymbol{\alpha}'_k$ for that group, each $a(\kappa_k^*, \boldsymbol{\alpha}'_k)$ is still computed independently of the other $p - 1$ adjustments and we are able to write $\kappa_k = \kappa_k^* + a(\kappa_k^*, \boldsymbol{\alpha}'_k)$ for each $k = 1, \dots, p$ as long as $a(\kappa_k^*, \boldsymbol{\alpha}'_k)$ exists. As before, this translates to $\mathbf{x}_i^T \boldsymbol{\beta} = \mathbf{x}_i^T \boldsymbol{\beta}^* + \mathbf{d}_i^T \mathbf{a}_i$ for each $i = 1, \dots, N$. Thus, (3.4) is satisfied for every $i = 1, \dots, N$ and the two models are equivalent. □

Consider a model with a single random intercept $U_i \sim \mathrm{N}(0, \sigma_i^2)$ for which $\boldsymbol{\alpha}_i$ is simply σ_i^2. When a group of observations is allowed to have more than one variance, Corollary 3.1.2 fails. Suppose, for example, one believes that a binary predictor has no impact on the conditional mean, but that it does have an impact on the random effects variance. In this case, $p = 1$ and $x_i = 1$ for all $i = 1, \ldots, N$, but σ_i^2 can take either of two possible values. Such a model could alternatively be thought of as having two random intercept terms with the random portion of the model expressed as $\mathbf{d}_i^T \mathbf{U}_i$, where $\mathbf{U}_i = (U_i, V_i)^T$ is a two-dimensional random vector and $\mathbf{d}_i = (1, 0)^T$ or $\mathbf{d}_i = (0, 1)^T$ depending on the value of the binary predictor. Using this approach, we write the conditional mean as

$$\mu_i^* = \mathrm{E}[Y_i | U_i = u, V_i = v] = h\big(\beta_0^* + z_i u + (1 - z_i)v + \mathbf{d}_i^T \mathbf{a}_i\big),$$

where $z_i \in \{0, 1\}$, $U_i \sim \mathrm{N}(0, \sigma^2)$, and $V_i \sim \mathrm{N}(0, \tau^2)$. To satisfy (3.4) for all $i = 1, \ldots, N$, we would need the following conditions to be satisfied:

1. $\beta_0 = \beta_0^* + a(\beta_0^*, \sigma^2)$;

2. $\beta_0 = \beta_0^* + a(\beta_0^*, \tau^2)$.

Since the adjustment $a(\cdot, \cdot)$ is generally not constant in its second argument, these two equations typically do not hold simultaneously unless the two variances, σ^2 and τ^2, are equal and the model corresponds to the constant variance case described earlier.

If, however, the binary predictor in the preceding example were also assumed to have an impact on the conditional mean and the data were divided into two distinct groups with two distinct variance parameters, then Corollary 3.1.2 would apply. We could write the conditional mean of such a model as

$$\mu_i^* = \mathrm{E}[Y_i | U_i = u, V_i = v] = h\big(\beta_0^* + \beta_1^* z_i + z_i u + (1 - z_i)v + \mathbf{d}_i^T \mathbf{a}_i\big),$$

where $z_i \in \{0, 1\}$, $U_i \sim \mathrm{N}(0, \sigma^2)$, and $V_i \sim \mathrm{N}(0, \tau^2)$. The adjustments for the two groups in this model would be $a(\beta_0^*, \tau^2)$ and $a(\beta_0^* + \beta_1^*, \sigma^2)$.

3.1.2 Matching Functional Forms

When a GLMM cannot be reduced to a one-way layout, whether or not a marginally interpretable GLMM is equivalent to an analogous conventional GLMM depends on the form of the adjustment, which itself depends on the choice of link function and random effects distribution. In general, if the functional form of $\mathbf{d}_i^T\mathbf{a}_i$ matches the form of the linear fixed effects predictor $\mathbf{x}_i^T\boldsymbol{\beta}$, then (3.4) holds for all $i = 1, \dots, N$ and the two models are equivalent. This is stated more formally as follows:

Proposition 3.2. *Consider two GLMMs, one with conditional mean given by (3.1) and one with conditional mean given by (3.2), each with p fixed effects parameters and q random effects. Assume the elements of the q-vector $\mathbf{d}_i$ are a subset of the elements of the p-vector $\mathbf{x}_i$ for all $i = 1, \dots, N$. Suppose that $f_{\mathbf{U}}$ is characterized by the parameters $\boldsymbol{\alpha}$ and is the same for both models, and that $f_{Y|\mathbf{U}}$ is from the same family of distributions for both models. If $\mathbf{d}_i^T\mathbf{a}_i$ exists for all $i = 1, \dots, N$ and can be written as a linear form of some subset of $\mathbf{x}_i$, then the two models are equivalent.*

Proof of Proposition 3.2: The fixed effects portions $\mathbf{x}_i^T\boldsymbol{\beta}$ and $\mathbf{x}_i^T\boldsymbol{\beta}^*$ of the two models each represent a linear combination of the p terms $x_{i,0}, \dots, x_{i,p-1}$. Suppose $\mathbf{d}_i^T\mathbf{a}_i$ can be written as a linear combination of r terms $z_{i,1}, \dots, z_{i,r}$, where $r \in \{1, \dots, p\}$ and $\{z_{i,1}, \dots, z_{i,r}\}$ is a subset of $\{x_{i,0}, \dots, x_{i,p-1}\}$. If we arrange the relevant subset of $\mathbf{x}_i$ and the corresponding subsets of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ into the r-vectors $\mathbf{z}_i$, $\boldsymbol{\gamma}$, and $\boldsymbol{\gamma}^*$, then for each $j = 1, \dots, r$ we can write the j^{th} term of $\mathbf{d}_i^T\mathbf{a}_i$ as $c_j z_{i,j}\gamma_j^*$ for some constant c_j. In turn, $z_{i,j}\gamma_j^*$ plus the j^{th} term of $\mathbf{d}_i^T\mathbf{a}_i$ is equal to $(1 + c_j)z_{i,j}\gamma_j^*$ for each $j = 1, \dots, r$. For each $k = 0, \dots, p-1$, we can therefore write $\beta_k = (1 + c_k)\beta_k^*$ for some constant c_k if β_k is one of the r elements of $\boldsymbol{\beta}$

contained in $\boldsymbol{\gamma}$, and $\beta_k = \beta_k^*$ otherwise. Thus, we are able to write $\boldsymbol{\beta}$ as a function of $\boldsymbol{\beta}^*$ such that (3.4) holds for all $i = 1, \ldots, N$, and the two models are equivalent. □

In some cases when the assumptions of Proposition 3.2 are not satisfied, it is possible to reparameterize the model so that the proposition applies. Consider, for example, a model with a log link, a single random intercept $U_i \sim \mathrm{N}(0, \sigma^2)$, and conditional mean given by

$$\mu_k^* = \mathrm{E}[Y_k | U_k = u] = \exp(\kappa_k^* + u + a_k),$$

where $k = 1, 2$. This mean could alternatively be expressed as

$$\mu_i^* = \mathrm{E}[Y_i | U_i = u] = \exp(z_i \kappa_1^* + (1 - z_i)\kappa_2^* + u + a_i) = \exp(\mathbf{x}_i^T \boldsymbol{\kappa}^* + u + a_i),$$

where $i = 1, \ldots, N$, $\boldsymbol{\kappa}^* = (\kappa_1^*, \kappa_2^*)^T$, $\mathbf{x}_i = (z_i, 1 - z_i)^T$, and $z_i \in \{0, 1\}$. Since we have a single random intercept, $d_i = 1$ for all $i = 1, \ldots, N$ and is therefore not a subset of $\mathbf{x}_i$. Further, from Corollary 2.2.2, $a_i = -\sigma^2/2$ for all $i = 1, \ldots, N$ and is therefore constant in z_i whereas $\mathbf{x}_i^T \boldsymbol{\kappa}^*$ is linear in z_i. Thus, Proposition 3.2 does not directly apply. Nonetheless, this model could be reparameterized to have conditional mean

$$\mu_i^* = \mathrm{E}[Y_i | U_i = u] = \exp(\beta_0^* + \beta_1^* z_i + u + a_i) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}^* + u + a_i),$$

where $i = 1, \ldots, N$, $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^T$, $\mathbf{x}_i = (1, z_i)^T$, and $z_i \in \{0, 1\}$. Now, $d_i = 1$ is in fact a subset of $\mathbf{x}_i$, both a_i and the β_0^* term in $\mathbf{x}_i^T \boldsymbol{\beta}^*$ are constant is z_i, and Proposition 3.2 is applicable. This leads to the following corollary:

Corollary 3.2.1. *If two GLMMs can be reparameterized in such a manner that all of the assumptions of Proposition 3.2 are satisfied, then the two models are equivalent.*

Proof of Corollary 3.2.1: If two models can be shown by Proposition 3.2 to be equivalent, then reparameterizations of those two equivalent models are also equivalent. □

Since the functional form of $\mathbf{d}_i^T\mathbf{a}_i$ depends on the choice of link function, we highlight situations wherein Proposition 3.2 does and does not apply for a few popular links.

Log Link

First, consider the model from the previous example with a log link and a single random intercept $U_i \sim \mathrm{N}(0, \sigma^2)$. Then, from Corollary 2.2.2, $a_i = -\sigma^2/2$ for all $i = 1, \ldots, N$, and (3.4) reduces to

$$\mathbf{x}_i^T\boldsymbol{\beta} = \mathbf{x}_i^T\boldsymbol{\beta}^* - \frac{\sigma^2}{2}.$$

Since the adjustment a_i is constant as a function of $\mathbf{x}_i$, equivalence holds as long as $\mathbf{x}_i$ contains an element that is constant across all $i = 1, \ldots, N$. Typically, the first element of $\mathbf{x}_i$ is always equal to one, representing a fixed intercept term. In such a case, we can satisfy (3.4) for all $i = 1, \ldots, N$ by defining $\beta_0 = \beta_0^* - \sigma^2/2$ and $\beta_j = \beta_j^*$ for $j = 1, \ldots, p-1$.

Now suppose a model with a log link contains q multivariate normal random effects $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, N$. Consider a model with a single continuous predictor such that $\mathbf{x}_i = (1, z_i)^T$ for all $i = 1, \ldots, N$, where $z_i \in \mathbb{R}$. Further, let $\mathbf{d}_i = \mathbf{x}_i$. It was shown in Section 2.3 that $\mathbf{d}_i^T\mathbf{a}_i = -\mathbf{d}_i^T\Sigma\mathbf{d}_i/2$ is a quadratic form in $\mathbf{d}_i$, which in this case means it is also a quadratic form in $\mathbf{x}_i$. Consequently, $\mathbf{x}_i^T\boldsymbol{\beta}^* + \mathbf{d}_i^T\mathbf{a}_i$ does not generally equal $\mathbf{x}_i^T\boldsymbol{\beta}$ for all $i = 1, \ldots, N$ because one is linear in z_i and the other is quadratic in z_i. The functional forms do not match and Proposition 3.2 does not apply.

It is, however, possible for a marginally interpretable GLMM with a log link and random slope to be equivalent to an analogous conventional GLMM. Consider a situation identical to the one described in the preceding paragraph, except now $\mathbf{x}_i = (1, z_i, z_i^2)^T$ while $\mathbf{d}_i = (1, z_i)^T$. In this case, $\mathbf{x}_i^T\boldsymbol{\beta}$, $\mathbf{x}_i^T\boldsymbol{\beta}^*$, and $\mathbf{d}_i^T\mathbf{a}_i$ are all quadratic in z_i, Proposition 3.2 applies, and the two models are equivalent.

Probit Link

Consider a model with a probit link and a normally distributed random intercept with variance σ^2. Proposition 2.3 tells us that $a_i = \big((1+\sigma^2)^{1/2} - 1\big)\mathbf{x}_i^T\boldsymbol{\beta}^*$. Here, Proposition 3.2 applies and, in the notation of the proof of Proposition 3.2, $c_j = \big((1+\sigma^2)^{1/2} - 1\big)$ for all $j = 0, \dots, p-1$. Thus, $\mathbf{x}_i^T\boldsymbol{\beta}^* + a_i = (1+\sigma^2)^{1/2}\mathbf{x}_i^T\boldsymbol{\beta}^*$ for all $i = 1, \dots, N$, and setting $\beta_j = (1+\sigma^2)^{1/2}\beta_j^*$ for each $j = 0, \dots, p-1$ satisfies (3.4) for every $i = 1, \dots, N$.

Now consider a probit model with multivariate normal random effects $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ that include random slopes. Specifically, let $\mathbf{d}_i = \mathbf{x}_i$. From Proposition 2.3, the adjustment has the form $\mathbf{d}_i^T\mathbf{a}_i = \big((1+\mathbf{d}_i^T\boldsymbol{\Sigma}\mathbf{d}_i)^{1/2} - 1\big)\mathbf{x}_i^T\boldsymbol{\beta}^* = \big((1+\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i)^{1/2} - 1\big)\mathbf{x}_i^T\boldsymbol{\beta}^*$, which is not necessarily linear in $\mathbf{x}_i$. Consequently, Proposition 3.2 does not always apply for a model with a probit link and normal random effects, and a marginally interpretable GLMM is not necessarily equivalent to a conventional GLMM in this context.

To emphasize the need for the elements of $\mathbf{d}_i$ to be a subset of $\mathbf{x}_i$ for all $i = 1, \dots, N$ in order for Proposition 3.2 to apply, continue to assume $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, but now suppose $\mathbf{x}_i = (1, z_i)^T$ and $\mathbf{d}_i = (1, w_i)^T$, where $z_i, w_i \in \mathbb{R}$ and each w_i is uncorrelated with each z_i. As in the preceding example, $\mathbf{d}_i^T\mathbf{a}_i = \big((1+\mathbf{d}_i^T\boldsymbol{\Sigma}\mathbf{d}_i)^{1/2} - 1\big)\mathbf{x}_i^T\boldsymbol{\beta}^*$, but since $\mathbf{d}_i$ is unrelated to $\mathbf{x}_i$, the adjustment is now linear in $\mathbf{x}_i$. We cannot, however, write $\boldsymbol{\beta} = c\boldsymbol{\beta}^*$ for any single constant c. For each $i = 1, \dots, N$, the appropriate constant would be $(1+\mathbf{d}_i^T\boldsymbol{\Sigma}\mathbf{d}_i)^{1/2}$, but this quantity varies with i and thereby prevents us from expressing $\boldsymbol{\beta}$ as a function of $\boldsymbol{\beta}^*$ in a manner that satisfies (3.4) for all $i = 1, \dots, N$.

Logit Link

When a GLMM has a logit link function and normal random effects, the adjustment $\mathbf{d}_i^T\mathbf{a}_i$ is nonlinear in $\mathbf{x}_i$ and Proposition 3.2 does not apply. The examples in Section 3.1.1

of a model with two binary predictors but no interaction between the two, and of a model with a single continuous predictor are two situations wherein equivalence fails to hold for a logistic-normal model.

Considering the case of a single random intercept U_i, if U_i follows the bridge distribution of Wang and Louis (2003), then a_i is linear in $\mathbf{x}_i$ and Proposition 3.2 applies. This situation is analogous to a model with a probit link and normal random intercept as there exists a constant $c \in \mathbb{R}$ such that $\boldsymbol{\beta} = c\boldsymbol{\beta}^*$ satisfies (3.4) for all $i = 1, \ldots, N$.

Complementary Log-Log Link

Much like for models with a logit link, equivalence generally only exists between a conventional model and a marginally interpretable model with a complementary log-log link when f_U is assumed to be the bridge density derived by Wang and Louis (2003).

3.2 Hypothesis Testing

In this section, we investigate the behavior of three classical large sample tests – the likelihood ratio test, the Wald test, and the score test – in the context of GLMMs. When a marginally interpretable GLMM and a conventional GLMM are equivalent (as defined in the previous section) under both the null hypothesis and the alternative hypothesis, likelihood ratio tests yield identical results under the two models. Wald tests and score tests, however, can lead to markedly different conclusions between the two models if one is not careful to focus the tests on the same quantity. Confidence intervals obtained from inverting these tests, including intervals obtained from inverting the likelihood ratio test, can also be discrepant in these situations, as will be shown through a few examples.

Suppose our model contains r parameters, denoted by the r-vector $\boldsymbol{\theta}$, and we want to make inference on those parameters. Unless otherwise indicated, our null hypothesis is $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and our alternative hypothesis is $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. We write the marginal likelihood of the data, previously expressed as the joint marginal density $f_{\mathbf{Y}}$, as $\mathcal{L}_N(\boldsymbol{\theta})$ to emphasize that we are treating it as a function of the unknown parameters. The subscript N indicates that this likelihood is based on N observations. The natural logarithm of $\mathcal{L}_N(\boldsymbol{\theta})$ is known as the *log-likelihood* and we denote it as $\ell_N(\boldsymbol{\theta})$ or as $\ell_N(\boldsymbol{\theta}|Y_1, \ldots, Y_N)$. The latter form is only used when it is convenient to emphasize the role of the data.

The *likelihood ratio test* (Neyman and Pearson, 1928a,b) is based on the ratio of the maximum likelihood under the null and alternative models. If we restate the null hypothesis as $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ and the alternative hypothesis as $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^C$, where $\boldsymbol{\Theta}_0$ represents the parameter space under the null model, and $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_0^C$ represents the unrestricted parameter space, then the *likelihood ratio* is given by

$$\lambda = \frac{\sup_{\boldsymbol{\Theta}_0} \mathcal{L}_N(\boldsymbol{\theta})}{\sup_{\boldsymbol{\Theta}} \mathcal{L}_N(\boldsymbol{\theta})} = \frac{\mathcal{L}_N(\hat{\boldsymbol{\theta}}_0)}{\mathcal{L}_N(\hat{\boldsymbol{\theta}})},$$

where $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ under the null hypothesis and $\hat{\boldsymbol{\theta}}$ is the unrestricted maximum likelihood estimate of $\boldsymbol{\theta}$. This leads to the test statistic

$$T_L = -2\log(\lambda) = -2\big(\ell_N(\hat{\boldsymbol{\theta}}_0) - \ell_N(\hat{\boldsymbol{\theta}})\big). \tag{3.8}$$

For independent and identically distributed observations, the statistic T_L asymptotically (as $N \to \infty$) follows a chi-squared distribution with r degrees of freedom under certain conditions when the null hypothesis is true (Wilks, 1938). For testing a single parameter, a likelihood ratio test rejects $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta \neq \theta_0$ when $T_L > \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ is the $100 \times (1 - \alpha)$ percentile of a χ^2_1 distribution and α is the level of the test.

The *Wald test* (Wald, 1943) is based on the asymptotic normality of the maximum likelihood estimator. The test statistic is given by

$$T_W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathrm{I}_N(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{3.9}$$

where $\mathrm{I}_N(\hat{\boldsymbol{\theta}})$ is the Fisher information matrix based on N observations. The i^{th} row and j^{th} column of this matrix is defined as

$$-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_N(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{3.10}$$

for $i, j = 1, \ldots, r$. For independent and identically distributed observations, the asymptotic distribution of T_W (as $N \to \infty$), like that of T_L, is a chi-squared distribution with r degrees of freedom (Wald, 1943). The Wald test is therefore asymptotically equivalent to the likelihood ratio test. In the univariate case, with null hypothesis $H_0 : \theta = \theta_0$ and alternative $H_1 : \theta \neq \theta_0$, the test statistic can be expressed as

$$Z_W = \frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})}, \tag{3.11}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ and $\mathrm{SE}(\hat{\theta})$ is its standard error. Here, the standard error of $\hat{\theta}$ is equal to the square root of the inverse of the observed Fisher information. That is, $\mathrm{SE}(\hat{\theta}) = \big(\mathrm{I}_N(\hat{\theta})\big)^{-1/2}$, where

$$\mathrm{I}_N(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\ell_N(\theta)|_{\theta=\hat{\theta}}.$$

When a univariate θ_i ($i = 1, \ldots, r$) is an element of an r-vector $\boldsymbol{\theta}$, $\mathrm{SE}(\hat{\theta}_i)$ is equal to the square root of the i^{th} diagonal element of $\big(\mathrm{I}_N(\hat{\boldsymbol{\theta}})\big)^{-1}$, where the elements of $\mathrm{I}_N(\hat{\boldsymbol{\theta}})$ are defined as in (3.10). For independent and identically distributed observations, the statistic Z_W converges in distribution (as $N \to \infty$) to a standard normal distribution, and its square is $Z_W^2 = T_W$. One could also use the statistic Z_W to conduct a one-sided test with null hypothesis $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$ by using the appropriate percentile of the standard normal distribution as the cutoff for the rejection region.

The *score test* (Rao, 1948) is based on the notion that the derivative of the log-likelihood is equal to zero at its maximum. Thus, the magnitude of the gradient of the log-likelihood under the null model gives an indication as to the optimality of the null parameter values. The test statistic is defined as

$$T_S = \big(\mathbf{S}(\boldsymbol{\theta_0})\big)^T \big(\mathrm{I}_N(\boldsymbol{\theta}_0)\big)^{-1} \big(\mathbf{S}(\boldsymbol{\theta_0})\big), \tag{3.12}$$

where $\mathbf{S}(\boldsymbol{\theta}_0)$ is the *score statistic* with i^{th} entry, $i = 1, \ldots, r$, given by

$$\frac{\partial}{\partial \theta_i} \ell_N(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{3.13}$$

and $\mathrm{I}_N(\cdot)$ is defined as in (3.10). Just like T_L and T_W, the statistic T_S asymptotically (as $N \to \infty$) follows a χ^2_r distribution when the data are independent and identically distributed (Rao, 1948). The score test therefore rejects H_0 when $T_S > \chi^2_{r,1-\alpha}$, where $\chi^2_{r,1-\alpha}$ is the $100 \times (1-\alpha)$ percentile of a χ^2_r distribution and α is the level of the test. For the univariate hypotheses $H_0\colon \theta = \theta_0$ and $H_1\colon \theta \neq \theta_0$, the test statistic can be expressed as

$$Z_S = S(\theta_0)\big(\mathrm{I}(\theta_0)\big)^{-1/2}, \tag{3.14}$$

and a standard normal distribution is a suitable reference distribution for large samples. To test a single element θ_i ($i = 1, \ldots, r$) of $\boldsymbol{\theta}$, we use the i^{th} element of $\mathbf{S}(\boldsymbol{\theta}_0)$, given by (3.13), and the square root of the i^{th} diagonal element of $\big(\mathrm{I}_N(\boldsymbol{\theta}_0)\big)^{-1}$ to compute Z_S. In some circles, namely the econometrics literature, the score test is better known as the *Lagrange multiplier test* (see Aitchison and Silvey, 1958; Silvey, 1959).

Although the likelihood ratio test is intuitively appealing because it selects the model for which the data are more plausible, it is also the most computationally intensive of these three tests because it requires one to fit both the null model and the unconstrained model. In contrast, the Wald test only requires one to fit the unconstrained model and the score test only requires one to fit the null model. In the context of testing regression parameters in a GLMM, the Wald test is most popular because it is customary to fit the unconstrained model and no additional estimation based on a constrained model is required to conduct a Wald test. For a formal discussion of likelihood ratio, Wald, and score tests, see Casella and Berger (2002, Section 10.3) and Lehmann and Romano (2005, Section 12.4). For intuition regarding the motivation for these tests, see Buse (1982).

In a GLMM, the parameter vector $\boldsymbol{\theta}$ consists of the fixed effects parameters $\boldsymbol{\beta}$ and the parameters $\boldsymbol{\alpha}$ that characterize the random effects distribution. We shall focus on inference for individual components of $\boldsymbol{\beta}$. In some cases, the procedures described here are also valid for testing elements of $\boldsymbol{\alpha}$, but issues arise, for example, when testing whether a variance component is equal to zero because zero lies on the boundary of the parameter space.

Results concerning the consistency and asymptotic normality of maximum likelihood estimators have historically been derived for independent data. In GLMMs, observations sharing the same realization of a random effect are correlated, and the data are not independent. In Section 3.2.1 we show how the consistency and asymptotic normality results derived for independent data extend to the GLMM context. In turn, we argue that the asymptotic distributions for the test statistics stated above for the case of independent data also apply when the model is a GLMM. In Section 3.2.2 we use simulated data to show that Wald tests and score tests of the null hypothesis $H_0 : \beta = 0$ or $H_0 : \beta^* = 0$ yield similar conclusions in large samples under the two parameterizations, but tests for the regression parameters in a conventional GLMM with a nonzero null value fail to hold their nominal level. This stems from the lack of equality between β and β^* when $\beta^* \neq 0$. Additionally, we show through simulation in Section 3.2.2 and through a series of examples in Section 3.2.3 that Wald tests and score tests under the two parameterizations do not always match in the small sample setting. The examples in Section 3.2.3 also emphasize the differences that arise in confidence intervals under the two model parameterizations.

3.2.1 Asymptotics

The validity of the likelihood ratio, Wald, and score tests follows from the consistency and asymptotic normality of the maximum likelihood estimator. Conditions for the consistency of maximum likelihood estimators have been given by several authors for a variety of situations. For example, Wald (1949) assumed that the model is correctly specified and that all observations are independent and identically distributed, and gave conditions under which the maximum likelihood estimator consistently estimates the true parameter value. White (1982) relaxed the assumption that the model is correctly specified, and gave conditions under which the maximum likelihood estimator consistently estimates the parameter value that minimizes the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the true model and the misspecified model. Here, misspecification means that the true underlying distribution of the data is not contained in the family of distributions from which the data are assumed to arise. For instance, if one assumes the data are normally distributed with unknown mean and variance, but the data actually follow a t-distribution, then the model is misspecified.

Neuhaus et al. (1994) extended the results of White (1982) to the case of mixed models for binary matched-pairs data, providing conditions for consistent estimation of the fixed effects parameters in such a model when the random effects distribution is misspecified in a particular fashion. That is, when the family of distributions to which the random effects distribution is assumed to belong does not contain the true random effects distribution. All other parts of the model, namely the link function and the mean structure, are assumed to be correctly specified; only the distribution of the random effects is incorrectly specified. Litière et al. (2007) considered a broader class of GLMMs and argued that the maximum likelihood estimator consistently estimates the fixed effects parameters when they are equal to zero, even when the random effects distribution is misspecified. This result relies on the number of clusters of observations – and thus the number of realizations of the random effects – going to infinity and only holds if the covariate associated with a random effect with misspecified distribution is uncorrelated with the covariate associated with the fixed effect being estimated. Neuhaus et al. (2013) showed that if correlation is present between the covariate of a fixed effect and the covariate of a random effect for which the shape of the distribution is misspecified, then the maximum likelihood estimator of the parameter for that fixed effect is not necessarily consistent.

We show that results concerning the consistency and asymptotic normalilty of the maximum likelihood estimator that were derived for a one-sample problem under the assumption of independent and identically distributed observations can be applied in the context of a GLMM. To do this, we begin with a one-sample problem and add complexity to the model until reaching a GLMM. Lehmann (1999, Chapter 7) considered a set of N random variates $Y_1, \ldots, Y_N$, with density $f_Y(y|\boldsymbol{\theta})$ depending on the parameter vector $\boldsymbol{\theta}$ of length r. We denote the parameter space for $\boldsymbol{\theta}$ by $\boldsymbol{\Theta}$ and assume that $f_Y(y|\boldsymbol{\theta})$ is either continuous in y or discrete with $f_Y(y|\boldsymbol{\theta}) = P(Y_i = y)$ for each $i = 1, \ldots, N$. In this setting, Lehmann (1999, Page 499) provided the following conditions for the existence of a consistent sequence of local maxima of the likelihood function $\mathcal{L}_N(\boldsymbol{\theta}) = \prod_{i=1}^{N} f_Y(y_i|\boldsymbol{\theta})$, presented here in a different order and with slightly different notation:

C1: There exists an open neighborhood of the true parameter value $\boldsymbol{\theta}_0$ that lies completely within the parameter space $\boldsymbol{\Theta}$;

C2: The observations $Y_1, \ldots, Y_N$ are independent and identically distributed;

C3: The parameters $\boldsymbol{\theta}$ are identifiable; that is, if $f_Y(y|\boldsymbol{\theta}_1) = f_Y(y|\boldsymbol{\theta}_2)$ then $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$;

C4: The set $A = \{y : f_Y(y|\boldsymbol{\theta}) > 0\}$ is independent of $\boldsymbol{\theta}$;

C5: For all $y \in A$, the partial derivatives $\frac{\partial}{\partial \theta_k} f_Y(y|\boldsymbol{\theta})$ exist for $k = 1, \ldots, r$;

C6: The partial derivatives of $\int f_Y(y|\boldsymbol{\theta})dy$ exist and can be obtained by differentiating under the integral sign.

Theorem 7.5.2 of Lehmann (1999, Page 501) further states that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, if it exists, is asymptotically normal. In addition to conditions C1-C4, this theorem requires the following three conditions, the first of which implies C5 and C6:

C7: For all $y \in A$, the partial third derivatives $\frac{\partial}{\partial \theta_i \partial \theta_j \partial \theta_k} f_Y(y|\boldsymbol{\theta})$ exist and are continuous for $i, j, k = 1, \ldots, r$, and the corresponding derivatives of $\int f_Y(y|\boldsymbol{\theta})dy$ exist and can be obtained by differentiating under the integral sign;

C8: If $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,r})^T$ denotes the true value of $\boldsymbol{\theta}$, then there exists a number c and a function $B_{ijk}(y)$, both depending on $\boldsymbol{\theta}_0$, such that $|\frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \ell_N(\boldsymbol{\theta})| \leq B_{ijk}(y)$ for all $\boldsymbol{\theta}$ with $\sum_{k=1}^{r} (\theta_k - \theta_{0,k})^2 < c$, where $\mathrm{E}_{\boldsymbol{\theta}_0}[B_{ijk}(Y)] < \infty$ for all $i, j, k = 1, \ldots, r$;

C9: The information matrix $\mathrm{I}_N(\boldsymbol{\theta})$ is positive definite and all of its elements are finite.

If conditions C1-C4 and C7-C9 are met, then

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathrm{N}_r\Big(\mathbf{0}, \big(\mathrm{I}_N(\boldsymbol{\theta}_0)\big)^{-1}\Big) \text{ as } N \to \infty,$$

where the elements of the matrix $\mathrm{I}_N(\boldsymbol{\theta}_0)$ are defined as in (3.10).

It follows directly from the asymptotic normality of the maximum likelihood estimator that the univariate Wald statistic Z_W given in (3.11) has an asymptotic standard normal distribution and that the multivariate Wald statistic T_W given in (3.9) has an asymptotic chi-squared distribution with r degrees of freedom. Conditions C1-C4 and C7-C9 are also sufficient to prove that the likelihood ratio test statistic T_L given by (3.8) and the score test statistic T_S given by (3.12) have an asymptotic chi-squared distribution with r degrees of freedom (Lehmann, 1999, Section 7.7). Further, the univariate score test statistic Z_S given by (3.14) has an asymptotic standard normal distribution. For independent observations $Y_1, \ldots, Y_N$ and a single parameter θ, the argument that Z_S is asymptotically standard normal relies on the fact that $S(\theta_0) = \frac{\partial}{\partial\theta}\ell_N(\theta|Y_1, \ldots, Y_N)|_{\theta=\theta_0} = \sum_{i=1}^{N} \frac{\partial}{\partial\theta}\ell_1(\theta|Y_i)|_{\theta=\theta_0}$, where the $\frac{\partial}{\partial\theta}\ell_1(\theta|Y_i)|_{\theta=\theta_0}$ are independent and identically distributed with mean zero and variance $\mathrm{I}_1(\theta_0)$. Here, $\mathrm{I}_1(\theta_0)$ represents the information contained in a single observation. Thus, by the standard central limit theorem,

$$\frac{1}{\sqrt{N}} S(\theta_0) = \sqrt{N}\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial\theta}\ell_1(\theta|Y_i)|_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathrm{N}\big(0, \mathrm{I}_1(\theta_0)\big) \text{ as } N \to \infty,$$

and $S(\theta_0)$ has an approximate $\mathrm{N}\big(0, \mathrm{I}_N(\theta_0)\big)$ distribution for large N. The variance of this distribution follows from the fact that the information contained in a set of independent observations is the sum of the information contained in the individual components of the set, which implies that $N\mathrm{I}_1(\theta_0) = \mathrm{I}_N(\theta_0)$.

Moving from a one-sample problem to a fixed effects regression setting, suppose each Y_i, for $i = 1, \ldots, N$, is associated with a set of p predictors $\mathbf{X}_i$, which we assume to be independent draws from a distribution with density $f_{\mathbf{X}}$. We treat $f_{\mathbf{X}}$ as known, meaning that this density does not depend on the unknown parameters $\boldsymbol{\theta}$. In this setting, we have N independent and identically distributed draws of $(\mathbf{X}_i, Y_i)$ with density $f_{\mathbf{X},Y}$ given by $f_{\mathbf{X},Y}(\mathbf{x}, y|\boldsymbol{\theta}) = f_{Y|\mathbf{X}}(y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x})$, and we assume that $\mathrm{E}[Y_i|\mathbf{X}_i = \mathbf{x}] = \mathbf{x}^T\boldsymbol{\theta}$. To extend the results concerning the maximum likelihood estimator to this situation, we must replace Y_i with $(\mathbf{X}_i, Y_i)$ and $f_Y(y|\boldsymbol{\theta})$ with $f_{\mathbf{X},Y}(\mathbf{x}, y|\boldsymbol{\theta})$ in conditions C1-C9. Since $\boldsymbol{\theta}$ is only tied to the Y_i and has no bearing on the $\mathbf{X}_i$, the conditions on $f_{\mathbf{X},Y}(\mathbf{x}, y|\boldsymbol{\theta})$ simplify to conditions on $f_{Y|\mathbf{X}}(y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta})$. Thus, to check that the conditions hold for $f_{\mathbf{X},Y}(\mathbf{x}, y|\boldsymbol{\theta})$

we need only verify that they hold for $f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathrm{x},\boldsymbol{\theta})$. For example, for checking C5,

$$\frac{\partial}{\partial\theta_k}f_{\mathbf{X},Y}(\mathrm{x},y|\boldsymbol{\theta})=\frac{\partial}{\partial\theta_k}f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathrm{x},\boldsymbol{\theta})f_{\mathbf{X}}(\mathrm{x})=f_{\mathbf{X}}(\mathrm{x})\frac{\partial}{\partial\theta_k}f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathrm{x},\boldsymbol{\theta}).$$

Hence, if the partial derivatives of $f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathrm{x})$ exist, then the partial derivatives of $f_{\mathbf{X},Y}(\mathrm{x},y|\boldsymbol{\theta})$ also exist. Similarly, for checking C8,

$$\begin{aligned}\frac{\partial}{\partial\theta_k}\ell_N(\boldsymbol{\theta})&=\frac{\partial}{\partial\theta_k}\sum_{i=1}^{N}\log\big(f_{\mathbf{X},Y}(\mathrm{x},y_i|\boldsymbol{\theta})\big)=\sum_{i=1}^{N}\frac{\partial}{\partial\theta_k}\log\big(f_{Y|\mathbf{X}}(y_i|\mathbf{X}=\mathrm{x},\boldsymbol{\theta})f_{\mathbf{X}}(\mathrm{x})\big)\\&=\sum_{i=1}^{N}\frac{\partial}{\partial\theta_k}\Big(\log\big(f_{Y|\mathbf{X}}(y_i|\mathbf{X}=\mathrm{x},\boldsymbol{\theta})\big)+\log\big(f_{\mathbf{X}}(\mathrm{x})\big)\Big)\\&=\sum_{i=1}^{N}\frac{\partial}{\partial\theta_k}\log\big(f_{Y|\mathbf{X}}(y_i|\mathbf{X}=\mathrm{x},\boldsymbol{\theta})\big).\end{aligned}$$

Again, we need only check the condition for $f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathrm{x},\boldsymbol{\theta})$ to ensure that it holds for $f_{\mathbf{X},Y}(\mathrm{x},y|\boldsymbol{\theta})$. Identifiability, required by C3, now relies on having a rich enough set of $\mathbf{X}_i$. For instance, if $\boldsymbol{\theta}=(\beta_0,\beta_1)^T$, $\mathbf{X}_i=(1,X_i)^T$, and $\mathrm{E}[Y_i|X_i=x]=\beta_0+\beta_1 x$, the parameters β_0 and β_1 are not identifiable if X_i has a degenerate distribution and assumes the same value for all $i=1,\ldots,N$. A fixed effects GLM is included in this regression setting. For a GLM, the mean structure is given by $\mathrm{E}[Y_i|\mathbf{X}_i=\mathrm{x}]=h(\mathrm{x}^T\boldsymbol{\theta})$ for inverse link function $h(\cdot)$. A nonlinear link function could make it more difficult to check conditions C1-C9, but the same basic results still apply.

Moving to a GLMM, continue to assume that we have observations $Y_1,\ldots,Y_N$ with corresponding covariates $\mathbf{X}_1,\ldots,\mathbf{X}_N$, but now suppose each observation is also associated with an independent realization U_i of a random effect with variance σ^2. The mean structure is now given by $\mathrm{E}[Y_i|\mathbf{X}_i=\mathrm{x},U_i=u]=h(\mathrm{x}^T\boldsymbol{\beta}+u)$, where $\boldsymbol{\beta}$ denotes the fixed effects parameters. The vector $\boldsymbol{\theta}$ of unknown parameters now consists of $\boldsymbol{\beta}$ and σ^2. For $i=1,\ldots,N$, we observe data $(\mathbf{X}_i,Y_i)$ from density $f_{\mathbf{X},Y|U}$, which we define as $f_{\mathbf{X},Y|U}(\mathrm{x},y|U,\boldsymbol{\theta})=f_{Y|\mathbf{X},U}(y|\mathbf{X}=\mathrm{x},U,\boldsymbol{\theta})f_{\mathbf{X}}(\mathrm{x})$. For maximum likelihood estimation, we maximize the marginal density $f_{\mathbf{X},Y}$, which can be obtained by integrating the conditional density $f_{\mathbf{X},Y|U}$ over the random effects density f_U. Checking conditions C1-C9 proceeds in the same manner as described above for a fixed effects regression problem. Since we only have one observation per realization of the random effect, the variance component σ^2 in this model is not identifiable. We discuss the importance of replication for consistent estimation of a random effects variance in greater detail in Section 3.5. Nonetheless, given a rich enough set of $\mathbf{X}_i$ the fixed effects parameters $\boldsymbol{\beta}$ are identifiable. Provided the conditions C1-C9 are met after replacing $f_Y(y|\boldsymbol{\theta})$ with $f_{Y|\mathbf{X}}(y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta})$, the maximum likelihood estimator for $\boldsymbol{\beta}$ is consistent and asymptotically normal.

For the random effects variance to be identifiable we require replication. That is, there must be realizations of the random effect that are tied to more than one observation. Suppose we have N observations that are clustered into n groups, each with m_i observations, where the m_i are independent draws from a distribution with probability mass function f_M. We denote these N observations by Y_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, and assume that each Y_{ij} corresponds to a p-vector of predictors, denoted $\mathbf{X}_{ij}$. We assume the $\mathbf{X}_{ij}$ are independent draws from a distribution with density $f_{\mathbf{X}}$ and that, given the U_i, m_i, and $\mathbf{X}_{ij}$, the Y_{ij} are conditionally independent with density $f_{\mathbf{Y}|M,\mathbf{X},U}$ and mean $\mathrm{E}[Y_{ij}|m_i, \mathbf{X}_{ij} = \mathbf{x}, U_i = u] = h(\mathbf{x}^T\boldsymbol{\beta} + u)$. Observations Y_{ij} sharing the same realization of the random effect U_i are not independent under this model. Thus, to satisfy condition C2, we must collect the Y_{ij} into independent clusters of correlated observations. We use $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ to denote the vector of m_i observations in cluster i and $\mathbf{X}_i$ to denote the $m_i \times p$ matrix with rows $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i}$. This leaves us with n independent and identically distributed units of the form $(m_i, \mathbf{X}_i, \mathbf{Y}_i)$. The joint marginal density for m_i, $\mathbf{X}_i$, and

$\mathbf{Y}_i$ is denoted $f_{M,\mathbf{X},\mathbf{Y}}$, and can be obtained by integrating $f_{M,\mathbf{X},\mathbf{Y}|U}$ over the distribution of the random effects, where $f_{M,\mathbf{X},\mathbf{Y}|U}$ is the product of $f_{\mathbf{Y}|M,\mathbf{X},U}$, $f_{\mathbf{X}|M}$, and f_M.

To extend the consistency and asymptotic normality results from earlier to this GLMM setting, we restate the conditions C1-C9 below as C1$'$-C9$'$. Some of these conditions match the earlier conditions exactly, while others require careful consideration of what constitutes a single independent unit. The modified conditions are as follows:

C1$'$: There exists an open neighborhood of the true parameter value $\boldsymbol{\theta}_0$ that lies completely within the parameter space $\boldsymbol{\Theta}$;

C2$'$: Units $(m_1, \mathbf{X}_1, \mathbf{Y}_1), \ldots, (m_n, \mathbf{X}_n, \mathbf{Y}_n)$ are independent and identically distributed;

C3$'$: The parameters $\boldsymbol{\theta}$ are identifiable;

C4$'$: The set $A = \{\mathbf{y} : f_{\mathbf{Y}|M,\mathbf{X}}(\mathbf{y}|m, \mathbf{X}, \boldsymbol{\theta}) > 0\}$ is independent of $\boldsymbol{\theta}$;

C5$'$: For all $\mathbf{y} \in A$, the partial derivatives $\frac{\partial}{\partial\theta_k} f_{\mathbf{Y}|M,\mathbf{X}}(\mathbf{y}|m, \mathbf{X}, \boldsymbol{\theta})$ exist for $k = 1, \ldots, r$;

C6$'$: The partial derivatives of $\int f_{\mathbf{Y}|M,\mathbf{X}}(\mathbf{y}|m, \mathbf{X}, \boldsymbol{\theta}) d\mathbf{y}$ exist and can be obtained by differentiating under the integral sign.

C7$'$: For all $\mathbf{y} \in A$, the partial third derivatives $\frac{\partial}{\partial\theta_i\partial\theta_j\partial\theta_k} f_{\mathbf{Y}|M,\mathbf{X}}(\mathbf{y}|m, \mathbf{X}, \boldsymbol{\theta})$ exist and are continuous, and the corresponding derivatives of $\int f_{\mathbf{Y}|M,\mathbf{X}}(\mathbf{y}|m, \mathbf{X}, \boldsymbol{\theta}) d\mathbf{y}$ exist and can be obtained by differentiating under the integral sign (for $i, j, k = 1, \ldots, r$);

C8$'$: If $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,r})^T$ denotes the true value of $\boldsymbol{\theta}$, then there exists a number c and a function $B_{ijk}(\mathbf{y})$, both depending on $\boldsymbol{\theta}_0$, such that $|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \ell_n(\boldsymbol{\theta})| \leq B_{ijk}(\mathbf{y})$ for all $\boldsymbol{\theta}$ with $\sum_{k=1}^{r} (\theta_k - \theta_{0,k})^2 < c$, where $\mathrm{E}_{\boldsymbol{\theta}_0}[B_{ijk}(\mathbf{Y})] < \infty$ for all $i, j, k = 1, \ldots, r$;

C9$'$: The information matrix $\mathrm{I}_n(\boldsymbol{\theta})$ is positive definite and all of its elements are finite.

The log-likelihood in C8$'$ and the information matrix in C9$'$ are based on n clusters of observations $(m_i, \mathbf{X}_i, \mathbf{Y}_i)$ as opposed to N individual observations Y_i as in C8 and C9. In these conditions, we use the conditional density $f_{\mathbf{Y}|M,\mathbf{X}}$ instead of the joint density $f_{M,\mathbf{X},\mathbf{Y}}$ because the parameters $\boldsymbol{\theta}$ are only tied to the $\mathbf{Y}_i$ for $i = 1, \ldots, n$. Thus, just as we conditioned on $\mathbf{X}_i$ in the fixed effects regression setting, we are able to condition on both $\mathbf{X}_i$ and m_i in this setting.

Analogous to the one-sample problem with independent observations, conditions C1$'$-C6$'$ are sufficient for consistency of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of the parameters $\boldsymbol{\theta}$ in a GLMM. Further, conditions C1$'$-C4$'$ along with C7$'$-C9$'$ are sufficient for asymptotic normality of $\hat{\boldsymbol{\theta}}$, for the test statistics T_L, T_W, and T_S given by (3.8), (3.9), and (3.12) to have an asymptotic chi-squared distribution with r degrees of freedom, and for the univariate test statistics Z_W and Z_S given by (3.11) and (3.14) to have an asymptotic standard normal distribution. The same arguments used by Lehmann (1999) apply in this context, except we now have n independent and identically distributed clusters of observations instead of N independent and identically distributed individual observations.

To see how one of these arguments extends from the one-sample problem to the GLMM, suppose we have a GLMM with n independent and identically distributed clusters of $m > 1$ observations. This is a special case of our more general formulation of the GLMM in which all n clusters contain the same number of observations and the predictors $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im}$ are identical for all $i = 1, \ldots, n$. If interest lies in a single parameter θ, then the same argument used earlier for the asymptotic distribution of the unvariate score test statistic Z_S shows that

$$\frac{1}{\sqrt{n}} S(\theta_0) = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ell_1(\theta|\mathbf{Y}_i)|_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathrm{N}\big(0, \mathrm{I}_1(\theta_0)\big) \text{ as } n \to \infty.$$

The subscript "1" in $\ell_1(\theta|\mathbf{Y}_i)$ and $\mathrm{I}_1(\theta_0)$ now indicates that the log-likelihood and information are based on a single cluster of observations instead of a single observation, and it is the number of clusters n that goes to infinity. It follows that $S(\theta_0)$ has an approximate $\mathrm{N}\big(0, \mathrm{I}_n(\theta_0)\big)$ distribution because $n\mathrm{I}_1(\theta_0) = \mathrm{I}_n(\theta_0)$. Typically, a univariate θ represents the i^{th} element of an r-vector $\boldsymbol{\theta}$. In such a case, $S(\theta_0)$ is the i^{th} entry of $\mathbf{S}(\boldsymbol{\theta}_0)$ and $\mathrm{I}_1(\theta_0)$ is the reciprocal of the i^{th} diagonal element of the inverse information matrix $\big(\mathrm{I}_1(\boldsymbol{\theta}_0)\big)^{-1}$.

In practice, $\mathrm{I}_n(\boldsymbol{\theta})$, which is the large-sample variance of the score statistic $\mathbf{S}(\boldsymbol{\theta})$, is estimated using the Hessian of the negative log-likelihood $-\ell_n(\boldsymbol{\theta})$. As the number of clusters n increases, the mean of the sampling distribution of this estimator, which we denote $\hat{\mathrm{I}}_n(\boldsymbol{\theta})$, should approach n times the expected information $\mathrm{I}_1(\boldsymbol{\theta})$ contained in a single cluster. As further evidence that the asymptotic results of Lehmann (1999) apply in the context of GLMMs, we derive the expected Fisher information $\mathrm{I}_1(\boldsymbol{\theta})$ for a specific GLMM and show that, on average, the Hessian of $-\ell_n(\boldsymbol{\theta})$ approaches n times this quantity as the number of clusters n grows large. Consider the case of n identically distributed pairs of observations Y_{ij}, where $i = 1, \ldots, n$ and $j = 1, 2$. We assume $Y_{ij}|U_i \sim \text{Bernoulli}(p_{ij})$ and $U_i \sim \mathrm{N}(0, \sigma^2)$, and express the conditional mean as $p_{ij} = \mathrm{E}[Y_{ij}|U_i = u] = h(\beta_0 + \beta_1 x_{ij} + u)$, with $x_{i1} = 0$ and $x_{i2} = 1$ for all $i = 1, \ldots, n$. Denoting $\eta_0 = \beta_0 + u$ and $\eta_1 = \beta_0 + \beta_1 + u$, the joint marginal density of $\mathbf{Y}$ is given as

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} \int \big(h(\eta_0)\big)^{y_{i1}} \big(1 - h(\eta_0)\big)^{1-y_{i1}} \big(h(\eta_1)\big)^{y_{i2}} \big(1 - h(\eta_1)\big)^{1-y_{i2}} f_U(u) du.$$

The marginal density for a single pair is

$$f_1\big((y_1, y_2)^T\big) = \int \big(h(\eta_0)\big)^{y_1} \big(1 - h(\eta_0)\big)^{1-y_1} \big(h(\eta_1)\big)^{y_2} \big(1 - h(\eta_1)\big)^{1-y_2} f_U(u) du,$$

where we drop the subscript i because the pairs are identically distributed. We express the vector of unknown parameters as $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)^T$ and denote the corresponding log-likelihood as $\ell_1(\boldsymbol{\theta}) = \log\big(f_1(\mathbf{y}|\boldsymbol{\theta})\big)$. A partial derivative of $\ell_1(\boldsymbol{\theta})$ is given by

$$\frac{\partial}{\partial \theta_k} \ell_1(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_k} \log\big(f_1(\mathbf{y}|\boldsymbol{\theta})\big) = \frac{\frac{\partial}{\partial \theta_k} f_1(\mathbf{y}|\boldsymbol{\theta})}{f_1(\mathbf{y}|\boldsymbol{\theta})},$$

and a partial second derivative is given by

$$\begin{aligned}\frac{\partial^2}{\partial\theta_k\partial\theta_j}\ell_1(\boldsymbol{\theta}) &= \frac{\partial}{\partial\theta_k}\frac{\frac{\partial}{\partial\theta_j}f_1(\mathbf{y}|\boldsymbol{\theta})}{f_1(\mathbf{y}|\boldsymbol{\theta})} \\ &=\bigg(f_1(\mathbf{y}|\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_k\partial\theta_j}\big(f_1(\mathbf{y}|\boldsymbol{\theta})\big) - \frac{\partial}{\partial\theta_k}\big(f_1(\mathbf{y}|\boldsymbol{\theta})\big)\frac{\partial}{\partial\theta_j}\big(f_1(\mathbf{y}|\boldsymbol{\theta})\big)\bigg)/\big(f_1(\mathbf{y}|\boldsymbol{\theta})\big)^2,\end{aligned} \tag{3.15}$$

where $k, j = 1, 2, 3$ index the elements of $\boldsymbol{\theta}$.

Noting that $(Y_1, Y_2)^T \in \{(0,0)^T, (0,1)^T, (1,0)^T, (1,1)^T\}$, the four possible expressions for the marginal density $f_1(\mathbf{y}|\boldsymbol{\theta})$ are

$$\begin{aligned} f_1\big((0,0)^T|\boldsymbol{\theta}\big) &= \int \big(1-h(\eta_0)\big)\big(1-h(\eta_1)\big)f_U(u)du; \\ f_1\big((0,1)^T|\boldsymbol{\theta}\big) &= \int \big(1-h(\eta_0)\big)h(\eta_1)f_U(u)du; \\ f_1\big((1,0)^T|\boldsymbol{\theta}\big) &= \int h(\eta_0)\big(1-h(\eta_1)\big)f_U(u)du; \\ f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int h(\eta_0)h(\eta_1)f_U(u)du. \end{aligned}$$

We shall show the form of the relevant first and second partial derivatives when $\mathbf{y} = (1,1)^T$. The form of these derivatives is similar for the other three possible values of $(y_1, y_2)^T$. For all $(y_1, y_2)^T$, we are able to differentiate under the integral sign because the integrand of $f_1(\mathbf{y}|\boldsymbol{\theta})$ is continuously differentiable with respect to β_0, β_1, and σ^2 for all $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 > 0$. Further, the integrand of $f_1(\mathbf{y}|\boldsymbol{\theta})$ converges to zero as $u \to \infty$ and as $u \to -\infty$. The partial first derivatives of $f_1(\mathbf{y}|\boldsymbol{\theta})$ are

$$\begin{aligned} \frac{\partial}{\partial\beta_0}f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int \big(h'(\eta_0)h(\eta_1) + h(\eta_0)h'(\eta_1)\big)f_U(u)du; \\ \frac{\partial}{\partial\beta_1}f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int h(\eta_0)h'(\eta_1)f_U(u)du; \\ \frac{\partial}{\partial\sigma^2}f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int h(\eta_0)h(\eta_1)\bigg(\frac{u^2}{2\sigma^4} - \frac{1}{2\sigma^2}\bigg)f_U(u)du, \end{aligned}$$

and the partial second derivatives are

$$\begin{aligned}
\frac{\partial^2}{(\partial\beta_0)^2} f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int \big(h''(\eta_0)h(\eta_1) + 2h'(\eta_0)h'(\eta_1) + h(\eta_0)h''(\eta_1)\big) f_U(u)du;\\
\frac{\partial^2}{\partial\beta_0\partial\beta_1} f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int \big(h'(\eta_0)h'(\eta_1) + h(\eta_0)h''(\eta_1)\big) f_U(u)du;\\
\frac{\partial^2}{(\partial\beta_1)^2} f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int h(\eta_0)h''(\eta_1) f_U(u)du;\\
\frac{\partial^2}{\partial\beta_0\partial\sigma^2} f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int \big(h'(\eta_0)h(\eta_1) + h(\eta_0)h'(\eta_1)\big)\left(\frac{u^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right) f_U(u)du;\\
\frac{\partial^2}{\partial\beta_1\partial\sigma^2} f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int h(\eta_0)h'(\eta_1)\left(\frac{u^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right) f_U(u)du;\\
\frac{\partial^2}{(\partial\sigma^2)^2} f_1\big((1,1)^T|\boldsymbol{\theta}\big) &= \int h(\eta_0)h(\eta_1)\left(\frac{u^2}{2\sigma^4} - \frac{1}{2\sigma^2} + \frac{\sigma^2 - 2u^2}{\sigma^2(u^2 - \sigma^2)}\right) f_U(u)du,
\end{aligned}$$

where $h'(\cdot)$ and $h''(\cdot)$ are the first two derivatives of $h(\cdot)$:

$$h'(\eta) = \frac{\exp(\eta)}{\big(1+\exp(\eta)\big)^2} \text{ and } h''(\eta) = \frac{\exp(\eta)\big(1-\exp(\eta)\big)}{\big(1+\exp(\eta)\big)^3}.$$

Inserting these expressions into (3.15) yields the second derivatives of $\ell_1(\boldsymbol{\theta})$ that comprise the information matrix $\mathrm{I}_1(\boldsymbol{\theta})$.

In Section 3.2.2 we describe a simulation study in which we generate data from a marginally interpretable model analogous to the conventional GLMM under consideration here, with $\beta_0^* = \mathrm{logit}(2/3) = 0.69$ and $\sigma^2 = 5$. This corresponds to $\beta_0 = 1.22$ in our current parameterization. Plugging these values for σ^2 and β_0 into the expressions given above, we can compute the expected information for a pair of observations. Table 3.1 shows the relevant second derivatives of $\ell_1(\boldsymbol{\theta})$ under the null hypothesis $H_0\colon \beta_1 = 0$. From this table, we see that the expected information matrix for a single pair of observations is

$$\mathrm{I}_1(\boldsymbol{\theta}) = \begin{bmatrix} 0.104 & 0.052 & -0.008 \\ 0.052 & 0.090 & -0.004 \\ -0.008 & -0.004 & 0.002 \end{bmatrix}. \tag{3.16}$$

In our simulation study, we generate 10,000 datasets with each of $n = 50$, $n = 100$, and $n = 300$ pairs of observations. For each dataset we estimate the information matrix

Table 3.1: Partial second derivatives of $\ell_1(\boldsymbol{\theta})$ for a logistic-normal model with pairs of correlated observations, a single binary predictor, $\beta_0 = 1.22$, $\beta_1 = 0$, and $\sigma^2 = 5$

	Response Pattern ($\mathbf{y}$)				Expected
Quantity	$(0,0)^T$	$(0,1)^T$	$(1,0)^T$	$(1,1)^T$	Value
$f_1(\mathbf{y}\|\boldsymbol{\theta})$	0.205	0.128	0.128	0.538	—
$\frac{\partial^2}{(\partial\beta_0)^2}\ell_1(\boldsymbol{\theta})$	-0.115	-0.128	-0.128	-0.089	-0.104
$\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ell_1(\boldsymbol{\theta})$	-0.058	-0.064	-0.064	-0.044	-0.052
$\frac{\partial^2}{(\partial\beta_1)^2}\ell_1(\boldsymbol{\theta})$	-0.103	-0.124	-0.124	-0.069	-0.090
$\frac{\partial^2}{\partial\beta_0\partial\sigma^2}\ell_1(\boldsymbol{\theta})$	0.057	0.021	0.021	-0.018	0.008
$\frac{\partial^2}{\partial\beta_1\partial\sigma^2}\ell_1(\boldsymbol{\theta})$	0.028	0.011	0.011	-0.009	0.004
$\frac{\partial^2}{(\partial\sigma^2)^2}\ell_1(\boldsymbol{\theta})$	-0.021	0.005	0.005	0.002	-0.002

by computing the Hessian of $-\ell_n(\boldsymbol{\theta})$. Across the 10,000 datasets for each sample size, the average estimated information matrix $\hat{\mathrm{I}}_n(\boldsymbol{\theta})$ should be approximately equal to $n\mathrm{I}_1(\boldsymbol{\theta})$, where $\mathrm{I}_1(\boldsymbol{\theta})$ is the expected information given in (3.16). Further, $\hat{\mathrm{I}}_n(\boldsymbol{\theta})$ and $n\mathrm{I}_1(\boldsymbol{\theta})$ should be more similar for larger values of n. We obtain the following estimates for $\frac{1}{n}\hat{\mathrm{I}}_n(\boldsymbol{\theta})$:

$$\frac{1}{50}\hat{\mathrm{I}}_{50}(\boldsymbol{\theta}) = \begin{bmatrix} 0.112 & 0.056 & -0.010 \\ 0.056 & 0.090 & -0.005 \\ -0.010 & -0.005 & 0.016 \end{bmatrix};$$

$$\frac{1}{100}\hat{\mathrm{I}}_{100}(\boldsymbol{\theta}) = \begin{bmatrix} 0.108 & 0.054 & -0.009 \\ 0.054 & 0.090 & -0.004 \\ -0.009 & -0.004 & 0.003 \end{bmatrix};$$

$$\frac{1}{300}\hat{\mathrm{I}}_{300}(\boldsymbol{\theta}) = \begin{bmatrix} 0.106 & 0.053 & -0.008 \\ 0.053 & 0.090 & -0.004 \\ -0.008 & -0.004 & 0.002 \end{bmatrix}.$$

As expected, the Hessian of the negative log-likelihood based on n clusters provides an accurate approximation of n times the information contained in a single pair of observations, with the accuracy improving as the sample size increases.

3.2.2 Simulation Study

To investigate the power of various testing procedures for the fixed effects parameters in both a marginally interpretable and a conventional GLMM, we performed a simulation study. In the results presented below, we simulated paired, binary outcome data with a single binary predictor and a single random intercept. This is a context in which the marginally interpretable and conventional GLMMs are equivalent. Further, it represents a low information situation as there are only two binary responses available to learn about each cluster. We simulated data from the marginally interpretable model with conditional mean

$$\mu_{ij}^* = \mathrm{E}[Y_{ij}|U_i = u] = h(\beta_0^* + \beta_1^* x_{ij} + u + a_{ij}),$$

where $i = 1, \ldots, n$, $j = 1, 2$, $h(\cdot)$ is the inverse logit function, $x_{i1} = 0$ for all $i = 1, \ldots, n$, and $x_{i2} = 1$ for all $i = 1, \ldots, n$. We drew each U_i from a $\mathrm{N}(0, \sigma^2)$ and each Y_{ij} given U_i from a $\mathrm{Bernoulli}(\mu_{ij}^*)$. We set $\beta_0^* = \mathrm{logit}(2/3) = 0.69$ and $\sigma^2 = 5$ for all datasets, and let $n \in \{50, 100, 300\}$ and $\beta_1^* \in \{-0.5, -0.4, \ldots, 0.5\}$. For each combination of n and β_1^* we generated 10,000 datasets, and then fit both a marginally interpretable GLMM and a conventional GLMM to each dataset. For conducting likelihood ratio and score tests with null hypothesis $H_0 : \beta_1^* = 0$ (or $H_0 : \beta_1 = 0$), we also fit both GLMMs excluding the treatment effect, meaning that β_0^* (or β_0) was the only fixed effect included in the model.

The models were fit via maximum likelihood estimation using techniques that will be described in Section 4.3. For a few datasets the algorithm used to fit the model failed to converge. When this occurred the dataset was discarded and a new one was generated to replace it. This occurred more often in the smaller datasets, but for the results reported here no more than three datasets were discarded for any simulation setting. Discarding three datasets out of 10,000 should not have introduced any substantial bias to the results.

For each simulated dataset, each model parameterization, and each testing procedure we conducted a hypothesis test with nominal level 0.05. For each simulation setting, we then empirically computed the power of the test by calculating the proportion of the 10,000 datasets for which the test rejected the null hypothesis. An upper bound on the standard error for this empirical power calculation, corresponding to an estimated power of 0.5, is 0.0071. Results are displayed as power curves as a function of the true parameter value β_1^*.

Our first set of results corresponds to tests of the null hypothesis $H_0 : \beta_1^* = 0$ versus the alternative hypothesis $H_1 : \beta_1^* \neq 0$. The left column of Figure 3.1 shows the empirical power for Wald tests based on the marginal and cluster-specific parameterizations of the model. For relatively small samples ($n = 50$ pairs of outcomes), the test based on the marginal parameterization is clearly more powerful, and the test based on the conventional parameterization fails to achieve its nominal level when the null hypothesis is true. As the sample size increases, the power curves for the two Wald tests become more similar, suggesting that the tests are equivalent in the large sample setting. Also included in each panel of Figure 3.1 is the corresponding power curve for a likelihood ratio test. Only one curve is drawn for each sample size because both the null and unconstrained model are equivalent under the two parameterizations in this situation and the likelihood ratio test is therefore the same for the two cases. Here, the likelihood ratio test corresponds closely to the Wald test based on the marginal parameterization.

The close correspondence between the power curves for the Wald tests under the two parameterizations for large samples is consistent with earlier findings reported by Neuhaus (1993) and Litière et al. (2007). Neuhaus (1993) studied regression parameters in conventional GLMMs and marginal models fit via GEE for clustered binary data, and showed that, asymptotically, the standard errors of the marginal parameters are attenuated by the

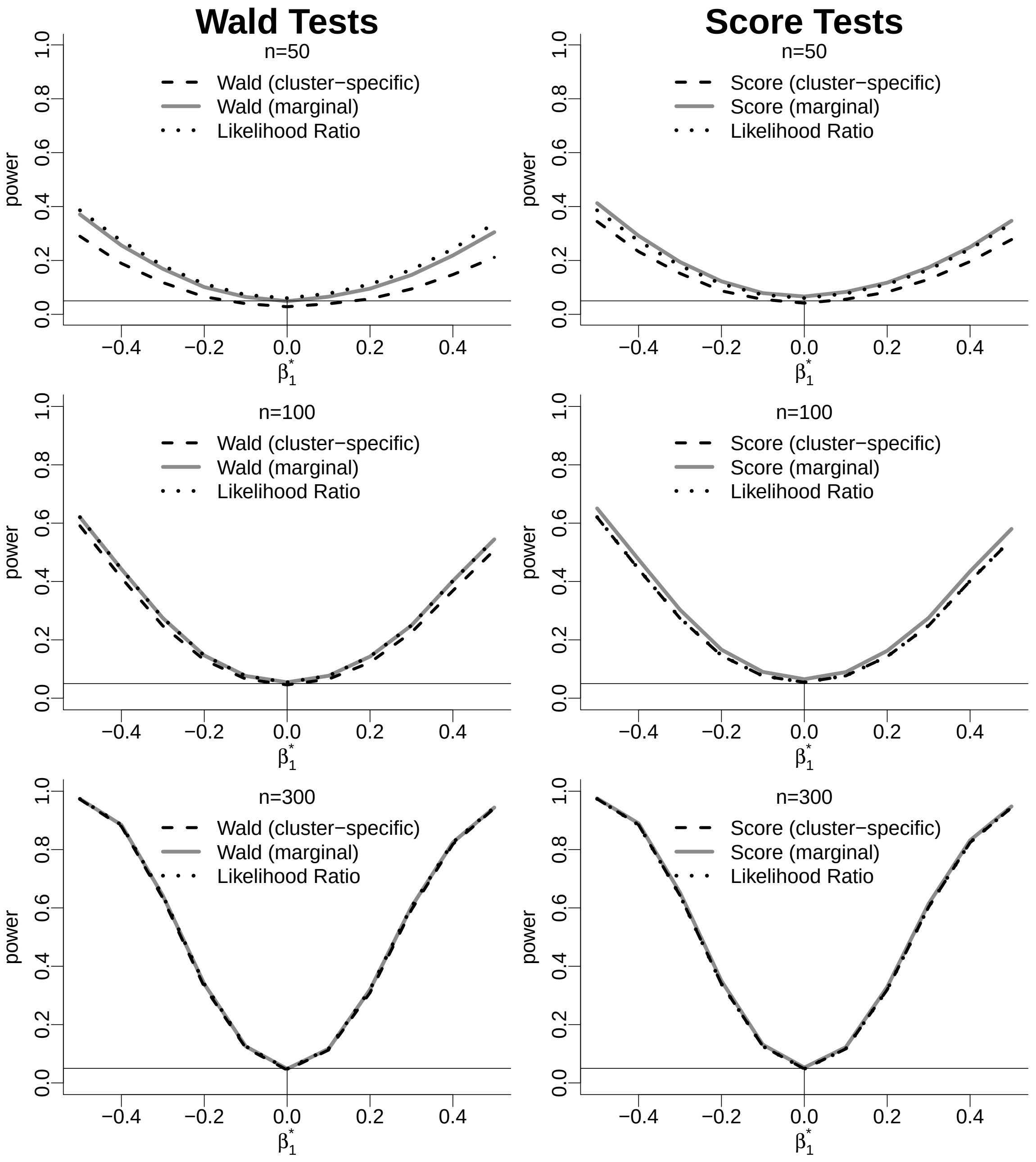

Figure 3.1: Empirical power curves for tests of the null hypothesis $H_0 : \beta_1^* = 0$; from top to bottom, the rows correspond to data with $n = 50$, $n = 100$, and $n = 300$ pairs of binary responses; Wald tests are shown in the left column, score tests are shown in the right column, and likelihood ratio tests are shown in both

same amount as the marginal parameters themselves when the true parameter value is zero. Thus, a Wald test of the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative $H_1 : \beta_1 \neq 0$ should yield the same result in large samples under both a marginal and a cluster-specific parameterization. Litière et al. (2007) studied the performance of Wald tests for regression parameters in GLMMs with misspecified random effects distributions, and proved that a Wald test of the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative $H_1 : \beta_1 \neq 0$ achieves its nominal level in large samples, even if the random effects distribution is misspecified in the sense that the family of distributions to which $f_{\mathbf{U}}$ is assumed to belong does not contain the true random effects distribution.

The right column of Figure 3.1 shows the empirical power for score tests conducted on the same datasets. The same likelhood ratio test power curves shown in the left column are included for reference. The discrepancy between the score tests based on the two different model parameterizations follows a pattern very similar to the discrepancy between the Wald tests. Namely, the test based on the marginal parameterization is noticeably more powerful for smaller sample sizes, but as the sample size increases the two tests become more similar. Also, the score tests are more powerful than the corresponding Wald tests when the number of pairs is relatively small.

The next set of results corresponds to tests of the null hypothesis $H_0 \colon \beta_1^* \leq 0.2$ versus the alternative hypothesis $H_1 : \beta_1^* > 0.2$. Figure 3.2 shows the empirical power for Wald tests based on the marginal and cluster-specific parameterizations of the model. In contrast to the tests with null hypothesis $H_0 \colon \beta_1^* = 0$ shown in Figure 3.1, the discrepancy between the empirical power curves based on the marginally interpretable model and the conventional model increases with sample size. Generally speaking, the test based on the marginal

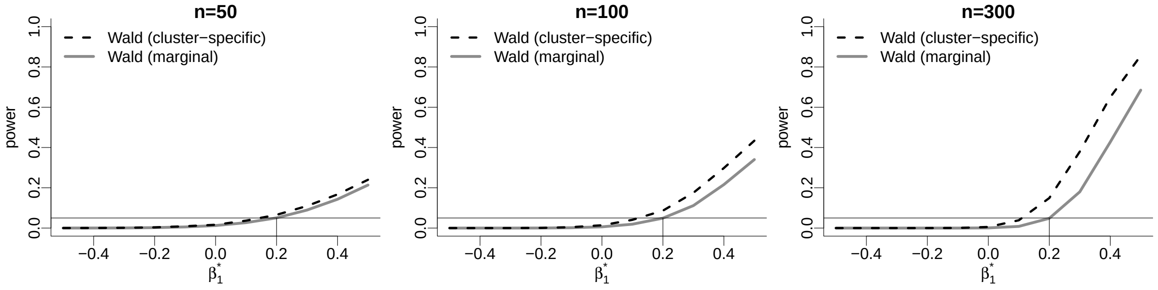

Figure 3.2: Empirical power curves for Wald tests of the null hypothesis $H_0 : \beta_1^* \leq 0.2$ corresponding to data with $n = 50$, $n = 100$, and $n = 300$ pairs of binary responses

parameterization holds its level near 0.05 whereas the test based on the cluster-specific parameterization is overpowered under the null hypothesis. The extent to which this test is overpowered increases with the sample size. This is due to the fact that $\beta_1 \neq \beta_1^*$. When $\beta_1^* = 0.2$, we have $\beta_1 = \beta_1^* + a$, where the adjustment a is greater than zero. Thus, the true value of β_1 is greater than 0.2 when $\beta_1^* = 0.2$. The test based on the cluster-specific parameter β_1 ultimately tests the wrong set of hypotheses, because $H_0 : \beta_1^* \leq 0.2$ and $H_0 \colon \beta_1 \leq 0.2$ are not the same. In this instance, if interest lies in the marginal effect, using a conventional GLMM leads to an inflated rate of Type I error (i.e., a greater likelihood of rejecting the null hypothesis when it is in fact true) because the cluster-specific parameter is pulled in the direction of the alternative hypothesis by the nonlinear inverse link function.

When the adjustment has the opposite sign and the cluster-specific parameter is pulled toward the null hypothesis, the test based on the conventional model tends to be underpowered. Consider, for example, testing the null hypothesis $H_0 : \beta_1^* \leq -0.2$ versus the alternative hypothesis $H_1 : \beta_1^* > -0.2$. We still have $\beta_1 = \beta_1^* + a$, but now the adjustment a is less than zero, which leads to fewer rejections of the null hypothesis when testing $H_0 \colon \beta_1^* \leq -0.2$ using the the cluster-specific parameter β_1. This is illustrated in Figure 3.3,

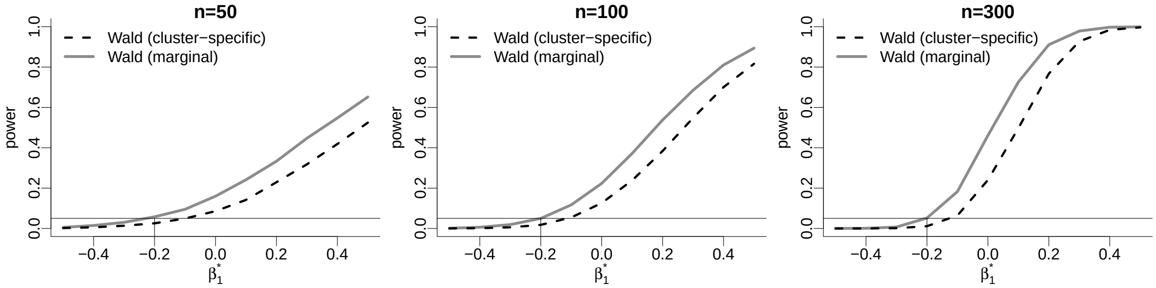

Figure 3.3: Empirical power curves for Wald tests of the null hypothesis $H_0 \colon \beta_1^* \leq -0.2$ corresponding to data with $n = 50$, $n = 100$, and $n = 300$ pairs of binary responses

which shows the empirical power for Wald tests of these hypotheses based on both model parameterizations. Once again, $H_0 \colon \beta_1 \leq -0.2$ is the wrong null hypothesis if interest lies in the marginal effect β_1^*.

3.2.3 Empirical Data Examples

The examples that follow demonstrate situations wherein a marginally interpretable model leads to conclusions that differ substantially from those of a conventional model and situations wherein the differences are relatively minor. Two key drivers of differences between the two model parameterizations are the heterogeneity in the population of interest and the number of clusters in the sample. In the first example, a teratological experiment on rats, the difference between the two sets of parameters is driven by the presence of a separate random effects variance for each of the two treatment groups. In the second example, a two-way crossover study of patients with cerebrovascular deficiency, the large discrepancy between the two types of parameters stems from a high degree of subject-to-subject variability. In the third example, which deals with migration patterns of the Common Cuckoo, we show that a discrepancy, albeit a more modest one, can arise even with a relatively small random effects variance when the number of clusters is small. The fourth example, a study of seed germination, demonstrates that the two model parameterizations can yield similar results when the random effects variance is small, and the fifth example, an investigation of roundworm in pigs, shows what happens when the number of clusters is relatively large. We also use these examples to emphasize that marginal inferences are typically of greater interest than cluster-specific or subject-specific inferences.

In all of the examples in this section, we assume a normally distributed random intercept and a logit link. All of the relevant data can be found in Appendix A. We fit the models via maximum likelihood estimation, using Gauss-Hermite quadrature to approximate the intractable integral in the marginal likelihood function. Computational details are provided in Chapter 4. The likelihood ratio test statistic reported is T_L as defined in (3.8) while the Wald and score statistics reported are the univariate versions Z_W and Z_S given by (3.11) and (3.14), respectively.

Rat Teratology

Weil (1970) fed one group of 16 pregnant rats, known as *dams*, a diet containing a chemical agent during pregnancy and lactation, while feeding a second group of 16 dams a control diet. We denote the number of pups in each of the 32 litters to survive four days from birth by m_i and the number of pups in each litter to survive the 21-day lactation period by Y_i, where $i = 1, \ldots, 32$. Interest lies in the proportion $p_i = Y_i/m_i$ of pups to survive 21 days among those alive after four days. Specifically, we want to determine if the chemical agent is associated with a lower (or higher) survival rate among the rat pups. The relevant data are summarized in Table A.3.

One approach to investigating the relationship between the survival rate and the chemical agent is to fit a logistic regression model that includes a fixed effect for treatment and a random effect for litter. Such a model assumes that $Y_i|U_i \sim \mathrm{Binomial}(m_i, \mathrm{E}[p_i|U_i])$. A conventional model parameterizes $\mathrm{E}[p_i|U_i = u]$ as

$$\mathrm{E}[p_i|U_i = u] = h(\beta_0 + \beta_1 x_i + u), \tag{3.17}$$

where $x_i = 1$ for litters exposed to the chemical agent, $x_i = 0$ for litters whose dams received the control diet, $U_i \sim \mathrm{N}(0, \sigma^2)$ is a random litter effect, and β_0 and β_1 are fixed effects. Although standard, using this conventional approach is misguided. The parameter β_1 in this model is cluster-specific; it depends on the random litter effects and provides information about the impact of the chemical agent on the survival rate *for a specific litter*. Ordinarily, interest lies in learning about the entire population, and this situation is no exception. Of greater interest in this study is the average impact of the chemical agent on the survival rate *across all litters in the population*. This effect can be studied using a marginally interpretable model that parameterizes $\mathrm{E}[p_i|U_i = u]$ as

$$\mathrm{E}[p_i|U_i = u] = h(\beta_0^* + \beta_1^* x_i + u + a_i). \tag{3.18}$$

For both of these models, we estimate the random effects variance σ^2 to be 1.81. Fixed effects parameter estimates are given in Table 3.2 along with 95% confidence intervals. The point estimates for the cluster-specific model are larger in magnitude and the corresponding confidence intervals are wider than for the marginally interpretable model. To determine if exposure to the chemical agent has a significant impact on rat pup survival, a relevant test would focus on the marginal parameter and test the null hypothesis $H_0 : \beta_1^* = 0$ versus the alternative hypothesis $H_1 : \beta_1^* \neq 0$. A more common, but less appropriate, test would focus on the cluster-specific parameter and use the hypotheses $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$. Table 3.2 also displays the results of a Wald test, likelihood ratio test, and score test for these hypotheses. Despite the differences in the point estimates obtained from the two models, all three tests yield similar results under each of the two parameterizations.

Table 3.2: Parameter estimates and corresponding standard errors, tests, and 95% confidence intervals for the fixed effects in the rat teratology study of Weil (1970) assuming a common random effects variance across the two treatment groups

Marginally Interpretable Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0^*)	2.03	0.39	Wald	5.16	<0.01	(1.26,2.80)
treatment (β_1^*)	-0.87	0.51	Wald	-1.72	0.09	(-1.86,0.12)
			LR	2.88	0.09	(-1.88,0.14)
			Score	-1.75	0.08	—

Cluster-Specific Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0)	2.63	0.48	Wald	5.44	<0.01	(1.68,3.57)
treatment (β_1)	-1.08	0.63	Wald	-1.73	0.08	(-2.31,0.15)
			LR	2.88	0.09	(-2.42,0.18)
			Score	-1.91	0.06	—

Earlier analyses of these data (see Liang and Hanfelt, 1994; Heagerty and Zeger, 2000; Wang and Louis, 2004) established that there is more between-litter heterogeneity in the treatment group than in the control group. We therefore refit the models given by (3.17) and (3.18) with a separate variance parameter for each of the two treatment groups to allow the random effects variance to depend on x_i. That is, we assume $U_i \sim \mathrm{N}(0, \sigma_{x_i}^2)$. For both models, we estimate the variance components as $\hat{\sigma}_0^2 = 0.20$ and $\hat{\sigma}_1^2 = 3.34$. The fixed effects parameter estimates are given in Table 3.3 along with corresponding hypothesis tests and 95% confidence intervals. It is evident that the choice between a marginal and cluster-specific parameterization of the model has a substantial impact on the results. Specifically, using the conventional approach, one would likely conclude that the treatment does not have a significant effect on rat pup survival. In contrast, using the marginally interpretable model, we find that the overall survival rate is significantly lower in litters whose dams were

Table 3.3: Parameter estimates and corresponding standard errors, tests, and 95% confidence intervals for the fixed effects in the rat teratology study of Weil (1970) assuming a different random effects variance for each of the two treatment groups

Marginally Interpretable Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0^*)	2.18	0.29	Wald	7.59	<0.01	(1.61,2.74)
treatment (β_1^*)	-1.09	0.47	Wald	-2.31	0.02	(-2.01,-0.17)
			LR	5.13	0.02	(-2.03,-0.15)
			Score	-2.59	0.01	—

Cluster-Specific Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0)	2.25	0.34	Wald	6.63	<0.01	(1.59,2.92)
treatment (β_1)	-0.58	0.65	Wald	-0.88	0.38	(–1.85,0.70)
			LR	0.72	0.39	(-1.90,0.89)
			Score	-0.91	0.36	—

exposed to the chemical agent than in litters whose dams were fed the control diet. Since the effect of the chemical agent on the overall survival rate is of key interest here, the correct statistical inference to draw is the population-based, marginal inference involving β_1^*. The magnitude of the difference between the two parameterizations in this case is due in part to the two treatment groups having different variances and, in turn, different adjustments in the marginally interpretable model. The null model in this situation contains one group but two variances and is therefore not equivalent under the two model parameterizations, as described in the first example after Corollary 3.1.2 in Section 3.1.1. Consequently, the likelihood ratio test differs between the two models in this case.

The Wald p-values and Wald confidence intervals reported in Table 3.3 are based on a standard normal reference distribution, which is the asymptotic distribution of the univariate Wald statistic used in this example. There is not, however, universal agreement that

Table 3.4: Wald tests and 95% confidence intervals for the treatment effect in the rat teratology study of Weil (1970) assuming a different random effects variance for each of the two treatment groups; results are reported for both a $\mathrm{N}(0,1)$ and a t_{30} reference distribution

	$\mathrm{N}(0,1)$		t_{30}	
Parameter	p-value	95% CI	p-value	95% CI
Marginal (β_1^*)	0.02	(-2.01,-0.17)	0.03	(-2.04,-0.13)
Cluster-Specific (β_1)	0.38	(-1.85,0.70)	0.39	(-1.91,0.76)

the standard normal is the appropriate distribution to use in the small sample setting (see Bolker et al., 2009). Table 3.4 displays analogous p-values and confidence intervals based on a t-distribution with 30 degrees of freedom. This alternative reference distribution is the default behavior of the `NLMIXED` procedure of SAS 9.4 (SAS Institute, Cary, NC) in this situation. Although the results are slightly different for the two distributions, the choice of reference distribution is of relatively minor importance in comparison to the choice of model parameterization (i.e., whether or not the model is marginally interpretable).

Two-Way Crossover Study

Jones and Kenward (1989) reported data from a two-way crossover study that included patients suffering from cerebrovascular deficiency and involved two treatment periods. In the first period, each patient was randomly assigned to receive either a placebo or an active drug. In the second period, patients who received the placebo in the first period were given the active drug and vice versa. At the end of each period, a cardiologist examined an electrocardiogram for each patient and determined it to be either normal ($Y = 1$) or abnormal ($Y = 0$). We focus on data, shown in Table A.4, for 67 subjects who were all treated at the same medical center. Of these subjects, 34 received the drug in the first period and the placebo in the second whereas the other 33 received the treatments in the opposite

Table 3.5: Parameter estimates for the fixed effects in the two-way crossover study of Jones and Kenward (1989), with corresponding standard errors, Wald tests, and 95% Wald confidence intervals

Marginally Interpretable Model

	Point	Standard		Wald test	
Parameter	Estimate	Error	Ratio	p-value	95% CI
intercept (β_0^*)	0.43	0.35	1.22	0.22	(-0.26,1.12)
treatment (β_1^*)	1.15	0.58	2.00	0.05	(0.02,2.28)
period (β_2^*)	0.17	0.51	0.34	0.74	(-0.82,1.16)
interaction (β_3^*)	-1.07	0.98	-1.10	0.27	(-2.99,0.85)

Subject-Specific Model

	Point	Standard		Wald test	
Parameter	Estimate	Error	Ratio	p-value	95% CI
intercept (β_0)	1.40	1.24	1.13	0.26	(-1.03,3.84)
treatment (β_1)	3.58	2.11	1.70	0.09	(-0.55,7.71)
period (β_2)	0.55	1.64	0.33	0.74	(-2.67,3.77)
interaction (β_3)	-3.32	3.27	-1.02	0.31	(-9.73,3.09)

order. We let $i = 1, \ldots, 67$ index the patients and $j = 1, 2$ index the periods. Predictors included in the model are treatment ($x_1 = 0$ for placebo, $x_1 = 1$ for drug), period ($x_2 = 0$ for period 1, $x_2 = 1$ for period 2), and the interaction between treatment and period. Since we have two observations for each subject, a random intercept for subject is included as well. Our marginally interpretable model has conditional mean

$$\mu_{ij}^* = \mathrm{E}[Y_{ij}|U_i = u] = h(\beta_0^* + \beta_1^* x_{1,ij} + \beta_2^* x_{2,ij} + \beta_3^* x_{1,ij} x_{2,ij} + u + a_{ij}),$$

where $h(\cdot)$ is the inverse logit function, $Y_{ij}|U_i \sim \mathrm{Bernoulli}(\mu_{ij}^*)$, and $U_i \sim \mathrm{N}(0, \sigma^2)$. The conventional model is analogous, but excludes the adjustment that makes the model marginally interpretable. Under both parameterizations the estimate of the random effects variance is $\hat{\sigma}^2 = 24.15$. The fixed effects parameter estimates, with corresponding standard errors, Wald tests, and 95% confidence intervals, are given in Table 3.5.

Table 3.6: Response groups in the crossover study of Jones and Kenward (1989), with corresponding parameterization of the logit of the marginal probability that $Y = 1$

Group	Treatment	Period	Logit of Marginal Probability
1	Placebo	first	β_0^*
2	Placebo	second	$\beta_0^* + \beta_2^*$
3	Active Drug	first	$\beta_0^* + \beta_1^*$
4	Active Drug	second	$\beta_0^* + \beta_1^* + \beta_2^* + \beta_3^*$

Note that there are four unique combinations of treatment and period, each with its own corresponding probability estimate. These are summarized in Table 3.6. Of primary interest in this study is whether the treatment has a significant effect on the probability of a normal electrocardiogram. Due to the interaction between treatment and period, we must average over the two periods to answer this question. In the first period, the treatment effect is represented by $(\beta_0^* + \beta_1^*) - (\beta_0^*) = \beta_1^*$ whereas in the second period it is represented by $(\beta_0^* + \beta_1^* + \beta_2^* + \beta_3^*) - (\beta_0^* + \beta_2^*) = \beta_1^* + \beta_3^*$. The mean of these two quantities is $\beta_1^* + \beta_3^*/2$. Consequently, we test the following hypotheses:

$$H_0 : \beta_1^* + \frac{1}{2}\beta_3^* = 0 \text{ vs. } H_1 : \beta_1^* + \frac{1}{2}\beta_3^* \neq 0.$$

These hypotheses consider whether the treatment has a significant effect on average across all subjects and periods. The analogous hypotheses from the conventional model,

$$H_0 : \beta_1 + \frac{1}{2}\beta_3 = 0 \text{ vs. } H_1 : \beta_1 + \frac{1}{2}\beta_3 \neq 0,$$

consider whether the treatment has a significant effect for a particular subject after controlling for period. Hypothesis tests and 95% confidence intervals corresponding to the quantities of interest are given in Table 3.7. Although one would likely conclude that the treatment effect is significant regardless of the model used, there are sizable differences

Table 3.7: Tests and 95% confidence intervals for the presence of a significant treatment effect after controlling for period in the crossover study of Jones and Kenward (1989)

Parameterization	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
marginal	0.62	0.24	Wald	2.54	0.01	(0.14,1.09)
			LR	6.78	0.01	(0.16,1.15)
			Score	3.18	<0.01	—
subject-specific	1.92	0.95	Wald	2.02	0.04	(0.06,3.78)
			LR	6.69	0.01	(0.40,4.79)
			Score	2.43	0.02	—

between the two models in terms of the estimated effect size, the length of the corresponding 95% confidence interval, and the strength of the evidence against the null hypothesis as summarized by the p-value. The magnitude of the difference in this scenario is driven by the large random effects variance ($\hat{\sigma}^2 = 24.15$).

To assess the efficacy of the drug across the entire target population, one should use the marginal parameterization. In this setting, using the conventional, subject-specific parameters leads to overstating both the effect size and its uncertainty when interest lies in the population-averaged, marginal effect. If the experiment were repeated on a different sample of subjects from the same population, the subject-specific parameter estimates would likely show more deviation from their counterparts in the original study than would the marginal parameter estimates. The subject-specific effects are, as the name suggests, tied to specific subjects; the marginal effects are better suited for population-based inference.

Bird Migration

Hewson et al. (2016) studied migration patterns of the Common Cuckoo, tracking 56 autumn migrations among 42 birds over a four-year period. These birds migrated each fall from the United Kingdom to Africa, taking either an eastern route or a western route. Both routes required the birds to cross the Sahara desert. Using trackers, Hewson et al. (2016) determined whether each bird survived the Sahara crossing. To avoid repeated measurements on the same birds, we consider only the first observed migration for each of the 42 Cuckoos in the sample. The observed migrations include 22 along the eastern route and 20 along the western route. The data are provided in Table A.5.

For studying the relationship between migration patterns and population decline, the scientific question of interest is whether the choice of route has a significant impact on the survival rate on average across all years. To answer this question, we fit a marginally interpretable model in which the probability μ_{ij}^* of surviving the Sahara ($Y = 1$ for survival, $Y = 0$ otherwise) as a function of the route taken ($x = 1$ for western route, $x = 0$ for eastern route) is given by

$$\mu_{ij}^* = \mathrm{E}[Y_{ij}|U_i = u] = h(\beta_0^* + \beta_1^* x_{ij} + u + a_{ij}),$$

where $h(\cdot)$ is the inverse logit function, $Y_{ij}|U_i \sim \text{Bernoulli}(\mu_{ij}^*)$, $U_i \sim \mathrm{N}(0, \sigma^2)$ is a random year effect, $i = 1, 2, 3, 4$ indexes the years, $j = 1, \ldots, m_i$ indexes bird within year, and $(m_1, m_2, m_3, m_4) = (5, 11, 13, 13)$. The U_i term in the model acknowledges that we expect variability in the survival rate by year because the conditions could be more dangerous for some years relative to others.

We estimate the random effects variance for this marginally interpretable model to be $\hat{\sigma}^2 = 0.50$. The fixed effects parameter estimates are provided in Table 3.8 along with corresponding 95% confidence intervals and tests of the null hypothesis $H_0 : \beta_1^* = 0$ versus the alternative $H_1 : \beta_1^* \neq 0$. These hypotheses are appropriate here because they target inference on the marginal parameter that is of interest to the study. Regardless of

Table 3.8: Parameter estimates for the fixed effects in the bird migration study of Hewson et al. (2016), with corresponding standard errors, tests, and 95% confidence intervals

Marginally Interpretable Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0^*)	2.49	0.90	Wald	2.76	0.01	(0.72,4.25)
route (β_1^*)	-2.19	1.04	Wald	-2.11	0.03	(-4.23,-0.16)
			LR	6.11	0.01	(-4.61,-0.39)
			Score	-2.38	0.02	—

Cluster-Specific Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0)	2.69	1.26	Wald	2.14	0.03	(0.23,5.15)
route (β_1)	-2.36	1.40	Wald	-1.69	0.09	(-5.10,0.37)
			LR	6.11	0.01	(-6.84,-0.39)
			Score	-2.38	0.02	—

the choice of test, we find a statistically significant effect of route upon the probability of survival when using the marginally interpretable model.

An inappropriate analysis would be to fit a conventional model that excludes the adjustment and does not have a marginal interpretation, and then focus inference on the cluster-specific parameter β_1. For example, one might test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Table 3.8 shows that this misguided model and test result in a much higher p-value when using the ubiquitous Wald test. Using this approach, we would falsely suggest that the route is not associated with the probability of survival because our model and test would not be calibrated to the marginal scientific question of interest.

Seed Germination

Crowder (1978) reported results from a 2×2 factorial experiment involving seed germination. Two types of seeds (*O. aegyptiaco* 73 and *O. aegyptiaco* 75) were covered in two different types of root extract (bean and cucumber). Each combination of seed type and extract was applied to five plates, except for the combination of *O. aegyptiaco* 75 and cucumber extract, which was applied to six plates. After a fixed duration, the number of germinated seeds, Y_i, and the total number of seeds, m_i, on each plate were counted, where $i = 1, \dots, 21$. These data can be found in Table A.6. We let $x_{1,i} = 1$ for *O. aegyptiaco* 75, $x_{1,i} = 0$ for *O. aegyptiaco* 73, $x_{2,i} = 1$ for cucumber extract, and $x_{2,i} = 0$ for bean extract. Defining $p_i = Y_i/m_i$, we fit a marginally interpretable model of the form

$$\mathrm{E}[p_i|U_i = u] = h(\beta_0^* + \beta_1^* x_{1,i} + \beta_2^* x_{2,i} + \beta_3^* x_{1,i} x_{2,i} + u + a_i),$$

where $U_i \sim \mathrm{N}(0, \sigma^2)$, $Y_i|U_i \sim \mathrm{Binomial}(m_i, \mathrm{E}[p_i|U_i])$, and $h(\cdot)$ is the inverse logit function. We also fit an analogous conventional GLMM that does not include the adjustment.

Fitting both models via maximum likelihood estimation, we estimate $\hat{\sigma}^2 = 0.06$. Parameter estimates for the fixed effects parameters are given in Table 3.9 along with corresponding hypothesis tests and confidence intervals. Unlike in the preceding examples, the two sets of estimates, tests, and intervals are virtually identical. The similarity between the two parameterizations in this context is due to the small magnitude of the random effects variance. As a consequence of the limited between-plate variability, the adjustment is also small in magnitude and there is little difference between the two models. This is especially clear for the test of whether there is an interaction effect. Under the null hypothesis $H_0 : \beta_3^* = 0$, the marginally interpretable GLMM is not equivalent to the conventional GLMM, and the likelihood ratio test is therefore not the same for the two parameterizations in this setting. However, the difference is so small that both the test statistic and the resulting p-value match to the second decimal place. Although the parameterization has little impact on estimation and inference, the marginal parameters, which focus on average

Table 3.9: Parameter estimates for the fixed effects in the seed germination study of Crowder (1978), with corresponding standard errors, tests, and 95% confidence intervals

Marginally Interpretable Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0^*)	-0.45	0.22	Wald	-2.04	0.04	(-0.87,-0.02)
seed (β_1^*)	-0.10	0.27	Wald	-0.35	0.73	(-0.63,0.44)
extract (β_2^*)	0.52	0.30	Wald	1.74	0.08	(-0.07,1.11)
interaction (β_3^*)	0.80	0.38	Wald	2.11	0.04	(0.06,1.54)
			LR	4.15	0.04	(0.03,1.59)
			Score	2.38	0.02	—

Cluster-Specific Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0)	-0.45	0.22	Wald	-2.03	0.04	(-0.89,-0.02)
seed (β_1)	-0.10	0.28	Wald	-0.35	0.73	(-0.64,0.45)
extract (β_2)	0.53	0.30	Wald	1.74	0.08	(-0.07,1.12)
interaction (β_3)	0.81	0.39	Wald	2.10	0.04	(0.06,1.57)
			LR	4.15	0.04	(0.03,1.63)
			Score	2.44	0.01	—

seed and extract effects across all plates, have a more convenient interpretation than the cluster-specific parameters, which focus on the seed and extract effects for a specific plate.

Swine Parasites

Larsen et al. (2000) reported data, which can be found in Table A.7, from a sample of pigs collected by Roepstorff et al. (1998). Fecal samples of 1,016 pigs from 108 pigsties were investigated to determine whether or not each pig was infected with roundworm. Of the 108 pigsties used to collect samples, 72 were conventional whereas 36 were *specific pathogen free* (SPF) and were therefore expected to be more sanitary. The goal of the study was to determine if roundworm occurred at a lower rate in SPF pigsties than in conventional

Table 3.10: Parameter estimates and corresponding standard errors, hypothesis tests, and 95% confidence intervals for the fixed effects in the pigsty data found in Larsen et al. (2000) assuming a common random effects variance across the two types of pigsties

Marginally Interpretable Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0^*)	-1.87	0.19	Wald	-9.65	<0.01	(-2.25,-1.49)
type (β_1^*)	-0.87	0.38	Wald	-2.27	0.02	(-1.63,-0.12)
			LR	5.43	0.02	(-1.66,-0.14)
			Score	-2.31	0.02	—

Cluster-Specific Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0)	-2.53	0.27	Wald	-9.38	<0.01	(-3.06,-2.00)
type (β_1)	-1.06	0.46	Wald	-2.31	0.02	(-1.96,-0.16)
			LR	5.43	0.02	(-2.02,-0.17)
			Score	-2.50	0.01	—

pigsties. We let Y_i be the number of infected pigs out of m_i in each pigsty, indexed by $i = 1, \ldots, 108$, and let $p_i = Y_i/m_i$ be the proportion of infected pigs in each pigsty. We include a fixed effect for the type of pigsty ($x_i = 1$ for SPF, $x_i = 0$ for conventional) and a random pigsty effect, fitting a marginally interpretable model of the form

$$\mu_i^* = \mathrm{E}[p_i|U_i = u] = h(\beta_0^* + \beta_1^* x_i + u + a_i),$$

where $h(\cdot)$ is the inverse logit function, $U_i \sim \mathrm{N}(0, \sigma^2)$, and $Y_i|U_i \sim \mathrm{Binomial}(m_i, \mu_i^*)$.

We fit both this marginally interpretable model and an analogous conventional model via maximum likelihood estimation, estimating $\hat{\sigma}^2 = 2.19$ in both cases. The fixed effects parameter estimates are summarized in Table 3.10. The point estimates and confidence intervals are strikingly different between the two parameterizations, but the hypothesis tests lead to identical conclusions. The discrepancy associated with the point estimates and

Table 3.11: Parameter estimates and corresponding standard errors, hypothesis tests, and 95% confidence intervals for the fixed effects in the pigsty data found in Larsen et al. (2000) assuming a different random effects variance for each type of pigsty

Marginally Interpretable Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0^*)	-1.88	0.19	Wald	-9.76	<0.01	(-2.26,-1.50)
type (β_1^*)	-0.82	0.42	Wald	-1.94	0.05	(-1.65,0.01)
			LR	3.49	0.06	(-1.64,0.04)
			Score	-1.92	0.06	—

Cluster-Specific Model

Parameter	Point Estimate	Standard Error	Test	Statistic	p-value	95% CI
intercept (β_0)	-2.50	0.27	Wald	-9.21	<0.01	(-3.04,-1.97)
type (β_1)	-1.20	0.63	Wald	-1.91	0.06	(-2.44,0.03)
			LR	4.77	0.03	(-2.80,-0.12)
			Score	-2.18	0.03	—

confidence intervals stems from a random effects variance that is large in comparison to the previous example. The tests, which have null value equal to zero, are expected to yield similar results because the sample size is relatively large.

Although there is little evidence to suggest differing amounts of heterogeneity among the two different types of pigsties, we also fit both a marginally interpretable model and a conventional model for which each treatment group had its own random effects variance. For both models, we estimated $\hat{\sigma}_1^2 = 2.71$ for the SPF pigsties and $\hat{\sigma}_0^2 = 2.05$ for the conventional pigsties. The fixed effects parameter estimates are summarized in Table 3.11. As in the single variance case, the point estimates and confidence intervals are strikingly dissimilar, but the Wald tests yield nearly identical conclusions. In this case, the likelihood ratio test differs between the two parameterizations because the null model (with $\beta_1^* = 0$) is not equivalent under the marginally interpretable model and the conventional GLMM.

Since interest lies in the average effect of pigsty type on roundworm incidence across all pigsties, the estimates and tests based on the marginal parameterization are more appropriate than those based on the cluster-specific parameterization.

3.3 Bayesian Inference

As an alternative to the classical procedures discussed in Section 3.2, one could instead conduct inference within a Bayesian framework. In contrast to the frequentist perspective, which treats each model parameter as a fixed but unknown value, the Bayesian perspective treats each model parameter as a quantity that follows an unknown probability distribution. Before observing data, one specifies a *prior distribution* with density $\pi_{\boldsymbol{\theta}}$ for the unknown parameters $\boldsymbol{\theta}$. After observing data $\mathbf{Y}$ from the density $f_{\mathbf{Y}|\boldsymbol{\theta}}$, *Bayes' Rule* is used to update the prior distribution and obtain a *posterior distribution* with density $\pi_{\boldsymbol{\theta}|\mathbf{Y}}$. Bayes' Rule states that

$$\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y}) = \frac{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{Y})} = \frac{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{\int f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Since the marginal density $f_{\mathbf{Y}}$ does not depend on the unknown parameters $\boldsymbol{\theta}$, the posterior density of $\boldsymbol{\theta}$ given $\mathbf{Y}$ is often expressed as

$$\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y}) \propto f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}).$$

The posterior distribution is the basis for inference in the Bayesian framework because it can be used to make probabilistic statements about the parameters. For example, if interested in whether a univariate parameter θ is greater than some value θ_0, one can determine the posterior probability that $\theta > \theta_0$ by computing the area under the posterior density curve for θ that lies above θ_0. That is,

$$P(\theta > \theta_0|\mathbf{Y}) = \int_{\theta_0}^{\infty} \pi_{\theta|\mathbf{Y}}(\theta|\mathbf{Y})d\theta.$$

This quantity is known as the *tail area above* θ_0 or as a *posterior predictive p-value* (see, for example, Gelman et al., 2004, Section 6.3), and is analogous to a one-sided p-value for a classical test with null hypothesis $H_0 : \theta \geq \theta_0$ and alternative hypothesis $H_1 : \theta < \theta_0$. Given a sample from the posterior distribution of θ, one can calculate the tail area above θ_0 by computing the proportion of values in the sample from $\pi_{\theta|\mathbf{Y}}$ that are greater than θ_0.

To formally test hypotheses in a Bayesian context, one could use Bayes factors. A *Bayes factor* represents a ratio of likelihoods under two models, and measures the multiplicative change from the prior odds to the posterior odds given the data. For a test of the simple hypotheses $H_0 \colon \theta = \theta_0$ and $H_1 \colon \theta = \theta_1$, we define the Bayes factor as

$$B = \frac{f_{\mathbf{Y}|\theta}(\mathbf{Y}|\theta_0)}{f_{\mathbf{Y}|\theta}(\mathbf{Y}|\theta_1)} = \frac{\pi_{\theta|\mathbf{Y}}(\theta_0|\mathbf{Y})/\pi_{\theta|\mathbf{Y}}(\theta_1|\mathbf{Y})}{\pi_\theta(\theta_0)/\pi_\theta(\theta_1)}.$$

This is simply a likelihood ratio; large values of B favor H_0 whereas small values of B favor H_1. For a test of the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta \neq \theta_0$, the Bayes factor can be computed using the Savage-Dickey density ratio, $B = \pi_{\theta|\mathbf{Y}}(\theta_0|\mathbf{Y})/\pi_\theta(\theta_0)$ (see Dickey, 1971; Verdinelli and Wasserman, 1995). As in the case with simple hypotheses, large values of B favor the null hypothesis and small values of B favor the alternative.

To illustrate inference in a Bayesian framework and show that model parameterization can influence one's conclusions in this setting as well, we revisit the rat teratology data of Weil (1970) that we introduced in Section 3.2.3. We again model the conditional mean as a function of a fixed treatment effect and a random litter effect, and allow the variance of the random litter effect to depend on the treatment. Indexing the litters by $i = 1, \ldots, 32$, we assume $U_i \overset{ind}{\sim} \mathrm{N}(0, \sigma^2_{x_i})$ and $Y_i|\boldsymbol{\beta}, U_i \overset{ind}{\sim} \mathrm{Binomial}(m_i, \mathrm{E}[p_i|\boldsymbol{\beta}, U_i])$, and model

$$\mathrm{E}[p_i|\boldsymbol{\beta}^*, U_i] = h(\beta^*_0 + \beta^*_1 x_i + U_i + a_i)$$

for the marginally interpretable GLMM and

$$\mathrm{E}[p_i|\boldsymbol{\beta}, U_i] = h(\beta_0 + \beta_1 x_i + U_i)$$

for the conventional GLMM. Here, $h(\cdot)$ is the inverse logit function. One change from our earlier model for these data is that we code $x_i = 1$ for exposure to the chemical agent and $x_i = -1$ (as opposed to $x_i = 0$) for the control diet. Since β_0^* and β_1^* are now treated as random quantities, it is more natural for the marginal means for the two treatment groups to be $h(\beta_0^* - \beta_1^*)$ and $h(\beta_0^* + \beta_1^*)$ rather than $h(\beta_0^*)$ and $h(\beta_0^* + \beta_1^*)$. When $x_i = -1$ for the control group, these two means are on an equal footing in terms of variance, whereas when $x_i = 0$ for the control group, the mean for the treatment group is inherently more variable. Our prior distributions for β_0, β_1, $\log(\sigma_0^2)$, and $\log(\sigma_1^2)$ are $\mathrm{N}(0, 2)$, $\mathrm{N}(0, 1)$, $\mathrm{N}(-1/2, 1)$, and $\mathrm{N}(-1/2, 1)$, respectively. Here, σ_0^2 is the variance parameter for the control group and σ_1^2 is the variance parameter for the treatment group. These priors were chosen such that the prior distribution for the expected survival rate for each of the two treatment groups is approximately uniform over the interval $(0, 1)$.

We sample from the posterior distribution of the unknown parameters, including the latent random variables U_i, using an MCMC algorithm that iteratively updates blocks of parameters using Metropolis steps. We first update $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, then $\boldsymbol{\alpha} = (\sigma_0^2, \sigma_1^2)^T$, and finally $\mathbf{U} = (U_1, \ldots, U_{32})^T$. A general description of this computational strategy is provided in Section 4.4. We executed our MCMC algorithm both with the adjustment included in the model and without it. Each chain was run for 1,010,000 steps, with the first 10,000 steps discarded as burn-in and every 100^{th} step thereafter retained for the final sample. This resulted in 10,000 draws from the posterior distribution for each model.

Table 3.12 displays posterior means and standard deviations for the parameters in both the marginally interpretable GLMM and the conventional GLMM. Differences between

Table 3.12: Posterior means of the unknown parameters in the model for the rat teratology data of Weil (1970) (with corresponding posterior standard deviations in parentheses)

Parameter	Marginally Interpretable GLMM	Conventional GLMM
β_0	1.62 (0.24)	1.92 (0.30)
β_1	-0.49 (0.23)	-0.38 (0.29)
σ_0	0.74 (0.30)	0.72 (0.28)
σ_1	1.55 (0.42)	1.58 (0.42)

these posterior means and the maximum likelihood estimates reported in Table 3.3 stem from the alternative coding of the treatment effect and from the prior distributions placed on the parameters. Figure 3.4 displays kernel density estimates based on the posterior samples for the two models. The tail area above zero for β_1^* in the marginally interpretable model, which corresponds to the marginal treatment effect, is 0.016. This is considerably less than the tail area of 0.088 for the cluster-specific treatment effect β_1 in the conventional GLMM. Thus, many would draw different conclusions about the importance of the treatment effect using the two different models. Indeed, Bayes factors for a test of no treatment effect ($H_0 \colon \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$), computed using the Savage-Dickey density ratio, come in at 0.45 for the marginally interpretable model and 1.37 for the conventional GLMM. This confirms the disparity, as the Bayes factor for the conventional model favors the null hypothesis whereas the Bayes factor for the marginally interpretable model does not.

Another approach for investigating the treatment effect is to compare the expected 21-day survival rates between the two treatment groups. For the marginally interpretable model, the expected proportion of rat pups to survive 21 days among those alive after four days (omitting the subscript i) is $\mathrm{E}[p|\boldsymbol{\beta}^*, \boldsymbol{\alpha}, x = 1] = h(\beta_0^* + \beta_1^*)$ for the treatment group and $\mathrm{E}[p|\boldsymbol{\beta}^*, \boldsymbol{\alpha}, x = -1] = h(\beta_0^* - \beta_1^*)$ for the control group. For the conventional GLMM the same expectation is $\mathrm{E}[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = 1] = \int h(\beta_0 + \beta_1 + u) f_U(u) du$ for the treatment group and $\mathrm{E}[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = -1] = \int h(\beta_0 - \beta_1 + u) f_U(u) du$ for the control group. Note that

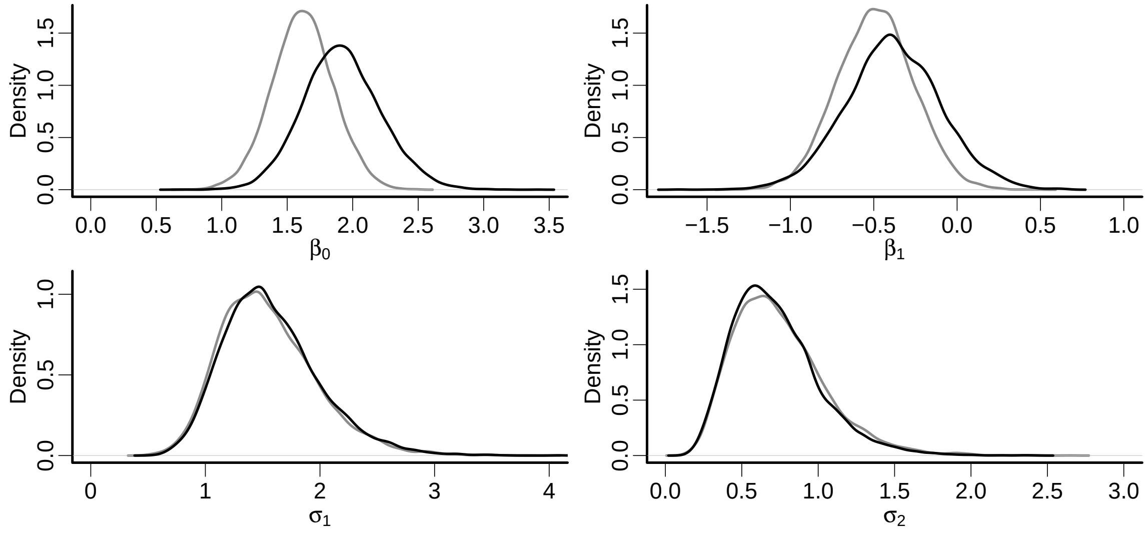

Figure 3.4: Kernel density estimates of the posterior densities for the unknown parameters in the model for the rat teratology data of Weil (1970); estimates obtained from the marginally interpretable model are in gray while those obtained from the conventional GLMM are in black

the parameters $\boldsymbol{\alpha}$ enter this expression through the random effects distribution f_U. Interest lies in whether or not the quantity $\mathrm{E}[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = 1] - \mathrm{E}[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = -1]$ (or the same quantity with $\boldsymbol{\beta}^*$ replacing $\boldsymbol{\beta}$) is nonzero. Kernel density estimates of the posterior density for this quantity under the two models are shown in Figure 3.5. The integral evaluation required for the conventional GLMM was accomplished using Monte Carlo integration. These densities are similar because the marginally interpretable and conventional models are equivalent in this setting as described in Section 3.1. Under both models, most of the posterior mass is below zero. The tail area above zero for the marginally interpretable model is 0.016, matching the tail area for β_1^*. The tail area above zero for the conventional GLMM is 0.014. This is close to the tail area for the marginally interpretable model, but contrasts sharply with the tail area of 0.088 for β_1. This demonstrates that the marginal parameter β_1^* targets inference on the quantity of interest, which is the difference between the two group means, whereas the cluster-specific parameter β_1 does not.

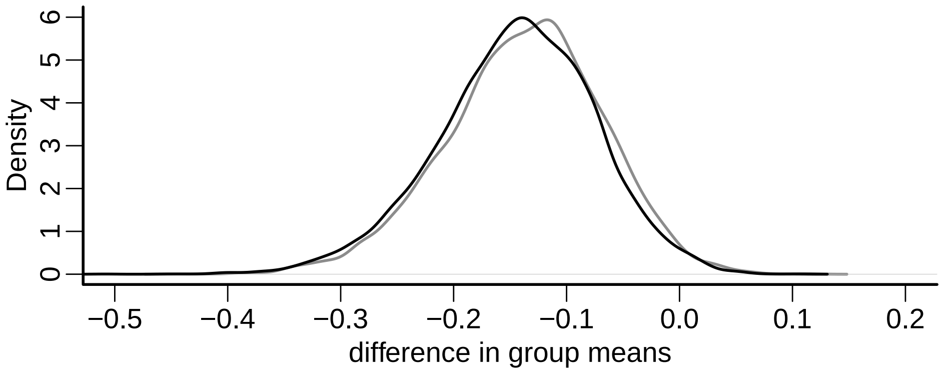

Figure 3.5: Kernel density estimates of the posterior density of the difference in the expected 21-day survival rate between the two treatment groups in the rat teratology study of Weil (1970) based on the marginally interpretable model (gray) and the conventional GLMM (black)

3.4 Stability of Fixed Effects Parameter Estimates

A fixed effect, as its name suggests, is a quantity that has a fixed value across an entire population. Whereas one might observe different realizations of a random effect in two samples from the same population, estimates of a fixed effect based on those two samples should be roughly the same. Due to differences among samples, some variation in the fixed effects parameter estimates is to be expected, especially if the sample size is small. For sufficiently large samples, estimates of a fixed effect should be fairly stable across different samples from the same population. Further, changes in the random effects distribution should not affect the magnitude of a fixed effect (provided the fixed effect is independent of the random effect). The extent to which this is true depends on the parameterization of the model. The marginal effects in a marginally interpretable GLMM represent population-level quantities whereas the cluster-specific effects in a conventional GLMM depend on the random effects, which could change from one sample to another. Estimates of the marginal

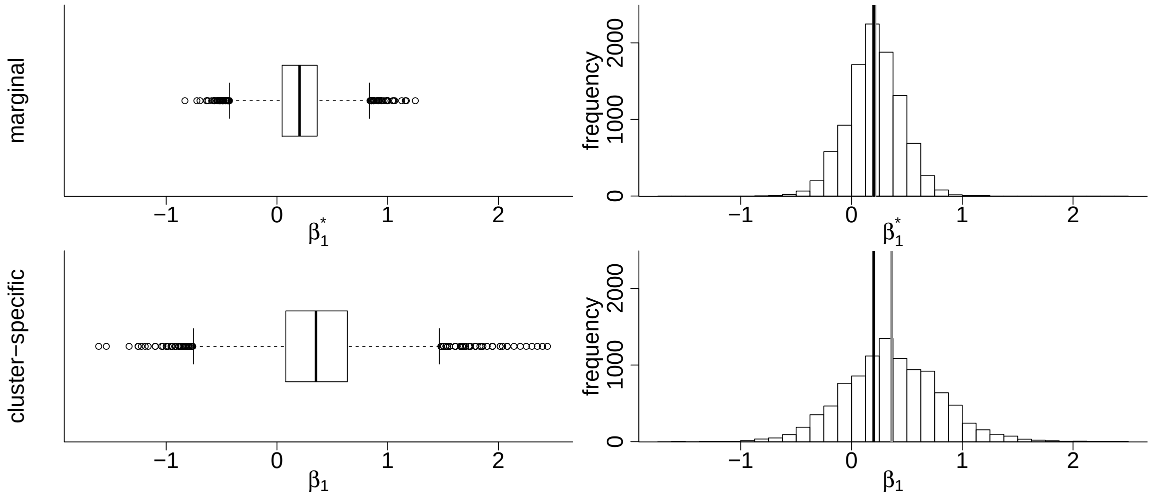

Figure 3.6: Boxplots and histograms showing the distributions of $\hat{\beta}_1^*$ and $\hat{\beta}_1$ across 10,000 simulated datasets with $\beta_1^* = 0.2$; the vertical black line in each histogram shows the true value of β_1^* while the vertical gray line shows the mean of the 10,000 estimates used to produce the histogram

parameters are therefore more stable across different samples from the same population than estimates of corresponding cluster-specific parameters.

To emphasize the relative stability of the marginal parameters in comparison to the cluster-specific parameters, we revisit the simulation study from Section 3.2.2. Figure 3.6 shows boxplots and histograms for the 10,000 estimates of β_1 and β_1^* obtained from the 10,000 datasets generated with 100 pairs of binary observations, $\sigma^2 = 5$, $\beta_0^* = \text{logit}(2/3)$, and $\beta_1^* = 0.2$. Each dataset can be viewed as an independent sample from a population with the same underlying parameter values. As expected, both $\hat{\beta}_1^*$ and $\hat{\beta}_1$ vary from dataset to dataset, but the estimates for β_1 exhibit greater variability than those for β_1^*. This suggests that cluster-specific parameter estimates are more sensitive to the differences among multiple samples from the same population than marginal parameter estimates. Figure 3.6 also shows that the mean of the 10,000 estimates for β_1 is 0.36 while the mean of the 10,000 estimates for β_1^* is 0.20, which matches the true underlying parameter value. The amount by which the mean for $\hat{\beta}_1$ deviates from 0.2 is driven by the random effects variance σ^2. If we were to run the simulation again with a different value for σ^2, the distribution of $\hat{\beta}_1$ would be shifted, whereas the distribution for $\hat{\beta}_1^*$ would remain centered at 0.2.

The additional variability in the cluster-specific parameter estimates is due in part to the dependence of β_1 on both σ^2 and β_0^*. Since different samples have different amounts of heterogeneity among clusters and different prevalences of $Y = 1$, estimates of σ^2 and β_0^* are likely to vary from sample to sample. Changes in these values alter the expected value of β_1, which adds to the variability in $\hat{\beta}_1$ across samples. Consider, for example, a drug trial conducted across several different clinics for which the response variable is a binary indicator of whether or not each patient's condition improved over the course of the study. The analysis involves fitting a logistic regression model for which the clinic effect is included as a random effect. If the study shows that the drug has a positive effect (relative to a control), an effort might be made to replicate the study at another set of clinics. If these clinics exhibit greater variability than the clinics in the original trial, then the cluster-specific parameter estimate, which is conditional on clinic, will suggest a different effect size even if the effect size for the population on average is the same for both trials.

Instability of the cluster-specific parameters across samples also occurs if there is a different underlying prevalence of a positive outcome for each of the two samples. Returning to the drug trial example, suppose that in the first trial 10% of patients see improvement in their condition whereas in the second 30% do. Even if the average effect of the drug (relative to the control) is the same for both samples, the estimated cluster-specific effect will be different. This phenomenon stems from the fact that the curvature of the logit function is not uniform over its domain and therefore has greater impact in some regions of the unit

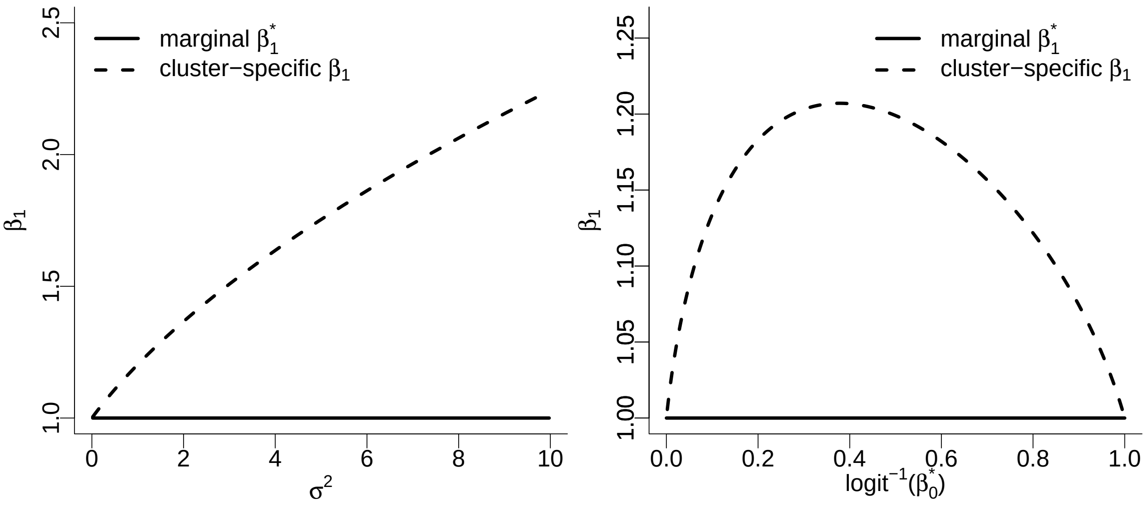

Figure 3.7: Plots depicting how β_1 varies with σ^2 (left panel, for $\beta_0^* = \text{logit}(0.3)$) and β_0^* (right panel, for $\sigma^2 = 1$); the solid line in each panel shows that $\beta_1^* = 1$ while the dashed curve shows how β_1 varies

interval than others. Thus, the shift from 0.1 in the first trial to 0.3 in the second trial leads to different estimates of β_1.

Figure 3.7 demonstrates how the value of β_1 varies with σ^2 and β_0^* when β_1^* remains unchanged. To produce this figure, we considered a model with a logit link, a normal random intercept $U \sim \text{N}(0, \sigma^2)$, and a single predictor $x \in \{0, 1\}$ representing the presence or absence of a treatment. The true value of the marginal parameter β_1^* is 1. Thus, on average, the treatment $x = 1$ corresponds to a multiplicative increase by a factor of 2.72 in the odds that $Y = 1$. In the left panel of Figure 3.7, we set $\beta_0^* = \text{logit}(0.3)$ and see that β_1 increases with σ^2 while β_1^* remains equal to one. In the right panel of Figure 3.7, we set $\sigma^2 = 1$ and see how β_1 varies with $\text{logit}^{-1}(\beta_0^*)$. Ultimately, regardless of the overall prevalence of $Y = 1$ and the variance of the random effects, the estimate of the marginal parameter β_1^* should be near one, whereas the estimate of the cluster-specific parameter

β_1 is more sensitive to changes in the underlying prevalence of $Y = 1$ and in the random effects variance σ^2 that arise from sample to sample.

In summary, the effect size for a cluster-specific effect varies from one sample to another based on the characteristics of those samples. As such, one might observe a smaller effect in a follow-up study after observing a large effect in an initial study even if the true underlying effect is the same. Such a difference might be construed as a failure to replicate the initial result, but in actuality is driven by the structure of the GLMM. A marginal effect should not change as much as a cluster-specific effect from one sample to another provided the samples are taken from the same population with the same true underlying effect size. Hence, marginal effects might be viewed as more reproducible than cluster-specific effects.

3.5 Consistent Estimation of the Random Effects Variance

Thus far, we have focused on inference for the fixed effects parameters in a GLMM. One can also conduct inference on the parameters that characterize the random effects distribution. Random effects are included in a mixed model as a means of introducing dependence and overdispersion into the model. For example, repeated measures on a subject may be systematically large, or count data may show extra-Poisson variation. The degree of dependence and overdispersion, at least in simple models, is determined by the distribution of the random effects, with the variance of this distribution being of particular importance. In order to consistently estimate the random effects variance, it is natural that one would require replication of the random effects. If each realization of a random effect is associated with just one observation, then there is no way to separate the variability in that random effect from the other variability in the data. As an example, consider the following hierarchical model:

$$\begin{aligned} Z_i &\sim F_{\sigma^2}; \\ Y_i|Z_i = z_i &\sim \text{Bernoulli}(z_i), \end{aligned}$$

where $i = 1, \ldots, N$ and F_{σ^2} is an arbitrary distribution on $(0, 1)$ with mean μ and variance σ^2. To obtain a marginal model for Y_i, one must integrate over Z_i. Regardless of the value of σ^2, the resulting marginal distribution for Y_i is $\text{Bernoulli}(\mu)$. No matter how many of these Bernoullis are collected, there is no replication tied to a single random effect and no information is obtained about σ^2. Hence, σ^2 cannot be estimated consistently.

Despite this intuition that the random effects variance cannot be estimated consistently in the absence of replication, Kim and Kim (2011) proved a surprising result for a conventional Bernoulli GLMM. Namely, they showed that the maximum likelihood estimator is strongly consistent for the random effects variance σ^2, even when there is no replication. We call this result the *Kim Paradox* and state it below in a form that, for simplicity of presentation, is not as general as the form presented in Kim and Kim (2011).

The Kim Paradox: *With no replication, one can estimate σ^2 consistently.* Let parameters β_0, $\beta_1 \neq 0$, and $\tau^2 > 0$ be fixed and known, and also let $X_i \sim \text{N}(0, \tau^2)$ and $U_i \sim \text{Uniform}(-c, c)$ ($c > 0$, $i = 1, 2, \ldots$), be independent sequences of random variables. Furthermore, let $h(\cdot) = \text{logit}^{-1}(\cdot)$, and define the conditionally independent sequence $Y_i|X_i = x_i, U_i = u_i \sim \text{Bernoulli}\big(h(\beta_0 + \beta_1 x_i + u_i)\big)$ for $i = 1, 2, \ldots$. Then $\widehat{\sigma}^2$, the maximum likelihood estimator of the random effects variance σ^2, is consistent.

Noting that the marginal mean $\text{E}[Y_i] = \text{E}[h(\beta_0 + \beta_1 x_i + U_i)]$ is the expectation of the conditional mean, the Kim Paradox arises from the fact that the marginal mean in a conventional GLMM is distorted by the random effects in such a fashion that there is a one-to-one mapping between $\mathrm{E}[Y_i]$ and the random effects variance σ^2. That is, for fixed values of β_0 and β_1, any $\mathrm{E}[Y_i] = \mathrm{E}[h(\beta_0 + \beta_1 x_i + U_i)]$ corresponds to a specific value of σ^2. This one-to-one mapping, along with a rich enough set of x_i, ensures that the marginal mean functions are identifiable and consistency of σ^2 follows.

A marginally interpretable Bernoulli GLMM of the form given by (3.2) and (3.3) resolves the Kim Paradox because the marginal mean in such a model is unaffected by changes in σ^2. Consequently, without replication the data contain no information about the random effects variance, and one cannot obtain a consistent estimator of σ^2. Thus, unlike with a conventional Bernoulli GLMM, one does not obtain a nonsensical consistency result with a marginally interpretable Bernoulli GLMM. This is stated more formally in the following proposition:

Proposition 3.3. *If $Y_i|U_i$ folllows a Bernoulli distribution and we have a marginally interpretable GLMM of the form given by (3.2) and (3.3) for which the random intercepts U_i, $i = 1, \ldots, N$, are independently distributed (i.e., each Y_i has its own unique U_i), then the marginal density of Y_i does not depend in any way on the distribution of U_i.*

Proof of Proposition 3.3: The marginal likelihood for each Y_i, $i = 1, \ldots, N$, is

$$\begin{aligned} f_Y(Y_i) &= \int f_{Y|U}(Y_i|U_i = u) f_U(u) du \\ &= \int \mathrm{E}[Y_i|U_i = u]^{Y_i} (1 - \mathrm{E}[Y_i|U_i = u])^{1-Y_i} f_U(u) du \\ &= \int h(\mathbf{x}_i^T \boldsymbol{\beta}^* + u + a_i)^{Y_i} \big(1 - h(\mathbf{x}_i^T \boldsymbol{\beta}^* + u + a_i)\big)^{1-Y_i} f_U(u) du. \end{aligned}$$

Because the model satisfies (3.3), if $Y_i = 1$ we have

$$f_Y(Y_i) = \int h(\mathbf{x}_i^T \boldsymbol{\beta}^* + u + a_i) f_U(u) du = h(\mathbf{x}_i^T \boldsymbol{\beta}^*),$$

whereas if $Y_i = 0$ we have

$$f_Y(Y_i) = \int \big(1 - h(\mathbf{x}_i^T \boldsymbol{\beta}^* + u + a_i)\big) f_U(u) du = 1 - h(\mathbf{x}_i^T \boldsymbol{\beta}^*).$$

Since $Y_i \in \{0, 1\}$, we can therefore write

$$f_Y(Y_i) = h(\mathbf{x}_i^T \boldsymbol{\beta}^*)^{Y_i} \big(1 - h(\mathbf{x}_i^T \boldsymbol{\beta}^*)\big)^{1-Y_i},$$

and f_Y is completely independent of σ^2, as required. □

Chapter 4: Computation

Parameter estimates for GLMMs have traditionally been obtained in a classical framework through maximum likelihood estimation. In many cases, the expression for the marginal likelihood includes an analytically intractable integral, which presents a computational challenge. This integral must be evaluated numerically or avoided by maximizing an analytical approximation of the likelihood instead of the true likelihood. Numerical integration techniques work well for models that have only one or two random effects associated with each observation, but become impracticable when the number of random effects is large. Methods that maximize an approximation of the likelihood work with higher-dimensional random effects, but do not always yield satisfactory parameter estimates. More recently, Bayesian approaches have been used to sample from the posterior distribution of the unknown parameters in GLMMs. These strategies typically carry a greater computational burden than corresponding frequentist approaches for simple models, but are more easily extended to models with many random effects. Bayesian GLMMs also provide a more natural measure of uncertainty for the parameters than the classical approach because one obtains a posterior distribution for the parameters and need not rely on the asymptotic properties of the maximum likelihood estimator to estimate standard errors based on the Fisher information.

The methodology used to fit a conventional GLMM can easily be adapted to fit a marginally interpretable GLMM. When evaluating the likelihood function, be it as part of an optimization procedure in a frequentist approach or for sampling from the posterior in a Bayesian approach, one must simply include the adjustment in the calculation. The expression for the adjustment, like the expression for the marginal likelihood, might involve an analytically intractable integral. In this chapter, we discuss numerical techniques for evaluating such integrals, and present an algorithm for accurately and efficiently computing the adjustment in a model with a logit link and normal random effects. We then review strategies for fitting GLMMs in both a classical framework and a Bayesian framework, and show how to incorporate the adjustment into these techniques to produce marginal parameter estimates. For the Bayesian setting, we introduce a method for improving mixing in an MCMC algorithm when the model includes a large number of latent random variables and present an example that demonstrates this technique.

4.1 Integral Approximations

When an integral cannot be evaluated analytically, one must resort to an approximation or a numerical integration technique. Common approaches include Laplace approximations, Monte Carlo integration, Gaussian quadrature, and adaptive quadrature. There also exist a number of algorithms tailored specifically to evaluating the logistic-normal integral, which arises when one has a model with a logit link and Gaussian random effects. This section provides an overview of these techniques.

4.1.1 Laplace Approximation

The *Laplace approximation* is an application of Laplace's method for integrals (e.g. de Bruijn, 1961, Chapter 4), which uses an asymptotic expansion of the integrand to obtain an analytical approximation of the integral. We define the Laplace approximation as

$$\int f(x)dx = \sqrt{2\pi}\sigma f(\mu), \tag{4.1}$$

where μ is the value at which $f(\cdot)$ attains its maximum and $\sigma = \xi^{-1/2}$ with ξ given by

$$\xi = -\frac{\partial^2}{\partial x^2} \log\big(f(x)\big)\bigg|_{x=\mu}. \tag{4.2}$$

Denoting $\ell(\cdot) = \log\big(f(\cdot)\big)$, the approximation (4.1) arises from a second-order Taylor series expansion of $\ell(\cdot)$ about μ. Specifically, one approximates

$$\ell(x) \approx \ell(\mu) + \ell'(\mu)(x-\mu) + \frac{1}{2}\ell''(\mu)(x-\mu)^2, \tag{4.3}$$

where $\ell'(\cdot)$ and $\ell''(\cdot)$ are the first two derivatives of $\ell(\cdot)$. If μ maximizes $\ell(\cdot)$, then $\ell'(\mu) = 0$ and the second term on the right-hand side of (4.3) disappears. We then approximate

$$\int \exp\big(\ell(x)\big)dx \approx \int \exp\Big(\ell(\mu) + \frac{1}{2}\ell''(\mu)(x-\mu)^2\Big)dx.$$

This quantity can be written as

$$\exp\big(\ell(\mu)\big)\frac{\sqrt{2\pi}}{\sqrt{-\ell''(\mu)}}\int \frac{1}{\sqrt{2\pi\big(-\ell''(\mu)\big)^{-1}}}\exp\Bigg(-\frac{1}{2\big(-\ell''(\mu)\big)^{-1}}(x-\mu)^2\Bigg)dx,$$

where the integral is equal to one because its integrand is a Gaussian probability density function. Thus, the approximation is

$$\int f(x)dx = \int \exp\big(\ell(x)\big)dx \approx \exp\big(\ell(\mu)\big)\frac{\sqrt{2\pi}}{\sqrt{-\ell''(\mu)}},$$

which is exactly (4.1) with $\sigma = \big(-\ell''(\mu)\big)^{-1/2} = \xi^{-1/2}$, where ξ is defined as in (4.2).

Laplace approximations were popularized by Tierney and Kadane (1986) and Tierney et al. (1989) for approximating posterior moments and marginal densities in a Bayesian context. The Laplace approximation also plays an important role in popular approximate likelihood methods for fitting GLMMs in a classical framework, which will be discussed in Section 4.3. More recently, Rue et al. (2009) proposed a method known as *integrated nested Laplace approximations* (INLA) that incorporates multiple Laplace approximations to obtain accurate approximations of posterior densities in Bayesian models with latent Gaussian processes.

The objective of INLA is to approximate the posterior marginal densities $\pi_{U|\mathbf{Y}}(U_i|\mathbf{Y})$ of the latent variables and $\pi_{\theta|\mathbf{Y}}(\theta_j|\mathbf{Y})$ of the model parameters, where i and j index the latent variables U_i and model parameters θ_j, respectively. We express these posteriors as

$$\begin{aligned}\pi_{U|\mathbf{Y}}(U_i|\mathbf{Y}) &= \int \pi_{U|\boldsymbol{\theta},\mathbf{Y}}(U_i|\boldsymbol{\theta},\mathbf{Y})\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta};\\ \pi_{\theta|\mathbf{Y}}(\theta_j|\mathbf{Y}) &= \int \pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta}_{-j},\end{aligned}$$

where $\boldsymbol{\theta}_{-j}$ is the vector of all elements of $\boldsymbol{\theta}$ except θ_j. INLA uses Laplace approximations equivalent to those used by Tierney and Kadane (1986) to approximate $\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y})$ as $\tilde{\pi}_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y})$ and $\pi_{U|\boldsymbol{\theta},\mathbf{Y}}(U_i|\boldsymbol{\theta},\mathbf{Y})$ as $\tilde{\pi}_{U|\boldsymbol{\theta},\mathbf{Y}}(U_i|\boldsymbol{\theta},\mathbf{Y})$. The approximation $\tilde{\pi}_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y})$ is then used to determine suitable evaluation points and weights for numerical evaluation of $\pi_{U|\mathbf{Y}}(U_i|\mathbf{Y})$ and $\pi_{\theta|\mathbf{Y}}(\theta_j|\mathbf{Y})$. For example, $\pi_{U|\mathbf{Y}}(U_i|\mathbf{Y})$ is approximated by

$$\tilde{\pi}_{U|\mathbf{Y}}(U_i|\mathbf{Y}) = \sum_k \tilde{\pi}_{U|\boldsymbol{\theta},\mathbf{Y}}(U_i|\boldsymbol{\theta}_k,\mathbf{Y})\tilde{\pi}_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}_k|\mathbf{Y})\Delta_k,$$

where k indexes the values of $\boldsymbol{\theta}$ at which $\tilde{\pi}_{U|\boldsymbol{\theta},\mathbf{Y}}(U_i|\boldsymbol{\theta},\mathbf{Y})\tilde{\pi}_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y})$ is evaluated and the Δ_k represent the corresponding weights. Rue et al. (2009) also discussed alternative approximations for $\pi_{U|\boldsymbol{\theta},\mathbf{Y}}(U_i|\boldsymbol{\theta},\mathbf{Y})$ that are more computationally efficient than the Laplace approximation but lead to approximations of $\pi_{U|\mathbf{Y}}(U_i|\mathbf{Y})$ that are less accurate.

4.1.2 Monte Carlo Integration

Rather than employing an analytical approximation, one could instead use a numerical procedure to evaluate an intractable integral. One such approach, introduced by Metropolis and Ulam (1949) and discussed in detail by Robert and Casella (2004, Chapter 3), is *Monte Carlo integration*. This strategy involves collecting many random draws of the integrand and computing their average. For an arbitrary function $p(\cdot)$ and a density function $f(\cdot)$, the classical Monte Carlo estimator approximates

$$\int p(x)f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} p(X_i),$$

where $X_1, \ldots, X_n$ represent n independent draws from the density f. By the strong law of large numbers, this estimator of the integral is unbiased and has variance $\mathrm{Var}\big(p(X)\big)/n$, where X is a random variable with density f. Increasing the number of random draws n from the density f results in a more precise estimator, but due to the stochastic nature of the estimator one can only compute probabilistic bounds for its error.

Several related methods exist for obtaining a more precise estimator. These variance reduction techniques for Monte Carlo integration are discussed, for example, by Givens and Hoeting (2013, Section 6.4). One popular approach is *importance sampling*, which rewrites the integral as

$$\int p(x)f(x)dx = \int \frac{p(x)f(x)}{q(x)} q(x)dx,$$

where $q(\cdot)$ is a density function. The density q should be easy to sample from and should ideally put higher density than f on "important" values of X. The corresponding importance sampling estimator, based on a sample $X_1, \ldots, X_n$ drawn from q, is

$$\int p(x)f(x)dx = \int \frac{p(x)f(x)}{q(x)} q(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i)f(X_i)}{q(X_i)}.$$

Provided that q appropriately weights values in the support of X, the importance sampling estimator is more precise than the classical Monte Carlo estimator based on the same number of random draws. Other variance reduction strategies include using *antithetic variables*, which are identically distributed but negatively correlated; using a *control variate*, which is a variable correlated with X that has known mean; and using the Rao-Blackwell theorem to condition on a sufficient statistic.

For multidimensional integrals, Monte Carlo integration requires one to draw random variates $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a higher-dimensional space. Accurate estimation of the integral requires draws from across this entire space. Completely random draws might fill the space inefficiently, and the number of draws required to accurately evaluate the integral grows quickly with the dimension of the integral. An alternative to the random draws that characterize Monte Carlo integration is to instead use deterministic draws that efficiently cover the space of interest. Such methods are known as *quasi-Monte Carlo* and can achieve comparable precision to standard Monte Carlo techniques with fewer draws from the density f (see Niederreiter, 1992, and the references therein). For a comparison of quasi-Monte Carlo, classical Monte Carlo, and Gauss-Hermite quadrature, see González et al. (2006).

4.1.3 Gaussian Quadrature

Another popular numerical integration technique, known as *Gaussian quadrature*, requires one to evaluate the integrand at a set of nodes $x_1, \ldots, x_n$ known as *quadrature points* or *abscissas* and then compute a weighted average of the function evaluations. The nodes and their correspoding weights, which are collectively known as a *quadrature rule*, are derived from the roots of orthogonal polynomials. We focus on Gauss-Hermite quadrature rules, which are derived from Hermite polynomials and are convenient for evaluating integrals that contain a normal density function. For a set of nodes $x_1, \ldots, x_n$ and weights $w_1, \ldots, w_n$, Gauss-Hermite quadrature approximates

$$\int f(x) \exp(-x^2) dx \approx \sum_{i=1}^{n} w_i f(x_i). \tag{4.4}$$

Appropriate quadrature points and weights can be found in a mathematical handbook such as Abramowitz and Stegun (1972) or can be computed with a formula (e.g. Golub and Welsch, 1969). Pinheiro and Bates (1995) characterize this form of Gaussian quadrature as a deterministic version of Monte Carlo integration because the integral is approximated by evaluating its integrand at a fixed number of points and then computing an average of those function evaluations. Unlike Monte Carlo integration, the points at which the integrand is evaluated are predetermined values as opposed to random draws from a distribution.

An alternative to traditional Gaussian quadrature is *adaptive quadrature*, which approximates the integral more efficiently by focusing the quadrature on an appropriate region of the integrand. This is accomplished by shifting and scaling the quadrature points. Suppose a function $f(\cdot)$ can be written as the product of a function $f^*(\cdot)$ and a normal density function. Adaptive Gauss-Hermite quadrature uses the fact that an integral of the form

$$\int f(x) dx = \int f^*(x) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{1}{2\sigma^2}(x - \mu)^2 \right) dx$$

can be written in the form given in (4.4) as

$$\begin{aligned} \int \frac{1}{\sqrt{\pi}} f^*(\mu + \sqrt{2}\sigma x) \exp(-x^2) dx &\approx \sum_{i=1}^{n} w_i \frac{1}{\sqrt{\pi}} f^*(\mu + \sqrt{2}\sigma x_i) \\ &= \sqrt{2}\sigma \sum_{i=1}^{n} w_i \exp(x_i^2) f(\mu + \sqrt{2}\sigma x_i), \end{aligned}$$

where the w_i and x_i ($i = 1, \ldots, n$) are the nodes and weights for traditional Gauss-Hermite quadrature (see Naylor and Smith, 1982). Thus, by selecting appropriate values for μ and

σ, one can transform the traditional quadrature points to ensure that the abscissas used for the approximation are located in a region of interest for the integrand $f(\cdot)$. Assuming $f^*(\cdot)$ is a density function, Naylor and Smith (1982) suggest choosing μ and σ to be estimates of the mean and standard deviation of $f^*(\cdot)$, while Liu and Pierce (1994) suggest choosing μ to be the mode of $f^*(\cdot)$ and σ to be $\xi^{-1/2}$, where ξ is defined as in (4.2) for a Laplace approximation. Pinheiro and Bates (1995) describe this approach as a deterministic version of importance sampling because, like importance sampling, it aims to evaluate the integrand at "important" values of x and then approximate the integral by computing an average of those function evaluations. Although adaptive Gauss-Hermite quadrature can achieve comparable accuracy to traditional Gauss-Hermite quadrature with fewer quadrature points, the gain in computational efficiency is tempered by the need to determine an appropriate transformation of the abscissas. If calculating μ and σ is nontrivial, the traditional approach could be more efficient than an adaptive approach. For further discussion of adaptive quadrature, see Rabe-Hesketh et al. (2002) and Pinheiro and Chao (2006).

When one uses Gauss-Hermite quadrature with a single quadrature point, the relevant node is $x_1 = 0$ and the corresponding weight is $w_1 = \sqrt{\pi}$. Thus, the adaptive quadrature approximation in this case is

$$\int f(x)dx \approx \sqrt{2\pi}\sigma f(\mu).$$

Using the approach of Liu and Pierce (1994), this approximation is equivalent to the Laplace approximation given by (4.1). As such, Gauss-Hermite quadrature can be viewed as a generalization of the Laplace approximation that is both more accurate and more computationally intensive. One can achieve an arbitrary level of accuracy with Gaussian quadrature by adding more quadrature points to the rule, but additional abscissas require additional function evaluations and come with additional computational expense.

As with Monte Carlo integration, Gaussian quadrature can become impracticable for high-dimensional integrals. If p integrals are nested within one another and n quadrature points are used to evaluate each integral, then a total of n^p function evaluations are required. For example, in two dimensions we have an inner integral and an outer integral. For each of the n quadrature points used to evaluate the outer integral, we must use n-point quadrature to evaluate the inner integral. Further, care must be taken to ensure that the inner integral is approximated accurately, as any error present in the evaluation of the inner integral will propagate to evaluation of the outer integral. Ultimately, Gauss-Hermite quadrature may be suitable for evaluating the marginal likelihood of a GLMM with only one or two random effects per observation, but for models with higher-dimensional random effects it is best to pursue other options.

4.1.4 Evaluating the Logistic-Normal Integral

A popular GLMM for binary response data assumes a logit link function and specifies the random effects distribution as a Gaussian distribution. Consequently, there is considerable interest in accurately and efficiently evaluating the *logistic-normal integral*. This is a typically intractable integral of the form

$$\int \frac{1}{1+e^{-z}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{1}{2\sigma^2}(z-\mu)^2 \right) dz. \tag{4.5}$$

All of the techniques described earlier in this section could be used to evaluate the logistic-normal integral. There also exist several methods tailored specifically to solving (4.5). For example, Crouch and Spiegelman (1990) proposed a trapezoidal quadrature rule that is more accurate than a corresponding Gauss-Hermite quadrature rule for evaluating (4.5). An alternative method, due to Monahan and Stefanski (1992), involves approximating the inverse logit function $h(\cdot)$ with a weighted mixture of normal distributions

$$h_k^*(z) = \sum_{i=1}^{k} p_{k,i}\Phi(zs_{k,i}),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution and the weights $p_{k,i}$ and $s_{k,i}$ are chosen to minimize the maximum approximation error over all values of z. This leads to to the integral approximation

$$\int h(z)\frac{1}{\sigma}\phi\biggl(\frac{z-\mu}{\sigma}\biggr)dz \approx \int h_k^*(z)\frac{1}{\sigma}\phi\biggl(\frac{z-\mu}{\sigma}\biggr)dz = \sum_{i=1}^{k} p_{k,i}\Phi\Biggl(\frac{\mu s_{k,i}}{(1+\sigma^2 s_{k,i}^2)^{1/2}}\Biggr). \quad (4.6)$$

Monahan and Stefanski (1992) provided values for the weights $p_{k,i}$ and $s_{k,i}$ up to $k = 8$ and demonstrated that when $k = 8$ the approximation (4.6) is within 2.1×10^{-9} of the true value of the integral for all μ and σ. One could use fewer than eight mixture weights to improve computational efficiency, but the increase in speed from using fewer weights is small relative to the corresponding loss of accuracy; we therefore recommend using $k = 8$.

More recently, Pirjol (2013) developed a recursive formula that provides an exact solution to the logistic-normal integral on a specific evenly spaced grid. Pirjol (2013) demonstrated that the integral

$$\varphi(\mu, \sigma^2) = \int \frac{1}{1+e^w}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\biggl(-\frac{1}{2\sigma^2}(w-\mu)^2\biggr)dw \quad (4.7)$$

satisfies the recursion

$$\varphi(\mu+\sigma^2, \sigma^2) = e^{-\mu-\frac{\sigma^2}{2}}\bigl(1-\varphi(\mu, \sigma^2)\bigr), \quad (4.8)$$

where $\varphi(0, \sigma^2) = 1/2$. Note that since

$$\frac{1}{1+e^w} + \frac{1}{1+e^{-w}} = 1, \quad (4.9)$$

the quantity $1 - \varphi(\mu, \sigma^2)$ is a logistic-normal integral. To see this, observe that

$$\begin{aligned} 1 - \varphi(\mu, \sigma^2) &= 1 - \int \frac{1}{1+e^w} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\biggl(- \frac{1}{2\sigma^2}(w-\mu)^2 \biggr) dw \\ &= \int \frac{1}{1+e^{-w}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\biggl(- \frac{1}{2\sigma^2}(w-\mu)^2 \biggr) dw, \end{aligned}$$

which is the form given by (4.5). It also follows from (4.9) that $1 - \varphi(\mu, \sigma^2) = \varphi(-\mu, \sigma^2)$.

We exploit the recursive result of Pirjol (2013) and combine it with the approximation of Monahan and Stefanski (1992) to develop a novel, hybrid approach for approximating the logistic-normal integral. To approximate the integral $1 - \varphi(\mu, \sigma^2)$ when $\mu > 0$, we first write $\mu = \mu^* + t\sigma^2$, where $\mu^* \in [0, \sigma^2)$ and t is a nonnegative integer. We then approximate $1 - \varphi(\mu^*, \sigma^2)$ using (4.6) with $k = 8$ mixture weights and apply the recursion (4.8) t times to obtain an approximation for $1 - \varphi(\mu, \sigma^2)$. When $\mu < 0$, the integral of interest is $1 - \varphi(\mu, \sigma^2) = \varphi(|\mu|, \sigma^2)$, and the approximation can still be handled as if the first argument of $\varphi(\cdot, \cdot)$ were positive. Denoting our approximation of $\varphi(\mu, \sigma^2)$ as $\tilde{\varphi}(\mu, \sigma^2)$, we define the error associated with this approximation as

$$\varepsilon(\mu, \sigma^2) = \varphi(\mu, \sigma^2) - \tilde{\varphi}(\mu, \sigma^2).$$

Pirjol (2013) showed that the error $\varepsilon(\mu, \sigma^2)$ is bounded by

$$|\varepsilon(\mu, \sigma^2)| \leq \exp\left(- \frac{1}{2\sigma^2}\mu^2 + \frac{1}{8}\sigma^2 \right) \sup_{z \in [0, \sigma^2)} |\varepsilon(z, \sigma^2)|,$$

which means that $\tilde{\varphi}(\mu, \sigma^2)$ is generally more accurate for larger values of μ and that the error associated with $\tilde{\varphi}(\mu, \sigma^2)$ is never worse than the maximum error of the Monahan-Stefanski approximation (4.6) over the range $[0, \sigma^2)$.

To assess the speed and accuracy of our hybrid approach we compared it to both 30-point Gauss-Hermite quadrature and to a direct application of (4.6). Specifically, for each of the 80 values of σ in the set $\{0.05, 0.10, \ldots, 4.00\}$ we evaluated the integral $1 - \varphi(\mu, \sigma^2)$

for 1,000 values of μ in each of the four intervals $[0, \sigma^2]$, $[\sigma^2, 2\sigma^2]$, $[2\sigma^2, 3\sigma^2]$, and $[3\sigma^2, 4\sigma^2]$ using the hybrid approach, the Monahan-Stefanski approximation, 30-point Gauss-Hermite quadrature, and 1,000-point Gauss-Hermite quadrature. This required 4,000 integral evaluations for each of the 80 values of σ and each method. These evaluations were completed on a Dual Quad Core Xeon 2.66 E5430 computer with 32 gigabytes of RAM. To ensure a fair comparison of speed, all four approaches were implemented using the `Rcpp` package in `R` (R Core Team, 2015; Eddelbuettel and François, 2011; Eddelbuettel, 2013). Gauss-Hermite quadrature with 1,000 quadrature points was treated as the *gold standard* to which the other three methods were compared to assess accuracy.

For each of the competing methods and each value of σ we computed the maximum "error" relative to 1,000-point quadrature within each of the four intervals for μ. Figure 4.1 summarizes the results of this accuracy assessment. Although 30-point Gauss-Hermite quadrature is the most accurate for small values of σ, the hybrid approach is the most accurate in the majority of cases. Notably, the hybrid approach, which combines the recursion of Pirjol (2013) with the approximation of Monahan and Stefanski (1992), clearly outperforms a direct application of the Monahan-Stefanski approximation.

The 320,000 integral evaluations required for the accuracy assessment took 2.1 seconds for the hybrid approach compared to 2.1 seconds for the direct application of the Monahan-Stefanski approximation, 2.2 seconds for 30-point Gauss-Hermite quadrature, and 19.4 seconds for 1,000-point Gauss-Hermite quadrature. Thus, the efficiency of the hybrid approach is comparable to that of the Monahan-Stefanski approach and slightly better than that of 30-point Gauss-Hermite quadrature. Further, 1,000-point quadrature is considerably less efficient than the other three methods. We therefore conclude that the hybrid approach offers the best tradeoff between accuracy and efficiency among the competing methods.

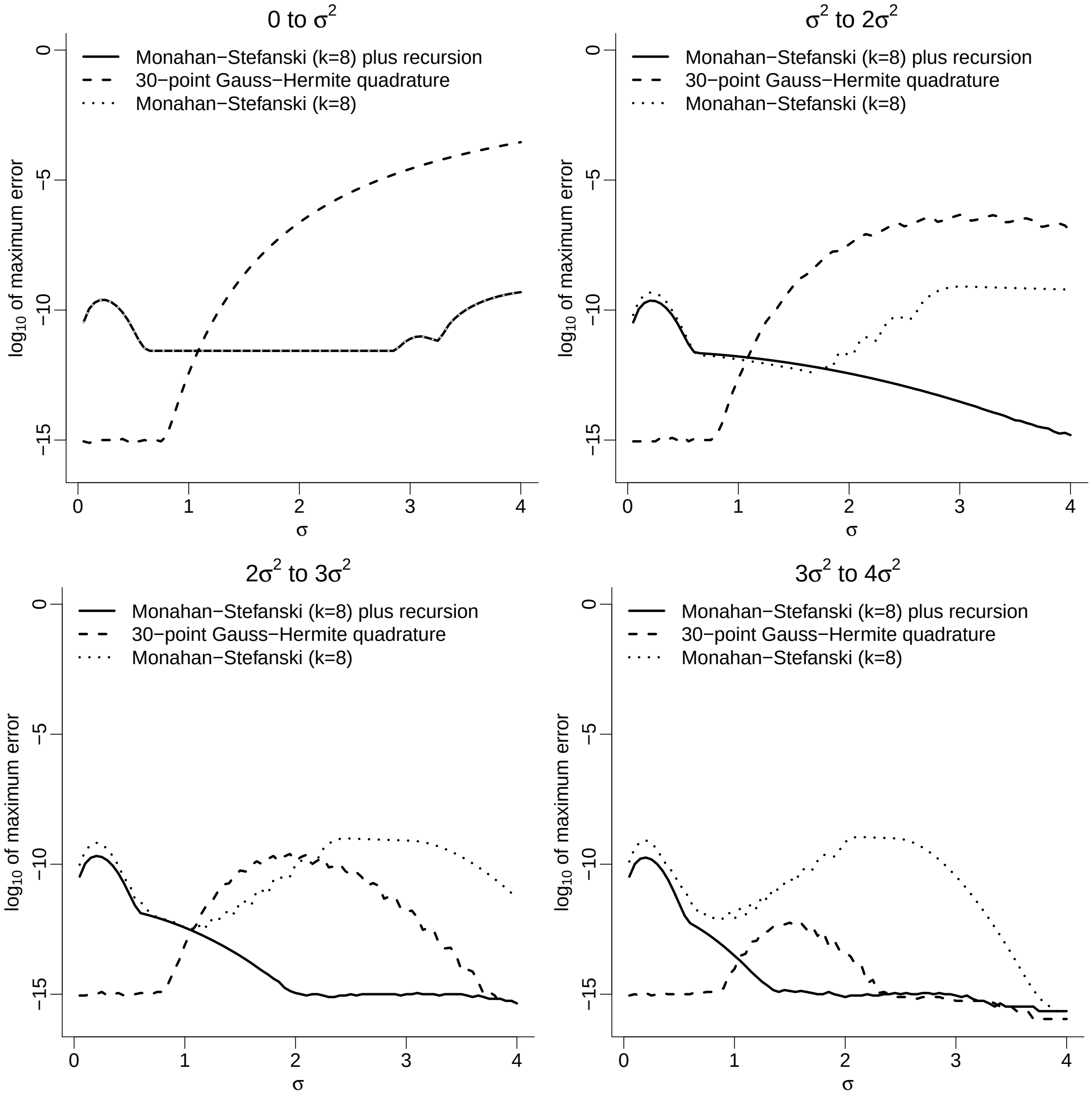

Figure 4.1: Maximum error relative to 1,000-point Gauss-Hermite quadrature for various approximations of the logistic-normal integral with μ in $[0, \sigma^2]$, $[\sigma^2, 2\sigma^2]$, $[2\sigma^2, 3\sigma^2]$, or $[3\sigma^2, 4\sigma^2]$; machine accuracy is approximately 10^{-16}, accounting for the floor in the plots

4.2 Computing the Adjustment

To compute the adjustment in a marginally interpretable GLMM, one must solve (2.2) for $\mathbf{d}_i^T \mathbf{a}_i$. The adjustment is a deterministic function of $\mathbf{x}_i^T \boldsymbol{\beta}$ and the parameters characterizing $f_{\mathbf{U}}$, which we denote by $\boldsymbol{\alpha}$. When fitting a marginally interpretable GLMM, the adjustment must be included in the calculation of the likelihood. Many model-fitting algorithms are iterative in nature, and at the point in the algorithm at which the adjustment must be included current estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will be available to plug into (2.2) to solve for $\mathbf{d}_i^T\mathbf{a}_i$. When the adjustment can be written in closed form, such as when the model has a probit link and Gaussian random effects or when the model has a log link, computing $\mathbf{d}_i^T\mathbf{a}_i$ is straightforward and has little impact on the amount of computation required to fit the model. When the integral in (2.2) is analytically intractable and the adjustment does not have a closed form, computing $\mathbf{d}_i^T\mathbf{a}_i$ is considerably more challenging and can add substantial computational expense to fitting the model.

The most common situation without a closed-form solution is a model with a logit link and Gaussian random effects. We now introduce an algorithm for accurate and efficient computation of $\mathbf{d}_i^T\mathbf{a}_i$ in such a model. We present the algorithm for the case of a univariate normal random intercept, but as a consequence of Proposition 2.1, the strategy described here also applies to a model with multivariate normal random effects. To simplify notation we denote $\kappa = \mathbf{x}_i^T\boldsymbol{\beta}$ and $a = \mathbf{d}_i^T\mathbf{a}_i$, and use $\varphi(\cdot,\cdot)$ as defined in (4.7). Further, suppose $g(\cdot) = \operatorname{logit}(\cdot)$ and $U_i \sim \mathrm{N}(0,\sigma^2)$. In this situation, (2.2) reduces to (2.4) and we can express the right-hand side of (2.4) as $1-\varphi(\kappa+a,\sigma^2)$. Thus, given κ and σ^2, the equation we must solve for a is

$$h(\kappa) = 1-\varphi(\kappa+a,\sigma^2). \tag{4.10}$$

Due to the symmetry of the problem, we need only consider the case of $\kappa > 0$. When $\kappa < 0$ the adjustment has the same magnitude but opposite sign as if $\kappa = |\kappa|$.

Several different techniques, such as binary segmentation or a Newton-Raphson algorithm, can be used to solve (4.10) for a. Any such technique requires evaluation of $\varphi(\kappa + a^*, \sigma^2)$ for several potential values a^* of the adjustment a. Although we can use the recursion (4.8) to calculate $\varphi(t\sigma^2, \sigma^2)$ exactly for any integer t, it is unlikely that the desired a will be such that $\kappa + a$ is an integer multiple of σ^2. Thus, we have need of an approximate numerical integration procedure. Since the function $\varphi(\cdot, \cdot)$ is decreasing in its first argument, we can use the recursion (4.8) to quickly identify an interval of length σ^2 in which $\kappa + a$ must reside. By narrowing our search for the correct value of a to such an interval we reduce the required number of evaluations of $\varphi(\kappa + a^*, \sigma^2)$.

Our algorithm for solving (4.10) includes the following steps. We start with $t = 0$ and increment t by one until $1 - \varphi(t\sigma^2, \sigma^2) \leq h(\kappa) < 1 - \varphi((t + 1)\sigma^2, \sigma^2)$. We use t^* to denote the value of t for which this inequality holds, and note that the value of a satisfying (4.10) must lie in the interval $[t^*\sigma^2 - \kappa, (t^* + 1)\sigma^2 - \kappa)$. We then employ binary segmentation, implemented using the the `uniroot` function in `R` (R Core Team, 2015), to search within this interval for the appropriate value of a. To evaluate $\varphi(\kappa + a^*, \sigma^2)$ for $\kappa + a^* \in [t^*\sigma^2, (t^* + 1)\sigma^2)$ we use the hybrid approach introduced in Section 4.1.4 that combines the approximation of Monahan and Stefanski (1992) with the recursive result of Pirjol (2013). Specifically, we use (4.6) to compute $\varphi(\kappa + a^* - t^*\sigma^2, \sigma^2)$ and then apply (4.8) t^* times to obtain $\varphi(\kappa + a^*, \sigma^2)$.

4.3 Frequentist Approaches to Model Fitting

In a classical framework, estimates for the unknown parameters in a GLMM are obtained through maximization of the marginal likelihood. We denote the vector of unknown parameters as $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})^T$, with $\boldsymbol{\beta}$ representing the fixed effects parameters and $\boldsymbol{\alpha}$ representing the parameters that characterize the random effects distribution. Given data $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$, the quantity we seek to maximize as a function of $\boldsymbol{\theta}$ is

$$f_{\mathbf{Y}}(\mathbf{Y}) = \int f_{\mathbf{Y}|\mathbf{U}}(\mathbf{Y}|\mathbf{U} = \mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}, \tag{4.11}$$

where $f_{\mathbf{U}}$ depends on $\boldsymbol{\alpha}$ and $f_{\mathbf{Y}|\mathbf{U}}$ depends on $\boldsymbol{\beta}$. When the link function is nonlinear, the integral in (4.11) is often analytically intractable and maximum likelihood estimation requires some sort of approximation.

A natural approach to maximizing the likelihood when the integral in (4.11) is intractable is to numerically evaluate the integral using one of the strategies described in Section 4.1 and then employ a numerical optimization technique; for example, a quasi-Newton method. While this approach can yield a high degree of accuracy, it can also be computationally expensive, especially when there are many random effects and the integration is over many dimensions. An alternative maximization strategy is to use the expectation-maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm assumes we have observed data $\mathbf{Y}$ and latent variables $\mathbf{U}$, which can be viewed as missing data, and finds optimal values of the unknown parameters $\boldsymbol{\theta}$ by iteratively performing two steps: an E-step and an M-step. In the E-step, one computes the expectation (with respect to $\mathbf{U}$) of the joint log-likelihood of $\mathbf{Y}$ and $\mathbf{U}$ given the observed data $\mathbf{Y}$. That is, one computes

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}\big[\log\big(f_{\mathbf{Y},\mathbf{U}}(\mathbf{Y},\mathbf{U}|\boldsymbol{\theta})\big)|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}\big],$$

where $\boldsymbol{\theta}^{(t)}$ is the current value for $\boldsymbol{\theta}$. In the M-step, one maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ and sets the appropriate value of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^{(t+1)}$ before returning to the E-step. This algorithm is often used for maximum likelihood estimation in linear mixed models, where the latent variables $\mathbf{U}$ represent the random effects. Unfortunately, for many GLMMs the quantity $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in the E-step is intractable. Wei and Tanner (1990) proposed an extension of the EM algorithm, called *Monte Carlo EM* (MCEM), that uses Monte Carlo integration to approximate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in the E-step. McCulloch (1997) and Booth and Hobert (1999) introduced versions of MCEM that are directly applicable to GLMMs, each with a different approach for numerically approximating $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

Rather than numerically approximating the integral in (4.11), one might instead choose to analytically approximate its integrand. This approximation serves to get the model into a form that allows one to use iterative procedures originally developed for linear mixed models to maximize the "likelihood" of the approximate model. Several related approaches employ this strategy. The most popular is arguably *penalized quasi-likelihood* (PQL), which is due to Breslow and Clayton (1993) and builds upon earlier approaches introduced by Stiratelli et al. (1984) and Green (1987). Breslow and Clayton (1993) use Laplace's method for integrals to approximate the log-likelihood of a GLMM and reduce the model to a form for which parameter estimation can be achieved using a Fisher scoring algorithm. Another widely used approach, known as *pseudo-likelihood estimation*, is due to Wolfinger and O'Connell (1993). This technique uses a first-order Taylor series expansion of the conditional mean to express the nonlinear mean structure $h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{U})$ as a linear function of $\boldsymbol{\beta}$ and $\mathbf{U}$, and yields parameter estimates similar to those obtained using PQL. Other closely related strategies are described by Schall (1991), Wolfinger (1993), and McGilchrist (1994). These approaches do not always yield satisfactory estimates. For instance, Breslow and Lin (1995) showed that PQL tends to yield biased estimates for binomial data with small cluster sizes, and that the bias tends to be larger when there is greater variability in the random effects. Several strategies have been introduced for correcting the bias that arises from these methods (e.g. Breslow and Lin, 1995; Kuk, 1995; Lin and Breslow, 1996).

Whereas the point estimates $\hat{\boldsymbol{\theta}}$ for the unknown parameters $\boldsymbol{\theta}$ in a classical GLMM are obtained via maximum likelihood estimation, corresponding standard errors are computed using the Fisher information matrix $\mathrm{I}_N(\hat{\boldsymbol{\theta}})$ as defined in (3.10). As discussed in Section 3.2.1, under certain conditions the maximum likelihood estimator of the parameters $\boldsymbol{\theta}$ in a GLMM are asymptotically normal with asymptotic covariance equal to the inverse of the Fisher information matrix. This matrix can be estimated by calculating the Hessian of the negative log-likelihood. To estimate the standard error of a particular element of $\hat{\boldsymbol{\theta}}$, one computes the square root of the appropriate diagonal element of $\left(\mathrm{I}_N(\hat{\boldsymbol{\theta}})\right)^{-1}$. Estimating the standard errors in this manner relies on the asymptotic efficiency of the maximum likelihood estimator and on the accuracy of the Hessian of the negative log-likelihood as an approximation of the expected Fisher information.

Frequentist methods for fitting GLMMs have been implemented in widely available software. The `GLIMMIX` procedure in SAS 9.4 (SAS Institute, Cary, NC) uses the pseudo-likelihood approach of Wolfinger and O'Connell (1993) by default, but the user is also able to obtain estimates based on approximating the integral in (4.11) using a Laplace approximation or adaptive Gauss-Hermite quadrature. The `NLMIXED` procedure in SAS 9.4 also allows one to fit a GLMM by using adaptive Gauss-Hermite quadrature to evaluate the likelihood. In `R` (R Core Team, 2015), there are numerous packages for fitting GLMMs. One popular package used for this purpose is `lme4` (Bates et al., 2015). The `glmer` function in this package uses a Laplace approximation for the integral in (4.11) by default, but includes adaptive Gauss-Hermite quadrature as an option. PQL is implemented in the `glmmPQL` function of the `MASS` package (Venables and Ripley, 2002). For the simulation study in Section 3.2.2 and the examples in Sections 2.3 and 3.2.3, we approximated the integral in (4.11) using Gauss-Hermite quadrature and then maximized the marginal likelihood using a quasi-Newton optimization technique implemented in the `optim` function of `R` (R Core Team, 2015).

4.4 Bayesian Approaches to Model Fitting

Rather than adopting a frequentist perspective and fitting GLMMs via maximum likelihood estimation, one might instead adopt a Bayesian perspective. As described in Section 3.3, Bayesian modeling focuses on estimation of the posterior distribution of the unknown parameters given the data. Starting with a prior density $\pi_{\boldsymbol{\theta}}$, one observes data $\mathbf{Y}$ from density $f_{\mathbf{Y}|\boldsymbol{\theta}}$ and then uses Bayes' Rule to obtain a posterior density $\pi_{\boldsymbol{\theta}|\mathbf{Y}}$, where $\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y}) \propto f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. The posterior distribution is central to Bayesian inference because it allows one to make probabilistic statements about the model parameters. In particular, the posterior distribution provides a natural concept of parameter uncertainty that is absent from most frequentist methods.

It is ordinarily difficult, or even impossible, to write the posterior density $\pi_{\boldsymbol{\theta}|\mathbf{Y}}$ in closed form. One option is to approximate the posterior density using a technique such as INLA. More commonly, MCMC is used to generate samples from the posterior distribution. The basic premise of MCMC is to sample from $\pi_{\boldsymbol{\theta}|\mathbf{Y}}$ by constructing a Markov chain that has this posterior as its limiting distribution. Thus, if one runs the Markov chain long enough, each step of the chain can be viewed as a sample from the target posterior. Common approaches to constructing such a Markov chain are the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) and the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). For a parameter vector $\boldsymbol{\theta}$ of length r, the Gibbs sampler involves iteratively updating each component θ_k ($k = 1, \ldots, r$) of $\boldsymbol{\theta}$ by drawing a value from the full conditional distribution of θ_k given $\boldsymbol{\theta}_{-k}$ and $\mathbf{Y}$, where $\boldsymbol{\theta}_{-k}$ is the vector of the $r - 1$ components of $\boldsymbol{\theta}$ besides θ_k. The Metropolis-Hastings algorithm provides an alternative approach when these full conditional distributions are difficult to sample from. For a Markov chain currently in state $\boldsymbol{\theta}^{(t)}$ this algorithm involves proposing a new state $\boldsymbol{\theta}^*$ from a transition density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ and then deciding whether to accept or reject the proposal. The acceptance probability for each proposal is

$$\min\left(1, \frac{\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}^*|\mathbf{Y})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}^{(t)}|\mathbf{Y})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right) = \min\left(1, \frac{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta}^*)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta}^{(t)})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(t)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right). \quad (4.12)$$

If the proposal is accepted one sets $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ and otherwise sets $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$. This process of proposing a new state and either accepting or rejecting the proposal is then repeated, but with the current state now $\boldsymbol{\theta}^{(t+1)}$. We refer to each update of $\boldsymbol{\theta}$ from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ as a *Metropolis step* and note that if the proposal density q is symmetric then $q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ and the acceptance probability reduces to

$$\min\left(1, \frac{\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}^*|\mathbf{Y})}{\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}^{(t)}|\mathbf{Y})}\right) = \min\left(1, \frac{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta}^*)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*)}{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\theta}^{(t)})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(t)})}\right).$$

For further discussion of these MCMC algorithms and other strategies for posterior simulation, see Gelman et al. (2004). In the context of GLMMs, Zeger and Karim (1991) introduced a Gibbs sampling algorithm and Gamerman (1997) proposed a more general Metropolis-Hastings approach.

For posterior simulation of a GLMM, the latent random variables $\mathbf{U}$ are treated as unknown parameters. Thus, in a GLMM we have $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathrm{U})^T$, where $\boldsymbol{\beta}$ is the vector of fixed effects parameters, $\boldsymbol{\alpha}$ is the vector of parameters characterizing the random effects distribution, and $\mathbf{U}$ is a vector containing every realization of each random effect in the model. Even when each observation corresponds to just one random effect, $\mathbf{U}$ is a vector and each element of $\mathbf{U}$ represents the realization of the random variable for a particular cluster of observations. If the data contain many clusters, then the random effects take on many different values and $\mathbf{U}$ is high-dimensional. In turn, the parameter vector $\boldsymbol{\theta}$ is high-dimensional. This presents a challenge for posterior simulation and will be addressed in

Section 4.4.1. For a marginally interpretable GLMM, evaluation of $f_{\mathbf{Y}|\boldsymbol{\theta}}$ requires calculation of the adjustment $\mathbf{d}^T\mathbf{a}$. Within each update of $\boldsymbol{\theta}$, the proposals $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ or the current values $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\alpha}^{(t)}$ can be used along with $\mathbf{x}$ and $\mathbf{d}$, which are treated as fixed and known, to compute the adjustment $\mathbf{d}^T\mathbf{a}$, which can in turn be included in evaluation of $f_{\mathbf{Y}|\boldsymbol{\theta}}$.

4.4.1 Improving Mixing in the Presence of Many Random Effects

A challenge associated with using MCMC to sample from a high-dimensional posterior density is poor mixing. Due to the large number of unknown parameters, the only proposals that tend to get accepted are those representing relatively small steps from the current state of the Markov chain. Consequently, there is substantial autocorrelation in the Markov chain and it is necesssary to run the algorithm for an exceedingly long time to generate a representative sample from the target posterior. One way to improve mixing is to sample the parameters in blocks, but this may not always be enough. In a GLMM, when there are many realizations of the random effects, the latent random variables $\mathbf{U}$ can dominate the likelihood and cause very few proposed $\boldsymbol{\beta}^*$ to be accepted. Specifically, since the model's mean structure is typically given by $h(\mathbf{x}^T\boldsymbol{\beta}+\mathbf{d}^T\mathbf{U})$, if $\mathbf{x}^T\boldsymbol{\beta}$ is large then $\mathbf{d}^T\mathbf{U}$ will generally be small, and vice versa. Thus, if the current state $\mathbf{U}^{(t)}$ of the latent variables is compatible with the current state $\boldsymbol{\beta}^{(t)}$ of the fixed effects parameters, it is difficult to find a proposal $\boldsymbol{\beta}^*$ that yields a greater value of the likelihood given $\mathbf{U}^{(t)}$. Thus, few $\boldsymbol{\beta}^*$ tend to get accepted, those $\boldsymbol{\beta}^*$ that do get accepted tend to represent small steps from the current state $\boldsymbol{\beta}^{(t)}$, and the chain mixes slowly. This is a common problem for fitting mixed models in a Bayesian framework (see, for example, Gelfand et al., 1995; Dunson and Herring, 2005).

To overcome the problem with slow mixing, we suggest the following solution. Along with each proposed $\boldsymbol{\beta}^*$, simultaneously propose random effects $\mathbf{U}^*$ that are consistent with the proposed $\boldsymbol{\beta}^*$. That is, given a proposal $\boldsymbol{\beta}^*$ for the fixed effects parameters, also propose random effects $\mathbf{U}^*$ in such a manner that the mean structure $h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{U})$ is unaffected and there is no net impact on the likelihood. If $\boldsymbol{\beta}^*$ is such that $\mathbf{x}^T\boldsymbol{\beta}^* > \mathbf{x}^T\boldsymbol{\beta}^{(t)}$, then define $\mathbf{U}^*$ such that $\mathbf{d}^T\mathbf{U}^* < \mathbf{d}^T\mathbf{U}^{(t)}$, whereas if $\boldsymbol{\beta}^*$ is such that $\mathbf{x}^T\boldsymbol{\beta}^* < \mathbf{x}^T\boldsymbol{\beta}^{(t)}$, then define $\mathbf{U}^*$ such that $\mathbf{d}^T\mathbf{U}^* > \mathbf{d}^T\mathbf{U}^{(t)}$. Either way, ensure that $\mathbf{x}^T\boldsymbol{\beta}^* + \mathbf{d}^T\mathbf{U}^* = \mathbf{x}^T\boldsymbol{\beta}^{(t)} + \mathbf{d}^T\mathbf{U}^{(t)}$. The decision to accept or reject the proposed $\boldsymbol{\beta}^*$ and $\mathbf{U}^*$ is then based entirely on the prior distribution for $\boldsymbol{\beta}$ and the random effects distribution assumed for $\mathbf{U}$. A formal update of $\mathbf{U}$ is still required after this simultaneous update of $\boldsymbol{\beta}$ and $\mathbf{U}$, but this strategy improves the acceptance rate for $\boldsymbol{\beta}$ and thereby facilitates faster mixing. We demonstrate this technique in the example that follows.

4.4.2 Epileptic Seizures Example

We illustrate the use of MCMC to sample from the posterior distribution of a GLMM, and the advantage of the strategy described in Section 4.4.1, with an application to data reported by Thall and Vail (1990) from a clinical trial of 59 epileptics conducted by Leppik et al. (1987). Each subject received either a placebo or the drug progabide and then made four successive follow-up visits to the clinic during which they reported the number of partial seizures they had suffered in the two-week period immediately preceding the visit. We denote these reported counts by Y_{ij}, where $i = 1, \ldots, 59$ indexes the subjects and $j = 1, 2, 3, 4$ indexes the visits. The relevant data can be found in Table A.8.

Thall and Vail (1990) used GEE to fit a marginal model to these data. They included as predictors the natural logarithm of one-fourth of the baseline count of partial seizures suffered by each patient in the eight-week period prior to treatment (denoted BASE$_i$), a treatment indicator (1 if progabide, 0 if placebo, denoted TRT$_i$), the interaction between

BASE$_i$ and TRT$_i$, the natural logarithm of the subject's age in years (denoted AGE$_i$), and a fourth-visit indicator (1 for the subject's fourth post-treatment visit, 0 otherwise, denoted VISIT4$_j$). Breslow and Clayton (1993) and Gamerman (1997) fit a GLMM to these data with the same fixed effects and also two independent random effects, specifying the conditional mean of Y_{ij} as

$$\begin{aligned}\mathrm{E}[Y_{ij}|\boldsymbol{\beta},\gamma_i,\delta_{ij}] = \exp\big(&\beta_0 + \beta_1 \times (\text{BASE}_i) + \beta_2 \times (\text{TRT}_i) + \beta_3 \times (\text{BASE}_i * \text{TRT}_i) + \\ &\beta_4 \times (\text{AGE}_i) + \beta_5 \times (\text{VISIT4}_j) + \gamma_i + \delta_{ij}\big),\end{aligned} \tag{4.13}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5)^T$ is the vector of fixed effects parameters, the $\gamma_i \overset{ind}{\sim} \mathrm{N}(0, \sigma^2)$ are random subject effects, and the $\delta_{ij} \overset{ind}{\sim} \mathrm{N}(0, \tau^2)$ are random effects for visit within subject. Further, conditional on the random effects γ_i and δ_{ij}, the reported seizure counts Y_{ij} are assumed to be independent observations from a $\mathrm{Poisson}(\mathrm{E}[Y_{ij}|\boldsymbol{\beta}, \gamma_i, \delta_{ij}])$.

We adopt a Bayesian approach and sample from a mixed model analogous to (4.13), but include an adjustment to ensure that the model is marginally interpretable. In light of the discussion in Section 2.2.1, the adjustment is simply $a_{ij} = -\sigma^2/2 - \tau^2/2$ for all i and j. We also code the treatment effect as TRT$_i = 1$ for progabide and as TRT$_i = -1$ for placebo to ensure that the two treatment groups are on the same footing in terms of variance. We assume a priori that $\boldsymbol{\beta}$, σ^2, and τ^2 are independent of one another, and place $\mathrm{N}_6(\mathbf{0}, 4\mathbf{I}_6)$, $\mathrm{N}(-1, 2)$, and $\mathrm{N}(-1, 2)$ prior distributions on $\boldsymbol{\beta}$, $\log(\sigma^2)$, and $\log(\tau^2)$, respectively. The priors on the variance components reflect our belief that there is little subject-to-subject and visit-to-visit variation, whereas the priors on the fixed effects parameters are meant to be noninformative while also not putting too much mass on unreasonably large values for the expected seizure count. We sample from the target posterior via MCMC by iteratively updating the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta})^T$, where $\boldsymbol{\beta}$ is the vector of fixed effects parameters, $\boldsymbol{\alpha} = (\sigma^2, \tau^2)^T$ includes the parameters characterizing the random effects distribution, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{59})^T$ includes the 59 latent variables associated with the subject random effect, and $\boldsymbol{\delta} = (\delta_{1,1}, \ldots, \delta_{59,4})^T$ includes the 236 latent variables associated with the visit random effect. Our target posterior is

$$\pi_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{Y}) \propto f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}|\sigma^2) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\tau^2) \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \pi_{\sigma}(\sigma^2) \pi_{\tau}(\tau^2),$$

where $f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta})$ is a product of Poisson densities given by

$$\begin{aligned} f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}) &= \prod_{i=1}^{59} \prod_{j=1}^{4} f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}, \sigma^2, \tau^2, \gamma_i, \delta_{ij}) \\ &= \prod_{i=1}^{59} \prod_{j=1}^{4} \frac{1}{Y_{ij}!} e^{-\left(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \gamma_i + \delta_{ij} - \frac{\sigma^2}{2} - \frac{\tau^2}{2}\right)} \left(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \gamma_i + \delta_{ij} - \frac{\sigma^2}{2} - \frac{\tau^2}{2}\right)^{Y_{ij}}, \end{aligned}$$

$f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}|\sigma^2)$ is a product of normal densities given by

$$f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}|\sigma^2) = \prod_{i=1}^{59} f_{\gamma}(\gamma_i|\sigma^2) = \prod_{i=1}^{59} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\gamma_i^2},$$

$f_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\tau^2)$ is a product of normal densities given by

$$f_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\tau^2) = \prod_{i=1}^{59} \prod_{j=1}^{4} f_{\delta}(\delta_{ij}|\tau^2) = \prod_{i=1}^{59} \prod_{j=1}^{4} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}\delta_{ij}^2},$$

and $\pi_{\boldsymbol{\beta}}$, π_{σ}, and π_{τ} are the prior densities for $\boldsymbol{\beta}$, σ^2, and τ^2, respectively.

A standard approach for sampling from this posterior involves using Metropolis steps to iteratively produce draws from the full conditional distributions of the unknown parameters. The conditional posterior of $\boldsymbol{\alpha} = (\sigma^2, \tau^2)^T$ given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, and $\mathbf{Y}$ is proportional to

$$f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}|\sigma^2) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\tau^2) \pi_{\sigma}(\sigma^2) \pi_{\tau}(\tau^2),$$

and the conditional posterior of $\boldsymbol{\beta}$ given $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, and $\mathbf{Y}$ is proportional to

$$f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}).$$

Further, for each $i = 1, \ldots, 59$, if we define $\boldsymbol{\gamma}_{-i}$ as all elements of $\boldsymbol{\gamma}$ except γ_i, then the conditional posterior of γ_i given $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma_{-i}}$, $\boldsymbol{\delta}$, and $\mathbf{Y}$ is proportional to

$$f_\gamma(\gamma_i|\sigma^2) \prod_{j=1}^{4} f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}, \sigma^2, \tau^2, \gamma_i, \delta_{ij}),$$

Since the γ_i are conditionally independent, $\boldsymbol{\gamma}_{-i}$ does not enter this expression for the full conditional of γ_i. Finally, for each $i = 1, \ldots, 59$ and $j = 1, 2, 3, 4$ the conditional posterior of δ_{ij} given $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}_{-(ij)}$, and $\mathbf{Y}$ is proportional to

$$f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}, \sigma^2, \tau^2, \gamma_i, \delta_{ij}) f_\delta(\delta_{ij}|\tau^2),$$

where $\boldsymbol{\delta}_{-(ij)}$ represents all elements of $\boldsymbol{\delta}$ except δ_{ij}. Since the δ_{ij} are conditionally independent, $\boldsymbol{\delta}_{-(ij)}$ does not enter this expression for the full conditional of δ_{ij}. These full conditional distributions lead naturally to an MCMC algorithm that iteratively performs the following steps:

Basic MCMC Algorithm

1. Propose $\boldsymbol{\alpha}^*$ given $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\gamma}^{(t)}$, and $\boldsymbol{\delta}^{(t)}$. With probability
$$\min\left(1, \frac{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^*, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(t)}|\sigma^{2*}) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}^{(t)}|\tau^{2*}) \pi_\sigma(\sigma^{2*}) \pi_\tau(\tau^{2*})}{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(t)}|\sigma^2_{(t)}) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}^{(t)}|\tau^2_{(t)}) \pi_\sigma(\sigma^2_{(t)}) \pi_\tau(\tau^2_{(t)})}\right)$$
set $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^*$. Otherwise, set $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)}$;

2. Propose $\boldsymbol{\beta}^*$ given $\boldsymbol{\alpha}^{(t+1)}$,$\boldsymbol{\gamma}^{(t)}$, $\boldsymbol{\delta}^{(t)}$. With probability
$$\min\left(1, \frac{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t+1)}), \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}) \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}) \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(t)})}\right)$$
set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^*$. Otherwise, set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$;

3. For $i = 1, \ldots, 59$, propose γ_i^* given $\boldsymbol{\beta}^{(t+1)}$, $\boldsymbol{\alpha}^{(t+1)}$, and $\boldsymbol{\delta}^{(t)}$. With probability
$$\min\left(1, \frac{f_\gamma(\gamma_i^*|\sigma^2_{(t+1)}) \prod_{j=1}^4 f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i^*, \delta_{ij}^{(t)})}{f_\gamma(\gamma_i^{(t)}|\sigma^2_{(t+1)}) \prod_{j=1}^4 f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i^{(t)}, \delta_{ij}^{(t)})}\right)$$
set $\gamma_i^{(t+1)} = \gamma_i^*$. Otherwise, set $\gamma_i^{(t+1)} = \gamma_i^{(t)}$;

4. For $i = 1, \ldots, 59$ and $j = 1, 2, 3, 4$, propose δ_{ij}^* given $\boldsymbol{\beta}^{(t+1)}$, $\boldsymbol{\alpha}^{(t+1)}$, and $\boldsymbol{\gamma}^{(t+1)}$. With probability

$$\min\left(1, \frac{f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i^{(t+1)}, \delta_{ij}^*) f_\delta(\delta_{ij}^*|\tau_{(t+1)}^2)}{f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i^{(t+1)}, \delta_{ij}^{(t)}) f_\delta(\delta_{ij}^{(t)}|\tau_{(t+1)}^2)}\right)$$

set $\delta_{ij}^{(t+1)} = \delta_{ij}^*$. Otherwise, set $\delta_{ij}^{(t+1)} = \delta_{ij}^{(t)}$.

For each step in this algorithm, we propose new values for the parameters given their current values. The proposed values are random draws from the following distributions:

$$\boldsymbol{\beta}^* \sim \mathrm{N}_6(\boldsymbol{\beta}^{(t)}, \mathbf{V});$$

$$\log(\sigma^{2*}) \sim \mathrm{N}\big(\log(\sigma_{(t)}^2), 0.16\big); \;\; \log(\tau^{2*}) \sim \mathrm{N}\big(\log(\tau_{(t)}^2), 0.0625\big);$$

$$\gamma_i^* \sim \mathrm{N}(\gamma_i^{(t)}, 0.25); \;\; \delta_{ij}^* \sim \mathrm{N}(\delta_{ij}^{(t)}, 0.25).$$

The covariance matrix $\mathbf{V}$ for proposing $\boldsymbol{\beta}^*$ is calculated using the weight matrix obtained from fitting an analogous fixed effects model with iteratively reweighted least squares. Specifically, we define

$$\mathbf{V} = \begin{bmatrix} 0.1735 & -0.0061 & 0.0068 & -0.0041 & -0.0476 & -0.0007 \\ -0.0061 & 0.0010 & -0.0002 & 0.0001 & 0.0011 & 0.0000 \\ 0.0068 & -0.0002 & 0.0061 & -0.0024 & -0.0019 & 0.0000 \\ -0.0041 & 0.0001 & -0.0024 & 0.0010 & 0.0012 & 0.0000 \\ -0.0476 & 0.0011 & -0.0019 & 0.0012 & 0.0136 & 0.0000 \\ -0.0007 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0030 \end{bmatrix}.$$

The general form of the acceptance probability for each set of proposals is given by (4.12), but since all of the proposal distributions are symmetric none of the acceptance probabilities are influenced by the proposal densities. Rather, the acceptance probability in each case reduces to a ratio of the full conditional with the proposed parameter values to the full conditional with the current parameter values.

Since our model includes 295 latent variables, this basic MCMC algorithm is plagued by the issues with slow mixing discussed in Section 4.4.1. We ran this algorithm for 2,100,000 steps, discarding the first 100,000 steps as burn-in. Only 12.9% of the proposals for $\boldsymbol{\beta}$ were accepted, and there is a high degree of autocorrelation in the Markov chains for the fixed effects parameters. Autocorrelation plots for $\beta_0, \ldots, \beta_5$ are shown in Figure 4.2. With the exception of β_5, all of these plots indicate a correlation greater than 0.1 at a lag of 200. To address this problem, we simultaneously propose $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$ to be consistent with each proposed $\boldsymbol{\beta}^*$. Specifically, for each $\boldsymbol{\beta}^*$ we also propose the following γ_i^* and δ_{ij}^* for each $i = 1, \ldots, 59$ and $j = 1, 2, 3, 4$:

$$\gamma_i^* = \gamma_i^{(t)} + \mathbf{x}_{i1}^T(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*), \tag{4.14}$$

$$\delta_{i4}^* = \delta_{i4}^{(t)} + (\mathbf{x}_{i4}^T - \mathbf{x}_{i1}^T)(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*), \text{ and } \delta_{ij}^* = \delta_{ij}^{(t)} \text{ for } j = 1, 2, 3. \tag{4.15}$$

This simultaneous proposal of $\boldsymbol{\beta}^*$, $\boldsymbol{\gamma}^*$, and $\boldsymbol{\delta}^*$ has no net impact on the likelihood. To make this clear, the conditional density of Y_{ij} given μ_{ij} is $\mathrm{Poisson}(\mu_{ij})$, where

$$\mu_{ij} = \mathrm{E}[Y_{ij}|\boldsymbol{\beta}, \gamma_i, \delta_{ij}] = \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma_i + \delta_{ij} + a).$$

Further, $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_{i3}$ because only the fourth-visit indicator varies within a subject. Defining γ_i^* and δ_{ij}^* as in (4.14) and (4.15), for $j = 1, 2, 3$

$$\begin{aligned}
&\mathbf{x}_{ij}^T\boldsymbol{\beta}^* + \gamma_i^* + \delta_{ij}^* + a \\
&= \mathbf{x}_{ij}^T\boldsymbol{\beta}^* + \left(\gamma_i^{(t)} + \mathbf{x}_{ij}^T(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*)\right) + \delta_{ij}^{(t)} + a \\
&= \mathbf{x}_{ij}^T\boldsymbol{\beta}^* - \mathbf{x}_{ij}^T\boldsymbol{\beta}^* + \mathbf{x}_{ij}^T\boldsymbol{\beta}^{(t)} + \gamma_i^{(t)} + \delta_{ij}^{(t)} + a \\
&= \mathbf{x}_{ij}^T\boldsymbol{\beta}^{(t)} + \gamma_i^{(t)} + \delta_{ij}^{(t)} + a,
\end{aligned}$$

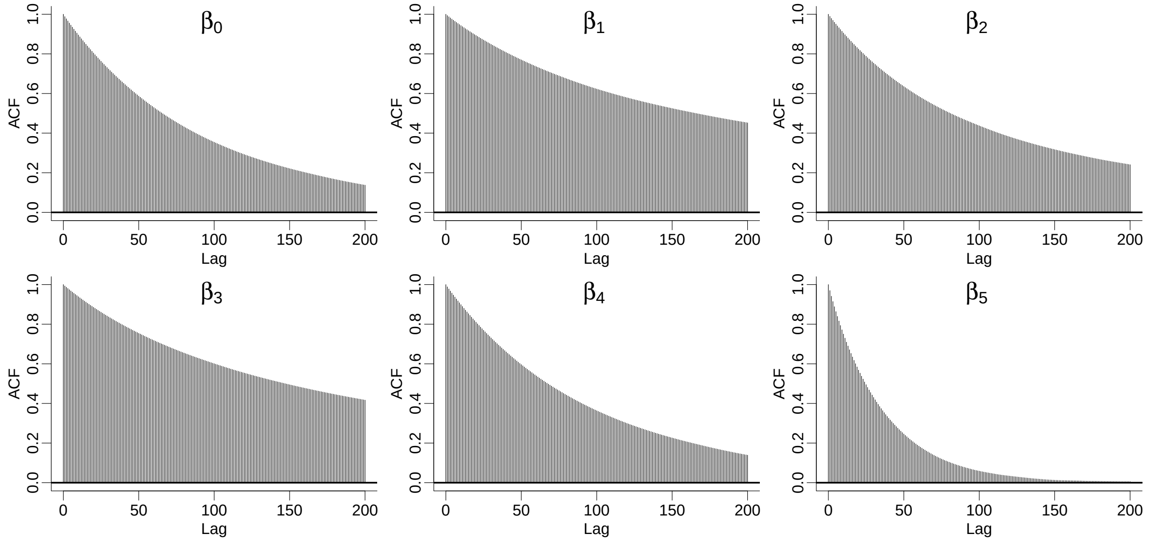

Figure 4.2: Autocorrelation plots for $\beta_0, \ldots, \beta_5$ for a basic MCMC algorithm for the epileptic seizures data found in Thall and Vail (1990)

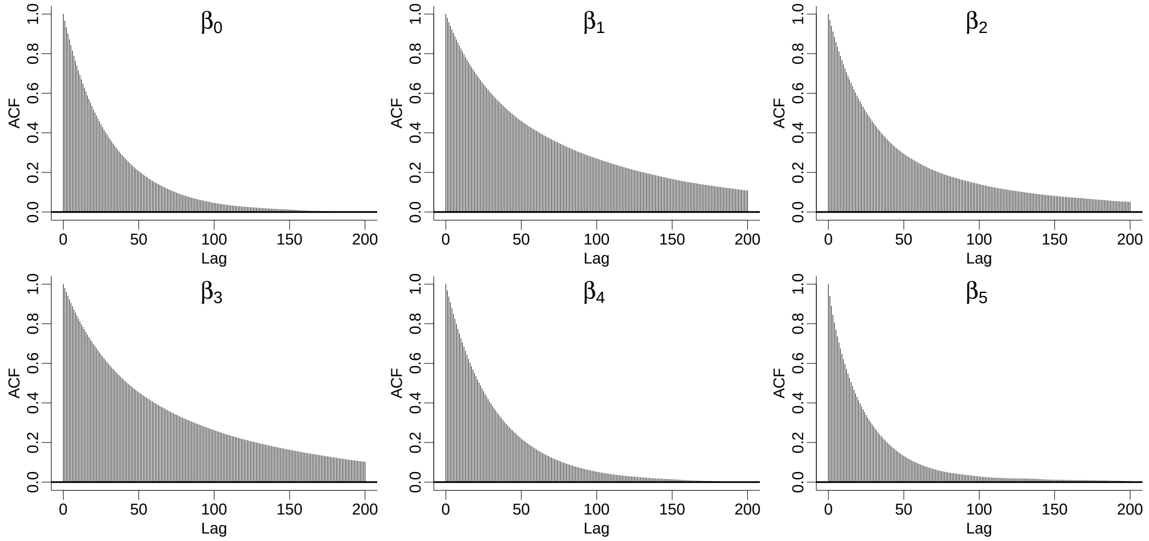

Figure 4.3: Autocorrelation plots for $\beta_0, \ldots, \beta_5$ for the modified MCMC algorithm for the epileptic seizures data found in Thall and Vail (1990)

and for $j = 4$

$$\begin{aligned}
&\mathbf{x}_{ij}^T\boldsymbol{\beta}^* + \gamma_i^* + \delta_{ij}^* + a \\
&= \mathbf{x}_{ij}^T\boldsymbol{\beta}^* + \big(\gamma_i^{(t)} + \mathbf{x}_{i1}^T(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*)\big) + \big(\delta_{ij}^{(t)} + (\mathbf{x}_{ij}^T - \mathbf{x}_{i1}^T)(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*)\big) + a \\
&= \mathbf{x}_{ij}^T\boldsymbol{\beta}^* + \mathbf{x}_{i1}^T(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*) - \mathbf{x}_{i1}^T(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*) + \mathbf{x}_{ij}^T(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*) + \gamma_i^{(t)} + \delta_{ij}^{(t)} + a \\
&= \mathbf{x}_{ij}^T\boldsymbol{\beta}^* - \mathbf{x}_{ij}^T\boldsymbol{\beta}^* + \mathbf{x}_{ij}^T\boldsymbol{\beta}^{(t)} + \gamma_i^{(t)} + \delta_{ij}^{(t)} + a \\
&= \mathbf{x}_{ij}^T\boldsymbol{\beta}^{(t)} + \gamma_i^{(t)} + \delta_{ij}^{(t)} + a.
\end{aligned}$$

Consequently, the conditional mean $\mathrm{E}[Y_{ij}|\boldsymbol{\beta}, \gamma_i, \delta_{ij}]$ is the same for both the current state and the proposed state. In turn, the conditional density $f_{\mathbf{Y}|\boldsymbol{\theta}}$ is also the same for both states. Noting that the conditional posterior of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})^T$ given $\boldsymbol{\alpha}$ and $\mathbf{Y}$ is proportional to

$$f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}|\sigma^2) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\tau^2) \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}),$$

the acceptance probability for $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^T$ depends entirely on $\pi_{\boldsymbol{\beta}}$, $f_{\boldsymbol{\gamma}}$, and $f_{\boldsymbol{\delta}}$ because $f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\gamma}^*, \boldsymbol{\delta}^*) = f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)})$. This joint proposal scheme leads to the following modified MCMC algorithm:

Modified MCMC Algorithm

1. Propose $\boldsymbol{\alpha}^*$ given $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\gamma}^{(t)}$, and $\boldsymbol{\delta}^{(t)}$. With probability
$$\min\left(1, \frac{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^*, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(t)}|\sigma^{2*}) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}^{(t)}|\tau^{2*}) \pi_\sigma(\sigma^{2*}) \pi_\tau(\tau^{2*})}{f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(t)}|\sigma^2_{(t)}) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}^{(t)}|\tau^2_{(t)}) \pi_\sigma(\sigma^2_{(t)}) \pi_\tau(\tau^2_{(t)})}\right)$$
set $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^*$. Else, set $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)}$;

2. Propose $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^T$ given $\boldsymbol{\alpha}^{(t+1)}$, with $\boldsymbol{\beta}^*$ being drawn from a proposal distribution and $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$ defined as in (4.14) and (4.15). With probability
$$\min\left(1, \frac{f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^*|\sigma^2_{(t+1)}) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}^*|\tau^2_{(t+1)}) \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(t)}|\sigma^2_{(t+1)}) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}^{(t)}|\tau^2_{(t+1)}) \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(t)})}\right)$$
set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^*$, $\boldsymbol{\gamma}' = \boldsymbol{\gamma}^*$, and $\boldsymbol{\delta}' = \boldsymbol{\delta}^*$. Else, set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$, $\boldsymbol{\gamma}' = \boldsymbol{\gamma}^{(t)}$, and $\boldsymbol{\delta}' = \boldsymbol{\delta}^{(t)}$;

3. For $i = 1, \ldots, 59$, propose γ_i^* given $\boldsymbol{\beta}^{(t+1)}$, $\boldsymbol{\alpha}^{(t+1)}$, and $\boldsymbol{\delta}'$. With probability

$$\min\left(1, \frac{f_\gamma(\gamma_i^*|\sigma^2_{(t+1)}) \prod_{j=1}^4 f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i^*, \delta'_{ij})}{f_\gamma(\gamma_i'|\sigma^2_{(t+1)}) \prod_{j=1}^4 f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i', \delta'_{ij})}\right)$$

set $\gamma_i^{(t+1)} = \gamma_i^*$. Else, set $\gamma_i^{(t+1)} = \gamma_i'$;

4. For $i = 1, \ldots, 59$ and $j = 1, 2, 3, 4$, propose δ_{ij}^* given $\boldsymbol{\beta}^{(t+1)}$, $\boldsymbol{\alpha}^{(t+1)}$, and $\boldsymbol{\gamma}^{(t+1)}$. With probability

$$\min\left(1, \frac{f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i^{(t+1)}, \delta_{ij}^*) f_\delta(\delta_{ij}^*|\tau^2_{(t+1)})}{f_{Y|\boldsymbol{\theta}}(Y_{ij}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \gamma_i^{(t+1)}, \delta'_{ij}) f_\delta(\delta'_{ij}|\tau^2_{(t+1)})}\right)$$

set $\delta_{ij}^{(t+1)} = \delta_{ij}^*$. Else, set $\delta_{ij}^{(t+1)} = \delta'_{ij}$.

The key difference between this modified algorithm and the basic algorithm is in Step 2. Here, in addition to updating $\boldsymbol{\beta}^{(t)}$ to $\boldsymbol{\beta}^{(t+1)}$, we update $\boldsymbol{\gamma}^{(t)}$ and $\boldsymbol{\delta}^{(t)}$ to the intermediate states $\boldsymbol{\gamma}'$ and $\boldsymbol{\delta}'$. Steps 3 and 4 then update $\boldsymbol{\gamma}'$ and $\boldsymbol{\delta}'$ to $\boldsymbol{\gamma}^{(t+1)}$ and $\boldsymbol{\delta}^{(t+1)}$ in a manner similar to Steps 3 and 4 in the basic MCMC algorithm.

Using the modified MCMC algorithm instead of the basic MCMC algorithm increases the acceptance rate for $\boldsymbol{\beta}$ from 12.9% to 50.7%. It also decreases the *integrated autocorrelation times* for the fixed effects parameters, as summarized in Table 4.1. The integrated autocorrelation time of a parameter provides a measure of the average number of iterations required to obtain approximately independent draws from the posterior distribution of that parameter. For $\beta_0, \ldots, \beta_4$, this quantity is more than three times larger using the basic MCMC algorithm versus using the modified MCMC algorithm. Thus, the strategy described in Section 4.4.1 successfully reduces the autocorrelation and allows us to obtain a representative sample from the target posterior with fewer steps of the Markov chain. As further illustration of the improved mixing, Figure 4.3 (displayed below Figure 4.2 for

Table 4.1: Integrated autocorrelation times (before thinning) for the fixed effects parameters in the marginally interpretable model for the epileptic seizures data found in Thall and Vail (1990) for both the basic MCMC algorithm and the modified MCMC algorithm

Parameter	Basic MCMC	Modified MCMC	Ratio
β_0	205.3	62.8	3.27
β_1	642.7	161.9	3.97
β_2	355.2	105.0	3.38
β_3	565.6	168.0	3.37
β_4	204.2	65.6	3.11
β_5	71.5	51.0	1.40

comparison) shows autocorrelation plots for $\beta_0, \ldots, \beta_5$ using the modified MCMC algorithm. Although some autocorrelation remains, even at a lag of 200, it is not as strong as the autocorrelation observed in Figure 4.2 for the basic MCMC algorithm.

We carried out the modified MCMC algorithm both for a marginally interpretable GLMM and a conventional GLMM with a conditional mean structure resembling (4.13). We ran each chain for 2,100,000 steps, discarding the first 100,000 steps as burn-in and retaining every 200^{th} step thereafter to obtain a final sample of 10,000 draws from the posterior distribution for each model. Table 4.2 displays posterior means and standard deviations for the unknown parameters in both the marginally interpretable model and the conventional GLMM. With the exception of the intercept β_0, the two sets of parameter estimates are virtually identical. These results are consistent with the discussion in Section 3.1.2 regarding the relationship between the marginal and subject-specific parameters for a model with a log link. They are also consistent with the claims of Breslow and Clayton (1993) and Ritz and Spiegelman (2004) that the slope parameters in this model, and any other model with a log link and a random intercept that is independent of the covariates in the model, have both a marginal and subject-specific interpretation. The intercept β_0 is

Table 4.2: Posterior means of the unknown parameters in the model for the epileptic seizures data found in Thall and Vail (1990) (with corresponding posterior standard deviations in parentheses)

Parameter	Marginally Interpretable GLMM	Conventional GLMM
β_0	-1.15 (1.09)	-1.29 (1.08)
β_1	1.04 (0.10)	1.04 (0.11)
β_2	-0.45 (0.21)	-0.45 (0.21)
β_3	0.16 (0.11)	0.16 (0.10)
β_4	0.33 (0.31)	0.32 (0.31)
β_5	-0.10 (0.09)	-0.10 (0.09)
σ	0.50 (0.07)	0.50 (0.07)
τ	0.37 (0.04)	0.37 (0.04)

greater for the marginally interpretable GLMM than for the conventional GLMM due to the tendency of the convex inverse link function to pull the marginal mean up.

Ultimately, we are able to obtain draws from the posterior distribution of the unknown parameters in our model, and can use these draws to make inference on the unknown parameters as described in Section 3.3. A challenge is presented by the poor mixing that results from the inclusion of many latent variables, but we are able to overcome this difficulty by modifying our proposal scheme using the strategy described in Section 4.4.1. Note that this challenge stems from the model being a GLMM, not from the model being marginally interpretable. Incorporating an adjustment to make the model marginally interpretable is not computationally difficult, and is advisable if one wants to make inference on population-level quantities.

Chapter 5: Discussion

We have defined a class of marginally interpretable GLMMs for which the marginal mean has a specific form after integration over the random effects distribution. Specifying a GLMM in this form yields fixed effects parameters that can be interpreted as the average effect of a covariate across the entire population. This is in contrast to a conventional GLMM, which yields fixed effects parameters that must be interpreted conditional on a specific realization of the random effects. The distinction here is between the average effect across the entire population and the effect for an average individual in the population. Due to Jensen's inequality, these two effects are not the same when the link function is nonlinear. By introducing an additive adjustment to the model that effectively shifts the location of the random effects distribution, we are able to counteract the curvature of the inverse link function and ensure that the marginal mean has the desired form.

Marginally interpretable GLMMs can be fit using techniques designed for conventional GLMMs with only minor modifications. Basically, one needs to include an appropriate adjustment in the expression for the conditional mean when evaluating the likelihood function. In some cases, the adjustment simply reparameterizes the model, meaning that the fixed effects parameters change but the fit of the model does not. In other cases, the adjustment fundamentally changes the structure of the model. We have derived the form of the adjustment for several popular link functions and have also provided an efficient algorithm for computing the adjustment in the commonly used logistic-normal model.

We have shown through several examples that inferences obtained from the two parameterizations of the GLMM (marginally interpretable and conventional) can be markedly different, even when the two versions of the model provide equal fit to the data. We argue that marginal effects are often of greater interest than cluster-specific effects, but acknowledge that cluster-specific effects could be useful in some settings. It is therefore advantageous to obtain marginal parameter estimates through a fully specified GLMM as opposed to a purely marginal model that only specifies the first two moments of the data. Unlike marginal models fit via GEE, marginally interpretable GLMMs have an underlying probabilistic model that can be used to make individual-level predictions in addition to marginal inferences. The likelihood function associated with a marginally interpretable GLMM also allows for model checking and model comparisons that are not possible with purely marginal models.

When the marginally interpretable and conventional formulations of the model are equivalent, as defined in Section 3.1, we note that there is only a discrepancy between hypothesis tests for the two models if one uses a Wald test or a score test. In these situations, a likelihood ratio test yields the same result whether a marginally interpretable GLMM or a conventional GLMM is fit. This reflects earlier findings in the literature (e.g. Jennings, 1986), which state that likelihood-based inferences are invariant under one-to-one transformations of the parameters. For the specific case of logistic regression, others (e.g. Hauck and Donner, 1977) have noted aberrant behavior by the Wald test under certain conditions and have suggested use of the likelihood ratio test instead. Thus, the fact that it is possible to conduct a likelihood ratio test in a GLMM framework is another advantage of using a fully specified model rather than a purely marginal model.

Marginally interpretable GLMMs can be used in many settings. One possible application is to *meta-analysis*. The objective of meta-analysis is to combine the results of several related studies into a single integrated analysis that provides information about the average effect of a treatment across all studies. A natural approach is to fit a model that includes random effects for the individual studies to account for heterogeneity across studies. Linear mixed models for meta-analysis are discussed, for example, by DerSimonian and Laird (1986), Berkey et al. (1995), and Stram (1996). These models often require a transformation of the response to satisfy (at least approximately) an underlying assumption of normality. An alternative approach is to perform a meta-analysis using a GLMM (e.g. Aitkin, 1999b; Platt et al., 1999; Turner et al., 2000). A conventional GLMM would yield parameter estimates that pertain to the fixed effects for a specific study, but such estimates of these effects are already available from the individual studies comprising the meta-analysis. The purpose of a meta-analysis is to investigate the average magnitude of an effect across a collection of studies. Marginally interpretable GLMMs focus on such population-averaged effects are therefore well-suited for this purpose.

Many of the examples we have provided relate to models with Gaussian random effects, in large part because it is common to assume that random effects follow a Gaussian distribution. However, a marginally interpretable GLMM does not require normal random effects and the techniques for fitting these models apply to a wide array of random effects distributions. One interesting class of random effects distributions consists of mixtures of normal distributions. Mixed models that represent the random effects distribution as a mixture of normals (see Magder and Zeger, 1996; Caffo et al., 2007; Komárek and Lesaffre, 2008) allow considerable flexibility in the shape of the random effects distribution and could be incorporated fairly easily into the framework of a marginally interpretable GLMM.

Whereas a mixture of normals can be viewed as a semiparametric distribution, even more flexible random effects distributions can be obtained in a nonparametric setting. There is an extensive literature devoted to GLMMs for which the random effects distribution is estimated in a nonparametric manner. These approaches aim to mitigate bias in the fixed effects parameter estimates that might arise from misspecification of the random effects distribution. As examples of this strategy, Follmann and Lambert (1989), Butler and Louis (1992), and Aitkin (1999a) treated the random effects distribution as a discrete mixture following the nonparametric maximum likelihood estimation approach of Laird (1978), whereas Chen et al. (2002) and Ghosh et al. (2007) assumed that the random effects distribution has a smooth density and used the semi-nonparametric maximum likelihood approach of Gallant and Nychka (1987). Additionally, Kleinman and Ibrahim (1998) and Antonelli et al. (2016), among others, used a Dirichlet process prior for the random effects distribution to fit GLMMs in a nonparametric Bayesian framework.

Fitting a marginally interpretable model with a nonparametric random effects distribution would require careful consideration of the computational strategy. The adjustment in such a model would not have a closed form, but could conceivably be calculated by placing appropriate constraints on the model's marginal mean. That being said, such nonparametric specification of the random effects distribution may not be necessary in a marginally interpretable GLMM. First, as discussed in Section 1.4, several studies have shown that bias in the fixed effects parameters due to misspecification of the random effects distribution is typically small. Further, by focusing on population-averaged effects instead of effects that are conditioned on the latent random variables, the marginally interpretable model reduces the dependence of the fixed effects parameters on the random effects distribution. This was demonstrated by Heagerty and Kurland (2001) for the closely related marginalized multilevel model. Thus, even though one could feasibly fit a marginally interpretable GLMM with a flexible, nonparametric random effects distribution, it may not necessarily be beneficial to do so.

Our focus has been on marginally interpretable models for which correlation is introduced through latent random variables. This corresponds to situations wherein dependence among the observations arises from the presence of groups or clusters in the data. An obvious extension is to more complex dependence structures; for example, time series data. Cox (1981) classified models for time series data into two categories, *observation driven* models and *parameter driven* models, which are separated by how correlation is introduced into the model. Observation driven models introduce dependence by conditioning on earlier observations and are referred to by Diggle et al. (2002) as *transition models*. We use Y_{it} to denote the response at time t for unit i, which is observed over time. A transition model relates the response Y_{it} to a set of covariates $\mathbf{x}_{it}$ conditional on all of the previously observed responses $Y_{i,1}, \ldots, Y_{i,t-1}$ (and also possibly the previously observed covariates $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,t-1}$) for that unit. In contrast, parameter driven models resemble GLMMs, but the latent random variables are assumed to comprise a random process and have a specific type of correlation structure.

Observation driven models for time series data are discussed, for example, by Zeger and Qaqish (1988) and Davis et al. (2003). The most common transition models are Markov models that assume the current response Y_{it} for unit i depends only on the most recent q

responses $Y_{i,t-q}, \ldots, Y_{i,t-1}$. For example, a first-order Markov model expresses the conditional mean of Y_{it} given the previously observed responses $Y_{i,1}, \ldots, Y_{i,t-1}$ as

$$\mathrm{E}[Y_{it}|Y_{i,1}, \ldots, Y_{i,t-1}] = \mathrm{E}[Y_{it}|Y_{i,t-1}] = h(\mathbf{x}_{it}^T\boldsymbol{\beta} + \gamma_{it}Y_{i,t-1}),$$

where $h(\cdot)$ is an inverse link function, $\boldsymbol{\beta}$ is a vector of fixed effects parameters, and each γ_{it} is a parameter characterizing the relationship between Y_{it} and $Y_{i,t-1}$. Much like in a mixed model, the parameters $\boldsymbol{\beta}$ in this Markov model are not equivalent to the parameters in a corresponding marginal model; they must be interpreted conditional on the value of the previous response. To obtain marginal parameters from a Markov model, Zeger and Qaqish (1988) suggested using an estimating equations approach. Azzalini (1994) proposed a version of a first-order Markov model with marginal parameters that can be fit using maximum likelihood estimation. Heagerty and Zeger (2000) called the model of Azzalini (1994) a *marginalized transition model* and Heagerty (2002) extended it to higher-order Markov models. For first-order Markov dependence, the model of Heagerty (2002) expresses the conditional mean as

$$\mathrm{E}[Y_{it}|Y_{i,t-1}] = h(\Delta_{it} + \gamma_{it}Y_{i,t-1}),$$

where Δ_{it} is defined implicitly such that the marginal mean satisifies $\mathrm{E}[Y_{it}] = h(\mathbf{x}_{it}^T\boldsymbol{\beta})$. This model is analogous to the marginalized multilevel model of Heagerty (1999) and Heagerty and Zeger (2000), and fits our definition of a marginally interpretable model. Heagerty (2002) showed that $\boldsymbol{\beta}$ and the γ_{it} are orthogonal to one another in this model, meaning that assumptions made about the dependence structure have no impact on estimation of the fixed effects parameters. Heagerty (2002) also provided details of a Newton-Raphson algorithm for fitting marginalized transition models via maximum likelihood estimation.

Parameter driven models for time series data more closely resemble the GLMMs that have been the focus of this dissertation. For a discussion of models of this type see Davis and Wu (2009) and the references therein. The conditional mean for such a model might be expressed as

$$\mathrm{E}[Y_{it}|U_t] = h(\mathbf{x}_{it}^T\boldsymbol{\beta} + U_t),$$

where the U_t comprise a first-order autoregressive process for which $U_t = \gamma U_{t-1} + \varepsilon_t$. Here, γ is an unknown parameter and, if we have an autoregressive Gaussian process, the ε_t are independent Gaussian random variables. Maximum likelihood approaches that employ MCEM have been used to fit this type of model to count data (Chan and Ledolter, 1995) and to binary outcome data (Klingenberg, 2008). Due to the conditioning on the latent random process, the parameters $\boldsymbol{\beta}$ in these models do not have a marginal interpretation. Zeger (1988) proposed fitting parameter driven models for time series count data using an estimating equations approach to model the marginal mean $\mathrm{E}[Y_{it}] = h(\mathbf{x}_{it}^T\boldsymbol{\beta})$. This is essentially a marginal model fit via GEE and, although it yields population-averaged parameter estimates, it is lacking as a formal statistical model because only the first two moments of the data are specified. The computational strategies described in Chapter 4 could be used to fit a marginally interpretable model in this context, but the dependence among the latent random variables adds complexity to the integration required to compute both the marginal likelihood and the adjustment. Nonetheless, a strategy could be developed, possibly using MCEM, that incorporates an adjustment into the expression for the conditional mean and thereby leads to parameters that have the desired marginal interpretation.

Another form of dependence to which the notion of a marginally interpretable model could be extended is spatial dependence. Much like data correlated in time, data correlated in space can be modeled using either an observation driven or a parameter driven approach. The observation driven approach includes autoregressive models such as those discussed by Besag (1974). Here, instead of conditioning on the most recent observations as in the time series setting, one conditions on observations that fall within a *neighborhood* of the observation of interest. The neighborhood could be defined based on distance, shared boundaries, or some other criterion. The marginalized transition model developed by Heagerty (2002) could conceivably be extended to this setting. A more natural extension of the marginally interpretable GLMM is to parameter driven models, which assume the correlation in the data arises from a latent spatial random process. Clayton and Kaldor (1987) and Diggle et al. (1998) introduced spatial models with latent random processes, but the parameters in their models have a conditional interpretation. Albert and McShane (1995) and McShane et al. (1997) discussed a GEE approach to obtaining marginal parameters from such models, while Yasui and Lele (1997) and Gotway and Wolfinger (2003) compared marginal and conditional approaches. As with time series data, the methods we developed to preserve the marginal mean in models with latent random variables should extend to models for spatial data with latent random processes. The correlation structure is even more complicated in the spatial setting, and the increase in computational complexity is nontrivial. For a spatial GLMM in which the latent process is a Gaussian process, INLA is a potential strategy for overcoming the computational challenges associated with fitting a marginally interpretable version of the model.

Our definition of a marginally interpretable GLMM requires a specific relationship between the conditional mean and the marginal mean such that the marginal mean is preserved after integrating out the random effects. Acknowledging that the marginal mean may not always be of interest, we can define marginally interpretable models in other mixed model settings. For example, in mixed effects quantile regression models (see Koenker, 2004;

Geraci and Bottai, 2014) or when modeling extremes (see Coles, 2001; Stephenson and Tawn, 2004) we can consider a definition for a marginally interpretable model based on relating certain conditional quantiles to corresponding marginal quantiles. Further research is needed to understand the form of the adjustments that arise in these settings.

In conclusion, marginally interpretable models have many desirable properties and also potentially wide applicability. Extending the notion of a marginally interpretable model beyond the GLMM framework that was focused on here is certainly possible and is an area for future research. The guiding principle of a marginally interpretable model is that components of the model that are included to account for correlation among observations (such as latent random variables) should not substantively change other parts of the model (namely the mean structure). Preserving the form of these other parts of the model is a sensible thing to do, and leads to more robust inferences that focus on quantities that are ordinarily of interest to the researcher.

References

Abramowitz, M. and Stegun, I. A., editors (1972). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. U.S. Government Printing Office, Washington, D.C., 10th edition.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis*, 47:639–653.

Aitchison, J. and Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, 29:813–828.

Aitkin, M. (1999a). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128.

Aitkin, M. (1999b). Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, 18:2343–2351.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Czáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Akademiai Kiadó, Budapest.

Albert, P. S. and McShane, L. M. (1995). A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. *Biometrics*, 51:627–638.

Antonelli, J., Trippa, L., and Haneuse, S. (2016). Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics. *Statistical Science*, 31:80–95.

Asmussen, S., Jensen, J. L., and Rojas-Nandayapa, L. (2016). On the Laplace transform of the lognormal distribution. *Methodology and Computing in Applied Probability*, 18:441–458.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, 81:767–775.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., and Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, 14:395–411.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36:192–236.

Boehm, L., Reich, B. J., and Bandyopadhyay, D. (2013). Bridging conditional and marginal inference for spatially referenced binary data. *Biometrics*, 69:545–554.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24:127–135.

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

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

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, 36:153–157.

Butler, S. M. and Louis, T. A. (1992). Random effects models with non-parametric priors. *Statistics in Medicine*, 11:1981–2000.

Caffo, B., An, M.-W., and Rohde, C. (2007). Flexible random intercept models for binary outcomes using mixture of normals. *Computational Statistics and Data Analysis*, 51:5220–5235.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, 90:242–252.

Chen, J., Zhang, D., and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics*, 3:347–360.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, New York, NY.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, 8:93–115.

Crouch, E. A. C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)exp(-t^2)dt$: Application to logistic-normal models. *Journal of the American Statistical Association*, 85:464–469.

Crowder, M. J. (1978). Beta-binomial Anova for proportions. *Journal of the Royal Statistical Society: Series C*, 27:34–37.

Davis, R. A., Dunsmuir, W. T. M., and Streett, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika*, 90:777–790.

Davis, R. A. and Wu, R. (2009). A negative binomial model for time series of counts. *Biometrika*, 96:735–749.

de Bruijn, N. G. (1961). *Asymptotic Methods in Analysis*. Dover, New York, NY.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42:204–223.

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

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C*, 47:299–350.

Dunson, D. B. and Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6:11–25.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York, NY.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40:1–18.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 51:309–317.

Follmann, D. A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84:295–300.

Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55:363–390.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7:57–68.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82:479–488.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24:461–479.

Ghosh, S., Das, K., and Congdon, P. (2007). Analysis of marginally specified semi-nonparametric models for clustered binary data. *Statistica Neerlandica*, 61:292–304.

Givens, G. H. and Hoeting, J. A. (2013). *Computational Statistics*. Wiley, Hoboken, NJ.

Golub, G. H. and Welsch, J. H. (1969). Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23:221–230.

González, J., Tuerlinckx, F., Boeck, P. D., and Cools, R. (2006). Numerical integration in logistic-normal models. *Computational Statistics and Data Analysis*, 51:1535–1548.

Gotway, C. A. and Wolfinger, R. D. (2003). Spatial prediction of counts and rates. *Statistics in Medicine*, 22:1415–1432.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B*, 46:149–192.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55:245–259.

Grömping, U. (1996). A note on fitting a marginal model to mixed effects log-linear regression data via GEE. *Biometrics*, 52:280–285.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hauck, W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72:851–853.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55:688–698.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, 58:342–351.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88:973–985.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15:1–19.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15:192–218.

Hewson, C. M., Thorup, K., Pearce-Higgins, J. W., and Atkinson, P. W. (2016). Population decline is linked to migration route in the Common Cuckoo. *Nature Communications*, 7:12296.

Hubbard, A. E., Ahern, J., Fleischer, N. L., der Laan, M. V., Lippman, S. A., Jewell, N., Bruckner, T., and Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 21:467–474.

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, 81:471–476.

Jewell, N. P. and Shiboski, S. C. (1990). Analysis of HIV infectivity based on partner studies. *Biometrics*, 46:1133–1150.

Jones, B. and Kenward, M. G. (1989). *Design and Analysis of Cross-Over Trials*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition.

Kim, Y. and Kim, D. (2011). Posterior consistency of random effects models for binary data. *Journal of Statistical Planning and Inference*, 141:3391–3399.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, 17:2579–2596.

Klingenberg, B. (2008). Regression models for binary time series with gaps. *Computational Statistics and Data Analysis*, 52:4076–4090.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91:74–89.

Komárek, A. and Lesaffre, E. (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis*, 52:3441–3458.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society: Series B*, 57:395–407.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

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

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

Larsen, K., Petersen, J. H., Budtz-Jørgensen, E., and Endahl, L. (2000). Interpreting parameters in the logistic regression model with random effects. *Biometrics*, 56:909–914.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19:219–228.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York, NY.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, NY, 3rd edition.

Leppik, I., Dreifuss, F., Porter, R., Bowman, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stackman, J., Graves, N., Sutula, T., Welty, T., Vickery, J., Brundage, R., Gates, J., Gumnit, R., and Gutierrez, A. (1987). A controlled study of progabide in partial seizures: Methodology and results. *Neurology*, 37:963–968.

Liang, K.-Y. and Hanfelt, J. (1994). On the use of the quasi-likelihood method in teratological experiments. *Biometrics*, 50:872–880.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society: Series B*, 54:3–40.

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

Lindsey, J. K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measures in clinical trials. *Statistics in Medicine*, 17:447–469.

Litière, S., Alonso, A., and Molenberghs, G. (2007). Type I and Type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63:1038–1044.

Litière, S., Alonso, A., and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*, 27:3125–3144.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81:624–629.

Magder, L. S. and Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*, 91:1141–1151.

Mancl, L. A. and Leroux, B. G. (1996). Efficiency of regression estimates for clustered data. *Biometrics*, 52:500–511.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition.

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

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, 26:388–402.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. Wiley, Hoboken, NJ, 2nd edition.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society: Series B*, 56:61–69.

McShane, L. M., Albert, P. S., and Palmatier, M. A. (1997). A latent process regression model for spatially correlated count data. *Biometrics*, 53:698–706.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341.

Miglioretti, D. L. and Heagerty, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics*, 5:381–398.

Monahan, J. F. and Stefanski, L. A. (1992). Normal scale mixture approximations to f*(z) and computation of the logistic-normal integral. In Balakrishnan, N., editor, *Handbook of the Logistic Distribution*, pages 529–540. Marcel Dekker, New York, NY.

Naylor, J. C. and Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society: Series C*, 31:214–225.

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

Neuhaus, J. M. (1993). Efficiency and tests of covariate effects with clustered binary data. *Biometrics*, 49:989–996.

Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79:755–762.

Neuhaus, J. M. and Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, 80:807–815.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59:25–35.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *The Canadian Journal of Statistics*, 22:139–148.

Neuhaus, J. M., McCulloch, C. E., and Boylan, R. (2011). A note on Type II error under random effects misspecification in generalized linear mixed models. *Biometrics*, 67:654–660.

Neuhaus, J. M., McCulloch, C. E., and Boylan, R. (2013). Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Statistics in Medicine*, 32:2419–2429.

Neyman, J. and Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A:175–240.

Neyman, J. and Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, 20A:263–294.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Parzen, M., Ghosh, S., Lipsitz, S., Sinha, D., Fitzmaurice, G. M., Mallick, B. K., and Ibrahim, J. G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *Annals of Applied Statistics*, 5:449–467.

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

Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15:58–81.

Pirjol, D. (2013). The logistic-normal integral and its generalizations. *Journal of Computational and Applied Mathematics*, 237:460–469.

Platt, R. W., Leroux, B. G., and Breslow, N. (1999). Generalized linear mixed models for meta-analysis. *Statistics in Medicine*, 18:643–654.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–839.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimates of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1–21.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44:50–57.

Ritz, J. and Spiegelman, D. (2004). Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, 13:309–323.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York, NY, 2nd edition.

Roepstorff, A., Nilsson, O., Oksanen, A., Gjerde, B., Richter, S. H., Örtenberg, E., Christensson, D., Martinsson, K. B., Bartlett, P., Nansen, P., Eriksen, L., Helle, O., Nikander, S., and Larsen, K. (1998). Intestinal parasites in swine in the Nordic countries: Prevalence and geographical distribution. *Veterinary Parasitology*, 76:305–319.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71:319–392.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727.

Schildcrout, J. S. and Heagerty, P. J. (2007). Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics*, 63:322–331.

Silvey, S. D. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics*, 30:389–407.

Stephenson, A. and Tawn, J. (2004). Bayesian inference for extremes: Accounting for the three extremal types. *Extremes*, 7:291–307.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971.

Stram, D. O. (1996). Meta-analysis of published data using a linear mixed-effects model. *Biometrics*, 52:536–544.

Swihart, B. J., Caffo, B. S., and Crainiceanu, C. M. (2014). A unifying framework for marginalised random-intercept models of correlated binary outcomes. *International Statistical Review*, 82:275–295.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46:657–671.

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

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84:710–716.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., and Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, 19:3417–3432.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, NY, 4th edition.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York, NY.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90:614–618.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20:595–601.

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

Wang, Z. and Louis, T. A. (2004). Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. *Biometrics*, 60:884–891.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704.

Weil, C. S. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology*, 8:177–182.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 1:60–62.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, 80:791–795.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243.

Yasui, Y. and Lele, S. (1997). A regression method for spatial disease rates: An estimating function approach. *Journal of the American Statistical Association*, 92:21–32.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, 75:621–629.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44:1019–1031.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77:642–648.

Appendix A: Data

Table A.1: Data on roadway fatalities and fuel consumption (in kilotonnes of petrol and diesel) in each of Scotland’s 29 mainland council areas between 2006 and 2011; retrieved from `statistics.gov.scot` on May 6, 2017

Council	Fatalities						Fuel Consumption					
Area	‘06	‘07	‘08	‘09	‘10	‘11	‘06	‘07	‘08	‘09	‘10	‘11
1	8	5	3	4	7	7	94	92	92	87	85	83
2	46	25	26	22	26	11	190	190	188	182	177	172
3	11	13	13	7	6	5	75	75	75	73	72	70
4	10	14	13	5	15	5	60	60	59	58	56	55
5	13	5	13	7	4	10	214	216	212	209	203	199
6	4	1	2	3	2	2	19	20	20	20	19	19
7	25	12	10	10	5	9	169	176	176	167	165	162
8	0	2	4	5	5	2	60	62	61	59	58	57
9	5	7	8	5	5	4	76	76	75	73	71	69
10	1	3	2	2	4	0	37	38	38	37	36	35
11	4	5	3	8	3	1	62	63	61	58	57	56
12	1	4	1	2	1	2	49	50	51	50	49	47
13	5	2	4	3	1	1	110	112	111	108	105	104
14	19	14	14	6	13	11	185	188	186	181	176	172
15	26	14	15	18	11	13	244	244	245	237	232	228
16	26	34	34	28	26	21	173	175	174	173	170	167
17	0	3	2	2	1	1	36	36	35	34	33	32
18	4	4	3	3	1	3	45	45	45	44	43	42
19	8	7	6	5	4	4	48	49	49	48	47	46
20	4	6	6	4	5	4	52	52	52	50	49	48
21	12	12	13	10	2	11	222	224	225	218	214	206
22	10	20	14	9	19	18	176	178	175	170	165	164
23	7	7	9	2	2	7	97	97	98	94	92	90
24	10	16	9	13	9	6	80	81	80	78	77	75
25	10	9	6	3	10	3	67	68	67	66	65	63
26	16	14	17	18	12	11	200	203	202	197	192	188
27	10	5	6	5	4	6	83	84	83	80	78	76
28	4	2	2	1	1	4	41	40	40	40	39	39
29	11	11	9	6	1	2	115	117	117	114	112	109

Table A.2: Population estimates for each of Scotland's 29 mainland council areas between 2006 and 2011; retrieved from `statistics.gov.scot` on May 8, 2017

Council	Population					
Area	2006	2007	2008	2009	2010	2011
1	209,620	212,460	214,020	217,020	219,730	222,460
2	241,180	244,390	246,840	249,020	251,430	253,650
3	112,500	113,540	114,490	114,830	115,410	116,200
4	90,870	90,790	89,910	89,450	88,620	88,930
5	452,060	456,040	458,520	463,240	469,940	477,940
6	49,540	50,600	51,190	51,290	51,330	51,500
7	149,780	150,370	151,010	151,160	151,100	151,410
8	143,370	143,700	144,290	145,170	146,060	147,200
9	120,450	120,950	121,590	122,110	122,410	122,690
10	105,590	105,050	104,940	104,960	104,920	105,000
11	93,850	95,560	97,470	98,340	99,140	99,920
12	89,750	89,840	89,870	89,980	90,410	90,810
13	151,090	152,320	153,290	154,210	155,140	156,250
14	357,260	358,750	360,050	361,410	362,610	365,300
15	568,480	571,760	576,200	581,620	586,500	593,060
16	220,780	224,000	226,980	228,750	230,730	232,730
17	82,320	82,110	82,000	81,670	81,510	81,220
18	80,000	80,370	81,540	81,900	82,360	83,450
19	90,780	91,440	92,830	93,170	93,690	93,470
20	136,790	137,420	137,910	137,830	137,790	138,090
21	328,740	331,170	333,290	335,160	336,280	337,720
22	139,390	141,140	143,130	144,370	145,600	146,850
23	171,270	171,860	172,640	173,020	173,700	174,700
24	110,860	112,200	113,360	113,590	113,690	113,880
25	112,100	112,380	112,610	112,490	112,600	112,980
26	308,450	310,380	311,320	312,180	313,180	313,900
27	88,090	88,430	88,540	88,690	89,550	90,330
28	91,420	91,370	91,190	91,080	90,800	90,610
29	167,110	169,470	171,380	173,040	174,090	175,300

Table A.3: Data from the rat teratology study of Weil (1970); each fraction represents the number of pups in a litter to survive 21 days out of those alive after four days

Control Diet							
13/13	12/12	9/9	9/9	8/8	8/8	12/13	11/12
9/10	9/10	8/9	11/13	4/5	5/7	7/10	7/10

Treatment Diet							
12/12	11/11	10/10	9/9	10/11	9/10	9/10	8/9
8/9	4/5	7/9	4/7	5/10	3/6	3/10	0/7

Table A.4: Data from the two-way crossover study reported by Jones and Kenward (1989); each cell displays the number of patients with the corresponding treatment sequence and response pattern

	Response Pattern				
Treatment Sequence	(1,1)	(0,1)	(1,0)	(0,0)	Total
Active Drug then Placebo	22	0	6	6	34
Placebo then Active Drug	18	4	2	9	33

Table A.5: Data from the bird migration study of Hewson et al. (2016); each cell displays the number of birds to survive the Sahara crossing as a fraction of the number of birds to take the corresponding route during the corresponding year

	Year				
Route	2011	2012	2013	2014	Total
East	3/3	7/9	5/5	5/5	20/22
West	2/2	0/2	5/8	5/8	12/20
Total	5/5	7/11	10/13	10/13	32/42

Table A.6: Data from the seed germination study of Crowder (1978)

			Seed Counts	
Plate	Seed	Extract	Germinated	Total
1	O75	Bean	10	39
2	O75	Bean	23	62
3	O75	Bean	23	81
4	O75	Bean	26	51
5	O75	Bean	17	39
6	O73	Bean	8	16
7	O73	Bean	10	30
8	O73	Bean	8	28
9	O73	Bean	23	45
10	O73	Bean	0	4
11	O75	Cucumber	5	6
12	O75	Cucumber	53	74
13	O75	Cucumber	55	72
14	O75	Cucumber	32	51
15	O75	Cucumber	46	79
16	O75	Cucumber	10	13
17	O73	Cucumber	3	12
18	O73	Cucumber	22	41
19	O73	Cucumber	15	30
20	O73	Cucumber	32	51
21	O73	Cucumber	3	7

Table A.7: Pigsty data found in Larsen et al. (2000); each fraction represents the number of pigs infected with roundworm out of all of the pigs in a pigsty

Specific Pathogen Free Pigsties

0/16	0/15	0/15	0/15	0/15	0/15	0/15	1/15	2/15	0/14	0/14	0/12
0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	1/10	1/10	2/10	2/10
0/9	1/7	1/7	0/5	0/5	0/5	0/5	1/5	1/5	4/5	1/3	0/1

Conventional Pigsties

0/15	0/15	0/15	0/15	0/15	1/15	1/15	1/15	1/15	2/15	3/15	3/15
3/15	4/15	4/15	6/15	2/11	0/10	0/10	0/10	0/10	0/10	0/10	0/10
0/10	0/10	0/10	0/10	1/10	1/10	1/10	1/10	3/10	3/10	4/10	0/9
0/9	0/9	0/9	1/9	2/9	5/9	9/9	0/8	0/8	2/8	3/8	0/7
1/7	1/7	1/7	1/7	2/7	3/7	0/6	0/6	0/6	0/5	0/5	0/5
1/5	1/5	1/5	1/5	0/4	0/4	3/4	0/3	0/2	0/2	0/1	1/1

Table A.8: Epileptic seizures data found in Thall and Vail (1990); each count represents the number of partial seizures reported by a patient during one of the post-treatment visits

Control Group						**Progabide Group**					
		Post-Treatment Visit						Post-Treatment Visit			
Age	Baseline	1st	2nd	3rd	4th	Age	Baseline	1st	2nd	3rd	4th
31	11	5	3	3	3	20	19	0	4	3	0
30	11	3	5	3	3	30	10	3	6	1	3
25	6	2	4	0	5	18	19	2	6	7	4
36	8	4	4	1	4	24	24	4	3	1	3
22	66	7	18	9	21	30	31	22	17	19	16
29	27	5	2	8	7	35	14	5	4	7	4
31	12	6	4	0	2	27	11	2	4	0	4
42	52	40	20	23	12	20	67	3	7	7	7
37	23	5	6	6	5	22	41	4	18	2	5
28	10	14	13	6	0	28	7	2	1	1	0
36	52	26	12	6	22	23	22	0	2	4	0
24	33	12	6	8	4	40	13	5	4	0	3
23	18	4	4	6	2	33	46	11	14	25	15
36	42	7	9	12	14	21	36	10	5	3	8
26	87	16	24	10	9	35	38	19	7	6	7
26	50	11	0	0	5	25	7	1	1	2	3
28	18	0	0	3	3	26	36	6	10	8	8
31	111	37	29	28	29	25	11	2	1	0	0
32	18	3	5	2	5	22	151	102	65	72	63
21	20	3	0	6	7	32	22	4	3	2	4
29	12	3	4	3	4	25	41	8	6	5	7
21	9	3	4	3	4	35	32	1	3	1	5
32	17	2	3	3	5	21	56	18	11	28	13
25	28	8	12	2	8	41	24	6	3	4	0
30	55	18	24	76	25	32	16	3	5	4	3
40	9	2	1	2	1	26	22	1	23	19	8
19	10	3	1	4	2	21	25	2	3	0	1
22	47	13	15	13	12	36	13	0	0	0	0
						18	76	11	14	9	8
						37	12	1	4	3	2
						32	38	8	7	9	4