AN ORTHODOX BLUP APPROACH TO GENERALIZED LINEAR MIXED MODELS

by

Renjun Ma

B.Sc., Wuhan University of Water Transportation Engineering, China, 1982

M.Sc., Wuhan University, China, 1987

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April, 1999

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date April 28, 1999

Abstract

We introduce a new class of generalized linear mixed models assuming Tweedie exponential dispersion model distributions for both the response and the random effects. This class of models accommodates a wide range of discrete, continuous and mixed data. By letting the random effects enter as weights as well as means in the conditional distributions, the variance matrix may be expressed as a sum of variance components. We consider an orthodox BLUP approach to parameter estimation and random effects prediction for this new class of models based on a predictor of the random effects that is truly best linear and unbiased, in contrast to the conventional BLUP which is the conditional mode. We obtain an optimal estimating equation based on the orthodox BLUP, which is solved by a modified Newton algorithm. This approach facilitates analysis of residuals and allows justification of asymptotic results under realistic conditions through standard estimating equation theory. An important feature of this approach is that the principal results depend only on the first and second moment assumptions of unobserved random effects. The common fitting algorithm based on orthodox BLUP enables us to study this new class of models as a single class, rather than as a collection of unrelated different models. This approach is illustrated with the analyses of seed germination data, epilepsy data and cake baking data. By means of asymptotic justifications, simulations and worked examples, we conclude that the orthodox BLUP approach is of practical value for analysis of clustered non-normal data.

Contents

List of Tables

List of Figures

Acknowledgements

Foremost I would like to thank my supervisor, Dr. Bent Jørgensen, for his excellent guidance, for his inspiration, for his constant encouragement and for his great patience.

I am very grateful to other members of my supervisory committee, Dr. John Petkau and Dr. Martin Puterman for their valuable comments and assistance. I am particularly indebted to Dr. John Petkau for his careful reading of the manuscript; his helpful discussion and constructive suggestions lead to the substantial improvement of the presentation of this thesis.

I wish to thank my program advisor, Dr. Jean Meloche, for helping me to adapt to the new environment.

I would like to thank Dr. Nancy Heckman for her invaluable advice and help, for her constant encouragement and support throughout my program. Many thanks go to Dr. Harry Joe for his useful advice and help throughout my program.

Many thanks go to Dr. Li Sun, Hongbin Zhang and Jeevanantham Rajeswaran for their valuable help with Latex and computer programming. I am very grateful to Christine Graham for her constant help and patience.

I express my gratitude to Department of Theoretical Statistics, Aarhus University for providing me help and access to research facility during my academic visit. I also acknowledge the support of the University of British Columbia through a University Graduate Fellowship.

Chapter 1

Introduction

In recent years much effort has been devoted to extending regression methodology to clustered non-normal data. A wide range of applications of generalized linear mixed models to clustered data has been investigated by many researchers (Breslow and Clayton 1993; Lee and Nelder 1996). Traditional generalized linear models (McCullagh and Nelder 1989) have unified regression analysis for a variety of independent responses, but have difficulties in providing valid inference for clustered, and hence correlated, data. Clustered data usually exhibit significant heterogeneity and overdispersion as well (cf. Engel 1987; McCullagh and Nelder 1989, p.124-126, 198-200). Such cluster effects are often modelled by incorporating random effects into generalized linear models.

The application of generalized linear mixed models to clustered data has attracted much attention since the seventies (Crowder 1978; Laird 1978). However, the early development of generalized linear mixed models focussed mainly on relatively simple cases such as random intercept models (Hinde 1982; Williams 1982; Breslow 1984; Stiratelli, Laird and Ware 1984; Anderson and Aitkin 1985; Brillinger and Preisler

1986). The modelling of more complicated problems has largely been hampered by the intractability of high-dimensional integrals involved in evaluating the likelihood. To avoid these numerical problems, numerous alternative approaches to the analysis of clustered data have been proposed recently. The most popular approaches are generalized estimating equations (Liang and Zeger 1986; Zeger, Liang and Albert 1988), Bayesian methods using computational techniques (Zeger and Karim 1991), penalized quasi-likelihood (Breslow and Clayton 1993) and maximum hierarchical likelihood (Lee and Nelder 1996) methods. We briefly review these approaches and discuss their advantages and disadvantages.

Generalized estimating equations (GEE) approach focuses on the marginal relationship between covariates and clustered responses, but only estimates the covariances as nuisance parameters by adopting a so-called "working correlation". This approach enjoys robustness against mis-specification of covariances, but sometimes suffers from a lack of efficiency due to such incorrect specification (Lipsitz et al. 1994; Fitzmaurice 1995). It is of limited use when the correlation structure is of primary interest. This approach mainly deals with marginal models instead of random effects models. On the other hand, in areas such as genetics, it is often the unobserved genetic effects which are of primary interest (Harville 1977; Clayton 1991; Karim and Zeger 1992). Robinson (1991) presented a variety of applications where the random effects themselves are of interest; therefore random effects models are more relevant in such situations.

As a general approach to complicated generalized linear mixed models, Zeger and Karim (1991) proposed to cast these models in a Bayesian framework and approximate maximum likelihood estimates using flat or diffuse priors. However, such approximations are often impossible because the posterior may not exist for such priors (Natarajan and McCulloch 1995; Hobert and Casella 1996). This problem of the posterior may not be detected when using computational techniques such as the Gibbs sampler; therefore misleading estimates may result.

On the other hand, recent non-Bayesian approaches to generalized linear mixed models have focussed on the explicit or implicit modification of the E-step in the EM algorithm (Dempster, Laird and Rubin 1977) due to the difficulty evaluating the conditional expectation of the random effects given the data. The most widely adopted technique is the generalization of the linear mixed model equations of Henderson et al. (1975), but with various approximations (Gilmour, Anderson and Rae 1984; Harville and Mee 1984; Schall 1991; Breslow and Clayton 1993; Wolfinger 1993; McGilchrist 1994; Lee and Nelder 1996). It can be shown that these approaches essentially modify the EM algorithm with the conditional expectation of the random effects given the data being replaced by the predictors based on the mode of the corresponding conditional distributions. These modal predictors are often referred to as the 'BLUP' (Schall 1991; McGilchrist 1994), although in general it is neither linear nor unbiased for non-normal distributions.

In contrast to the modal predictor approaches, Bayesian approaches enjoy great flexibility in modelling the data with normal or non-normal random effects distributions, but are computationally intensive. The computational time required is sufficiently long as to possibly discourage fitting several different models (Karim and Zeger 1992). This drawback may pose serious problems in practice because it is generally difficult to justify a particular distribution for the unobserved random effects. Data will often point with almost equal emphasis at several different models and it is important that we recognize these models and their possible different interpretations. On the other hand, the modal predictor approaches enable us to explore several different models with reasonable computing time. The penalized quasi-likelihood (PQL) approach deals with models with approximate multivariate normal random effects. The maximum hierarchical likelihood (MHL) approach widens the choice of random effects distributions to include conjugate distributions (George et al. 1987), but fails to model the dependence structure of the random effects. The introduction of flexible, yet tractable distribution classes to model both the distributional shape and dependence structure of random effects is certainly needed to facilitate appropriate inference.

The implementation of the Bayesian approach is relatively straightforward in contrast to the requirement for manipulation of large matrices for modal predictor approaches when there are a large number of random effects; however the assessment of convergence of computational techniques such as the Gibbs sampler remains an area of debate (Glifford 1993; Smith and Roberts 1993). On the other hand, non-Bayesian approaches provide a natural framework for model checking, but are forced to rely on asymptotics. In fact, the existing approaches to generalized linear mixed models mainly concentrate on the model fitting part, but to a large extent ignore another important ingredient of the modelling process, the model checking component (Lee and Nelder 1996).

The justifications of these modal predictor approaches were generally intuitive until Breslow and Clayton (1993) presented some ad hoc justifications for their penalized quasi-likelihood approach and Lee and Nelder (1996) provided rigorous justification for their maximum hierarchical likelihood approach. The asymptotic results for quite general settings are obtained under certain conditions by Nelder and Lee (1996), but with respect to large cluster sizes with fixed number of clusters. This raised concerns about more practical situations where the number of clusters is large, but with relatively small cluster sizes (Clayton 1996; Engel and Keen 1996). In addition, the estimating equations based on the modal preditors are generally biased (Breslow and Lin 1995; Lin and Breslow 1996a).

This thesis considers generalized linear mixed models based on the class of Tweedie exponential dispersion model distributions (Jørgensen 1987a) for both the response and the random effects. This gives a very flexible class of models which includes various combinations of Poisson, normal, gamma, inverse Gaussian, compound Poisson and extreme stable and positive stable distributions. In the context of clustered data, the hierarchical structure is very clear so that the modeled covariance structure should clearly reflect those hierarchies. The dominant tradition in accounting for dependence between or within clusters is to explicitly incorporate random effects into a monotonic transform of the conditional mean ignoring the dispersion components. This approach generally does not lead to variance components decomposition structure for covariance matrix of the response. By incorporating correlated or uncorrelated random effects into both mean and dispersion components, the covariance matrix of our model possesses an interpretable variance components decomposition.

The novelty of our approach lies in the introduction of a new unbiased estimating equation, which is based on a modification of the EM algorithm where conditional expectations are approximated by an *orthodox BLUP*, in the context of generalized linear mixed models. An orthodox BLUP is defined as the best linear unbiased predictor in the literal sense, (cf. Brockwell and Davis 1991 p. 64). The unbiased estimating equations introduced via the orthodox BLUP lead to consistent estimators for both regression and dispersion parameters under practical conditions where the number of clusters is large. The estimating equation for the regression parameters is optimal in the sense of Godambe (1976). While the parametric nature of our models facilitates residual analysis, our estimating procedure also allows a semi-parametric interpretation of the models. Our approach does not require manipulation of large matrices; therefore is computationally simpler than the modal predictor approaches. This approach is applicable to a wide range of clustered discrete, continuous and mixed data.

The organization of this thesis is as follows. In Chapter 2, besides a brief introduction of Tweedie exponential dispersion model distributions and some estimating function results, we compare the orthodox BLUP and modal predictors and present a few data examples to motivate our study. In Chapter 3, we propose a class of nested random effect models and derive their moment structures. The prediction of random effects and the consistency properties of the orthodox BLUP are discussed in Chapters 4. In Chapter 5, we discuss estimation for both regression and dispersion parameters as well as asymptotic properties of these parameter estimators. An outline of the residual analysis and computational procedure is presented in Chapter 6. We address in detail a so-called conventional model and its relationship with other models in Chapter 7. Illustrative examples and simulations are presented in Chapters 8 and 9, respectively. We present a discussion in Chapter 10.

Chapter 2

Preliminaries

2.1 Generalized linear models with random effects

2.1.1 Definitions

We study generalized linear models with random effects based on the class of Tweedie exponential dispersion model distributions. Here we first define exponential dispersion models. A random variable Y is said to follow reproductive exponential dispersion model $ED(\mu, \sigma^2)$ if its probability density functions can be written in the form:

$$p(y; \phi, \eta) = a(y; \phi) \ \exp\{\phi[y \ \eta - \kappa(\eta)]\}, \tag{2.1}$$

where $\mu = \mathrm{E}[Y]$ and $\sigma^2 = 1/\phi$. Let $\kappa'(\cdot)$, the first derivative of $\kappa(\cdot)$, be denoted by $\tau(\cdot)$. $V(\mu) = \tau'\{\tau^{-1}(\mu)\}$ is called the variance function. Further,

$$d(y; \mu) = 2[y\{\tau^{-1}(y) - \tau^{-1}(\mu)\} - \kappa\{\tau^{-1}(y)\} + \kappa\{\tau^{-1}(\mu)\}] \tag{2.2}$$

is known as the (unit) deviance function.

The distribution of $Z = \phi Y$ is called an additive exponential dispersion model, denoted by $Z \sim ED^*(\eta, \phi)$. The justification of terminology 'reproductive' and 'additive' can be found in Jørgensen (1997, p11).

Now we can define a generalized linear model with random effects. Let $\mathbf{Y} = (Y_{11}, ..., Y_{1n_1}, ..., Y_{m1}, ..., Y_{mn_m})^T$ be an $n = \sum_{i=1}^m n_i$-dimensional vector of observed responses. Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ be a p-dimensional vector of fixed effects and $\mathbf{U} = (U_1, ..., U_m)^T$ be a m-dimensional vector of random effects.

We suppose that, given $\mathbf{U} = \mathbf{u}$, $Y_{11}, ..., Y_{1n_1}, ..., Y_{m1}, ..., Y_{mn_m}$ are conditionally independent and

$$Y_{ij}|\mathbf{U} = \mathbf{u} \sim ED(\mu_{ij}^{\mathbf{u}}, \sigma^2), \tag{2.3}$$

Let $\boldsymbol{\mu}^{\mathbf{u}} = (\mu_{11}^{\mathbf{u}}, ..., \mu_{1n_1}^{\mathbf{u}}, ..., \mu_{m1}^{\mathbf{u}}, ..., \mu_{mn_m}^{\mathbf{u}})^\top$ and let $g(\cdot)$ denote the link function. We denote $(g(\mu_{11}^{\mathbf{u}}), ..., g(\mu_{1n_1}^{\mathbf{u}}), ..., g(\mu_{m1}^{\mathbf{u}}), ..., g(\mu_{mn_m}^{\mathbf{u}}))^\top$ by $g(\boldsymbol{\mu}^{\mathbf{u}})$. Suppose further that

$$g(\boldsymbol{\mu}^{\mathbf{u}}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \tag{2.4}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are two known matrices. Then

$$\mathrm{E}[Y_{ij}|\mathbf{U} = \mathbf{u}] = \mu_{ij}^{\mathbf{u}}. \tag{2.5}$$

The expectation and covariance matrix of $\mathbf{U}$ are denoted by the following equations:

$$\mathrm{E}[\mathbf{U}] = \boldsymbol{\eta} \text{ and } \mathrm{Var}[\mathbf{U}] = \mathbf{D}(\boldsymbol{\gamma}), \tag{2.6}$$

where $\mathbf{D}(\boldsymbol{\gamma})$ is a $q \times q$ covariance matrix depending on an unknown vector of variance components $\boldsymbol{\gamma}$.

The expectations of **U** or transformed **U** are often assumed to be known. The transformed **U** is also often called random effects.

2.1.2 Tweedie exponential dispersion models

Many exponential dispersion models have variance functions that are asymptotically of the Tweedie form, leading to a general convergence theorem with the Tweedie models as limiting distributions (Jørgensen et al. 1994). For this reason, Tweedie models occupy a central position among exponential dispersion models. Now we define Tweedie exponential dispersion models. We call (2.1) a reproductive Tweedie exponential dispersion model, denoted by $\mathrm{Tw}_p(\mu, \sigma^2)$, if $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \sigma^2\mu^p$. Here $p = 0, 2, 3$ and $1 < p < 2$ correspond to well known normal, gamma, inverse Gaussian and compound Poisson distributions respectively. The case $p = 1$ with $\sigma^2 = 1$ corresponds to Poisson distributions. In fact, using usual notations, we have

$$\mathrm{Tw}_1(\mu, 1) = \text{ Poisson}(\mu),$$

$$\mathrm{Tw}_2(\mu, \sigma^2) = \text{ Gamma}(\mu, \sigma^2),$$

and

$$\mathrm{Tw}_3(\mu, \sigma^2) = \text{ Inverse Gaussian}(\mu, \sigma^2).$$

Furthermore Tweedie exponential dispersion models posses the following scale transformation property:

$$c\mathrm{Tw}_p(\mu, \sigma^2) = \mathrm{Tw}_p(c\mu, c^{2-p}\sigma^2). \tag{2.7}$$

A complete list of Tweedie exponential dispersion models is given in Table 2.1 where S, Ω and Θ denote the support, mean parameter domain and canonical parameter domain of the Tweedie exponential dispersion model, respectively (Jørgensen

Table 2.1: Summary of Tweedie exponential dispersion models.

Distributions	p	S	Ω	Θ
Extreme stable	$p<0$	$\mathbf{R}$	$\mathbf{R}_+$	$\mathbf{R}_0$
Normal	$p=0$	$\mathbf{R}$	$\mathbf{R}$	$\mathbf{R}$
[Do not exist]	$0<p<1$	—	$\mathbf{R}_+$	$\mathbf{R}_0$
Poisson	$p=1$	$\mathbf{N}_0$	$\mathbf{R}_+$	$\mathbf{R}$
Compound Poisson	$1<p<2$	$\mathbf{R}_0$	$\mathbf{R}_+$	$\mathbf{R}_-$
Gamma	$p=2$	$\mathbf{R}_+$	$\mathbf{R}_+$	$\mathbf{R}_-$
Positive stable	$2<p<3$	$\mathbf{R}_+$	$\mathbf{R}_+$	$-\mathbf{R}_0$
Inverse Gaussian	$p=3$	$\mathbf{R}_+$	$\mathbf{R}_+$	$-\mathbf{R}_0$
Positive stable	$p>3$	$\mathbf{R}_+$	$\mathbf{R}_+$	$-\mathbf{R}_0$
Extreme stable	$p=\infty$	$\mathbf{R}$	$\mathbf{R}$	$\mathbf{R}_-$

Notation: $-\mathbf{R}_0=(-\infty,0]$

1987a).

To ease the derivation of the estimation function for regression parameter for our random effects models later, we rewrite the density of the Tweedie exponential dispersion models $\mathrm{Tw}_p(\mu,\sigma^2)$ as follows:

$$f_p(y;\mu,\sigma^2)=\begin{cases} c_p(y;\sigma^2)\exp\left\{\frac{1}{\sigma^2}\left(\frac{y\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}\right)\right\} & \text{if } p\neq 1,2,\\ c_2(y;\sigma^2)\exp\left\{-\frac{1}{\sigma^2}\left(\frac{y}{\mu}+\log\mu\right)\right\} & \text{if } p=2,\\ c_1(y)\exp\left\{y\log\mu-\mu\right\} & \text{if } p=1,\end{cases}$$

where the explicit expressions for $c_p(y;\sigma^2)$ are given by Jørgensen (1987). The fact that $c_p(y;\sigma^2)$ do not depend on μ is crucial to the derivation of our unbiased estimating function in Chapter 5. The exact expressions for $c_p(y;\sigma^2)$ are immaterial in the derivation of the unbiased estimating function, thus omitted. For more details about

Tweedie exponential dispersion model, see Jørgensen (1987, 1996).

2.2 Prediction of random effects

The prediction of random effects plays an important role in random effects models. It is especially useful in the identification of outliers.

The distinction between Bayesian and non-Bayesian approaches to generalized linear models with random effects is clear. Bayesian approaches use posterior means or modes as point estimators for both parameters and random effects. The regression coefficients and the random effect variance are assumed to be random vectors and treated symmetrically with the observed responses and unobserved random effects.

On the other hand, non-Bayesian approaches such as penalized quasi-likelihood approach of Breslow and Clayton (1993) and the maximum h-likelihood estimation approach of Lee and Nelder (1996) predict random effects using the mode of the conditional distribution of the random effects given the data. We concentrate on making comparisons of these two approaches with our approach.

2.2.1 Likelihoods

Before we make comparisons, we need to introduce the concept of the *partially observed joint (log) likelihood* as follows.

Let $\mathbf{Y}$ be the response and $\mathbf{U}$ be the unobserved random effects. We assume that both the conditional distribution of $\mathbf{Y}$ given $\mathbf{U}$ and the distribution of $\mathbf{U}$ are parametric. The partially observed joint log likelihood with only the response being observed is defined as

$$\ell(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma};\mathbf{Y},\mathbf{U}) = \ell(\boldsymbol{\beta},\boldsymbol{\alpha};\mathbf{Y}|\mathbf{U}) + \ell(\boldsymbol{\gamma};\mathbf{U}),$$

where $\ell(\boldsymbol{\beta},\sigma^2;\mathbf{Y}|\mathbf{U})$ is the logarithm of the conditional density of $\mathbf{Y}$ given $\mathbf{U}$ with $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ being the regression and dispersion parameters respectively, and $\ell(\boldsymbol{\gamma};\mathbf{U})$ is that for $\mathbf{U}$ with parameter $\boldsymbol{\gamma}$. Thus the *marginal likelihood* is

$$\int \exp[\ell(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma};\mathbf{Y},\mathbf{U})]d\mathbf{U}.$$

In Breslow and Clayton's paper, the partially observed joint likelihood is approximated by the *penalized quasi-likelihood* defined as

$$q(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma};\mathbf{Y},\mathbf{U}) = \frac{1}{2\sigma^2}\Sigma_{i=1}^{t}\Sigma_{j=1}^{n_i} d_{ij}(Y_{ij};\mu_{ij}^{\mathbf{u}}) - \frac{1}{2}\mathbf{u}^{\top}\mathbf{D}^{-1}(\boldsymbol{\gamma})\mathbf{u}, \qquad (2.8)$$

where $d_{ij}(\cdot;\cdot)$ denotes the deviance function defined in Section 2.1.1.

2.2.2 Penalized quasi-likelihood and h-likelihood approaches

Breslow and Clayton consider the model in (2.3), but assume $\mathbf{U}$ follows, at least approximately, a multivariate normal distribution with mean $\mathbf{0}$. They use solutions $(\hat{\boldsymbol{\beta}},\hat{\mathbf{u}})$, for fixed $\boldsymbol{\gamma}$, from the following equations

$$\frac{\partial q(\mathbf{y},\boldsymbol{\beta},\mathbf{u},\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}} = \mathbf{0}, \quad \frac{\partial q(\mathbf{y},\boldsymbol{\beta},\mathbf{u},\boldsymbol{\gamma})}{\partial\mathbf{u}} = \mathbf{0}$$

as the estimators for $(\boldsymbol{\beta},\mathbf{u})$. That is, they obtain estimators through maximizing an approximation of the partially observed joint likelihood. They adopted the Fisher's scoring algorithm to obtain the solutions. The multivariate normality assumption is convenient to incorporate correlation structure among random effects, but quite restrictive to model the distributional shape.

On the other hand, Lee and Nelder consider the model in (2.3), that is

$$Y_{ij}|\mathbf{U} = \mathbf{u} \sim ED(\mu_{ij}^{\mathrm{u}}, \sigma^2),$$

but with

$$g(\boldsymbol{\mu}^{\mathrm{u}}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{v},$$

where $\mathbf{v}$ is a monotonic transformed variable of $\mathbf{u}$, denoted by $\mathbf{v} = \mathbf{v}(\mathbf{u})$, and the distribution of random effects is assumed appropriately.

They call the partially observed joint log likelihood of ($\mathbf{Y}$,$\mathbf{V}$) the h-likelihood, denoted by $\ell(\mathbf{Y}, \mathbf{V}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$. The solution $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}})$, for fixed $\boldsymbol{\gamma}$, from the following equations:

$$\frac{\partial \ell(\mathbf{y}, \boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = \mathbf{0}, \quad \frac{\partial \ell(\mathbf{y}, \boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\gamma})}{\partial \mathbf{v}} = \mathbf{0} \tag{2.9}$$

is used as the estimator for $(\boldsymbol{\beta}, \mathbf{V})$. They also adopted the Fisher's scoring algorithm to obtain the solutions. Lee and Nelder's method is feasible when the distribution of the random effect is conjugate to that of the observed response.

2.2.3 Comparison for different predictors

To better compare our approach with other current approaches, we make a comparison of conditional expectation, conditional mode (the mode of conditional distribution of $\mathbf{U}$ given $\mathbf{Y}$) and the orthodox BLUP.

When both $\mathbf{Y}$ and $\mathbf{U}$ are normally distributed, the conditional mode of $\mathbf{U}$ given $\mathbf{Y}$ equals the corresponding conditional expectation which is also orthodox BLUP.

Hence Lee and Nelder approximated the conditional expectation by the maximum h-likelihood estimates (MHLEs), that is, the solutions of equation (2.9). In fact, these solutions are equivalent to the modes of the conditional distribution of $\mathbf{U}$ given $\mathbf{Y}$ since

$$\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}; \mathbf{Y}, \mathbf{U}) = \ell_1(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}; \mathbf{U}|\mathbf{Y}) + \ell_2(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}; \mathbf{Y}).$$

Breslow and Clayton also adopted this approach in their penalized quasi-likelihood method, but they differentiated penalized quasi-likelihood which is an approximation of the partially observed joint likelihood. That is, they used an approximation of the conditional mode to be the predictor of random effects.

Clearly the mode is neither necessarily a good approximation to the corresponding mean when the conditional distribution is not approximately symmetric about its mode such as in the case of normal, nor is it easy to evaluate its departure from the conditional mean. Lee and Nelder actually consider the partially observed joint likelihood of the response and a monotonic transform of original random effects. They claimed that their MHLE predictors of random effects are invariant with respect to monotone transformation of the random effects $\mathbf{U}$. However the location of the mode depends on the dominating measure, therefore estimates on the original scale of random effects are preferred. Thus we suggest approximating the conditional expectation by orthodox BLUP on the original scale of random effects where appropriate.

The orthodox BLUP is defined as a linear unbiased predictor of $\mathbf{U}$ given $\mathbf{Y}$ which minimizes the mean square distance between the random effects $\mathbf{U}$ and their predictor within the class of linear functions of $\mathbf{Y}$. We call it 'orthodox' BLUP to distinguish it from the modal predictor.

Comparing with the conditional expectation which is the best unbiased predictor for $\mathbf{U}$ among all functions of $\mathbf{Y}$, the orthodox BLUP is truly the best unbiased predictor for $\mathbf{U}$ among all linear functions of $\mathbf{Y}$. We will demonstrate that the orthodox BLUP is often a good predictor for $\mathbf{U}$ though the conditional expectation is generally a non-linear function of $\mathbf{Y}$. Unlike the conditional mode, the mean square distance between random effects and the corresponding orthodox BLUP can usually be evaluated easily.

2.3 Estimating functions

Our orthodox BLUP approach is based on estimating functions. Some results on estimating functions will be used repeatedly later. We briefly summarize them here.

2.3.1 Unbiased estimating functions

Suppose $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ are independent random vectors. Let us consider the estimation for the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ based on estimating functions of the form $\boldsymbol{\psi}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^m \boldsymbol{\psi}_i(\boldsymbol{\theta}; \mathbf{Y}_i) = \sum_{i=1}^m \boldsymbol{\psi}_i(\boldsymbol{\theta})$, where $\boldsymbol{\psi}_i, i = 1, \ldots, m$ are unbiased estimating functions, that is

$$\mathrm{E}_{\boldsymbol{\theta}} \boldsymbol{\psi}_i(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{0},$$

where $\boldsymbol{\psi}(\boldsymbol{\theta}; \mathbf{Y})$ is of the same dimension as the dimension of $\boldsymbol{\theta}$.

We also assume that $\boldsymbol{\psi}(\boldsymbol{\theta}; \mathbf{Y})$ is regular, (Jørgensen and Labouriau 1994; McLeish and Small, 1988). Let the density function of $\mathbf{Y}$ and the range of $\mathbf{Y}$ be denoted by

$p(\mathbf{Y};\boldsymbol{\theta})$ and $\mathcal{X}$, respectively. We define a regular estimating function as follows:

Definition 2.1

An estimating function $\boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y})$ is said to be regular if the following conditions are satisfied for all $\boldsymbol{\theta}$ in the parameter space:

1. *The support of $\mathbf{Y}$ does not depend on $\boldsymbol{\theta}$;*
2. $E_{\boldsymbol{\theta}}\boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}) = \mathbf{0}$;
3. *The partial derivative $\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j}$ exists for almost every $\mathbf{Y}$ in $\mathcal{X}$, $j = 1, \ldots, q$;*
4. *The order of integration and differentiation may be interchanged as follows:*

$$\frac{\partial}{\partial \theta_j} \int_{\mathcal{X}} \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}) p(\mathbf{Y};\boldsymbol{\theta}) d\mathbf{Y} = \int_{\mathcal{X}} \frac{\partial}{\partial \theta_j} \left\{ \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}) p(\mathbf{Y};\boldsymbol{\theta}) \right\} d\mathbf{Y},$$

   $j = 1, \ldots, q$.

5. $\mathbf{S}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y})}{\partial \boldsymbol{\theta}}$ *is a nonsingular $q \times q$ matrix;*
6. $\mathbf{V}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y})\boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y})^{\top}]$ *is a $q \times q$ positive-definite matrix.*

$\mathbf{S}(\boldsymbol{\theta})$ and $\mathbf{V}(\boldsymbol{\theta})$ are called the sensitivity and variability matrices, respectively. Note that we do not assume that $\boldsymbol{\psi}_i(\boldsymbol{\theta};\mathbf{Y})$s are regular. Actually the corresponding sensitivity and variability matrices for $\boldsymbol{\psi}_i(\boldsymbol{\theta};\mathbf{Y})$ are often singular in practice, due to the presence of categorical covariates. However, we assume that $\boldsymbol{\psi}_i(\boldsymbol{\theta};\mathbf{Y})$ satisfies all above conditions except the non-singularity requirement of its sensitivity and variability matrices.

In the estimating function approach one considers estimators which can be expressed as solutions of the following estimating equation:

$$\boldsymbol{\psi}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^{m} \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \mathbf{0}.$$

By standard asymptotic theory for estimating functions, we may show that under certain regularity conditions, the sequence of roots, $\hat{\boldsymbol{\theta}}^{(m)}$, associated with the estimating function $\sum_{i=1}^{m} \boldsymbol{\psi}_i(\boldsymbol{\theta})$, is consistent for $\boldsymbol{\theta}$ and asymptotically normal. Specifically we have (Artes and Jørgensen 1998):

Lemma 2.1

$$\sqrt{m}(\hat{\boldsymbol{\theta}}^{(m)} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \bar{\mathbf{J}}^{-1}(\boldsymbol{\theta})\right) \ \textit{as } m \to \infty, \tag{2.10}$$

where $\bar{\mathbf{J}}(\boldsymbol{\theta})$ can be expressed in terms of the *sensitivity* and *variability* matrices for cluster i, defined by $\mathbf{S}_i(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left\{\frac{\partial \boldsymbol{\psi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right\}$, and $\mathbf{V}_i(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\psi}_i(\boldsymbol{\theta})\boldsymbol{\psi}_i^\top(\boldsymbol{\theta})\right\}$ respectively:

$$\bar{\mathbf{J}}(\boldsymbol{\theta}) = \lim_{m\to\infty} \left\{m^{-1}\sum_{i=1}^{m} \mathbf{S}_i(\boldsymbol{\theta})\right\}\left\{m^{-1}\sum_{i=1}^{m} \mathbf{V}_i(\boldsymbol{\theta})\right\}^{-1}\left\{m^{-1}\sum_{i=1}^{m} \mathbf{S}_i(\boldsymbol{\theta})\right\}^\top \tag{2.11}$$

The asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$ is therefore given by the inverse of the *Godambe information matrix* defined by

$$\mathbf{J}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\theta})\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{S}(\boldsymbol{\theta})^\top, \tag{2.12}$$

where $\mathbf{S}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbf{S}_i(\boldsymbol{\theta})$ and $\mathbf{V}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbf{V}_i(\boldsymbol{\theta})$.

The estimating function is a generalization of the score function $U(\boldsymbol{\theta})$:

$$U(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}},$$

where $\ell(\mathbf{Y};\boldsymbol{\theta}) = \log p(\mathbf{Y};\boldsymbol{\theta})$ is the log likelihood.

The sensitivity and variability matrices for the score function have following relationship:

$$\mathbf{S}(\boldsymbol{\theta}) = -\mathbf{V}(\boldsymbol{\theta}),$$

which does not hold for regular estimating functions in general.

The estimator and Godambe information matrix for the score function are the maximum likelihood estimator, denoted by MLE, and the Fisher information matrix, denoted by $I(\boldsymbol{\theta})$, respectively. Among all regular estimating functions, the score function is optimal in the sense that the estimator associated with the score function attains the minimum asymptotic variance among estimators associated with all regular estimating functions. To state this result precisely, let the Godambe information matrix for any given regular estimating function $\boldsymbol{\psi}$ be denoted by $J_{\boldsymbol{\psi}}(\boldsymbol{\theta})$. Then (Jørgensen and Labouriau 1994)

$$J_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\theta}) - I^{-1}(\boldsymbol{\theta}),$$

is nonnegative-definite for all $\boldsymbol{\theta}$ in the parameter space.

However, in the context of generalized linear mixed models, the full score function for both the response and the random effects is not available since the random effects are unobserved. Thus we consider estimating functions other than the score function. We can also consider the optimality property, but within a more restricted class of estimating functions. In the next section, we state an optimality result within a certain linear class.

2.3.2 Optimal estimating functions

Crowder (1986, 1987) proved, under some regularity conditions, the following

Lemma 2.2

$$\boldsymbol{\psi}_{opt}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbf{S}_i(\boldsymbol{\theta})\mathbf{V}_i^{-1}(\boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{\theta}) \tag{2.13}$$

is an optimal estimating function within the class of all linear estimating functions of the form

$$\sum_{i=1}^{m} \mathbf{Q}_i(\boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{\theta}), \tag{2.14}$$

where $\mathbf{Q}_i(\boldsymbol{\theta})$ *is a constant weights matrix of appropriate dimensions. More specifically,* $\mathbf{Q}_i(\boldsymbol{\theta})$ *is a matrix function of parameter* $\boldsymbol{\theta}$*, but does not involve* $\mathbf{Y}$*.*

The solution $\widehat{\boldsymbol{\theta}}$ from the estimating equation $\boldsymbol{\psi}_{opt}(\boldsymbol{\theta}) = \mathbf{0}$ is then asymptotically normal with the asymptotic mean $\boldsymbol{\theta}$ and asymptotic variance given by the inverse of

$$\mathbf{J}_{opt}(\boldsymbol{\theta}) = \sum_{i=1}^{m} -\mathbf{S}_i(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbf{V}_i(\boldsymbol{\theta}) = \sum_{i=1}^{m} \ J_i(\boldsymbol{\theta}).$$

That is, the Godambe information for the optimal estimating equation is the sum of the Godambe informations for each i.

2.3.3 Nuisance parameter case

Suppose that $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^{\top}, \boldsymbol{\theta}_{(2)}^{\top})^{\top}$ where $\boldsymbol{\theta}_{(1)}$ and and $\boldsymbol{\theta}_{(2)}$ are taken as the parameter of interest and the nuisance parameter, respectively. We partition $\boldsymbol{\psi}_i(\boldsymbol{\theta})$ into $(\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\theta}), \boldsymbol{\psi}_i^{(2)}(\boldsymbol{\theta}))$, where $\boldsymbol{\psi}_i^{(1)}$ and $\boldsymbol{\psi}_i^{(2)}$ are of the same dimensions as those of $\boldsymbol{\theta}_{(1)}$ and $\boldsymbol{\theta}_{(2)}$, respectively. Let $\boldsymbol{\psi}^{(k)} = \sum_{i=1}^{m} \boldsymbol{\psi}_i^{(k)}$ $k = 1, 2$. Then the asymptotic covariance matrix for $\widehat{\boldsymbol{\theta}}_{(1)}$ is given by (2.15) if $\mathrm{E}_{\boldsymbol{\theta}} \frac{\partial \boldsymbol{\psi}^{(1)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(2)}} = \mathbf{0}$.

Lemma 2.3

$$\mathit{Asymptotic\ Var}(\widehat{\boldsymbol{\theta}}_{(1)}) = \mathbf{S}_{11}^{-1}(\boldsymbol{\theta})\mathbf{V}_{11}(\boldsymbol{\theta})\mathbf{S}_{11}^{-1}(\boldsymbol{\theta})^{\top}, \tag{2.15}$$

where $\mathbf{S}_{kl}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \frac{\partial \psi^{(k)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(l)}^{\top}}$ *and* $\mathbf{V}_{kl}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left\{ \psi^{(k)}(\boldsymbol{\theta}) \psi^{(l)}(\boldsymbol{\theta})^{\top} \right\}$ $k, l = 1, 2.$

Knudsen (1998) obtained this result in his unpublished thesis. Note that since $\mathbf{S}_{12}(\boldsymbol{\theta}) = \mathbf{0}$, his proof is straightforward, and is based on the following equation:

$$\begin{pmatrix} \mathbf{S}_{11}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{S}_{21}(\boldsymbol{\theta}) & \mathbf{S}_{22}(\boldsymbol{\theta}) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{V}_{11}(\boldsymbol{\theta}) & \mathbf{V}_{12}(\boldsymbol{\theta}) \\ \mathbf{V}_{21}(\boldsymbol{\theta}) & \mathbf{V}_{22}(\boldsymbol{\theta}) \end{pmatrix} \begin{pmatrix} \mathbf{S}_{11}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{S}_{21}(\boldsymbol{\theta}) & \mathbf{S}_{22}(\boldsymbol{\theta}) \end{pmatrix}^{-\top}$$

$$= \begin{pmatrix} \mathbf{S}_{11}^{-1}(\boldsymbol{\theta}) \mathbf{V}_{11}(\boldsymbol{\theta}) \mathbf{S}_{11}^{-1}(\boldsymbol{\theta})^{\top} & * \\ * & * \end{pmatrix}.$$

Clearly the upper left block of the right hand side, $\mathbf{S}_{11}^{-1}(\boldsymbol{\theta}) \mathbf{V}_{11}(\boldsymbol{\theta}) \mathbf{S}_{11}^{-1}(\boldsymbol{\theta})^{\top}$, is the asymptotic covariance of $\widehat{\boldsymbol{\theta}}_{(1)}$ since the right hand side is the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$.

The asymptotic covariance of $\widehat{\boldsymbol{\theta}}_{(1)}$ will, in general, be affected by both the variance of the nuisance parameter estimator and the variability for the estimating function for the nuisance parameter. However, under nuisance parameter insensitivity, that is, $\mathrm{E}_{\boldsymbol{\theta}} \frac{\partial \psi^{(1)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(2)}} = \mathbf{0}$, this result tells us that the asymptotic covariance matrix of the estimator for the parameter of interest will be affected by the variability and sensitivity of the estimating function for the parameter of interest, but not by the remaining parts of the variability and sensitivity matrices for the estimating function.

2.4 Data examples

In this section, we describe some clustered data examples where over-dispersion and heterogeneity may exist. We will analyze these data sets in Chapter 8.

2.4.1 Epilepsy data

Thall and Vail (1990) presented longitudinal data (see Table A.2) from a clinical trial of 59 epileptics who were randomized to a new anti-epileptic drug progabide (Trt=1) or a placebo (Trt=0). Baseline data available at the start of the trial included the number of epileptic seizures during the 8-week period prior to randomization and the patient's age. A multivariate response variable consisted of the counts of seizures during the 2-week periods before each of four visits to the clinic.

Count data are traditionally modelled by the Poisson distribution; however the Poisson assumption that $\mathrm{E}(Y) = \mathrm{Var}(Y)$ is generally inconsistent with the empirical behaviour for clustered data. A set of count data is called over-dispersed or under-dispersed if $\mathrm{Var}(Y) \geq \mathrm{E}(Y)$ or $\mathrm{Var}(Y) \leq \mathrm{E}(Y)$, respectively. Over-dispersion is very common in biological data. An over-dispersion diagnostic plot for count data relative to Poisson regression models (Lambert and Roeder, 1995) is displayed for the epilepsy data in Figure 2.1. The graph is convex if over-dispersion exists with respect to the corresponding generalized linear model. The clear convexity of the over-dispersion diagnostic graph for the epilepsy data with respect to the Poisson generalized linear model may indicate the existence of over-dispersion.

Subject (random) effects are often incorporated into Poisson models to account for heterogeneity, over-dispersion relative to Poisson models, and dependence among the repeated measurements within the same subject. It is natural to introduce a second level of random effects at familial level to account for familial aggregation of epilepsy (Paik, Tsai and Ottman 1994); however we do not have such familial grouping information for this data set. A second level of random effects is often introduced to

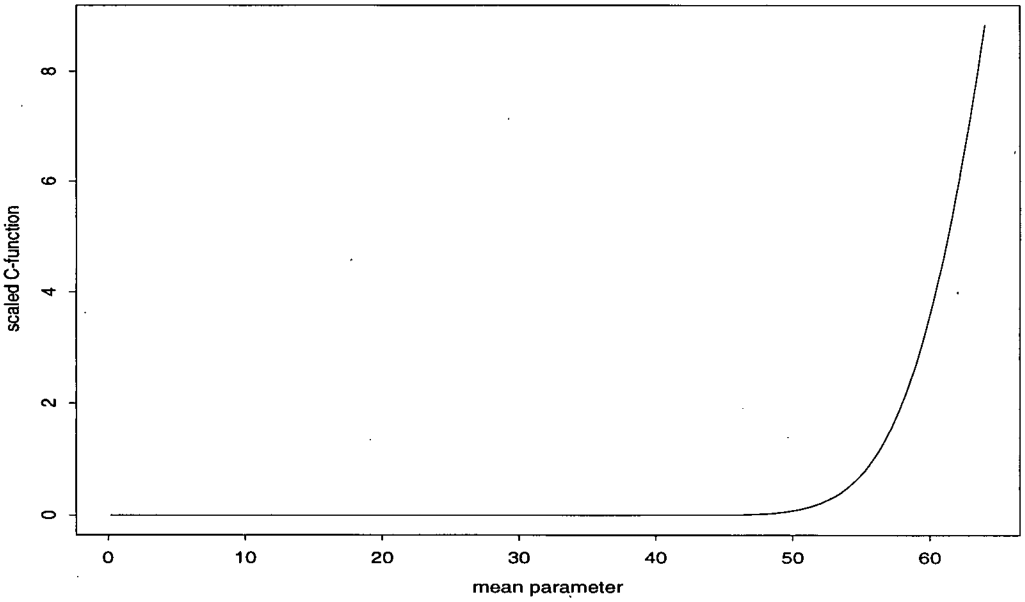


Figure 2.1: Overdispersion diagnostic plot for epilepsy data.

account for uncontrollable covariates at each visit (Lee and Nelder 1996). We will thus consider generalized linear models with two levels of random effects in next section. Such two levels of random effects arises from many practical situations (Hedeker and Gibbons 1994; Goldstein 1995).

2.4.2 Seed germination data

Crowder (1978) presented data (see Table A.1) on the proportion of seeds that germinated on each of 21 plates arranged according to a 2×2 factorial layout by seed variety and type of root extract. In particular, he presented the total number of seeds on each plate and the number of seeds germinated on that plate (Table A.1). He noted that there is between-plate heterogeneity of proportions. Figure 2.2 shows that these proportions vary from 0 to 0.83 among plates. Furthermore, the variability among clustered binary responses also often exceeds what would be expected due to binomial variation alone. It is natural to account for such heterogeneity and overdispersion by means of random effects models.

2.4.3 Cake baking data

Cochran and Cox (1957) presented data (see Table A.3) from an experiment in baking chocolate cakes. Three recipes were tested and each recipe was replicated 15 times, giving total of 45 batches of cake mix. Each batch was divided into six cakes and these were baked at six different temperatures. After baking, a breaking angle was measured as the response of the experiment. Firth and Harris (1991) found the residual plots based on the analysis of variance revealed heterogeneity between batches. This can also be seen from the boxplot of the responses by batches in Figure 2.3.

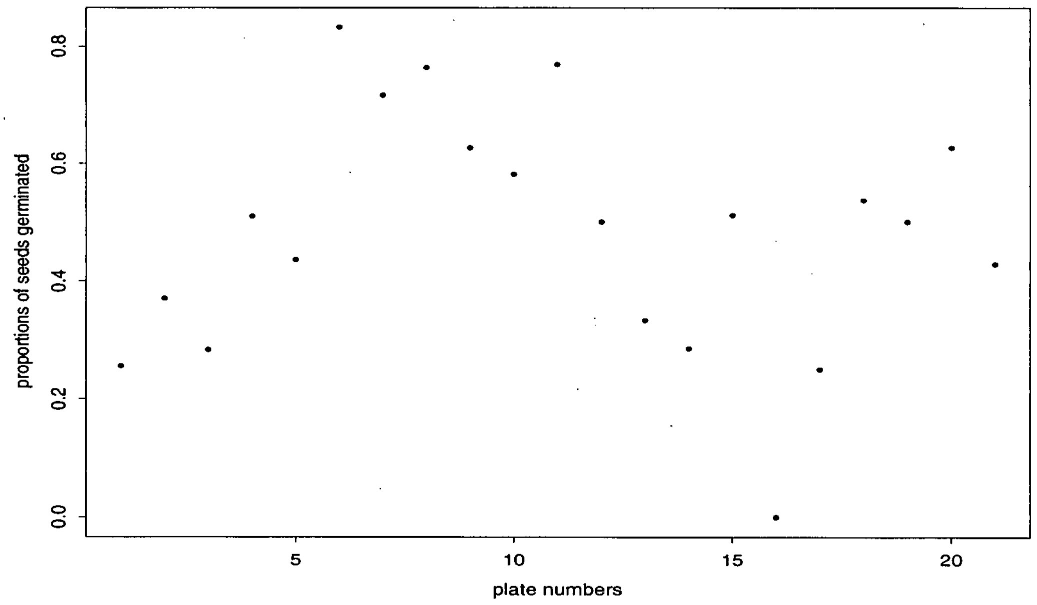


Figure 2.2: Scatter plot of proportions of seeds germinated versus plates.

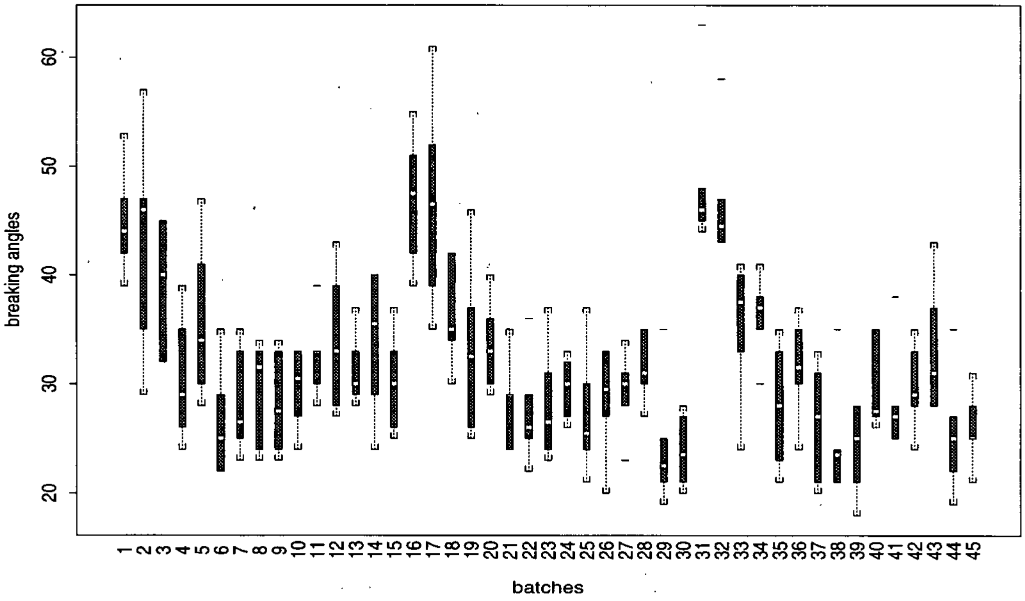


Figure 2.3: Boxplots of cake baking data by batches.

Chapter 3

Tweedie mixed models

In this chapter, we introduce a class of models with nested random effects based on the class of Tweedie exponential dispersion model distributions. We call these models Tweedie mixed models. When the conditional responses follow a specific distribution, say, the Poisson or gamma, we call it a Poisson mixed model or gamma mixed model, respectively. For a model with Poisson responses and gamma random effects distributions, we call it Poisson-gamma model. We will also discuss the covariance structure of the model.

3.1 Models

In this section, we consider three-level hierarchical models where each model is composed of m independent clusters indexed by i. Within each cluster i, there are J_i correlated sub-clusters indexed by (i, j). Then within each sub-cluster (i, j) there are n_{ij} correlated observations. Let the vector of observations be denoted by $\mathbf{Y} = (Y_{111}, ..., Y_{11n_{11}}, ..., Y_{mJ_m1}, ..., Y_{mJ_mn_{mJ_m}})^\top$. Then Y_{ijk} represents the kth observation in sub-cluster (i, j). One such hierarchy would be a multi-center longitudinal clinical trial involving repeated measurements within patients and patients nested within study centers. Another example is teaching method evaluation involving high scool students within classes and classes nested within schools. (Goldstein 1995; Gray et al. 1995). Y_{ijk} denotes the kth measurement for patient j taken in the ith medical center for the former example, whereas the represents the test score for student k in class j of the ith school.

Denote the cluster, sub-cluster and observation covariates by $\mathbf{z}_i, \mathbf{z}_{ij}$ and $\mathbf{z}_{ijk}$, respectively. We assume that there exist cluster and sub-cluster specific random effects. The random effects for the cluster i and sub-cluster (i, j) are denoted by U_i and U_{ij}, respectively. Thus the vector of the random effects can be written as $\mathbf{U} = (U_1, ..., U_m, U_{11}, ..., U_{mJ_m})^\top = (U_*, U_{**})$, where U_* and U_{**} stand for $(U_1, ..., U_m)^\top$ and $(U_{11}, ..., U_{mJ_m})^\top$, for short. We assume further that, given the random effects, the responses are independent and follow certain Tweedie distributions. Specifically:

A1) Given $\mathbf{U} = \mathbf{u}$, $Y_{111}, ..., Y_{11n_{11}}, ..., Y_{ij1}, ..., Y_{ijn_{ij}}, ..., Y_{mJ_m1}, ..., Y_{mJ_mn_{mJ_m}}$ are conditionally independent, and the conditional distribution of Y_{ijk}, given $\mathbf{U} = \mathbf{u}$, depends on u_{ij} only which is

$$\begin{aligned} Y_{ijk}|\mathbf{U} = \mathbf{u} &\sim \mathrm{Tw}_p(\mu_{ijk}u_{ij}, \rho^2 {u_{ij}}^{1-p}) \\ &= u_{ij}\mathrm{Tw}_p\left(\mu_{ijk}, \frac{\rho^2}{u_{ij}}\right), \end{aligned} \tag{3.1}$$

where $\mu_{ijk} = \exp(\mathbf{z}_{ijk}^\top \boldsymbol{\beta}_{(3)})$. For case $p = 1$, namely the Poisson distribution, $\rho^2 = 1$.

Furthermore we assume that, given the cluster level random effects, the sub-cluster level random effects are independent and follow Tweedie distributions as follows:

A2) Given $\mathbf{U}_* = \mathbf{u}_*$, $U_{11}, ..., U_{mJ_m}$ are conditionally independent, and the conditional distribution of U_{ij}, given $\mathbf{U}_* = \mathbf{u}_*$, depends on u_i only which is

$$\begin{aligned} U_{ij}|\mathbf{U}_* = \mathbf{u}_* &\sim \mathrm{Tw}_q(\mu_{ij}u_i, \omega^2 {u_i}^{1-q}) \\ &= u_i\mathrm{Tw}_q\left(\mu_{ij}, \frac{\omega^2}{u_i}\right). \end{aligned} \tag{3.2}$$

where $\mu_{ij} = \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}_{(2)})$.

A3) Finally we assume that the cluster level random effects are independent with

$$U_i \sim \mathrm{Tw}_r(\mu_i, \sigma^2), \tag{3.3}$$

where $\mu_i = \exp(\mathbf{z}_i^\top \boldsymbol{\beta}_{(1)})$.

Let $\mathbf{x}_{ijk}^\top = (\mathbf{z}_i^\top, \mathbf{z}_{ij}^\top, \mathbf{z}_{ijk}^\top)$. Then the full covariate matrix is $\mathbf{X} = (\mathbf{x}_{111}, \ldots, \mathbf{x}_{mJ_m n_{mJ_m}})^\top$. The Tweedie distributions Tw_r, Tw_q and Tw_p are called the level 1, level 2 and level 3 distributions, respectively. To avoid non-positive random effects, $r \geq 2$ and $q \geq 2$ are required.

Now we discuss some implications of the model assumptions. First, note that (3.1) interprets U_{ij} as weights. Note also that the dispersion component takes a special form $\rho^2 U_{ij}^{1-p}$ for the assumed level 3 Tweedie distribution in assumption A1). This special form will enable us to obtain unbiased estimating equation for the regression parameters later. It also leads to variance component structure of the covariance matrix which will be shown in the next section. This special dispersion component form also gives the linear structure of the conditional variance of the level 3 distributions. Thus we have the following *double linearity*:

Double linearity

$$\mathrm{E}[Y_{ijk}|\mathbf{U}] = \mu_{ijk}U_{ij}, \qquad \mathrm{Var}[Y_{ijk}|\mathbf{U}] = \rho^2\mu_{ijk}^p U_{ij}.$$

The first statement is clear from the assumption, whereas the second statement is immediately verified as follows:

$$\mathrm{Var}[Y_{ijk}|\mathbf{U}] = \rho^2 U_{ij}^{1-p}\mu_{ijk}^p U_{ij}^p = \rho^2\mu_{ijk}^p U_{ij}.$$

That is, both the conditional mean and the conditional variance of the level 3 distribution is linear in the second level random effects. In fact, for any given p, the double linearity is equivalent to the assumption A1). Similarly, for any given q, A2) is equivalent to the following double linearity of the level 2 distributions.

$$\mathrm{E}[U_{ij}|\mathbf{U}_*] = \mu_{ij}U_i, \qquad \mathrm{Var}[U_{ij}|\mathbf{U}_*] = \omega^2\mu_{ij}^q U_i.$$

We discuss some decompositions of the models. We consider the structure of the conditional mean. In the literature, the modeling of within cluster dependence is usually focused on incorporating random effects into a monotonic transform of the conditional mean assuming constant dispersion components. For multiplicative models, it is usually defined as

$$\mathrm{E}[Y_{ijk}|\mathbf{U},\mathbf{V}] = \mu_{ijk}U_iV_{ij}.$$

This implies that

$$\mathrm{Var}[Y_{ijk}|\mathbf{U}] = \omega^2 V(\mu_{ijk}U_iV_{ij})$$

where $\mathbf{V}(\cdot)$ is the variance function for the conditional distribution of Y_{ijk}, given $(\mathbf{U},\mathbf{V})$. In addition, U_i and V_{ij} are often taken as independent, but the covariances of the responses in general do not then exhibit a variance component structure (Morton 1987; Thall and Vail 1990; Firth and Harris 1991; Lee and Nelder 1996). In our model assumptions, the conditional expectation of $\mathbf{Y}$ given $\mathbf{U}$ is linear in the random effects. To compare our model assumptions with others, we give the following *multiplicative decomposition* for the conditional expectation of $\mathbf{Y}$ given $\mathbf{U}$.

Multiplicative decomposition

$$\mathrm{E}[Y_{ijk}|\mathbf{U}] = \mu_{ijk}U_{ij} = \mu_{ijk}U_i\frac{U_{ij}}{U_i} = \mu_{ijk}U_iV_{ij},$$

where $V_{ij} = \frac{U_{ij}}{U_i}$ and $\mathrm{Cov}[U_i, V_{ij}] = 0$. The latter can be shown as follows:

$$\begin{aligned} \mathrm{E}[\frac{U_{ij}}{U_i}|\mathbf{U}_*] &= \mathrm{E}\left\{\frac{\mathrm{E}(U_{ij}|\mathbf{U}_*)}{U_i}\right\} \\ &= \mathrm{E}\left\{\frac{\mu_{ij}U_i}{U_i}\right\} = \mu_{ij}, \end{aligned}$$

therefore

$$\begin{aligned} \mathrm{Cov}[U_i, V_{ij}] &= \mathrm{E}\left(U_i\frac{U_{ij}}{U_i}\right) - \mathrm{E}(U_i)\mathrm{E}\left(\frac{U_{ij}}{U_i}\right) \\ &= \mathrm{E}\left(U_{ij}\right) - \mu_i\mu_{ij} = 0. \end{aligned}$$

Additive decomposition

Besides the above multiplicative decomposition, the response of this model also possesses the following *additive decomposition* with three uncorrelated, but dependent components as follows:

$$Y_{ijk} = (Y_{ijk} - \mu_{ijk}U_{ij}) + (\mu_{ijk}U_{ij} - \mu_{ij}\mu_{ijk}U_i) + \mu_{ij}\mu_{ijk}U_i. \tag{3.4}$$

This additive decomposition will facilitate residual analysis. The two sides of the equation are clearly equal. The three components on the right-hand side can be easily shown to be uncorrelated through the covariance structure described in the next section.

Tweedie mixed models with one level of random effects

A Tweedie mixed model with one level of random effects is a special case of the Tweedie mixed model with two levels of random effects by setting $\omega^2 = 0$ and $J_i = 1$ for all i.

3.2 Covariance structure

The derivations of both our parameter estimators and random effects predictors are based on the moment structure of the model. Thus we investigate the moment structure of the model here.

3.2.1 Derivation

We begin our investigation with the moments of the random effects. The intra-dependence within clusters are clearly reflected by the covariance structure described below. The derivations of the following moment expressions are straightforward using the conditioning technique. Using Kronecker notation $\delta_{(s,i)}$ being 1 if $s = i$, 0 otherwise, the covariance structure can be expressed in the following way:

$$\mathrm{E}[U_i] = \mu_i \text{ and } \mathrm{Cov}[U_s, U_i] = \delta(s,i)\sigma^2\mu_i^r.$$

$$\mathrm{E}[U_{ij}|U_*] = \mu_{ij}U_i \text{ and } \mathrm{Var}[U_{ij}|U_*] = \omega^2\mu_{ij}^q U_i.$$

$$\mathrm{Cov}[U_s, U_{ij}] = \delta(s,i)\sigma^2\mu_i^r\mu_{ij},$$

$$\mathrm{E}[U_{ij}] = \mu_i\mu_{ij}, \tag{3.5}$$

$$\mathrm{Cov}[U_{st}, U_{ij}] = \delta(s,i)\left\{\sigma^2\mu_i^r\mu_{ij}\mu_{it} + \delta(t,j)\omega^2\mu_i\mu_{ij}^q\right\}, \tag{3.6}$$

$$\mathrm{Cov}[U_s, Y_{ijk}] = \delta(s,i)\sigma^2\mu_i^r\mu_{ij}\mu_{ijk},$$

$$\mathrm{Cov}[U_{st}, Y_{ijk}] = \delta(s,i)\left\{\sigma^2\mu_i^r\mu_{ij}\mu_{it} + \delta(t,j)\omega^2\mu_i\mu_{ij}^q\right\}\mu_{ijk}, \tag{3.7}$$

$$\mathrm{E}[Y_{ijk}|\mathbf{U}] = \mu_{ijk}U_{ij}.$$

$$\mathrm{Var}[Y_{ijk}|\mathbf{U}] = \rho^2\mu_{ijk}^pU_{ij}.$$

$$\mathrm{E}[Y_{ijk}] = \mu_i\mu_{ij}\mu_{ijk} = \exp(\mathbf{x}_{ijk}^\top\boldsymbol{\beta}), \tag{3.8}$$

therefore the link function is log.

$$\begin{aligned}\mathrm{Cov}[Y_{stl}, Y_{ijk}] &= \delta(s,i)\left\{\sigma^2\mu_i^r\mu_{ij}\mu_{it}\mu_{ijk}\mu_{itl}\right.\\ &+ \delta(t,j)[\omega^2\mu_i\mu_{ij}^q\mu_{ijk}\mu_{ijl}\\ &+ \left.\delta(l,k)\rho^2\mu_i\mu_{ij}\mu_{ijk}^p]\right\}.\end{aligned} \tag{3.9}$$

Since all the derivations of these moment structures are similar, we show, as an example, the derivations of covariance of the responses. Given $\mathbf{U}$, $Y_{111}, \ldots, Y_{11n_{11}}, \ldots, Y_{ij1}$, $\ldots, Y_{ijn_{ij}}, \ldots, Y_{mJ_m1}, \ldots, Y_{mJ_mn_{mJ_m}}$ are conditionally independent, so we have $\mathrm{Cov}[Y_{stl}, Y_{ijk}|\mathbf{U}] = 0$ if $(s,t,l) \neq (i,j,k)$, thus

$$\begin{aligned}\mathrm{Cov}[Y_{stl}, Y_{ijk}|\mathbf{U}] &= \delta(s,i)\delta(t,j)\delta(l,k)\mathrm{Var}(Y_{ijk}|\mathbf{U})\\ &= \delta(s,i)\delta(t,j)\delta(l,k)\rho^2\mu_{ijk}^pU_{ij}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{Cov}[Y_{stl}, Y_{ijk}] &= \mathrm{E}\left\{\mathrm{Cov}[Y_{stl}, Y_{ijk}|\mathbf{U}]\right\} + \mathrm{Cov}[\mathrm{E}(Y_{stl}|\mathbf{U}), \mathrm{E}(Y_{ijk}|\mathbf{U})]\\ &= \delta(s,i)\delta(t,j)\delta(l,k)\mathrm{E}[\mathrm{Var}(Y_{ijk}|\mathbf{U})]\end{aligned}$$

$$\begin{aligned} & +\mathrm{Cov}[\mu_{stl}U_{st}, \mu_{ijk}U_{ij}] \\ = \; & \delta(s,i)\delta(t,j)\delta(l,k)\rho^2\mu_{ijk}^{p}\mathrm{E}(U_{ij}) \\ & +\mu_{stl}\mu_{ijk}\mathrm{Cov}[U_{st}, U_{ij}]. \end{aligned} \tag{3.10}$$

The proof is completed by plugging (3.5) and (3.6) into the last equation.

Now we return to (3.4) and verify that the three components on the right-hand side are uncorrelated. The verifications are very similar so we only show that the first two components are uncorrelated. It follows from (3.6) and (3.7) that

$$\mathrm{Cov}[Y_{ijk}, U_{ij}] = \mu_{ijk}\mathrm{Var}(U_{ij});$$

hence

$$\begin{aligned} \mathrm{Cov}[Y_{ijk} - \mu_{ijk}U_{ij}, \mu_{ijk}U_{ij}] &= \mu_{ijk}\mathrm{Cov}[Y_{ijk}, U_{ij}] - \mu_{ijk}^2\mathrm{Cov}[U_{ij}, U_{ij}] \\ &= \mu_{ijk}^2\mathrm{Var}(U_{ij}) - \mu_{ijk}^2\mathrm{Var}(U_{ij}) = 0. \end{aligned}$$

Similarly we have

$$\mathrm{Cov}[Y_{ijk} - \mu_{ijk}U_{ij}, \mu_{ij}\mu_{ijk}U_i] = 0.$$

So the result follows immediately. Finally, note that

$$\mathrm{Corr}[Y_{ijk}, Y_{stl}] = \frac{\mathrm{Cov}[Y_{ijk}, Y_{stl}]}{\sqrt{\mathrm{Var}(Y_{ijk})}\sqrt{\mathrm{Var}(Y_{stl})}}$$

clearly depends on the mean parameters.

3.2.2 Matrix expressions

To facilitate the derivations of the quantities of interest in the rest of the thesis, it is desirable to express the moment structure in matrix form. After introducing some matrix notation, we will state the covariance matrices between random effects and the responses in matrix form. The matrix form will be useful in the derivation of the orthodox BLUP predictors of the random effects. We will then concentrate on the derivation of the inverse of the covariance matrix of the response since this inverse plays a key role in all stages of the derivation of the orthodox BLUP approach.

Let us first introduce some matrix notation. Let $\mathbf{Y}_i$ denote the responses corresponding to the ith cluster, that is, $(Y_{i11}, \ldots, Y_{i1n_{i1}}, \ldots, Y_{iJ_i1}, \ldots, Y_{iJ_in_{iJ_i}})^\top$. Let $\boldsymbol{\mu}_{i*} = (\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{iJ_i})$, $\boldsymbol{\mu}_{ij*} = (\boldsymbol{\mu}_{ij1}, \ldots, \boldsymbol{\mu}_{ijn_{ij}})^\top$, $\boldsymbol{\mu}_{i**} = (\boldsymbol{\mu}_{i1*}^\top, \ldots, \boldsymbol{\mu}_{iJ_i*}^\top)$ and $\boldsymbol{\nu}_{ij*} = (\mu_i\mu_{ij}\mu_{ij1}, \ldots, \mu_i\mu_{ij}\mu_{ijn_{ij}})^\top$. Then

$$\mathrm{E}(\mathbf{Y}_i) = \boldsymbol{\nu}_i = (\boldsymbol{\nu}_{i1*}^\top, \ldots, \boldsymbol{\nu}_{iJ_i*}^\top)^\top.$$

In addition, for any $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{a}^r$ is defined as $(a_1^r, \ldots, a_n^r)$. Then we have

$$\mathrm{Cov}[U_i, \mathbf{Y}_i] = \sigma^2\mu_i^{r-1}\boldsymbol{\nu}_i^\top, \tag{3.11}$$

$$\mathrm{Cov}[U_{ij}, \mathbf{Y}_i] = \sigma^2\mu_i^{r-1}\mu_{ij}\boldsymbol{\nu}_i^\top + \omega^2\mu_{ij}^{q-1}(\mathbf{0}^\top, \ldots, \boldsymbol{\nu}_{ij*}^\top, \ldots, \mathbf{0}^\top). \tag{3.12}$$

Now we derive an explicit expression for the inverse of $\mathrm{Var}(\mathbf{Y})$. Note first that

$$\mathrm{Var}^{-1}(\mathbf{Y}) = \begin{pmatrix} \mathrm{Var}^{-1}(\mathbf{Y}_1) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathrm{Var}^{-1}(\mathbf{Y}_m) \end{pmatrix}$$

since different clusters are independent. Thus it is enough to derive the inverse of $\mathrm{Var}(\mathbf{Y}_i)$. The inversion of $\mathrm{Var}(\mathbf{Y}_i)$ can be further simplified by noting that $\mathrm{Var}(\mathbf{Y}_i)$ is a patterned matrix which reflects the hierarchy. To find a simple expression for

$\text{Var}(\mathbf{Y}_i)$, we define

$$\mathbf{A}_{ij} = \rho^2 \mu_i \mu_{ij} \begin{pmatrix} \mu_{ij1}^p & & 0 \\ & \ddots & \\ 0 & & \mu_{ijn_{ij}}^p \end{pmatrix}_{n_{ij} \times n_{ij}} + \omega^2 \mu_i \mu_{ij} \boldsymbol{\mu}_{ij*} \boldsymbol{\mu}_{ij*}^\top.$$

Then

$$\text{Var}(\mathbf{Y}_i) = \begin{pmatrix} \mathbf{A}_{i1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{A}_{iJ_i} \end{pmatrix} + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top.$$

Therefore $\text{Var}(\mathbf{Y}_i)$ can be expressed as a positive-definite matrix plus a matrix product between a column vector and a row vector. Thus we can invert $\text{Var}(\mathbf{Y}_i)$ using the following well-known formula (cf. Rao 1973)

$$(\mathbf{A} + \mathbf{a}\mathbf{b}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{b}^\top \mathbf{A}^{-1}}{1 + \mathbf{b}^\top A \mathbf{a}}, \tag{3.13}$$

where $\mathbf{a}$ and $\mathbf{b}$ are vectors and $\mathbf{A}$ is invertible. With

$$\mathbf{A}_i = \begin{pmatrix} \mathbf{A}_{i1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{A}_{iJ_i} \end{pmatrix},$$

$\mathbf{a} = \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i$ and $\mathbf{b} = \boldsymbol{\nu}_i$, we obtain

$$\text{Var}(\mathbf{Y}_i)^{-1} = \left(\mathbf{A}_i + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top\right)^{-1} = \mathbf{A}_i^{-1} - \frac{\sigma^2 \mu_i^{r-2} \mathbf{A}_i^{-1} \boldsymbol{\nu}_i (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)^\top}{1 + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i^\top (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)}, \tag{3.14}$$

where

$$\mathbf{A}_i^{-1} = \begin{pmatrix} \mathbf{A}_{i1}^{-1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{A}_{iJ_i}^{-1} \end{pmatrix}.$$

Now

$$\mathbf{A}_i^{-1}\boldsymbol{\nu}_i = \begin{pmatrix} \mathbf{A}_{i1}^{-1}\boldsymbol{\nu}_{i1*} \\ \vdots \\ \mathbf{A}_{iJ_i}^{-1}\boldsymbol{\nu}_{iJ_i*} \end{pmatrix},$$

where each block $\mathbf{A}_{ij}$ is again of the form (3.13). Applying the same matrix inversion formula to the $\mathbf{A}_{ij}$ gives

$$\mathbf{A}_{ij}^{-1} = \frac{1}{\rho^2\mu_i\mu_{ij}} \begin{pmatrix} \frac{1}{\mu_{ij1}^p} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\mu_{ijn_{ij}}^p} \end{pmatrix}_{n_{ij}\times n_{ij}} - \frac{\omega^2\mu_{ij}^{q-1}w_{ij}}{\rho^2\mu_i\mu_{ij}}\mu_{ij*}^{1-p}(\mu_{ij*})^{1-p}, \qquad (3.15)$$

where $w_{ij} = 1/(\rho^2 + \omega^2\mu_{ij}^{q-1}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p})$. Thus we have

$$\begin{aligned} \mathbf{A}_{ij}^{-1}\boldsymbol{\nu}_{ij*} &= \frac{1}{\rho^2}\mu_{ij*}^{1-p} - \frac{\omega^2\mu_{ij}^{q-1}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p}w_{ij}}{\rho^2}\mu_{ij*}^{1-p} \\ &= w_{ij}\boldsymbol{\mu}_{ij*}^{1-p}, \end{aligned}$$

Plugging the last two equalities into (3.14) gives the explicit formula for the inverse of the covariance matrix of the responses. We will write out this explicit expression with simpler notation later. This explicit expression will be very useful in our discussion of random effects prediction and parameter estimation in the following chapters.

Chapter 4

Orthodox BLUP predictors of random effects

In this chapter, we study the orthodox BLUP of random effects. Robinson (1991) presented a variety of applications where the prediction of random effects is of interest. These applications arise from areas such as sports, genetic study, quality and management, time series, image analysis, geostatistics, actuarial science and small-area estimation. The prediction of random effects is also found to be very useful in the identification of outliers. Fellner (1986) discussed this topic with some examples. He found that looking at the predictors of random effects is often more sensitive for detecting outliers than merely looking at residuals.

We derive the explicit expressions for the orthodox predictors of random effects. Let us begin with the general expressions for orthodox BLUP.

4.1 Orthodox BLUP

Let $\mathbf{U}$ and $\mathbf{Y}$ be random vectors with finite second moments. We define the orthodox BLUP of $\mathbf{U}$ given $\mathbf{Y}$ by:

$$\widehat{\mathbf{U}} = \mathrm{E}(\mathbf{U}) + \mathrm{Cov}(\mathbf{U}, \mathbf{Y})\mathrm{Var}^{-1}(\mathbf{Y})\,(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))\,. \tag{4.1}$$

The mean squared distance between $\widehat{\mathbf{U}}$ and $\mathbf{U}$ can be evaluated through the following equation (Harvey 1981; Jørgensen et al. 1996b):

$$\begin{aligned} \mathrm{Var}(\widehat{\mathbf{U}} - \mathrm{U}) &= \mathrm{E}[(\widehat{\mathbf{U}} - \mathbf{U})(\widehat{\mathbf{U}} - \mathrm{U})^{\top}] \\ &= \mathrm{Var}(\mathbf{U}) - \mathrm{Cov}(\mathbf{U}, \mathbf{Y})\mathrm{Var}^{-1}(\mathbf{Y})\mathrm{Cov}(\mathbf{Y}, \mathbf{U}). \end{aligned} \tag{4.2}$$

This mean squared distance can be further decomposed into the sum of the mean squared distance between orthodox BLUP of $\mathbf{U}$ given $\mathbf{Y}$ and the conditional expectation of $\mathbf{U}$ given $\mathbf{Y}$ and mean squared distance between the conditional expectation of $\mathbf{U}$ given $\mathbf{Y}$ and $\mathbf{U}$ as follows:

$$\begin{aligned} \mathrm{E}[(\widehat{\mathbf{U}} - \mathbf{U})(\widehat{\mathbf{U}} - \mathbf{U})^{\top}] &= \mathrm{E}[(\mathrm{E}(\widehat{\mathbf{U}} - \mathrm{E}(\mathbf{U}|\mathbf{Y}))(\widehat{\mathbf{U}} - \mathrm{E}(\mathbf{U}|\mathbf{Y}))^{\top}] \\ &\quad + \mathrm{E}[(\mathrm{E}(\mathbf{U}|\mathbf{Y}) - \mathbf{U})(\mathrm{E}(\mathbf{U}|\mathbf{Y}) - \mathbf{U})^{\top}]. \end{aligned} \tag{4.3}$$

In addition, we have the following two desirable orthogonality properties concerning the orthodox BLUP:

$$\mathrm{Cov}[\widehat{\mathbf{U}} - \mathbf{U}, \widehat{\mathbf{U}}] = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}[\widehat{\mathbf{U}} - \mathbf{U}, \mathbf{Y}] = \mathbf{0}. \tag{4.4}$$

That is, the residuals between the random effects and their orthodox BLUP predictors are orthogonal to both the response and the predictors. These orthogonality properties will be repeatedly used later.

The first two moments of the orthodox BLUP are as follows:

$$\begin{aligned}\mathrm{E}(\widehat{\mathbf{U}}) &= \mathrm{E}(\mathbf{U}),\\ \mathrm{Var}(\widehat{\mathbf{U}}) &= \mathrm{Cov}(\mathbf{U},\mathbf{Y})\mathrm{Var}^{-1}(\mathbf{Y})\mathrm{Cov}(\mathbf{Y},\mathbf{U}).\end{aligned} \tag{4.5}$$

The first equation follows immediately from (4.1) since $\mathrm{E}[\mathbf{Y}-\mathrm{E}(\mathbf{Y})]=\mathbf{0}$. The second statement can be easily verified by using (4.4) as follows:

$$\begin{aligned}\mathrm{Var}(\mathbf{U}) &= \mathrm{Var}(\mathbf{U}-\widehat{\mathbf{U}}+\widehat{\mathbf{U}})\\ &= \mathrm{Var}(\mathbf{U}-\widehat{\mathbf{U}})+\mathrm{Var}(\widehat{\mathbf{U}})+2\mathrm{Cov}(\mathbf{U}-\widehat{\mathbf{U}},\widehat{\mathbf{U}})\\ &= \mathrm{Var}(\mathbf{U}-\widehat{\mathbf{U}})+\mathrm{Var}(\widehat{\mathbf{U}})+\mathbf{0}.\end{aligned} \tag{4.6}$$

The verification is completed by noting (4.2). Comparing (4.2) with (4.5), we have

$$\mathrm{Var}(\widehat{\mathbf{U}}) \le \mathrm{Var}(\mathbf{U}).$$

Thus the variance of the orthodox BLUP predictor of random effects is generally smaller than that of the random effects. Thus the orthodox BLUP predictor is also referred to as the shrinkage predictor of random effects.

4.2 Random effects predictors

Explicit expressions for the orthodox BLUP predictors of random effects $\mathbf{U}$ given $\mathbf{Y}$ can be derived from (4.1). Since different clusters are independent, we derive the random effects predictors from the following two formulae:

$$\widehat{U}_i = \mathrm{E}(U_i) + \mathrm{Cov}(U_i,\mathbf{Y}_i)\mathrm{Var}(\mathbf{Y}_i)^{-1}\left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right), \tag{4.7}$$

and

$$\widehat{U}_{ij} = \mathrm{E}(U_{ij}) + \mathrm{Cov}(U_{ij},\mathbf{Y}_i)\mathrm{Var}(\mathbf{Y}_i)^{-1}\left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right). \tag{4.8}$$

Let us first derive the explicit expression for the orthodox BLUP predictor $\widehat{U}_i$. It is more convenient to derive the expression using a matrix form based on (3.11), (3.12) and (3.14) as follows:

$$\begin{aligned}
\widehat{U}_i &= \mu_i + \sigma^2\mu_i^{r-1}\boldsymbol{\nu}_i^\top\left\{\mathbf{A}_i^{-1} - \frac{\sigma^2\mu_i^{r-2}\mathbf{A}_i^{-1}\boldsymbol{\nu}_i(\mathbf{A}_i^{-1}\boldsymbol{\nu}_i)^\top}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}\right\}(\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \mu_i + \sigma^2\mu_i^{r-1}\left\{1 - \frac{\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}\right\}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \mu_i + \sigma^2\mu_i^{r-1}\frac{1}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \frac{\mu_i + \sigma^2\mu_i^{r-1}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\mathbf{Y}_i}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}. \qquad (4.9)
\end{aligned}$$

Noting that

$$\begin{aligned}
\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1} = (\mathbf{A}_i^{-1}\boldsymbol{\nu}_i)^\top &= \left((\mathbf{A}_{i1}^{-1}\boldsymbol{\nu}_{i1*})^\top, \ldots, (\mathbf{A}_{iJ_i}^{-1}\boldsymbol{\nu}_{iJ_i*})^\top\right) \\
&= \left(w_{i1}(\boldsymbol{\mu}_{i1*}^{1-p})^\top, \ldots, w_{iJ_i}(\boldsymbol{\mu}_{iJ_i*}^{1-p})^\top\right), \qquad (4.10)
\end{aligned}$$

we have

$$\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i = \mu_i\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p}.$$

Hence

$$\widehat{U}_i = \frac{\mu_i + \sigma^2\mu_i^{r-1}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ijk}^{1-p}Y_{ijk}}{1+\sigma^2\mu_i^{r-1}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p}}, \qquad (4.11)$$

where $w_{ij} = 1/(\rho^2 + \omega^2\mu_{ij}^{q-1}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p})$ as defined in (3.15).

Thus the orthodox BLUP predictor for each cluster i is a linear function of responses within the cluster. A reasonable predictor of U_i should be nonnegative since U_i is. The orthodox BLUP predictor $\widehat{U}_i$ is clearly nonnegative when the responses are. In fact, we mainly concentrate on models with nonnegative responses in this thesis.

Similarly we can derive an explicit expression for $\widehat{U}_{ij}$. We introduce a notation $\mathbf{e}_{ij}$ which denotes $(\mathbf{0}^\top, \ldots, \boldsymbol{\nu}_{ij*}^\top, \ldots, \mathbf{0}^\top)^\top$. We have

$$\operatorname{Cov}[U_{ij}, \mathbf{Y}_i] = \sigma^2 \mu_i^{r-1} \mu_{ij} \boldsymbol{\nu}_i^\top + \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top. \tag{4.12}$$

Thus

$$\begin{aligned}
\widehat{U}_{ij} &= (\sigma^2 \mu_i^{r-1} \mu_{ij} \boldsymbol{\nu}_i^\top + \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top) \operatorname{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top \operatorname{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top \left\{ \mathbf{A}_i^{-1} - \frac{\sigma^2 \mu_i^{r-2} \mathbf{A}_i^{-1} \boldsymbol{\nu}_i (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)^\top}{1 + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i^\top (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)} \right\} (\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top \mathbf{A}_i^{-1} \left\{ \mathbf{E} - \frac{\sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)^\top}{1 + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i^\top (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)} \right\} (\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} (\mathbf{A}_i^{-1} \mathbf{e}_{ij})^\top \left\{ \mathbf{Y}_i - \boldsymbol{\nu}_i - \frac{\sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)^\top}{1 + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i^\top (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)} (\mathbf{Y}_i - \boldsymbol{\nu}_i) \right\} \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} (\mathbf{A}_i^{-1} \mathbf{e}_{ij})^\top \left\{ \mathbf{Y}_i - \boldsymbol{\nu}_i \left(1 + \frac{\sigma^2 \mu_i^{r-2} (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)^\top (\mathbf{Y}_i - \boldsymbol{\nu}_i)}{1 + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i^\top (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)} \right) \right\} \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} (\mathbf{A}_i^{-1} \mathbf{e}_{ij})^\top \left\{ \mathbf{Y}_i - \boldsymbol{\nu}_i \frac{1 + \sigma^2 \mu_i^{r-2} (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)^\top \mathbf{Y}_i}{1 + \sigma^2 \mu_i^{r-2} \boldsymbol{\nu}_i^\top (\mathbf{A}_i^{-1} \boldsymbol{\nu}_i)} \right\} \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} (\mathbf{A}_i^{-1} \mathbf{e}_{ij})^\top (\mathbf{Y}_i - \frac{1}{\mu_i} \boldsymbol{\nu}_i \widehat{U}_i).
\end{aligned} \tag{4.13}$$

Plugging $\mathbf{A}_i^{-1} \mathbf{e}_{ij} = (\mathbf{0}^\top, \ldots, w_{ij} (\mu_{ij*}^{1-p})^\top, \ldots, \mathbf{0}^\top)^\top$ into (4.13), we obtain

$$\begin{aligned}
\widehat{U}_{ij} &= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} w_{ij} \sum_{k=1}^{n_{ij}} \mu_{ijk}^{1-p} (Y_{ijk} - \mu_{ij} \mu_{ijk} \widehat{U}_i) \\
&= \rho^2 w_{ij} \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} w_{ij} \sum_{k=1}^{n_{ij}} \mu_{ijk}^{1-p} Y_{ijk},
\end{aligned} \tag{4.14}$$

where the second expression is obtained from the identity

$$\rho^2 w_{ij} = 1 - \omega^2 \mu_{ij}^{q-1} w_{ij} \sum_{k=1}^{n_{ij}} \mu_{ijk}^{2-p}.$$

The first expression for $\widehat{U}_{ij}$ shows that $\widehat{U}_{ij}$ is $\mu_{ij}\widehat{U}_i$ adjusted by a linear function of the responses within sub-cluster (i, j). The second expression shows that $\widehat{U}_{ij}$ is also nonnegative when the responses are.

4.3 Mean squared distance

To evaluate the distance between random effects predictors and the corresponding random effects, we study the diagonal elements of $\mathrm{Var}(\widehat{\mathbf{U}} - \mathbf{U})$. These mean squared distances form the basis of our discussion of consistency of random effects predictors in next section.

Using the notation $c(i) = \mathrm{E}(\widehat{U}_i - U_i)^2$ and $c(ij) = \mathrm{E}(\widehat{U}_{ij} - U_{ij})^2$, it follows from (4.6) that

$$\begin{aligned}\mathrm{Var}(U_i) &= \mathrm{Var}(U_i - \widehat{U}_i + \widehat{U}_i)\\ &= \mathrm{Var}(U_i - \widehat{U}_i) + \mathrm{Var}(\widehat{U}_i) + 2\mathrm{Cov}[U_i - \widehat{U}_i, \widehat{U}_i]\\ &= c(i) + \mathrm{Var}(\widehat{U}_i) + 0.\end{aligned}$$

Hence

$$\begin{aligned}c(i) &= \mathrm{Var}(U_i) - \mathrm{Var}(\widehat{U}_i)\\ &= \sigma^2\mu_i^r - \mathrm{Var}(\widehat{U}_i).\end{aligned}\tag{4.15}$$

Similarly, we have

$$\begin{aligned}c(ij) &= \mathrm{Var}(U_{ij}) - \mathrm{Var}(\widehat{U}_{ij})\\ &= \sigma^2\mu_i^r\mu_{ij}^2 + \omega^2\mu_i\mu_{ij}^q - \mathrm{Var}(\widehat{U}_{ij}).\end{aligned}\tag{4.16}$$

We derive explicit expressions for $c(i)$ and $c(ij)$ via $\mathrm{Var}(\widehat{\mathbf{U}})$. The explicit expression for $\mathrm{Var}(\widehat{U}_i)$ can be derived from (4.5) as follows:

$$\begin{aligned}
\mathrm{Var}(\widehat{U}_i) &= \mathrm{Cov}(U_i, \mathbf{Y}_i)\mathrm{Var}(\mathbf{Y}_i)^{-1}\mathrm{Cov}(\mathbf{Y}_i, U_i) \\
&= \sigma^2\mu_i^{r-1}\boldsymbol{\nu}_i^\top \left\{ \mathbf{A}_i^{-1} - \frac{\sigma^2\mu_i^{r-2}\mathbf{A}_i^{-1}\boldsymbol{\nu}_i(\mathbf{A}_i^{-1}\boldsymbol{\nu}_i)^\top}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i} \right\} \sigma^2\mu_i^{r-1}\boldsymbol{\nu}_i \\
&= (\sigma^2\mu_i^{r-1})^2\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i \left\{ 1 - \frac{\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i} \right\} \\
&= \frac{(\sigma^2\mu_i^{r-1})^2\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}.
\end{aligned}$$

Thus

$$\begin{aligned}
c(i) &= \sigma^2\mu_i^r - \frac{(\sigma^2\mu_i^{r-1})^2\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i} \\
&= \sigma^2\mu_i^r \left\{ 1 - \frac{\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i} \right) \\
&= \frac{\sigma^2\mu_i^r}{1+\sigma^2\mu_i^{r-2}\boldsymbol{\nu}_i^\top\mathbf{A}_i^{-1}\boldsymbol{\nu}_i}.
\end{aligned}$$

Hence we have the following proposition:

Proposition 4.1

$$c(i) = \frac{\sigma^2\mu_i^r}{1+\sigma^2\mu_i^{r-1}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p}}.$$

Now we derive the explicit expression of $c(ij)$ based on (4.16). As the derivation of a concise expression for $\mathrm{Var}(\widehat{U}_{ij})$ using the matrix form in (4.5) becomes much more complicated, instead we derive $\mathrm{Var}(\widehat{U}_{ij})$ directly from (4.4). To ease the derivation, we first rewrite $\widehat{U}_i$ as

$$\widehat{U}_i = \frac{c(i)}{\mu_i}\sum_{l=1}^{J_i}\sum_{k=1}^{n_{il}} w_{il}\mu_{il}\mu_{ilk}^{1-p}Y_{ilk} + \text{ constant.}$$

Also note that

$$\sigma^2\mu_i^{r-1}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p} = \frac{\sigma^2\mu_i^r}{c(i)} - 1.$$

Furthermore, it follows from (4.4) that $\mathrm{Cov}(\widehat{U}_{ij}, \widehat{U}_{ij}) = \mathrm{Cov}(\widehat{U}_{ij}, U_{ij})$. Now we have

$$\begin{aligned}
\mathrm{Var}(\widehat{U}_{ij}) &= \mathrm{Cov}(\widehat{U}_{ij}, U_{ij}) \\
&= \mathrm{Cov}(\rho^2 w_{ij}\mu_{ij}\widehat{U}_i + \omega^2\mu_{ij}^{q-1}w_{ij}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{1-p}Y_{ijk}, U_{ij}) \\
&= \rho^2 w_{ij}\mu_{ij}\mathrm{Cov}(\widehat{U}_i, U_{ij}) + \omega^2\mu_{ij}^{q-1}w_{ij}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{1-p}\mathrm{Cov}(Y_{ijk}, U_{ij}). \qquad (4.17)
\end{aligned}$$

But

$$\begin{aligned}
\mathrm{Cov}(\widehat{U}_i, U_{ij}) &= \mathrm{Cov}\left(\frac{c(i)}{\mu_i}\sum_{l=1}^{J_i}\sum_{k=1}^{n_{il}} w_{il}\mu_{il}\mu_{ilk}^{1-p}Y_{ilk} + \text{ const}, U_{ij}\right) \\
&= \frac{c(i)}{\mu_i}\sum_{l=1}^{J_i}\sum_{k=1}^{n_{il}} w_{il}\mu_{il}\mu_{ilk}^{1-p}\mathrm{Cov}(Y_{ilk}, U_{ij}) \\
&= \frac{c(i)}{\mu_i}\sum_{l=1}^{J_i}\sum_{k=1}^{n_{il}} w_{il}\mu_{il}\mu_{ilk}^{2-p}\left\{\sigma^2\mu_i^r\mu_{ij}\mu_{il} + \delta_{(l,j)}\omega^2\mu_i\mu_{ij}^q\right\} \\
&= c(i)\mu_{ij}\left\{\sigma^2\mu_i^{r-1}\sum_{l=1}^{J_i}\sum_{k=1}^{n_{il}} w_{il}\mu_{il}\mu_{ilk}^{2-p} + \omega^2\mu_{ij}^{q-1}w_{ij}\sum_{k=1}^{n_{il}}\mu_{ijk}^{2-p}\right\} \\
&= c(i)\mu_{ij}\left\{\frac{\sigma^2\mu_i^r}{c(i)} - 1 + 1 - \rho^2 w_{ij}\right\} \\
&= \sigma^2\mu_i^r\mu_{ij} - \rho^2 c(i)\mu_{ij}w_{ij}, \qquad (4.18)
\end{aligned}$$

and

$$\begin{aligned}
\omega^2\mu_{ij}^{q-1}w_{ij}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{1-p}\mathrm{Cov}(Y_{ijk}, U_{ij}) &= (\sigma^2\mu_i^r\mu_{ij}^2 + \omega^2\mu_i\mu_{ij}^q)\omega^2\mu_{ij}^{q-1}w_{ij}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p} \\
&= (\sigma^2\mu_i^r\mu_{ij}^2 + \omega^2\mu_i\mu_{ij}^q)(1 - \rho^2 w_{ij}). \qquad (4.19)
\end{aligned}$$

Plugging (4.18) and (4.19) into (4.17), after some simplification we have

$$\mathrm{Var}(\widehat{U}_{ij}) = \sigma^2\mu_i^r\mu_{ij}^2 + \omega^2\mu_i\mu_{ij}^q - \rho^2 w_{ij}\left(\omega^2\mu_i\mu_{ij}^q + \rho^2 c(i)w_{ij}\mu_{ij}^2\right).$$

Therefore we have

Proposition 4.2

$$c(ij) = \rho^2 w_{ij} \left(\omega^2 \mu_i \mu_{ij}^q + \rho^2 c(i) w_{ij} \mu_{ij}^2 \right).$$

This expression does not involve σ^2 and μ_{ijk} explicitly, but implicitly via $c(i)$ and w_{ij}. In the expressions for both $c(i)$ and $c(ij)$, the μ_{ijk}s appear only in the form of $\sum_{k=1}^{n_{ij}} \mu_{ijk}^{2-p}$, that is the average of μ_{ijk}^{2-p} within sub-clusters.

4.4 Consistency

In this section, we draw consistency results about random effects predictors based on the mean squared distances derived in last section. According to Chebyshev's inequality, we have $c(i) \to \mathbf{0}$ implies $\widehat{U}_i \xrightarrow{P} U_i$, and $c(ij) \to \mathbf{0}$ implies $\widehat{U}_{ij} \xrightarrow{P} U_{ij}$. To draw our consistency results based on this inequality, we need to find upper bounds for $c(i)$ and $c(ij)$. Since oversimplified upper bounds do not result in desirable consistency results, we derive the upper bounds through a sequence of inequalities. After some algebra, we can show that $c(i)$ is bounded above as follows:

Lemma 4.1

$$c(i) \leq \frac{\sigma^2 \mu_i^r \left(\rho^2 + \omega^2 min_j(n_{ij}) max_j(\mu_{ij}^{q-1}) min_{j,k}(\mu_{ijk}^{2-p}) \right)}{\rho^2 + \omega^2 min_j(n_{ij}) \; max_j(\mu_{ij}^{q-1}) min_{j,k}(\mu_{ijk}^{2-p}) + \sigma^2 min_j(n_{ij}) J_i \; mu_i^{r-1} min_j(\mu_{ij}) min_{j,k}(\mu_{ijk}^{2-p})}.$$

Proof

We start our investigation with w_{ij}.

$$\begin{aligned} w_{ij} &= \frac{1}{\rho^2 + \omega^2 \mu_{ij}^{q-1} \sum_{k=1}^{n_{ij}} \mu_{ijk}^{2-p}} \\ &\leq \frac{1}{\rho^2 + \omega^2 \min_j(n_{ij}) \min_j(\mu_{ij}^{q-1}) \min_{jk}(\mu_{ijk}^{2-p})}. \end{aligned} \tag{4.20}$$

Also we have

$$\begin{aligned}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ijk}^{2-p} &= \frac{\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p}}{\rho^2+\omega^2\mu_{ij}^{q-1}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p}} \\ &= \frac{1}{\frac{\rho^2}{\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p}}+\omega^2\mu_{ij}^{q-1}} \\ &\geq \frac{1}{\frac{\rho^2}{\min_j(n_{ij})\min_{jk}(\mu_{ijk}^{2-p})}+\omega^2\max_j(\mu_{ij}^{q-1})} \\ &= \frac{\min_j(n_{ij})\min_{jk}(\mu_{ijk}^{2-p})}{\rho^2+\omega^2\min_j(n_{ij})\max_j(\mu_{ij}^{q-1})\min_{jk}(\mu_{ijk}^{2-p})},\end{aligned}$$

hence

$$\sigma^2\mu_i^{r-1}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p} \geq \frac{\sigma^2\mu_i^{r-1}J_i\min_j(\mu_j^{q-1})\min_j(n_{ij})\min_{jk}(\mu_{ijk}^{2-p})}{\rho^2+\omega^2\min_j(n_{ij})\max_j(\mu_{ij}^{q-1})\min_{jk}(\mu_{ijk}^{2-p})}.$$

Now we have

$$\begin{aligned}c(i) &= \frac{\sigma^2\mu_i^r}{1+\sigma^2\mu_i^{r-1}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p}} \\ &\leq \frac{\sigma^2\mu_i^r\left(\rho^2+\omega^2\min_j(n_{ij})\max_j(\mu_{ij}^{q-1})\min_{j,k}(\mu_{ijk}^{2-p})\right)}{\rho^2+\omega^2\min_j(n_{ij})\max_j(\mu_{ij}^{q-1})\min_{j,k}(\mu_{ijk}^{2-p})+\sigma^2\min_j(n_{ij})J_i\mu_i^{r-1}\min_j(\mu_{ij})\min_{j,k}(\mu_{ijk}^{2-p})}.\end{aligned}$$

□

The derivation of an upper bound for $c(ij)$ is straightforward. Note that Proposition 4.1 immediately implies that

$$c(i) \leq \sigma^2\mu_i^r.$$

In addition, we have

$$\rho^2 w_{ij} = \frac{\rho^2}{\rho^2+\omega^2\mu_{ij}^{q-1}\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p}} \quad \leq \quad 1.$$

Thus

$$\begin{aligned} c(ij) &= \rho^2 w_{ij} \left(\omega^2 \mu_i \mu_{ij}^q + \rho^2 c(i) w_{ij} \mu_{ij}^2 \right) \\ &\leq \rho^2 w_{ij} \left(\omega^2 \mu_i \mu_{ij}^q + \sigma^2 \mu_i^r \mu_{ij}^2 \right). \end{aligned}$$

It follows from (4.20) that $c(ij)$ is bounded above as follows:

Lemma 4.2

$$c(ij) \leq \frac{\rho^2 (\sigma^2 \mu_i^r \mu_{ij}^2 + \omega^2 \mu_i \mu_{ij}^q)}{\rho^2 + \omega^2 min_j \mu_{ij}^{q-1} min_{j,k} (\mu_{ijk}^{2-p})}.$$

The following 'small dispersion asymptotics' (Jørgensen 1987b) is an immediate consequence of Lemma 4.1 and 4.2:

Proposition 4.3

1. *$\widehat{U}_i \xrightarrow{P} U_i$ as $\sigma^2 \to 0$ or $(\omega^2, \rho^2) \to (0, 0)$;*
2. *$\widehat{U}_{ij} \xrightarrow{P} U_{ij}$ as $\rho^2 \to 0$ or $(\sigma^2, \omega^2) \to (0, 0)$.*

For large sample asymptotics, clearly we have

Proposition 4.4

$$\widehat{U}_i \xrightarrow{P} U_i \text{ as } J_i \to \infty \text{ and } \widehat{U}_{ij} \xrightarrow{P} U_{ij} \text{ as } min_j(n_{ij}) \to \infty,$$

if all μ_i, μ_{ij} and μ_{ijk} are contained in a compact set not containing zero.

This result distinguishes the consistency conditions for $\widehat{U}_i$ from those for $\widehat{U}_{ij}$. The former is implied by a large number of sub-clusters within a certain cluster, while the latter is implied by large sizes of the sub-clusters. This result matches our intuition.

The above results can be expressed in a more delicate form derived from Chebyshev's inequality as follows:

Proposition 4.5

$$\widehat{U}_i = U_i + O_P(\sigma^2) \quad \textit{and} \quad \widehat{U}_i = U_i + O_P(\omega^2 + \rho^2);$$
$$\widehat{U}_{ij} = U_{ij} + O_P(\rho^2) \quad \textit{and} \quad \widehat{U}_{ij} = U_{ij} + O_P(\sigma^2 + \omega^2);$$
$$\widehat{U}_i = U_i + O_P(\tfrac{1}{\sqrt{J_i}}) \quad \textit{and} \quad \widehat{U}_{ij} = U_{ij} + O_P(\tfrac{1}{\sqrt{n_{ij}}}).$$

under the same conditions.

It follows from (4.3) that $c(i) \to 0$ and $c(ij) \to 0$ also imply $\widehat{U}_i \xrightarrow{P} \mathrm{E}(U_i|Y)$ and $\widehat{U}_{ij} \xrightarrow{P} \mathrm{E}(U_{ij}|Y)$, respectively. Thus $\widehat{U}_i \xrightarrow{P} \mathrm{E}(U_i|Y)$ and $\widehat{U}_{ij} \xrightarrow{P} \mathrm{E}(U_{ij}|Y)$ are clearly implied by the conditions stated above. Similarly, we have

Proposition 4.6

$$\widehat{U}_i = E(U_i|Y) + O_P(\sigma^2) \quad \textit{and} \quad \widehat{U}_i = E(U_i|Y) + O_P(\omega^2 + \rho^2);$$
$$\widehat{U}_{ij} = E(U_{ij}|Y) + O_P(\rho^2) \quad \textit{and} \quad \widehat{U}_{ij} = E(U_{ij}|Y) + O_P(\sigma^2 + \omega^2);$$
$$\widehat{U}_i = E(U_i|Y) + O_P(\tfrac{1}{\sqrt{J_i}}) \quad \textit{and} \quad \widehat{U}_{ij} = E(U_{ij}|Y) + O_P(\tfrac{1}{\sqrt{n_{ij}}}).$$

A more delicate discussion of the consistency of the random effects predictors would involve the relationships among the quantities $\mu_i, \mu_{ij}, \mu_{ijk}, p, q, r, n_{ij}$. Some further results will be discussed in Section 7.1.1.

Chapter 5

Parameter estimation

5.1 Estimation of regression parameters

We begin our discussion on the estimation for the regression parameters with the case of known dispersion parameters. The inclusion of unknown dispersion parameters will be discussed in next section.

5.1.1 Estimated score function

As the score function is optimal among all regular estimating functions, we begin our investigation with the *partially observed score function* defined below:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_{(1)}} = \sum_{i=1}^{m} \mathbf{z}_i \frac{\mu_i^{1-r}}{\sigma^2}(U_i - \mu_i),$$

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_{(2)}} = \sum_{i=1}^{m} \sum_{j=1}^{J_i} \mathbf{z}_{ij} \frac{\mu_{ij}^{1-q}}{\omega^2}(U_{ij} - U_i \mu_{ij}),$$

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_{(3)}} = \sum_{i=1}^{m} \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} \mathbf{z}_{ijk} \frac{\mu_{ijk}^{1-p}}{\rho^2}(Y_{ijk} - U_{ij} \mu_{ijk}).$$

The marginal score function for the response **Y** is obtained by taking expectation of the above partially observed score function with respect to the conditional distribution of the unobserved random effects **U** given the responses **Y**. Since the partially observed score function is linear in the unobserved random effects **U**, the marginal score function for the response **Y** is obtained by replacing unobserved random effects **U** in the partially observed score function by the expectations of the unobserved random effects **U** given the responses **Y**, denoted by E(**U**|**Y**). To find the maximum likelihood estimate, in principle, we can then solve the marginal score equation obtained by setting the marginal score function to zero.

However, a closed form expression for E(**U**|**Y**) is generally difficult or impossible to obtain. Note that E(**U**|**Y**) is the best unbiased predictor of unobserved random effects **U** given the response **Y** and we showed that the orthodox BLUP predictor, the best linear unbiased predictor of the unobserved random effects **U**, generally converges to E(**U**|**Y**) in probability. Thus we approximate E(**U**|**Y**) by the orthodox BLUP predictor of the unobserved random effects in the marginal score function as follows:

$$\boldsymbol{\psi}^{(1)}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathbf{z}_i \frac{\mu_i^{1-r}(\boldsymbol{\beta})}{\sigma^2} \left(\widehat{U}_i(\boldsymbol{\beta}) - \mu_i(\boldsymbol{\beta})\right) = \sum_{i=1}^{m} \boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta}), \tag{5.1}$$

$$\boldsymbol{\psi}^{(2)}(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{J_i} \mathbf{z}_{ij} \frac{\mu_{ij}^{1-q}(\boldsymbol{\beta})}{\omega^2} \left(\widehat{U}_{ij}(\boldsymbol{\beta}) - \widehat{U}_i(\boldsymbol{\beta})\mu_{ij}(\boldsymbol{\beta})\right) = \sum_{i=1}^{m} \boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta}), \tag{5.2}$$

$$\boldsymbol{\psi}^{(3)}(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} \mathbf{z}_{ijk} \frac{\mu_{ijk}^{1-p}(\boldsymbol{\beta})}{\rho^2} \left(Y_{ijk} - \widehat{U}_{ij}(\boldsymbol{\beta})\mu_{ijk}(\boldsymbol{\beta})\right) = \sum_{i=1}^{m} \boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta}). \tag{5.3}$$

The three functions in (5.1),(5.2) and (5.3) are clearly linear functions of the responses as the orthodox BLUP predictors of the random effects are. We call $\boldsymbol{\psi}(\boldsymbol{\beta})$ as defined below the *estimated score function* (based on the orthodox BLUP):

$$\begin{aligned}
\boldsymbol{\psi}(\boldsymbol{\beta}) &= (\boldsymbol{\psi}^{(1)}(\boldsymbol{\beta})^\top, \boldsymbol{\psi}^{(2)}(\boldsymbol{\beta})^\top, \boldsymbol{\psi}^{(3)}(\boldsymbol{\beta})^\top)^\top \\
&= (\sum_{i=1}^{m} \boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta})^\top, \sum_{i=1}^{m} \boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})^\top, \sum_{i=1}^{m} \boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta})^\top)^\top \\
&= \sum_{i=1}^{m} (\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta})^\top, \boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})^\top, \boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta})^\top)^\top \\
&= \sum_{i=1}^{m} \boldsymbol{\psi}_i(\boldsymbol{\beta}),
\end{aligned}$$

where $\boldsymbol{\psi}_i(\boldsymbol{\beta})$ corresponds to the estimated score function for the ith independent cluster. Clearly $\boldsymbol{\psi}_i(\boldsymbol{\beta})$ is an unbiased estimating function as the orthodox BLUP predictors of the random effects are unbiased.

If we treat the unobserved random effects as 'missing data', the EM algorithm can then be used to obtain the maximum likelihood estimate. This algorithm iterates between an E-step, which involves evaluating the conditional expectations of unobserved random effects given the response in the marginal score function using current parameter values, and an M-step, which involves obtaining updated parameter estimates by solving the marginal score equation. The orthodox BLUP approach is equivalent to approximating the E-step by replacing the conditional expectations by orthodox BLUP predictors and approximating the M-step by solving

$$\sum_{i=1}^{m} \boldsymbol{\psi}_i(\boldsymbol{\beta}) = \mathbf{0}.$$

The roots of this equation are then used as regression parameter estimates.

5.1.2 Standard errors of regression parameter estimators

Letting $\hat{\boldsymbol{\beta}}^{(m)}$ denote the sequence of roots of $\sum_{i=1}^{m} \boldsymbol{\psi}_i(\boldsymbol{\beta}) = \mathbf{0}$, it follows from Lemma 2.1 that $\hat{\boldsymbol{\beta}}^{(m)}$ is consistent for $\boldsymbol{\beta}$ and asymptotically normal as $m \to \infty$. Specifically, the asymptotic variance is given by the inverse of the Godambe information matrix:

$$\mathbf{J}(\boldsymbol{\beta}) = \mathbf{S}(\boldsymbol{\beta})\mathbf{V}(\boldsymbol{\beta})^{-1}\mathbf{S}(\boldsymbol{\beta})^{\top}, \tag{5.4}$$

where the sensitivity matrix $\mathbf{S}(\boldsymbol{\beta})$ and the variability matrix $\mathbf{V}(\boldsymbol{\beta})$ are given by:

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathbf{S}_i(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathrm{E}_{\boldsymbol{\beta}} \left\{ \frac{\partial \boldsymbol{\psi}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\top}} \right\},$$

$$\mathbf{V}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathbf{V}_i(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathrm{E}_{\boldsymbol{\beta}} \left\{ \boldsymbol{\psi}_i(\boldsymbol{\beta})\boldsymbol{\psi}_i^{\top}(\boldsymbol{\beta}) \right\}.$$

From now on, we simply denote $\hat{\boldsymbol{\beta}}^{(m)}$ by $\hat{\boldsymbol{\beta}}$.

An analogue of Wald's test is available for testing the hypothesis $H_0 : \boldsymbol{\beta}_{(1)} = \mathbf{0}$, where $\boldsymbol{\beta}_{(1)}$ is a sub-vector of $\boldsymbol{\beta}$. The test statistic is:

$$W = \widehat{\boldsymbol{\beta}}_{(1)}^{\top} \left\{ \mathbf{J}^{11}(\widehat{\boldsymbol{\beta}}) \right\}^{-1} \widehat{\boldsymbol{\beta}}_{(1)},$$

where $\mathbf{J}^{11}(\widehat{\boldsymbol{\beta}})$ is the corresponding block of the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$. Asymptotically, this statistic follows a $\chi^2(k)$-distribution, where k is the size of the sub-vector $\widehat{\boldsymbol{\beta}}_{(1)}$.

5.1.3 Newton scoring algorithm

The orthodox BLUP approach is actually a linearized EM algorithm. Instead of using this approximate EM algorithm, we adopt a more efficient algorithm, the *Newton* *scoring algorithm*, introduced by Jørgensen et al. (1996a) to solve the estimating equation $\boldsymbol{\psi}(\boldsymbol{\beta}) = \mathbf{0}$.

The Newton scoring algorithm is defined as the Newton algorithm applied to the equation $\boldsymbol{\psi}(\boldsymbol{\beta}) = \mathbf{0}$, but with the derivative of $\boldsymbol{\psi}(\boldsymbol{\beta})$ replaced by its expectation $\mathbf{S}(\boldsymbol{\beta})$. The resulting algorithm gives the following updated value for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} - \mathbf{S}^{-1}(\boldsymbol{\beta})\boldsymbol{\psi}(\boldsymbol{\beta}). \tag{5.5}$$

Clearly the sensitivity and variability matrices are crucial in this estimation procedure. If $\mathbf{S}_{ij}$ denotes $\frac{\partial \boldsymbol{\psi}^{(i)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{(j)}^{\top}}$ and $\mathbf{V}_{ij}$ denotes $\mathrm{E}(\boldsymbol{\psi}^{(i)}(\boldsymbol{\beta})\boldsymbol{\psi}^{(j)}(\boldsymbol{\beta})^{\top})$, then the sensitivity and variability matrices can be expressed as follows:

$$\mathbf{S}(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \mathbf{S}_{23} \\ \mathbf{S}_{31} & \mathbf{S}_{32} & \mathbf{S}_{33} \end{pmatrix},$$

$$\mathbf{V}(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} \end{pmatrix}.$$

Note that $\mathbf{S}(\boldsymbol{\beta})$ is apparently of asymmetric form. For example, $\mathbf{S}_{13} = \frac{\partial \boldsymbol{\psi}^{(1)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{(3)}^{\top}}$, whereas $\mathbf{S}_{31} = \frac{\partial \boldsymbol{\psi}^{(3)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{(1)}^{\top}}$. Actually $\mathbf{S}(\boldsymbol{\beta})$ can be shown to be symmetric. However a block-by-block direct derivation of these matrix blocks from (5.1), (5.2) and (5.3) would be complicated by the nonlinearity of $\boldsymbol{\beta}$ in $\hat{\mathbf{U}}$ (Jørgensen et al. 1996b, 1996c). Instead, we introduce a concise matrix expression for $\boldsymbol{\psi}(\boldsymbol{\beta})$ in the next section. This expression not only facilitates derivation of the sensitivity and variability matrices, but also helps to clarify the relationships between the sensitivity and variability matrices.

5.1.4 Optimality

After introducing a matrix expression for $\boldsymbol{\psi}(\boldsymbol{\beta})$, we will study the relationships between the sensitivity and variability matrices and the optimality of the estimated score function. The concept of optimality here is in the sense that the estimated score function attains the minimum asymptotic covariance for $\hat{\boldsymbol{\beta}}$ among a certain linear class discussed below.

Recall that the full covariate matrix is $\mathbf{X}$ with $\mathbf{x}_{ijk}^{\top} = (\mathbf{z}_i^{\top}, \mathbf{z}_{ij}^{\top}, \mathbf{z}_{ijk}^{\top})$. Let $\mathbf{X}_i$ denote the sub-matrix of $\mathbf{X}$ corresponding to the ith cluster. We state the following results:

Theorem 5.1 *For Tweedie mixed models, the estimated score function, sensitivity and variability matrices can be expressed as follows:*

1. $\boldsymbol{\psi}(\boldsymbol{\beta}) = \mathbf{X}^{\top} diag\left(E(\mathbf{Y})\right) \, Var^{-1}(\mathbf{Y}) \left(\mathbf{Y} - E(\mathbf{Y})\right),$

2. $\mathbf{V}(\boldsymbol{\beta}) = Var\left(\boldsymbol{\psi}(\boldsymbol{\beta})\right) = \mathbf{X}^{\top} diag\left(E(\mathbf{Y})\right) \, Var(\mathbf{Y})^{-1} diag\left(E(\mathbf{Y})\right) \mathbf{X},$

3. $\mathbf{J}(\boldsymbol{\beta}) = -\mathbf{S}(\boldsymbol{\beta}) = \mathbf{V}(\boldsymbol{\beta}).$

The first statement gives a *global matrix expression* for the estimated score function. This expression can be viewed as a multivariate version of the quasi-score function first proposed by Wedderburn (1974) . The second statement gives the expression for the variability matrix. The third statement for the sensitivity matrix is similar to that for the Fisher information matrix. Both the derivation and computing efforts are greatly eased by the relationship given in the second and the third statements. Clearly now the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by the inverse of the variability matrix alone.

The key statement in Theorem 5.1 is the first statement; the second and third statements are immediate consequences. We will leave the derivation of $\boldsymbol{\psi}(\boldsymbol{\beta})$ to a separate section since it is long and technical. Now we show the second statement.

$$\begin{aligned}
\mathbf{V}(\boldsymbol{\beta}) &= \mathrm{Var}\left(\boldsymbol{\psi}(\boldsymbol{\beta})\right) \\
&= \mathrm{Var}\left(\mathbf{X}^\top \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathrm{Var}^{-1}(\mathbf{Y}) \left(\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right)\right) \\
&= \mathbf{X}^\top \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathrm{Var}^{-1}(\mathbf{Y}) \mathrm{Var}(\mathbf{Y}) \mathrm{Var}^{-1}(\mathbf{Y}) \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathbf{X} \\
&= \mathbf{X}^\top \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathrm{Var}^{-1}(\mathbf{Y}) \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathbf{X}.
\end{aligned}$$

The third statement can also be easily shown based on the first statement. Noting that

$$\frac{\partial \mathrm{E}(\mathbf{Y})}{\partial \boldsymbol{\beta}^\top} = \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathbf{X}$$

and $\mathrm{E}_{\boldsymbol{\beta}}\left(\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right) = \mathbf{0}$, we have

$$\begin{aligned}
\mathbf{S}(\boldsymbol{\beta}) &= \mathrm{E}_{\boldsymbol{\beta}}\left\{\frac{\partial \boldsymbol{\psi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}\right\} \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left\{\frac{\partial\left(\mathbf{X}^\top \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathrm{Var}(\mathbf{Y})^{-1}\right)}{\partial \boldsymbol{\beta}^\top} \left(\left(\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right) \otimes \mathbf{E}\right)\right\} \\
&\quad + \mathrm{E}_{\boldsymbol{\beta}}\left\{\left(\mathbf{X}^\top \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathrm{Var}(\mathbf{Y})^{-1}\right) \frac{\partial\left(\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right)}{\partial \boldsymbol{\beta}^\top}\right\} \\
&= \frac{\partial\left(\mathbf{X}^\top \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathrm{Var}(\mathbf{Y})^{-1}\right)}{\partial \boldsymbol{\beta}^\top} \left(\mathrm{E}_{\boldsymbol{\beta}}\left(\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right) \otimes \mathbf{E}\right) \\
&\quad + \left(\mathbf{X}^\top \mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right) \mathrm{Var}(\mathbf{Y})^{-1}\right) \left(-\frac{\partial \mathrm{E}(\mathbf{Y})}{\partial \boldsymbol{\beta}^\top}\right)
\end{aligned}$$

$$\begin{aligned} &= \mathbf{0} - \mathbf{X}^\top \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathrm{Var}(\mathbf{Y})^{-1}\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathbf{X} \\ &= -\mathbf{V}(\boldsymbol{\beta}), \end{aligned}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{E}$ is an identity matrix whose order equals the size of $\boldsymbol{\beta}$.

Now we can derive the explicit expression for $\mathbf{S}(\boldsymbol{\beta})$. Note that

$$\begin{aligned} \mathbf{S}(\boldsymbol{\beta}) &= -\mathbf{V}(\boldsymbol{\beta}) \\ &= -\mathbf{X}^\top \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathrm{Var}(\mathbf{Y})^{-1}\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathbf{X} \\ &= -(\mathbf{X}_1^\top, \ldots, \mathbf{X}_m^\top)\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathrm{Var}(\mathbf{Y})^{-1}\mathrm{diag}\,(\mathrm{E}(\mathbf{Y})) \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{pmatrix} \\ &= -\sum_{i=1}^{m} \mathbf{X}_i^\top \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\mathrm{diag}\,(\mathrm{E}(\,Y_i))\,\mathbf{X}_i. \end{aligned} \tag{5.6}$$

Rewrite $\mathrm{Var}^{-1}(\mathbf{Y}_i)$ as

$$\mathrm{Var}^{-1}(\mathbf{Y}_i) = \begin{pmatrix} \mathbf{A}_{i1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{A}_{iJ_i} \end{pmatrix} - \frac{c(i)}{\mu_i^2} \begin{pmatrix} w_{i1}\boldsymbol{\mu}_{i1*}^{1-p} \\ \vdots \\ w_{iJ_i}\boldsymbol{\mu}_{iJ_i*}^{1-p} \end{pmatrix} (w_{i1}(\boldsymbol{\mu}_{i1*}^{1-p})^\top, \ldots, w_{iJ_i}(\boldsymbol{\mu}_{iJ_i*}^{1-p})^\top),$$

where $\mathbf{A}_{ij}^{-1}$ is given in (3.15). Plugging this expression for $\mathrm{Var}^{-1}(\mathbf{Y}_i)$ into (5.6), we have

$$\begin{aligned} \mathbf{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{m} c(i)(\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p}\mathbf{x}_{ijk})(\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ij}\mu_{ijk}^{2-p}\mathbf{x}_{ijk})^\top \\ &\quad + \sum_{i=1}^{m} \frac{\omega^2\mu_i^2}{\rho^2} \sum_{j=1}^{J_i} w_{ij}\mu_{ij}^{q+1}(\sum_{k=1}^{n_{ij}} \mu_{ijk}^{2-p}\mathbf{x}_{ijk})(\sum_{k=1}^{n_{ij}} \mu_{ijk}^{2-p}\mathbf{x}_{ijk})^\top \end{aligned}$$

$$-\sum_{i=1}^{m}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}}\frac{1}{\rho^2}\mu_{ijk}^{2-p}\mathbf{x}_{ijk}\mathbf{x}_{ijk}^{\top}. \tag{5.7}$$

In next section, we show that

$$\boldsymbol{\psi}_i(\boldsymbol{\beta}) = \mathbf{X}_i^{\top}\text{diag}\left(\text{E}(\mathbf{Y}_i)\right)\text{Var}(\mathbf{Y}_i)^{-1}\left(\mathbf{Y}_i - \text{E}(\mathbf{Y}_i)\right).$$

This immediately implies that the results of Theorem 5.1 hold for each cluster. Applying Lemma 2.2 to the estimated score functions for each cluster, we may directly show the optimality of the estimated score function among the class of all linear estimating functions of the form $\sum_{i=1}^{m}\mathbf{Q}_i(\boldsymbol{\beta})\boldsymbol{\psi}_i(\boldsymbol{\beta})$ if the variability matrices $\mathbf{V}_i(\boldsymbol{\beta})$s are nonsingular; however, these variability matrices are often singular in practice. On the other hand, $\sum_{i=1}^{m}\mathbf{Q}_i(\boldsymbol{\beta})\boldsymbol{\psi}_i(\boldsymbol{\beta})$ is a subclass of the class of all linear estimating functions of the form $\sum_{i=1}^{m}\mathbf{P}_i(\boldsymbol{\beta})\left(\mathbf{Y}_i - \text{E}(\mathbf{Y}_i)\right)$ since $\boldsymbol{\psi}_i(\boldsymbol{\beta})$ is a linear function of $\mathbf{Y}_i - \text{E}(\mathbf{Y}_i)$. Applying Lemma 2.2 to the latter linear class, we have

$$\begin{aligned}\boldsymbol{\psi}(\boldsymbol{\beta}) &= \sum_{i=1}^{m}\mathbf{X}_i^{\top}\text{diag}\left(\text{E}(\mathbf{Y}_i)\right)\text{Var}(\mathbf{Y}_i)^{-1}\left(\mathbf{Y}_i - \text{E}(\ Y_i)\right)\\ &= -\sum_{i=1}^{m}\text{E}_{\boldsymbol{\beta}}\left\{\frac{\partial\left(\mathbf{Y}_i - \text{E}(\mathbf{Y}_i)\right)}{\partial\boldsymbol{\beta}^{\top}}\right\}\text{Var}(\mathbf{Y}_i)^{-1}\left(\mathbf{Y}_i - \text{E}(\mathbf{Y}_i)\right)\end{aligned}$$

is the optimal estimating function among the linear class.

5.1.5 Derivation of matrix form of $\boldsymbol{\psi}(\boldsymbol{\beta})$

We will derive the matrix expression for $\boldsymbol{\psi}(\boldsymbol{\beta})$ in terms of $\boldsymbol{\psi}_i(\boldsymbol{\beta})$. Furthermore we will deal with $\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta})$, $\boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})$ and $\boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta})$ separately. The basic technique of handling $\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta})$ and $\boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})$ is to convert (5.1) and (5.2) into explicit linear functions of $\mathbf{Y}_i - \text{E}(\mathbf{Y}_i)$ directly. We will then deal with $\boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta})$ by converting (5.3) into explicit linear function of $\mathbf{Y}_i - \text{E}(\mathbf{Y}_i)$ indirectly due to the technical convenience.

The derivation of $\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta})$ is quite straightforward. First rewrite $\widehat{U}_i$ as

$$\widehat{U}_i = \mu_i + \sigma^2 \mu_i^{r-1} \mathrm{E}(\mathbf{Y}_i)^\top \mathrm{Var}(\mathbf{Y}_i)^{-1} \left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right).$$

Thus we have

$$\begin{aligned}
\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta}) &= \mathbf{z}_i \frac{\mu_i^{1-r}}{\sigma^2} (\widehat{U}_i - \mu_i) \\
&= \mathbf{z}_i \frac{\mu_i^{1-r}}{\sigma^2} (\sigma^2 \mu_i^{r-1} \mathrm{E}(\mathbf{Y}_i)^\top \mathrm{Var}(\mathbf{Y}_i)^{-1} \left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right) \\
&= \mathbf{z}_i \mathrm{E}(\mathbf{Y}_i)^\top \mathrm{Var}(\mathbf{Y}_i)^{-1} \left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right) \\
&= (\mathbf{z}_i, \ldots, \mathbf{z}_i) \mathrm{diag}\left(\mathrm{E}(\mathbf{Y}_i)\right) \mathrm{Var}(\mathbf{Y}_i)^{-1} \left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right).
\end{aligned}$$

Similarly, we may deal with $\boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})$ by noting that

$$\begin{aligned}
\widehat{U}_{ij} &= (\sigma^2 \mu_i^{r-1} \mu_{ij} \boldsymbol{\nu}_i^\top + \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top) \mathrm{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \boldsymbol{\nu}_i) \\
&= \mu_{ij} \widehat{U}_i + \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top \mathrm{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \boldsymbol{\nu}_i).
\end{aligned}$$

Thus we obtain

$$\widehat{U}_{ij} - \mu_{ij} \widehat{U}_i = \omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top \mathrm{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)).$$

Plugging this formula in $\boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})$, we have

$$\begin{aligned}
\boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta}) &= \sum_{j=1}^{J_i} \mathbf{z}_{ij} \frac{\mu_{ij}^{1-q}}{\omega^2} (\widehat{U}_{ij} - \widehat{U}_i \mu_{ij}) \\
&= \sum_{j=1}^{J_i} \mathbf{z}_{ij} \frac{\mu_{ij}^{1-q}}{\omega^2} \left(\omega^2 \mu_{ij}^{q-1} \mathbf{e}_{ij}^\top \mathrm{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i))\right) \\
&= \sum_{j=1}^{J_i} \mathbf{z}_{ij} \mathbf{e}_{ij}^\top \mathrm{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)) \\
&= (\sum_{j=1}^{J_i} \mathbf{z}_{ij} \mathbf{e}_{ij}^\top) \mathrm{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)) \\
&= (\mathbf{z}_{i1} \boldsymbol{\nu}_{i1*}^\top, \ldots, \mathbf{z}_{iJ_i} \boldsymbol{\nu}_{iJ_i}^\top) \mathrm{Var}(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)) \\
&= (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iJ_i}, \ldots, \mathbf{z}_{iJ_i}) \mathrm{diag}\left(\mathrm{E}(\mathbf{Y}_i)\right) \mathrm{Var}(\mathbf{Y}_i)^{-1} \left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right).
\end{aligned}$$

A direct derivation of $\boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta})$ would be much more complicated than those of $\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta})$ and $\boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})$. Instead, we introduce a new matrix notation to ease the derivation. Let

$$\mathbf{U}_{i*}\boldsymbol{\mu}_{i**} = (U_{i1}\mu_{i11}, \ldots, U_{i1}\mu_{i1n_{i1}}, \ldots, U_{iJ_i}\mu_{iJ_i1}, \ldots, U_{iJ_i}\mu_{iJ_in_{iJ_i}})^\top.$$

This new vector is the same size as the longer vector $\boldsymbol{\mu}_{i**}$ with components $U_{ij}\mu_{ijk}$s. That is, each component μ_{ijk} of $\boldsymbol{\mu}_{i**}$ is matched by a random effect component U_{ij} within the same sub-cluster. We may call it a nested product between $\mathbf{U}_{i*} = (U_{i1}, \ldots, U_{iJ_i})$ and $\boldsymbol{\mu}_{i**} = (\mu_{i11}, \ldots, \mu_{iJ_in_{iJ_i}})$. This nested product notation is especially convenient in the generalization of Tweedie mixed models with two levels of nested random effects models to higher levels. Similarly we have

$$\widehat{\mathbf{U}}_{i*}\boldsymbol{\mu}_{i**} = (\widehat{U}_{i1}\mu_{i11}, \ldots, \widehat{U}_{i1}\mu_{i1n_{i1}}, \ldots, \widehat{U}_{iJ_i}\mu_{iJ_i1}, \ldots, \widehat{U}_{iJ_i}\mu_{iJ_in_{iJ_i}})^\top.$$

With this notation, we can express $\boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta})$ simply in terms of $\mathbf{Y}_i - \widehat{\mathbf{U}}_{i*}\boldsymbol{\mu}_{i**}$. Before doing so, we note:

$$\widehat{\mathbf{U}}_{i*}\boldsymbol{\mu}_{i**} = \mathrm{E}(\mathbf{Y}_i) + \mathrm{Cov}(\mathbf{U}_{i*}\boldsymbol{\mu}_{i**}, \mathbf{Y}_i)\mathrm{Var}(\mathbf{Y}_i)^{-1}\left(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)\right),$$

Now we can rewrite the covariance of the responses in terms of that between $\widehat{\mathbf{U}}_{i*}\boldsymbol{\mu}_{i**}$ and $\mathbf{Y}_i$ as follows:

$$\mathrm{Var}(\mathbf{Y}_i) \quad = \quad \mathrm{Cov}(\mathbf{U}_{i*}\boldsymbol{\mu}_{i**}, \mathbf{Y}_i) + \begin{pmatrix} \rho^2\mu_{i11}^{p-1}\nu_{i11} & & 0 \\ & \ddots & \\ 0 & & \rho^2\mu_{iJ_in_{iJ_i}}^{p-1}\nu_{iJ_in_{iJ_i}} \end{pmatrix}$$

$$= \quad \mathrm{Cov}(\mathbf{U}_{i*}\boldsymbol{\mu}_{i**}, \mathbf{Y}_i) + \begin{pmatrix} \rho^2\mu_{i11}^{p-1} & & 0 \\ & \ddots & \\ 0 & & \rho^2\mu_{iJ_in_{iJ_i}}^{p-1} \end{pmatrix} \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,.$$

Hence $\mathbf{Y}_i - \widehat{\mathbf{U}}_{i*}\boldsymbol{\mu}_{i**}$ has the following simple expression:

$$\begin{aligned} \mathbf{Y}_i - \widehat{\mathbf{U}}_{i*}\boldsymbol{\mu}_{i**} &= \mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i) - \mathrm{Cov}(\mathbf{U}_{i*}\boldsymbol{\mu}_{i**}, \mathbf{Y}_i)\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)) \\ &= \{\mathrm{Var}(\mathbf{Y}_i) - \mathrm{Cov}(\mathbf{U}_{i*}\boldsymbol{\mu}_{i**}, \mathbf{Y}_i)\}\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)) \\ &= \begin{pmatrix} \rho^2\mu_{i11}^{p-1} & & 0 \\ & \ddots & \\ 0 & & \rho^2\mu_{iJ_in_{iJ_i}^{p-1}} \end{pmatrix} \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i))\,. \end{aligned}$$

Now we have

$$\begin{aligned} \boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta}) &= \sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} \mathbf{z}_{ijk}\frac{\mu_{ijk}^{1-p}}{\rho^2}(Y_{ijk} - \widehat{U}_{ij}\mu_{ijk}) \\ &= (\frac{\mu_{i11}^{1-p}}{\rho^2}\mathbf{z}_{i11}, \ldots, \frac{\mu_{iJ_in_{iJ_i}}^{1-p}}{\rho^2}\mathbf{z}_{iJ_in_{iJ_i}})(\mathbf{Y}_i - \widehat{\mathbf{U}}_{i*}\boldsymbol{\mu}_{i**}) \\ &= (\mathbf{z}_{i11}, \ldots, \mathbf{z}_{iJ_in_{iJ_i}})\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i))\,. \end{aligned}$$

These equations for $\boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta})$, $\boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta})$ and $\boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta})$ give

$$\begin{aligned} \boldsymbol{\psi}_i(\boldsymbol{\beta}) &= \begin{pmatrix} \boldsymbol{\psi}_i^{(1)}(\boldsymbol{\beta}) \\ \boldsymbol{\psi}_i^{(2)}(\boldsymbol{\beta}) \\ \boldsymbol{\psi}_i^{(3)}(\boldsymbol{\beta}) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{z}_i & \ldots & \mathbf{z}_i \\ \mathbf{z}_{i1} & \ldots & \mathbf{z}_{iJ_i} \\ \mathbf{z}_{i11} & \ldots & \mathbf{z}_{iJ_in_{iJ_i}} \end{pmatrix} \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i)) \end{aligned}$$

$$\begin{aligned} &= (\mathbf{x}_{i11}, \ldots, \mathbf{x}_{iJ_in_{iJ_i}})\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(Y_i - \mathrm{E}(\mathbf{Y}_i)) \\ &= \mathbf{X}_i^\top\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(\mathbf{Y}_i - \mathrm{E}(\mathbf{Y}_i))\,. \end{aligned} \tag{5.8}$$

Now Theorem 5.1 follows immediately. Noting that $\mathrm{Var}^{-1}(\mathbf{Y})$ is block diagonal, we have

$$\begin{aligned} \boldsymbol{\psi}(\boldsymbol{\beta}) &= \sum_{i=1}^{m} \boldsymbol{\psi}_i(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{m} \mathbf{X}_i^\top\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}_i))\,\mathrm{Var}(\mathbf{Y}_i)^{-1}\,(\mathbf{Y}_i - \mathrm{E}(\,Y_i)) \\ &= (\mathbf{X}_1^\top, \ldots, \mathbf{X}_m^\top)\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathrm{Var}(\mathbf{Y})^{-1} \begin{pmatrix} \mathbf{Y}_1 - \mathrm{E}(\mathbf{Y}_1) \\ \vdots \\ \mathbf{Y}_m - \mathrm{E}(\mathbf{Y}_m) \end{pmatrix} \\ &= \mathbf{X}^\top\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathrm{Var}^{-1}(\mathbf{Y})\,(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))\,. \end{aligned}$$

5.2 Estimation of dispersion parameters

We now discuss the situation when the dispersion parameters are unknown. We estimate unknown dispersion parameters using the *adjusted Pearson estimator* (Jørgensen et al. 1996b); that is, the Pearson estimator adjusted by bias correction.

5.2.1 Adjusted Pearson estimators

Recall (4.16), or equivalently

$$\sigma^2 = \frac{c(i)}{\mu_i^r} + \frac{\mathrm{E}(\hat{U}_i - \mu_i)^2}{\mu_i^r}. \tag{5.9}$$

We may thus estimate σ^2 by the following adjusted Pearson estimator:

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m} \frac{(\hat{U}_i - \mu_i)^2}{\mu_i^r} + \frac{1}{m}\sum_{i=1}^{m} \frac{c(i)}{\mu_i^r}. \tag{5.10}$$

To obtain an unbiased estimator for ω^2, we consider

$$\begin{aligned}
\mathrm{E}(\widehat{U}_{ij} - \mu_{ij}\widehat{U}_i)^2 &= \mathrm{Var}(\widehat{U}_{ij}) + \mu_{ij}^2 \mathrm{Var}(\widehat{U}_i) \\
&\quad -2\mu_{ij}\mathrm{Cov}(\widehat{U}_i, \widehat{U}_{ij}) \\
&= \mathrm{Var}(\widehat{U}_{ij}) + \mu_{ij}^2 \mathrm{Var}(\widehat{U}_i) \\
&\quad -2\mu_{ij}\mathrm{Cov}(\widehat{U}_i, U_{ij}), \qquad (5.11)
\end{aligned}$$

where the last equality follows from (4.4).

Rewrite (4.15) and (4.16) as

$$\mathrm{Var}(\widehat{U}_i) = \sigma^2\mu_i^r - c(i), \qquad (5.12)$$

and

$$\mathrm{Var}(\widehat{U}_{ij}) = \sigma^2\mu_i^r\mu_{ij}^2 + \omega^2\mu_i\mu_{ij}^q - c(ij). \qquad (5.13)$$

Plugging (5.12), (5.13) and (4.18) into (5.11), after some simplification we obtain

$$\omega^2\mu_i\mu_{ij}^q = \mathrm{E}(\widehat{U}_{ij} - \mu_{ij}\widehat{U}_i)^2 + c(i)\mu_{ij}^2 + c(ij) - 2\rho^2 c(i) w_{ij}\mu_{ij}^2. \qquad (5.14)$$

Thus we may estimate ω^2 as follows:

$$\begin{aligned}
\hat{\omega}^2 &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{(\widehat{U}_{ij} - \mu_{ij}\widehat{U}_i)^2}{\mu_i\mu_{ij}^q} \\
&\quad + \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{c(i)\mu_{ij}^2 + c(ij) - 2\rho^2 c(i) w_{ij}\mu_{ij}^2}{\mu_i\mu_{ij}^q}. \qquad (5.15)
\end{aligned}$$

The dispersion parameter ρ^2 can be estimated similarly. Noting that (4.4) implies $\mathrm{Cov}(Y_{ijk}, \widehat{U}_{ij}) = \mathrm{Cov}(Y_{ijk}, U_{ij})$, we have

$$\begin{aligned}
\mathrm{E}(Y_{ijk} - \widehat{U}_{ij}\mu_{ijk})^2 &= \mathrm{Var}(Y_{ijk}) + \mu_{ijk}^2\mathrm{Var}(\widehat{U}_{ij}) - 2\mu_{ijk}\mathrm{Cov}(Y_{ijk}, \widehat{U}_{ij}) \\
&= \mathrm{Var}(Y_{ijk}) + \mu_{ijk}^2\mathrm{Var}(\widehat{U}_{ij}) - 2\mu_{ijk}\mathrm{Cov}(Y_{ijk}, U_{ij})
\end{aligned}$$

$$\begin{aligned} &= (\mathrm{Var}(Y_{ijk}) - \mu_{ijk}\mathrm{Cov}(Y_{ijk}, U_{ij})) \\ &\quad +\mu_{ijk}^2 \left(\mathrm{Var}(\widehat{U}_{ij}) - \mathrm{Var}(U_{ij})\right) \\ &= \rho^2 \mu_i \mu_{ij} \mu_{ijk}^p + c(ij)\mu_{ijk}^2. \qquad (5.16) \end{aligned}$$

Therefore ρ^2 can be estimated as follows:

$$\begin{aligned} \hat{\rho}^2 &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\frac{(Y_{ijk} - \widehat{U}_{ij}\mu_{ijk})^2}{\mu_i\mu_{ij}\mu_{ijk}^p} \\ &\quad +\frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\frac{c(ij)\mu_{ijk}^{2-p}}{\mu_i\mu_{ij}}. \qquad (5.17) \end{aligned}$$

Similar to the REML estimator, we may also consider a small sample degree of freedom correction. For example, we may replace m by m minus the number of regression parameters estimated; however the latter correction is suggested on an intuitive basis.

5.2.2 Asymptotic properties

Let $\boldsymbol{\xi}$ denote $(\sigma^2, \omega^2, \rho^2)^\top$. Let

$$\phi_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{(\hat{U}_i - \mu_i)^2}{\mu_i^r} - \frac{c(i)}{\mu_i^r} - \sigma^2,$$

$$\phi_i^{(2)}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{J_i}\sum_{j=1}^{J_i}\left\{\frac{(\widehat{U}_{ij} - \mu_{ij}\widehat{U}_i)^2}{\mu_i\mu_{ij}^q} + \frac{c(i)\mu_{ij}^2 + c(ij) - 2\rho^2 c(i) w_{ij}\mu_{ij}^2}{\mu_i\mu_{ij}^q}\right\} - \omega^2,$$

and

$$\phi_i^{(3)}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\left\{\frac{(Y_{ijk} - \widehat{U}_{ij}\mu_{ijk})^2}{\mu_i\mu_{ij}\mu_{ijk}^p} + \frac{c(ij)\mu_{ijk}^{2-p}}{\mu_i\mu_{ij}}\right\} - \rho^2.$$

Let

$$\boldsymbol{\phi}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{i=1}^{m}\frac{1}{m}\left(\phi_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\xi}), \phi_i^{(2)}(\boldsymbol{\beta}, \boldsymbol{\xi}), \phi_i^{(3)}(\boldsymbol{\beta}, \boldsymbol{\xi})\right)^\top = \sum_{i=1}^{m}\frac{1}{m}\boldsymbol{\phi}_i(\boldsymbol{\beta}, \boldsymbol{\xi}).$$

If we replace $(\hat{\sigma^2}, \hat{\omega^2}, \hat{\rho^2})$ in (5.10), (5.15) and (5.17) by $(\sigma^2, \omega^2, \rho^2)^\top$, clearly these three equations can be rewritten as an unbiased estimating equation

$$\sum_{i=1}^{m} \frac{1}{m} \boldsymbol{\phi}_i(\boldsymbol{\beta}, \boldsymbol{\xi}) = \mathbf{0},$$

where $\boldsymbol{\phi}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ is unbiased estimating function. Together with

$$\boldsymbol{\psi}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{i=1}^{m} \boldsymbol{\psi}_i(\boldsymbol{\beta}) = \mathbf{0},$$

we obtain a set of unbiased estimating equations. Applying Lemma 2.1 again, we obtain that the sequence of roots, $(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\xi}}^\top)^\top$, is consistent for $(\boldsymbol{\beta}^\top, \boldsymbol{\xi}^\top)^\top$ and asymptotically normal with mean $(\boldsymbol{\beta}^\top, \boldsymbol{\xi}^\top)^\top$, as $m \to \infty$.

Note that

$$\begin{aligned} \mathrm{E}\frac{\partial \boldsymbol{\psi}(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} &= \mathrm{E}\left\{ \frac{\partial \left(\mathbf{X}^\top \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\, \mathrm{Var}^{-1}(\mathbf{Y})\,(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))\right)}{\partial \boldsymbol{\xi}} \right\} \\ &= \frac{\partial \left(\mathbf{X}^\top \mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\, \mathrm{Var}^{-1}(\mathbf{Y})\right)}{\partial \boldsymbol{\xi}} \mathrm{E}\,(\mathbf{Y} - \mathrm{E}(\mathbf{Y})) \\ &= \mathbf{0}. \end{aligned}$$

Taking $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ as parameter of interest and nuisance parameter respectively and applying Lemma 2.15 to $(\boldsymbol{\psi}(\boldsymbol{\beta}, \boldsymbol{\xi})^\top, \boldsymbol{\phi}(\boldsymbol{\beta}, \boldsymbol{\xi})^\top)^\top = \mathbf{0}$, we conclude that the asymptotic variance for $\hat{\boldsymbol{\beta}}$ is still given by $\mathbf{S}^{-1}(\boldsymbol{\beta})\mathbf{V}(\boldsymbol{\beta})\mathbf{S}^{-1}(\boldsymbol{\beta}) = -\mathbf{S}^{-1}(\boldsymbol{\beta})$.

In the literature of generalized linear mixed models, the asymptotic variance of $\hat{\boldsymbol{\beta}}$ for the unknown dispersion parameters case is generally taken as that obtained for the known dispersion parameters case, but with unknown dispersion parameters being naively replaced by their corresponding estimators. The usage of this naive plug-in estimator generally needs to be more cautious since it ignores the additional variability stemming from the need to estimate the dispersion parameters. Lee and Nelder (1996) proved that this information loss is asymptotically negligible for h-likelihood. Breslow and Clayton (1993) and McGilchrist (1994) did some simulations to investigate this problem for their methods. Now we showed that the asymptotic variance of $\hat{\boldsymbol{\beta}}$ for our orthodox BLUP approach is exactly the same as the naive plug-in estimator.

This result coincides with the noted observation from simulation studies for penalized quasi-likelihood method (Breslow and Lin 1995). That is, inference about regression parameters is mainly affected by the bias, rather than the variance, of the dispersion parameter estimators.

5.2.3 Heterogeneity

In practice, heterogeneity is often substantial across clusters. To account for the heterogeneity in the marginal distributions of the responses, we may allow different dispersion parameters for different clusters. One of the choices would be to assume the similar models as those in Section 3.1, but with dispersion parameters $(\sigma^2, \omega^2, \rho^2)$ being replaced by cluster specific dispersion parameters $(\sigma^2, \omega_i^2, \rho_i^2)$. All the previous derivations of orthodox BLUP random effects predictors and estimated score function remain valid if we replace $(\sigma^2, \omega^2, \rho^2)$ by $(\sigma^2, \omega_i^2, \rho_i^2)$. The dispersion parameters can now be estimated through the following equations:

$$\begin{aligned} \hat{\omega}_i^2 &= \frac{1}{J_i}\sum_{j=1}^{J_i} \frac{(\widehat{U}_{ij} - \mu_{ij}\widehat{U}_i)^2}{\mu_i \mu_{ij}^q} \\ &+ \frac{1}{J_i}\sum_{j=1}^{J_i} \frac{c(i)\mu_{ij}^2 + c(ij) - 2\rho^2 c(i) w_{ij} \mu_{ij}^2}{\mu_i \mu_{ij}^q}, \end{aligned} \tag{5.18}$$

and

$$\begin{aligned}\hat{\rho}_i^2 &= \frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\frac{(Y_{ijk}-\widehat{U}_{ij}\mu_{ijk})^2}{\mu_i\mu_{ij}\mu_{ijk}^p}\\ &\quad+\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\frac{c(ij)\mu_{ijk}^{2-p}}{\mu_i\mu_{ij}}.\end{aligned}\tag{5.19}$$

That is, the appropriate quantities are now averaged only within each cluster to account for the heterogeneity. Therefore consistency of dispersion parameter estimators $(\hat{\omega}_i^2, \hat{\rho}_i^2)$ for large number of clusters is not available for this situation.

In addition, dispersion parameters $(\sigma^2, \omega_j^2, \rho_j^2)$ often appear in the literature when the designs are balanced with respect to the number of sub-clusters, that is, $J_1 = \ldots = J_m = J$. Such dispersion parameters can be estimated as follows:

$$\begin{aligned}\omega_j^2 &= \frac{1}{m}\sum_{i=1}^{m}\frac{(\widehat{U}_{ij}-\mu_{ij}\widehat{U}_i)^2}{\mu_i\mu_{ij}^q}\\ &\quad+\frac{1}{m}\sum_{i=1}^{m}\frac{c(i)\mu_{ij}^2+c(ij)-2\rho^2c(i)w_{ij}\mu_{ij}^2}{\mu_i\mu_{ij}^q},\end{aligned}\tag{5.20}$$

and

$$\begin{aligned}\rho_j^2 &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\frac{(Y_{ijk}-\widehat{U}_{ij}\mu_{ijk})^2}{\mu_i\mu_{ij}\mu_{ijk}^p}\\ &\quad+\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\frac{c(ij)\mu_{ijk}^{2-p}}{\mu_i\mu_{ij}}.\end{aligned}\tag{5.21}$$

The index j here often represents time in longitudinal data settings where the dispersion parameter estimators are obtained by averaging the appropriate quantities across time. These two estimators are consistent for large number of clusters.

Chapter 6

Residual analysis and computational procedure

In this chapter, we briefly discuss residual analysis and computational issues.

6.1 Residual analysis

Residual analysis for both the responses and the random effects is an important ingredient of our approach. Note that the marginal residuals of the responses can be decomposed into three uncorrelated residual components as follows:

$$Y_{ijk} - \mu_i \mu_{ij} \mu_{ijk} = (Y_{ijk} - \mu_{ijk} U_{ij}) + (U_{ij} - \mu_{ij}) \mu_{ijk} U_i + (U_i - \mu_i) \mu_{ij} \mu_{ijk}. \qquad (6.1)$$

These three components actually correspond to the residuals for the three levels of distributional assumptions given in the model. Thus we may check those distributional assumptions via estimated residuals $Y_{ijk} - \mu_{ijk}\widehat{U}_{ij}$, $\hat{U}_{ij} - \mu_{ij}\widehat{U}_i$ and $\hat{U}_i - \mu_i$. All three estimated residuals have zero mean and can be standardized to have unit variance. Actually the variances of the three estimated residuals are already available from (5.16), (5.14) and (5.9). We now define the appropriate types of standardized (estimated) residuals for the three levels of distributions.

Level 1:

$$r^{(1)} = \frac{\widehat{U}_i - \mu_i}{\sqrt{\sigma^2 \mu_i^r - c(i)}}. \tag{6.2}$$

Level 2:

$$r_c^{(2)} = \frac{\widehat{U}_{ij} - \mu_{ij}\hat{U}_i}{\sqrt{\omega^2 \mu_i \mu_{ij}^q - c(i)\mu_{ij}^2 - c(ij) + 2\rho^2 c(i) w_{ij} \mu_{ij}^2}}. \tag{6.3}$$

Level 3:

$$r_c^{(3)} = \frac{Y_{ijk} - \mu_{ijk}\widehat{U}_{ij}}{\sqrt{\rho^2 \mu_i \mu_{ij} \mu_{ijk}^p + c(ij)\mu_{ijk}^2}}. \tag{6.4}$$

The level 2 and 3 residuals, $r_c^{(2)}$ and $r_c^{(3)}$, are residuals for conditional distributions, so we will call them *conditional residuals.* The level 1 residual $r^{(1)}$ is the residual for the first level random effects distribution. We can also check the marginal distributional assumptions of the response and the second level random effects through the following *marginal residuals*:

$$r^{(2)} = \frac{\widehat{U}_{ij} - \mu_i \mu_{ij}}{\sqrt{\sigma^2 \mu_i^r \mu_{ij}^2 + \omega^2 \mu_i \mu_{ij}^q - c(ij)}},$$

$$r^{(3)} = \frac{Y_{ijk} - \mu_i \mu_{ij} \mu_{ijk}}{\sqrt{\sigma^2 \mu_i^r \mu_{ij}^2 \mu_{ijk}^2 + \omega^2 \mu_i \mu_{ij}^q \mu_{ijk}^2 + \rho^2 \mu_i \mu_{ij} \mu_{ijk}^p}}.$$

The basic idea in residual analysis is then to use plots of standardized residuals in much the same way as in standard generalized linear models. Plots of standardized residuals against each covariate are useful for detecting nonlinearity of the data relative to the model. To check the log link assumption, we plot $\log Y_{ijk}$ against the log fitted values $\log \mu_{ijk}\widehat{U}_{ijk}$. Ideally this plot should show a horizontal linear relationship; curvature or other unusual shape would indicate inadequacy of the log link assumption.

As in generalized linear models, we may check the form of the variance functions and hence the distributional forms by plotting the standardized residuals against the corresponding fitted values. In order to check the variance function of the distribution for the random effects, we thus plot the standardized residuals against log fitted values.

6.2 Computational procedure

Initial values for regression parameter estimates are taken as the regression parameter estimates obtained from standard generalized linear models techniques assuming independent responses. We take initial random effects predictions for $\hat{U}_i$ and $\hat{U}_{ij}$ as the average of the responses within cluster i divided by the average of all responses and the average of the responses within sub-cluster (i, j) divided by the average of all responses, respectively. That is

$$\widehat{U}_i^{(0)} = \frac{\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} Y_{ijk}}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} Y_{ijk}},$$

and

$$\widehat{U}_{ij}^{(0)} = \frac{\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} Y_{ijk}}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} Y_{ijk}}.$$

The initial dispersion parameter estimates are calculated from Pearson estimators via (5.10), (5.15) and (5.17), but omitting the bias-correction terms. More specifically, we have

$$\widehat{\sigma}^2_{(0)} = \frac{1}{m}\sum_{i=1}^{m}\frac{(\hat{U}_i - \mu_i)^2}{\mu_i^r},$$

and

$$\widehat{\omega}^2_{(0)} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{(\widehat{U}_{ij} - \mu_{ij}\widehat{U}_i)^2}{\mu_i\mu_{ij}^q},$$

and

$$\widehat{\rho}^2_{(0)} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\frac{(Y_{ijk} - \widehat{U}_{ij}\mu_{ijk})^2}{\mu_i\mu_{ij}\mu_{ijk}^p}.$$

The algorithm then iterates between updating the regression parameter estimates via the Newton scoring algorithm, updating random effect predictors via the orthodox BLUP and updating dispersion parameter estimates via the adjusted Pearson estimators.

Chapter 7

Conventional, semiparametric and binomial mixed models

In the literature of multiplicative random effects models, the marginal means of the random effects are usually taken to be 1 (Morton 1987; Thall and Vail 1990; Firth and Harris 1991; Lee and Nelder 1996). Imposing these conventional constraints on the Tweedie mixed models is equivalent to taking $\mu_i = 1$ and $\mu_{ij} = 1$ for all i and j. We thus call a Tweedie mixed model with such contraints a conventional Tweedie mixed model. After discussing the conventional Tweedie mixed models in the following section, we will discuss mixed models beyond the Tweedie family which are related to the conventional Tweedie mixed models. The first model is a semiparametric mixed model with only first and second monents assumptions about the random effects; the second one is a mixed model assuming log-normally distributed random effects; whereas the last model involves binomial distributions for the conditional responses.

7.1 Conventional Tweedie mixed models

To be more specific, we give the assumptions for the conventional Tweedie mixed models:

B1) Given $\mathbf{U} = \mathbf{u}$, $Y_{111}, ..., Y_{11n_{11}}, ..., Y_{ij1}, ..., Y_{ijn_{ij}}, ..., Y_{mJ_m1}, ..., Y_{mJ_mn_{mJ_m}}$ are conditionally independent, and the conditional distribution of Y_{ijk}, given $\mathbf{U} = \mathbf{u}$, depends on u_{ij} only as

$$\begin{aligned} Y_{ijk}|\mathbf{U} = \mathbf{u} &\sim \mathrm{Tw}_p(\mu_{ijk}u_{ij}, \rho^2 {u_{ij}}^{1-p}) \\ &= u_{ij}\mathrm{Tw}_p(\mu_{ijk}, \frac{\rho^2}{u_{ij}}), \end{aligned}$$

where $\mu_{ijk} = \exp(\mathbf{x}_{ijk}^\top \boldsymbol{\beta})$.

B2) Given $\mathbf{U}_* = \mathbf{u}_*$, $U_{11}, ..., U_{mJ_m}$ are conditionally independent, and the conditional distribution of U_{ij}, given $\mathbf{U}_* = \mathbf{u}_*$, depends on u_i only, which is

$$\begin{aligned} U_{ij}|\mathbf{U}_* = \mathbf{u}_* &\sim \mathrm{Tw}_q(u_i, \omega^2 {u_i}^{1-q}) \\ &= u_i\mathrm{Tw}_q(1, \frac{\omega^2}{u_i}). \end{aligned}$$

B3)

$$U_1, ..., U_m \quad iid \quad \mathrm{Tw}_r(1, \sigma^2).$$

The derived random effects predictors, estimated score function and adjusted Pearson estimators for the more general models in Section 3.1 remain valid with the constraints $\mu_i = \mu_{ij} = 1$. Since the detailed expressions for the random effects predictors, estimated score function and adjusted Pearson estimators will be useful to discuss related models beyond the Tweedie family, we will present these expressions in the following sections.

7.1.1 Covariance and variance function

The first and second moments of the second level random effects possess the following simple structure

$$\mathrm{E}[U_{ij}] = 1,$$

and

$$\mathrm{Cov}[U_{st}, U_{ij}] = \begin{cases} \sigma^2 + \omega^2 & \text{if } (s,t) = (i,j) \\ \sigma^2 & \text{if } s = i \text{ and } t \neq j \\ 0 & \text{otherwise.} \end{cases}$$

The covariance between the random effects and responses is simply given by

$$\mathrm{Cov}[U_{st}, Y_{ijk}] = \delta(s,i)\left\{\sigma^2 + \delta(t,j)\omega^2\right\}\mu_{ijk}.$$

The first and second moments of the response for the conventional Tweedie mixed models are then given by

$$\mathrm{E}[Y_{ijk}] = \mu_{ijk},$$

and

$$\mathrm{Cov}[Y_{stl}, Y_{ijk}] = \begin{cases} \rho^2\mu_{ijk}^p + (\sigma^2 + \omega^2)\mu_{ijk}^2 & \text{if } (s,t,l) = (i,j,k) \\ (\sigma^2 + \omega^2)\mu_{ijk}\mu_{ijl} & \text{if } (s,t) = (i,j) \text{ and } l \neq k \\ \sigma^2\mu_{ijk}\mu_{itl} & \text{if } s = i \text{ and } t \neq j \\ 0 & \text{otherwise.} \end{cases}$$

The covariance of the response displays an interpretable variance components structure which clearly shows the variance contribution from each level of random effects.

The marginal variance of Y_{ijk} for conventional Tweedie mixed models has a simple form as follows:

$$\mathrm{Var}(Y_{ijk}) = \rho^2 \mu_{ijk}^p + (\sigma^2 + \omega^2)\mu_{ijk}^2. \tag{7.1}$$

The case $p = 1$ and $p = 2$ which correspond to Poisson-Tweedie models and gamma-Tweedie models are especially interesting. The marginal variances of these two models are

$$\mathrm{Var}(Y_{ijk}) = \mu_{ijk} + (\sigma^2 + \omega^2)\mu_{ijk}^2,$$

and

$$\mathrm{Var}(Y_{ijk}) = (\rho^2 + \sigma^2 + \omega^2)\mu_{ijk}^2,$$

respectively. Therefore the marginal variance functions of these two models coincide with those of the negative binomial and gamma distributions, respectively. It is known that the marginal distribution of Y_{ijk} of the conventional Poisson-gamma models with only one level of random effects ($\omega^2 = 0$) follows a negative binomial distribution. This negative binomial distribution is frequently adopted to account for overdispersion relative to Poisson model (Venables and Ripley 1994). It is unclear, in general, if the marginal distributions of the responses of the Poisson-Tweedie and gamma-Tweedie models follow the negative binomial and gamma distributions, respectively. It is known that the distribution of a random variable is uniquely characterized by its variance function within the exponential dispersion models (Jørgensen 1997); however, the marginal distribution of Y_{ijk} may not follow an exponential dispersion model (Jørgensen 1987). Hence whether the marginal distributions of the responses of these two models follow the negative binomial and gamma distributions, respectively, remains an open question.

Finally, we give an expression for the correlation between the responses:

$$\mathrm{Corr}(Y_{ijk}, Y_{stl}) = \frac{\delta(s,i)\left\{\sigma^2\mu_{ijk}\mu_{itl} + \delta(t,j)[\omega^2\mu_{ijk}\mu_{ijl} + \delta(l,k)\rho^2\mu_{ijk}^p]\right\}}{\sqrt{\rho^2\mu_{ijk}^p + (\sigma^2+\omega^2)\mu_{ijk}^2}\sqrt{\rho^2\mu_{stl}^p + (\sigma^2+\omega^2)\mu_{stl}^2}}.$$

Clearly the correlation matrix still depends on both the regression and the dispersion parameters for the conventional Tweedie mixed models.

7.1.2 Random effects predictors

The random effects predictors for conventional Tweedie mixed models also have simple expressions as follows:

$$\widehat{U}_i = \frac{1 + \sigma^2 \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ij}\mu_{ijk}^{1-p} Y_{ijk}}{1 + \sigma^2 \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ij}\mu_{ijk}^{2-p}}, \tag{7.2}$$

where $w_{ij} = 1/(\rho^2 + \omega^2 \sum_{k=1}^{n_{ij}} \mu_{ijk}^{2-p})$, and

$$\widehat{U}_{ij} = \rho^2 w_{ij}\widehat{U}_i + \omega^2 w_{ij} \sum_{k=1}^{n_{ij}} \mu_{ijk}^{1-p} Y_{ijk}. \tag{7.3}$$

The conventional Poisson-gamma model coincides with the conjugate Poisson-gamma model studied by Lee and Nelder (1996) when there is only one level of the random effects. For this model both the orthodox BLUP predictors and MHLEs for the random effects are exactly the conditional expectations of the random effects given the responses; therefore both orthodox BLUP and MHLE estimation procedure lead to maximum likelihood estimates for $\boldsymbol{\beta}$ when dispersion parameters are known. In general, the orthodox BLUP predictors and MHLEs are not identical. Lee and Nelder (1996) presented a class of conjugate hierarchical generalized linear models with one level of random effects, and provided an explicit expression for MHLEs of the random effects. As with any multiplicative conjugate hierarchical generalized linear models within this class, a direct calculation shows that the MHLEs for the random effects are linear functions of the response. However, the orthodox BLUP predictors of the random effects are the best linear unbiased predictor for Tweedie mixed models.

The estimates for $\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{U}})$ and $\text{Var}(\hat{\mathbf{U}} - \mathbf{U})$ are useful in making inferences about realized or sample values of $\mathbf{U}$ (Harville 1976). Lee and Nelder (1996) mentioned that they have not found consistent estimates for these quantities except in few special cases. They further conjectured that consistent estimates may not exist under their regularity condition: the number of clusters is fixed when the number of observations increases. We have not obtained estimates for $\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{U}})$ for the orthodox BLUP approach yet, but exact expressions for $\text{Var}(\hat{\mathbf{U}} - \mathbf{U})$ are available below:

$$\text{Cov}[\hat{U}_i - U_i, \hat{U}_s - U_s] = \delta_{(s,i)} c(i)$$

$$\text{Cov}[\hat{U}_i - U_i, \hat{U}_{st} - U_{st}] = \delta_{(s,i)} \rho^2 c(i) w_{it}$$

$$\text{Cov}[\hat{U}_{ij} - U_{ij}, \hat{U}_{st} - U_{st}] = \delta_{(s,i)} \left\{ \rho^2 w_{ij} [\delta_{(t,j)} \omega^2 + \rho^2 c(i) w_{it}] \right\}.$$

These are clearly consistent estimates for $\text{Var}(\hat{\mathbf{U}} - \mathbf{U})$ when the parameter estimators are plugged in.

Besides the general small dispersion and large sample asymptotics established in Section 4.4, we give another small dispersion asymptotics result:

$$c(i) \leq \frac{\sigma^2(\rho^2 + \omega^2 \text{min}_j(n_{ij}) \text{min}_{jk}(\mu_{ijk}^{2-p}))}{\rho^2 + \omega^2 \text{min}_j(n_{ij}) \text{min}_{jk}(\mu_{ijk}^{2-p}) + \sigma^2 \text{min}_j(n_{ij}) \text{min}_{jk}(\mu_{ijk}^{2-p}) J_i}, \tag{7.4}$$

and

$$c(ij) \le \frac{\rho^2(\omega^2+\sigma^2)}{\rho^2+\omega^2\min_j(n_{ij})\min_{jk}(\mu_{ijk}^{2-p})}. \tag{7.5}$$

Obviously, when $p \neq 2$, we also have

$$\hat{U}_{ij} \xrightarrow{P} U_{ij} \text{ as } \min_{jk}(\mu_{ijk}^{2-p}) \to \infty.$$

For Poisson-Tweedie ($p = 1$) and compound Poisson-Tweedie ($1 < p < 2$) models, $\min_{jk}(\mu_{ijk}^{2-p}) \to \infty$ is equivalent to all $\mu_{ijk} \to \infty$. For positive stable Tweedie models ($p > 2$), including the inverse Gaussian, $\min_{jk}(\mu_{ijk}^{2-p}) \to \infty$ is equivalent to all $\mu_{ijk} \to 0$.

Clearly when $p = 2$, the same large sample asymptotics conclusion as in Section 4.4 holds without any restrictions on μ_{ijk}s. Actually, the large sample asymptotics conclusion upon the second level random effects now holds for any p under much relaxed conditions such as $\min_j(n_{ij})\min_{jk}(\mu_{ijk}^{2-p}) \to \infty$ as $\min_j(n_{ij}) \to \infty$. A sufficient condition is that $\min_{jk}(\mu_{ijk}^{2-p}) \ge c\log(\min_j(n_{ij}))/\min_j(n_{ij})$ for a positive constant c. That is, the only restriction is that μ_{ijk} should not tend to zero too fast for the Poisson and compound Poisson cases, whereas $\min_{jk}(\mu_{ijk}^{2-p})$ should not tend to infinity too fast for positive stable case.

Actually, for conventional Tweedie mixed models, only the behavior of $\hat{U}_{ij}$ affects the orthodox BLUP estimation procedure. This will become apparent in Section 7.2.

7.1.3 Parameter estimation

The estimated score function now reduces to

$$\boldsymbol{\psi}(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} \mathbf{x}_{ijk}\frac{\mu_{ijk}^{1-p}}{\rho^2}(Y_{ijk} - \hat{U}_{ij}\mu_{ijk}), \tag{7.6}$$

whereas the matrix form of the estimated score function is still given by

$$\psi(\boldsymbol{\beta}) = \mathbf{X}^{\top}\mathrm{diag}\left(\mathrm{E}(\mathbf{Y})\right)\mathrm{Var}^{-1}(\mathbf{Y})\left(\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right).$$

In addition, the expression for sensitivity matrix $\mathbf{S}(\boldsymbol{\beta})$ also has the following simplified version:

$$\begin{aligned}\mathbf{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{m} c(i)(\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ijk}^{2-p}\mathbf{x}_{ijk})(\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}} w_{ij}\mu_{ijk}^{2-p}\mathbf{x}_{ijk})^{\top} \\ &\quad + \sum_{i=1}^{m}\sum_{j=1}^{J_i}\frac{\omega^2 w_{ij}}{\rho^2}(\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p}\mathbf{x}_{ijk})(\sum_{k=1}^{n_{ij}}\mu_{ijk}^{2-p}\mathbf{x}_{ijk})^{\top} \\ &\quad - \sum_{i=1}^{m}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}}\frac{1}{\rho^2}\mu_{ijk}^{2-p}\mathbf{x}_{ijk}\mathbf{x}_{ijk}^{\top}. \end{aligned} \qquad (7.7)$$

7.1.4 Adjusted Pearson estimators

The adjusted Pearson estimators have simpler forms:

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}(\widehat{U}_i - 1)^2 + \frac{1}{m}\sum_{i=1}^{m} c(i), \qquad (7.8)$$

$$\hat{\omega}^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\left\{(\widehat{U}_{ij} - \widehat{U}_i)^2 + c(ij) + c(i) - 2\rho^2 c(i) w_{ij}\right\}, \qquad (7.9)$$

$$\hat{\rho}^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\left\{\frac{(y_{ijk} - \mu_{ijk}\widehat{U}_{ij})^2}{\mu_{ijk}^{p}} + c(ij)\mu_{ijk}^{2-p}\right\}. \qquad (7.10)$$

7.2 Random effects beyond Tweedie family

For the general Tweedie mixed models defined in Section 3.1, the estimation procedure of the orthodox BLUP approach depends on the specific parametric forms of the random effects distributions among Tweedie family of distributions through

μ_i, μ_{ij}, q and r. Since we have $\mu_i = \mu_{ij} = 1$ for conventional Tweedie mixed models, this dependence of the estimation procedure on the specific parametric random effects distributional forms forms is then via q and r only. However, for conventional Tweedie mixed models, q and r actually do not enter the estimation procedure in either the estimated score function, the sensitivity matrix, the random effects predictors or the adjusted Pearson estimators. That is, the estimation procedure of the orthodox BLUP approach to conventional Tweedie mixed models is robust against misspecification of the random effects distributions within the Tweedie family.

This robustness indicates that the orthodox BLUP approach to conventional Tweedie mixed models depends on the random effects only via the first and second moments of the second level random effects. Therefore we can extend the orthodox BLUP approach to deal with conventional models with only the first and second moment assumptions about the second level of random effects. These models are usually referred as semiparametric random effects models with nonparametric random effects.

7.2.1 Nonparametric random effects

To discuss this semiparametric model in more detail, we assume that

C1) Given $\mathbf{U} = \mathbf{u}$, $Y_{111}, ..., Y_{11n_{11}}, ..., Y_{ij1}, ..., Y_{ijn_{ij}}, ..., Y_{mJ_m1}, ..., Y_{mJ_mn_{mJ_m}}$ are conditionally independent, and the conditional distribution of Y_{ijk}, given $\mathbf{U} = \mathbf{u}$, depends on u_{ij} only as

$$\begin{aligned} Y_{ijk}|\mathbf{U} = \mathbf{u} &\sim Tw_p(\mu_{ijk}u_{ij}, \rho^2 {u_{ij}}^{1-p}) \\ &= u_{ij}\mathrm{Tw}_p(\mu_{ijk}, \frac{\rho^2}{u_{ij}}), \end{aligned} \tag{7.11}$$

where $\mu_{ijk} = \exp(\mathbf{x}_{ijk}^\top \boldsymbol{\beta})$. For case $p = 1$, namely Poisson distribution, $\rho^2 = 1$.

C2) The random effects $U_{11}, \ldots, U_{mJ_m}$ are positive random variables with

$$\mathrm{E}(U_{ij}) = 1,$$

and

$$\mathrm{Cov}[U_{st}, U_{ij}] = \delta_{(s,i)} \left\{ \sigma^2 + \delta_{(t,j)} \omega^2 \right\}.$$

Here we make no further parametric assumptions on the random effects. In addition, we do not explicitly assume any first level random effects.

Note that the random effects $U_{11}, \ldots, U_{mJ_m}$ have the same first and second moments as in conventional Tweedie mixed models; therefore it follows from assumption C1) that $\mathrm{Cov}[U_{st}, Y_{ijk}]$, $\mathrm{E}[Y_{ijk}|\mathbf{U}]$ and $\mathrm{Cov}[Y_{stl}, Y_{ijk}]$ are also the same as in the conventional Tweedie mixed models since the derivation based on the conditioning techniques in Section 3.1 did not involve the Tweedie distributional assumption of U_{ij}. These moments then imply that the orthodox BLUP predictors of the random effects U_{ij}, the estimated score function and the sensitivity matrix have exactly the same expressions as in conventional Tweedie mixed models.

The adjusted Pearson estimator for ρ^2 also has the same expression as in conventional Tweedie mixed models. However the adjusted Pearson estimators for ω^2 and σ^2 for conventional Tweedie mixed models involve $\widehat{U}_i$ although we do not assume the existence of $\widehat{U}_i$. On the other hand, in conventional Tweedie mixed models $\widehat{U}_i$ is a linear function of the responses and the adjusted Pearson estimators are moment estimators; therefore the estimating equation induced by the adjusted Pearson estimators for ω^2 and σ^2 remain valid if we use the linear expression for $\widehat{U}_i$ as an intermediate quantity to estimate ω^2 and σ^2 in the current semiparametric models. The difference is that the linear expression for $\widehat{U}_i$ does not necessarily have a parametric interpretation beyond the Tweedie family.

In summary, the estimating equations associated with the estimated score function and adjusted Pearson estimators given in (7.6), (7.8), (7.9) and (7.10) are still unbiased estimating equations for both the regression and dispersion parameters in these semiparametric models. The regression and dispersion parameter estimators are also consistent asymptotically normal with the inverse of $-\mathbf{S}(\boldsymbol{\beta})$ being the asymptotic variance of $\boldsymbol{\beta}$ under the same regularity conditions.

As the expression for $\mathrm{Var}(\widehat{\mathbf{U}} - \mathbf{U})$ in (4.2)

$$\begin{aligned} c(ij) &= \mathrm{Var}(\widehat{U}_{ij} - U_{ij}) \\ &= \mathrm{E}(U_{ij}) - \mathrm{Cov}(U_{ij}, \mathbf{Y}_i)\mathrm{Var}^{-1}(\mathbf{Y}_i)\mathrm{Cov}(\mathbf{Y}_i, U_{ij}) \end{aligned}$$

depends on the first and second moments only, the expression for $c(ij)$ will not be changed. Thus both small dispersion and large sample consistency results $\widehat{U}_{ij}$ in conventional Tweedie mixed models can be extended to these semiparametric models under the same regularity conditions.

In conclusion, besides its parametric interpretation with the Tweedie random effects distributions, the orthodox BLUP approach to generalized linear mixed models is robust against misspecification of the random effects distributions.

7.2.2 Log-normal random effects

An interesting example of non-Tweedie distributions for U_{ij} in the conventional Tweedie mixed models setting is the log-normal. Let V_{ij}s be log-normal random variables

$$\log V_{ij} \sim N\left(0, \tau^2\right),$$

with $\tau^2 = \log(1 + \sigma^2 + \omega^2)$ and covariance structure

$$\operatorname{Cov}[V_{st}, V_{ij}] = \frac{\delta_{(s,i)}\left\{\sigma^2 + \delta_{(t,j)}\omega^2\right\}}{1 + \sigma^2 + \omega^2}.$$

Clearly $U_{ij} = \frac{V_{ij}}{\sqrt{1+\sigma^2+\omega^2}}$ satisfies

$$\mathrm{E}(U_{ij}) = 1,$$

and

$$\operatorname{Cov}[U_{st}, U_{ij}] = \delta_{(s,i)}\left\{\sigma^2 + \delta_{(t,j)}\omega^2\right\}.$$

Assume C1) in the last section holds, then we would have

$$\begin{aligned}\log \mathrm{E}(Y_{ijk}|\mathbf{U}) &= \mathbf{x}_{ijk}^{\top}\boldsymbol{\beta} + \log U_{ij}\\ &= \mathbf{x}_{ijk}^{\top}\boldsymbol{\beta} - \frac{1}{2}\log(1 + \sigma^2 + \omega^2) + \log V_{ij}.\end{aligned}$$

where, under log link, the term $\frac{1}{2}\log(1+\sigma^2+\omega^2)$ affects only the value of the intercept.

Now the conditional expectations of the responses given the random effects can be expressed as a linear combination of regression parameters plus the normal random effects. This example shows the connection between generalized linear models with Tweedie random effects distributions and those with the normal random effects distribution. Note that the covariance structure assumptions are made about the log-normal random effects instead of the normal random effects; therefore, unlike Tweedie random effects cases, the partially observed joint log likelihood for both the response and log-normal random effects is not explicitly given here.

7.2.3 Binomial mixed models

Binomial mixed models are usually handled by logistic-normal and binomial-beta models because of their analytic tractability. While our orthodox BLUP approach to Tweedie mixed models is not directly applicable to this case since the binomial distribution does not belong to the Tweedie family, we can deal with binomial mixed models via the conventional Poisson-Tweedie models based on the the following well-known relationship between the binomial and Poisson random variables:

Suppose that Y_1 and Y_2 are independent with

$$Y_k \sim \text{Poisson}(\mu_k),$$

then

$$Y_1 + Y_2 \sim \text{Poisson}(\mu_1 + \mu_2),$$

and

$$Y_1 | Y_1 + Y_2 = N \sim \text{binomial}(N, \mu_1/(\mu_1 + \mu_2)).$$

(cf. McCullagh and Nelder 1989).

Suppose that there are m pairs of observations, (R_i, N_i) $i = 1, 2, \ldots, m$, R_i being the number of successes in N_i trials. Consider the following paired Poisson-Tweedie models:

$$\begin{cases} Y_{i1} = R_i | \mathbf{U} = \mathbf{u} \sim \text{Poisson}\left(u_{i1} \exp[\mathbf{x}_i^\top (\boldsymbol{\alpha} + \boldsymbol{\beta})]\right) \\ Y_{i2} = N_i - R_i | \mathbf{U} = \mathbf{u} \sim \text{Poisson}\left(u_{i2} \exp[\mathbf{x}_i^\top \boldsymbol{\alpha}]\right), \end{cases}$$

where $Y_{11}, Y_{12}, \ldots, Y_{m1}, Y_{m2}$ are independent given $\mathbf{U} = \mathbf{u}$. Assume further that the random effects $\mathbf{U} = (U_1, ..., U_m, U_{11}, ..., U_{mJ_m})^\top$ satisfy assumption B2) and B3) for conventional Tweedie mixed models with $J_i = 2$ for all i as usual. Then we have

$$Y_{i1}|Y_{i1}+Y_{i2}=N_i,\mathbf{U} \quad \sim \quad \text{binomial}(N_i,p_i),$$

where

$$p_i=\frac{U_{i1}\exp[\mathbf{x}_i^\top(\boldsymbol{\alpha}+\boldsymbol{\beta})]}{U_{i1}\exp[\mathbf{x}_i^\top(\boldsymbol{\alpha}+\boldsymbol{\beta})]+U_{i2}\exp[\mathbf{x}_i^\top\boldsymbol{\alpha}]}.$$

Hence

$$\text{logit } p_i=\mathbf{x}_i^\top\boldsymbol{\beta}+\text{logit}\frac{U_{i1}}{U_{i1}+U_{i2}}.$$

Clearly $\sigma^2=\omega^2=\mathbf{0}$ corresponds to binomial models without the random effects and $\sigma^2=\mathbf{0}$ is equivalent to one level random effects models.

The binomial-beta model is the special case with $\sigma^2=\mathbf{0}$ and $U_{ij}\sim \text{Gamma}(1,\omega^2)$ $j=1,2$. This is because $V_i=\frac{U_{i1}}{U_{i1}+U_{i2}}$ is known to follow a symmetric beta distribution if $U_{ij}\sim\text{Gamma}(1,\omega^2)$ $j=1,2$. Asymmetric beta distributed random effects can easily be incorporated by allowing different dispersion parameters ω_j^2 for U_{ij} $j=1,2$. Lee and Nelder (1996) also derived the binomial-beta model from a pair of Poisson-gamma models with the common covariates $\mathbf{x}_i^\top$, but different regression parameters $\boldsymbol{\beta}_j$ for each of the two groups $(J=1,2)$. Under logit link, they have

$$\text{logit } p_i=\mathbf{x}_i^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)+\text{logit}\frac{U_{i1}}{U_{i1}+U_{i2}},$$

with $\boldsymbol{\beta}_1-\boldsymbol{\beta}_2$ corresponding to our $\boldsymbol{\beta}$ here. But they derived their estimating equations for MHLEs directly from the binomial-beta model.

We derive our binomial mixed models based on paired Poisson-Tweedie models with the common regression parameter $(\boldsymbol{\alpha}^\top,\boldsymbol{\beta}^\top)^\top$, but different covariates for each of the two groups, that is, $(\mathbf{x}_i^\top,\mathbf{x}_i^\top)$ for R_i and $(\mathbf{x}_i^\top,\mathbf{0})$ for N_i-R_i. This approach enables us to obtain the fixed effects estimate $\boldsymbol{\beta}$ as well as its standard error and the random effects predictors $\hat{v}_i$ for binomial mixed models by fitting the paired Poisson-Tweedie models. Clearly $\hat{\boldsymbol{\beta}}$ is a consistent estimate for $\boldsymbol{\beta}$ as $m \to \infty$ since $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is shown to be a consistent estimate for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in Section 5.2.2.

There are many different approaches to Poisson mixed models in the literature. The paired Poisson mixed models trick provides a way of handling binomial mixed models via Poisson mixed models. The extension of this trick to handle multinomial mixed models via Poisson mixed models is straightforward. Such multinomial data examples can be found in McCullagh and Nelder (1989).

Chapter 8

Data analysis

Illustrative examples for the orthodox BLUP approach to analyzing data of different types based on conventional Tweedie mixed models are presented in this chapter.

We will concentrate mainly on the analysis of the epilepsy data to illustrate the use of variations of the models. We will also analyze the seed germination data and cake baking data to illustrate the use of orthodox BLUP approach to handle binomial and continuous data.

8.1 Count data

In this section, we illustrate the orthodox BLUP approach to count data with analyses of the epilepsy data described in Section 2.4.1. Preliminary analysis indicated that the counts were much lower during the fourth visit so Breslow and Clayton (1993) introduced a linear trend covariate Visit (coded (-0.3,-0.1,0.1,0.3)). To facilitate visual understanding of the data, Breslow (1996) plotted the logarithm of the seizure counts reported at baseline (visit 0) and at each of the four visits as in Figure 8.1. Counts at the four follow-up visits were increased by 0.5 when taking log-transform

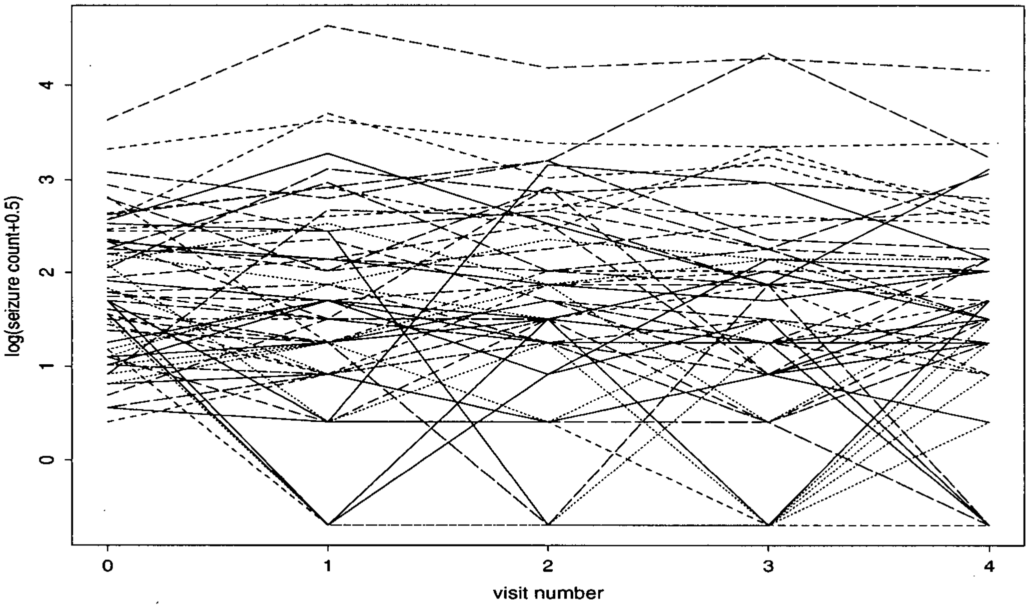


Figure 8.1: Profile plot of log transformed epilepsy seizure counts.

to avoid infinities. The baseline counts were divided by four since the baseline counts were measured over eight weeks instead of two weeks. The baseline seizure counts are less variable because the period over which they were measured was four times as long as for each of the subsequent visits. Patient 207 had exceptionally high counts at baseline count and all subsequent visits.

8.1.1 Overview of the previous analyses

Breslow and Clayton (1993) reanalyzed the epilepsy data using the penalized quasi-likelihood approach, whereas Lee and Nelder (1996) also reanalyzed the same data based on hierarchical generalized linear models. To make a systematic comparison of these analyses, we summarize their models as follows.

Lee and Nelder (1996) considered a model where the response Y given the random effects $(\mathbf{U}, \mathbf{V})$ follows a Poisson distribution and

$$Y_{ij}|\mathbf{U} = \mathbf{u}, \mathbf{V} = \mathbf{v} \quad \sim \quad \text{Poisson}(\mu_{ij} u_i v_{ij}) \; i = 1, \ldots, 59, \; j = 1, 2, 3, 4, \tag{8.1}$$

where U_is and V_{ij}s are all mutually independent having $\mathrm{E}(U_i) = \mathrm{E}(V_{ij}) = 1$ and $\mathrm{Var}(U_i) = \sigma^2$ and $\mathrm{Var}(V_{ij}) = \sigma_j^2$. Lee and Nelder assume that U_i and V_{ij} follow gamma distributions.

Thall and Vail (1990) considered a similar model, their model 22, but with V_{ij} replaced by V_j with known first two moments only. Breslow and Clayton (1993) also considered models similar to Lee and Nelder's models. Since these approaches yield very similar results, we list the results of Lee and Nelder in Table 8.1 for later comparisons in this section.

Model GLM in Table 8.1 is the standard Poisson regression model. Model HGLM 1 has $\sigma_j = 0$ for all $j = 1, 2, 3, 4$, namely the model of random subject effects only. Model HGLM 2 has $V_{ij} = V_i \times \text{Visit}_j$ where Visit_j is the jth component of covariate Visit for $j = 1, 2, 3, 4$. The model HGLM 3 is the one with $\sigma_1 = \sigma_2 = \sigma_3$ and $\sigma_4 = 0$. Lee and Nelder set $\sigma_4 = 0$ since it is reported to tend to zero when fitting. By the way, the estimated GLM regression coefficient for covariate Visit was reported as -0.29 in their paper which appears to be a typographical error.

Table 8.1: HGLM parameter estimates for the epilepsy data.

	GLM		HGLM 1		HGLM 2		HGLM 3	
Parameter	est.*	s.e.*	est.	s.e.	est.	s.e.	est.	s.e.
Constant	−2.80	0.41	−1.35	1.20	−1.32	1.20	−1.21	1.20
Base	0.95	0.04	0.88	0.13	0.89	0.13	0.89	0.12
Trt	−1.34	0.16	−0.89	0.39	−0.86	0.39	−0.86	0.38
Base.Trt	0.56	0.06	0.34	0.20	0.33	0.20	0.32	0.19
Age	0.90	0.12	0.51	0.36	0.49	0.36	0.46	0.35
Visit	−0.08	0.10	−0.29	0.10	−0.16	0.16	−0.28	0.14
σ^2			0.27		0.26		0.22	
σ_1^2					0.40		0.15	
σ_4^2					0		0	

$*$ est. and s.e. represent estimates and standard errors respectively.

8.1.2 Poisson-Tweedie models

Our reanalysis of these data is based on conventional two level Poisson-Tweedie models. The fixed effects are logarithm of a quarter of baseline seizure counts (Base), logarithm of age (Age), Trt and Visit. Since there is no repetition for each visit, the third index is thus omitted. We denote y_{ij} the seizure count for patient i on visit j; therefore the random effects u_i and u_{ij} are at patient and visit levels respectively. The formulation is as follows:

$$Y_{ij}|\mathbf{U} = \mathbf{u} \quad \sim \quad \text{Poisson}(\mu_{ij}u_{ij}),$$

whereas

$$U_{ij}|\mathbf{U}_* = \mathbf{u}_* \quad \sim \quad \text{Tw}_q(1, \omega_{ij}^2 u_i), \tag{8.2}$$

and

$$U_1, \ldots, U_{59} \quad \text{i i d} \quad \text{Tw}_r(1, \sigma^2),$$

where $j = 1, 2, 3, 4$ and $i = 1, \ldots, 59$. The conditional independences corresponding to assumptions B1) and B2) given in Section 7.1 are assumed to hold. As explained in Chapter 7, this model can also be regarded as a semiparametric model.

We consider four models with different assumptions for the dispersion parameters ω_{ij}^2. Model($\omega^2 = 0$), Model(ω^2), Model(ω_j^2) and Model(ω_i^2) assume $\omega_{ij}^2 =$ 0, ω^2, ω_j^2 and ω_i^2 in (8.2), respectively. Thus Model($\omega^2 = 0$) is the same one level model as HGLM 1, but without the gamma distributional assumption for the random effects. Model(ω^2) is the model with equal ω^2 for all (i, j) corresponding to HGLM 3. Model(ω_j^2), namely the model with distinct ω_j^2 for different visits, has the same covariance structure of the response as that of model 22 proposed by Thall and Vail through crossed random effects design. The last model, Model(ω_i^2), has distinct ω_i^2 for each subject.

We start with the Model(ω^2), including all possible two way interaction terms among the covariates. We removed all those interaction terms except the interaction term Base.Trt from the model based on the Wald test and residual analysis. The p-value of Wald test for this last interaction term is slightly larger than 5%, but we retained it in all our models to allow a systematic comparison. That is, we take Base, Age, Trt, Visit and Base.Trt as the covariates. We fit the four models and present the results in Table 8.2.

Table 8.2: Parameter estimates for the epilepsy data.

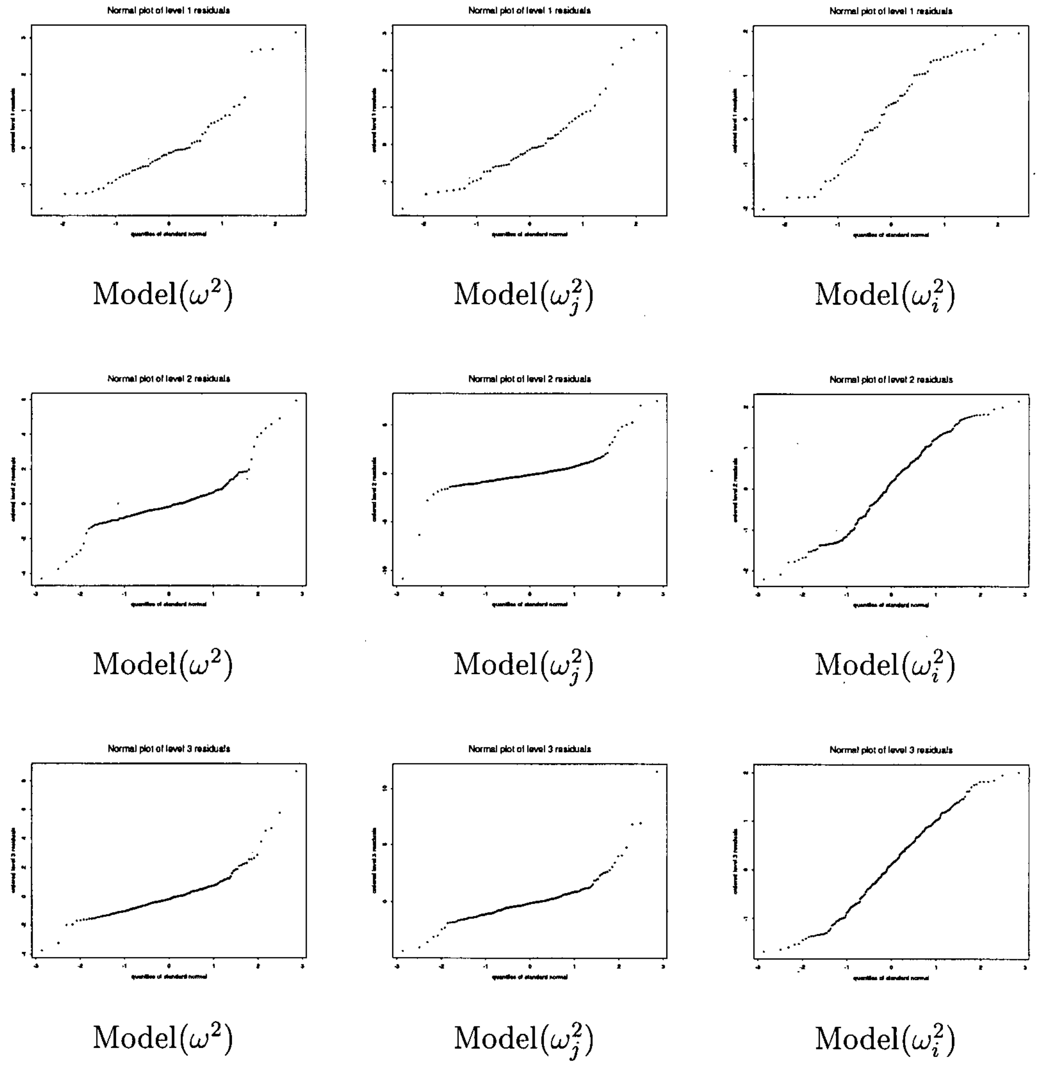
	Model($\omega^2 = 0$)		Model(ω^2)		Model(ω_j^2)		Model(ω_i^2)	
Parameter	est.*	s.e.*	est.	s.e.	est.	s.e.	est.	s.e.
Constant	−1.35	1.22	−1.35	1.22	−1.37	1.16	−1.11	0.72
Base	0.88	0.14	0.88	0.14	0.87	0.13	0.94	0.07
Trt	−0.89	0.41	−0.89	0.42	−0.91	0.40	−0.49	0.28
Base.Trt	0.34	0.21	0.34	0.21	0.35	0.20	0.03	0.13
Age	0.51	0.36	0.51	0.36	0.52	0.34	0.36	0.21
Visit	−0.22	0.10	−0.22	0.22	−0.28	0.23	−0.35	0.18
σ^2	0.28		0.19		0.16		0.007	
ω_1^2			0.36		0.33		0.03	
ω_2^2					0.32		0.08	
ω_3^2					0.67		0.54	
ω_4^2					0.25		12.9	

$*$ est. and s.e. represent estimates and standard errors respectively.

8.1.3 Model checking

We have done model checking through residual analyses. The normal plots for Model(ω^2), Model(ω_j^2) and Model(ω_i^2) are displayed in the first, second and third columns, respectively of Figure 8.2. The plots in the three rows correspond to the level 1, 2 and 3 residuals defined in Section 6.1. As expected, Model(ω_i^2) exhibits the least curvature since there are many more dispersion parameters estimated. Unlike Model(ω_i^2), the dispersion parameter estimators are known to be consistent for the Model(ω^2) and Model(ω_j^2). There is not much difference between the normal plots for Model(ω^2) and Model(ω_j^2).

We plot the standardized residuals against fitted values for the level 2 and level

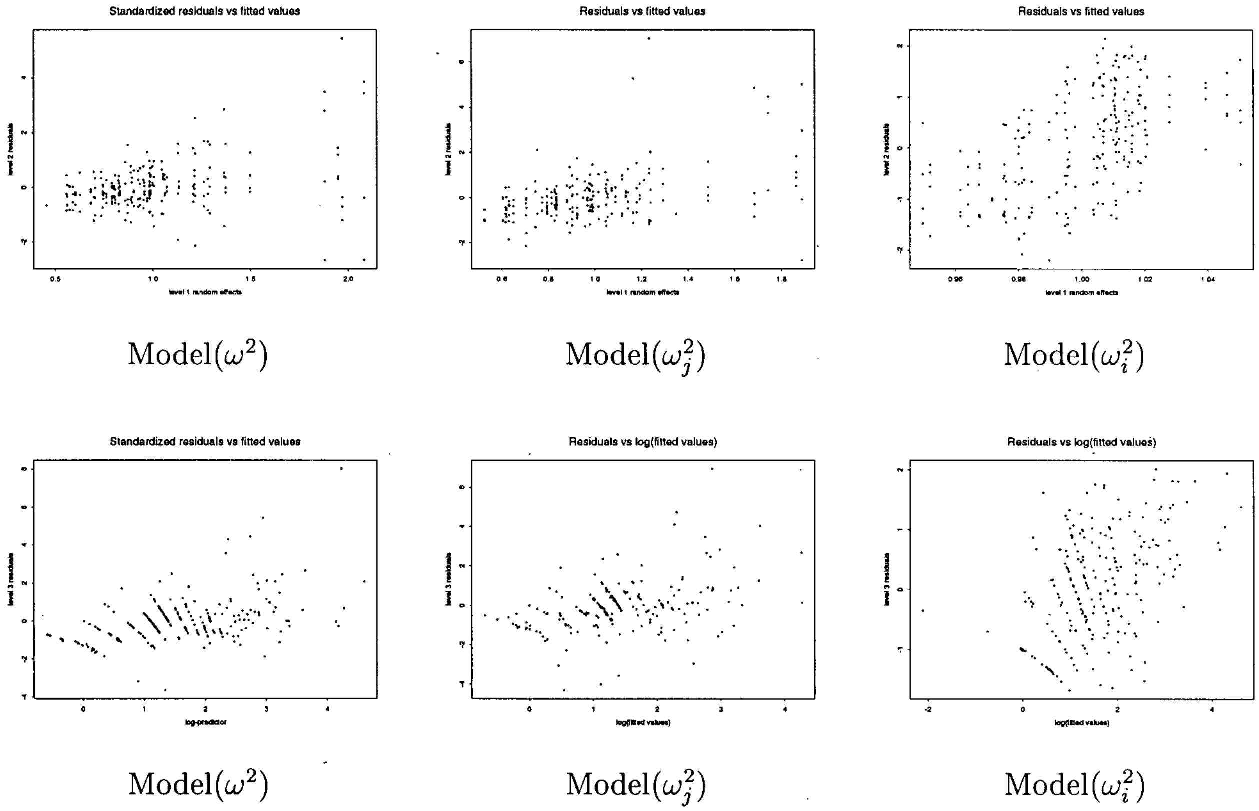


Model(ω^2) Model(ω_j^2) Model(ω_i^2)

Model(ω^2) Model(ω_j^2) Model(ω_i^2)

Model(ω^2) Model(ω_j^2) Model(ω_i^2)

Figure 8.2: Normal plots of level 1, 2 and 3 residuals for epilepsy data.



Model(ω^2) Model(ω_j^2) Model(ω_i^2)

Model(ω^2) Model(ω_j^2) Model(ω_i^2)

Figure 8.3: Scatter plots of level 2 and 3 residuals for epilepsy data.

3 distributions in Figure 8.3. The level 1 residual plot is omitted since the fitted values are fixed at 1. The residual plots for these three models are arranged in the same fashion as the normal plots. The residuals for the level 3 distribution from the three models are approximately around zero, but display distinct lines due to the discreteness of the Poisson distribution. The residuals of Model(ω^2) for the level 2 distribution show a slightly megaphone shape, whereas the residuals of Model(ω_i^2) for the level 2 distribution exhibit an upward trend. The residuals of Model(ω_j^2) for the level 2 distribution show less pattern than the other two.

Based on the residual plots, Model(ω_j^2) appears to fit slightly better than Model(ω^2). But the inferences about regression parameters are not significantly different for these two models. The dispersion parameter estimators are shown to be consistent for these two models as well. On the other hand, the normal plots for Model(ω_i^2) look less curved.

8.1.4 Comparison of different approaches

Thall and Vail (1990) found their model 22 had the best fit among their fitted models. The numerical results from Breslow and Clayton (1993) were reported as similar to those obtained from this model 22. Analyses from the hierarchical generalized linear models and penalized quasi-likelihood are also very similar except for the intercept estimators. The regression parameter estimates and the standard errors obtained from Model($\omega^2 = 0$) and Model(ω^2) are quite similar to those in their counterpart from HGLMs except our standard error estimates are slightly larger. Thus all different approaches yield essentially the same results about the regression parameters although different random effects distributional assumptions were made.

The conclusions based on Model($\omega^2 = 0$), Model(ω^2) and Model(ω_j^2) are similar to those previous studies. The interaction between the treatment and baseline is retained in the models since the approximate p-values for this interaction is just slightly larger than $\alpha = 0.05$. As Thall and Vail indicated, this means the predicted mean seizure rate for the treatment group is either higher or lower than that for the placebo group, accordingly as the baseline count does or does not exceed a critical threshold (minus the ratio of the regression parameter for Trt to the regression coefficient for the interaction term). This seems to suggest that the new drug progabide may be contraindicated for patients with high seizure rates. However they pointed out this suggestion should be treated with caution since it is based on a single dataset and a particular family of models.

In fact, the interaction between the treatment and baseline is statistically insignificant in Model(ω_i^2). Based on the Wald test, the p-value of dropping the interaction term from Model(ω_i^2) is as high as 0.8. That is, the contraindication of the new drug progabide for patients with high seizure rates disappears after accounting properly for the marginal heterogeneity among patients. Dropping this interaction term, the p-value for the treatment effect in this model becomes extremely small. That is, the effectiveness of this new drug progabide is statistically significant.

Thall and Vail (1990) focused their analyses on the regression parameters in the log-linear models for the marginal seizure rates and the covariance parameters in various patterned covariance matrices. They carefully examined the marginal residuals comparing the observed and fitted counts of each subject at each visit, and identified patients 207, 135, 225, 227 and 112 as 'outliers'.

The primary interest of Breslow and Clayton (1993) is systematic identification of patients who had extreme levels, or extreme degrees of change over time, in their seizure rates. They identified patient 135 as having marked improvement over time after an initially high seizure rate, and identified patients 227, 225 and 112 as having the highest overall count levels relative to expectation based on the covariates. They also identified patient 232 as having low and zero counts. However Lee and Nelder were more cautious in claiming outliers and indicated a single outlier, patient 227.

The marginal heterogeneity among patients is substantial, even after accounting for the treatment and baseline variables. In all models except Model(ω_i^2), the dispersion parameters are assumed to remain the same across subjects. That is, the marginal variance only changes with the marginal mean in a systematic way. However, in biological studies, due to the genetic, environmental or other unknown factors, the marginal heterogeneity across subjects often exceeds what could be explained by the mean change. We plot the sample variances s_i^2 against sample means $\bar{y}_i$ across subjects in Figure 8.4 to study the mean-to-variance relationship. The boxplots of the responses for different subjects are also displayed in Figure 8.4 to serve this purpose. As indicated by (7.1), the variance is a monotone function of the mean for Tweedie mixed models with equal dispersion parameters. Thus the two plots in Figure 8.4 should reflect this monotone pattern if the dispersion parameters do not vary across subjects. These two plots do show a kind of monotone pattern, but not very closely.

To account for the irregular marginal heterogeneity, we considered Model(ω_i^2) where different dispersion parameters ω_i^2 are allowed across patients. This model seems to have effectively captured this irregular heterogeneity. The minimum, first quartile, third quartile and the maximum of $\widehat{\omega}_i^2$s are displayed in Table 8.2 labeled as ω_j^2 for $j = 1, 2, 3, 4$. The estimated ω_i^2s range from 0.03 to 12.9 corresponding to patients 213 and 225 respectively. Model(ω_i^2) also seems to be very sensitive in detecting outliers. The scatter plot of $\widehat{\omega}_i^2$s in Figure 8.5 identifies patients 135, 227, 112, 207 and 225 with large $\widehat{\omega}_i^2$. These patients match outliers reported by Thall and Vail (1990) and Breslow and Clayton (1993).

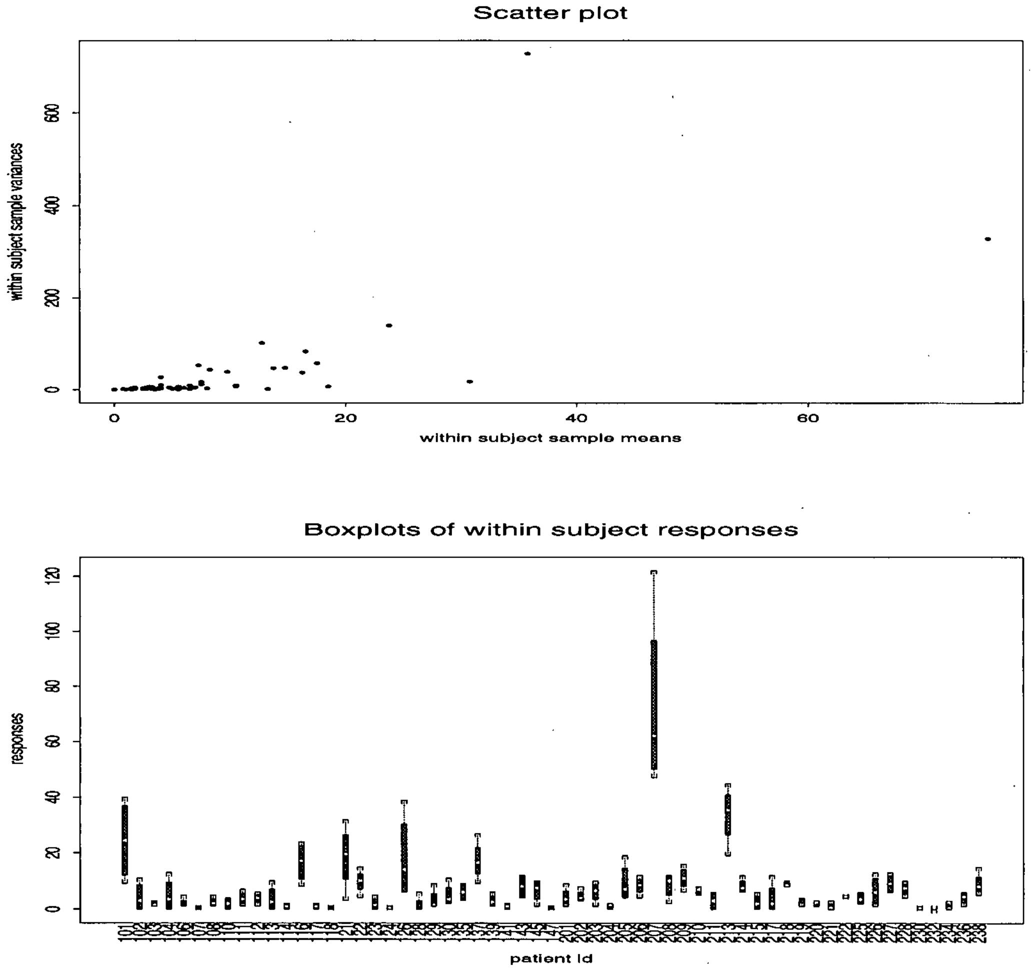


Figure 8.4: Heterogeneity plots for epilepsy data: (a) scatter plot of within subject sample variances versus sample means; (b) boxplots of within subject responses for all patients.

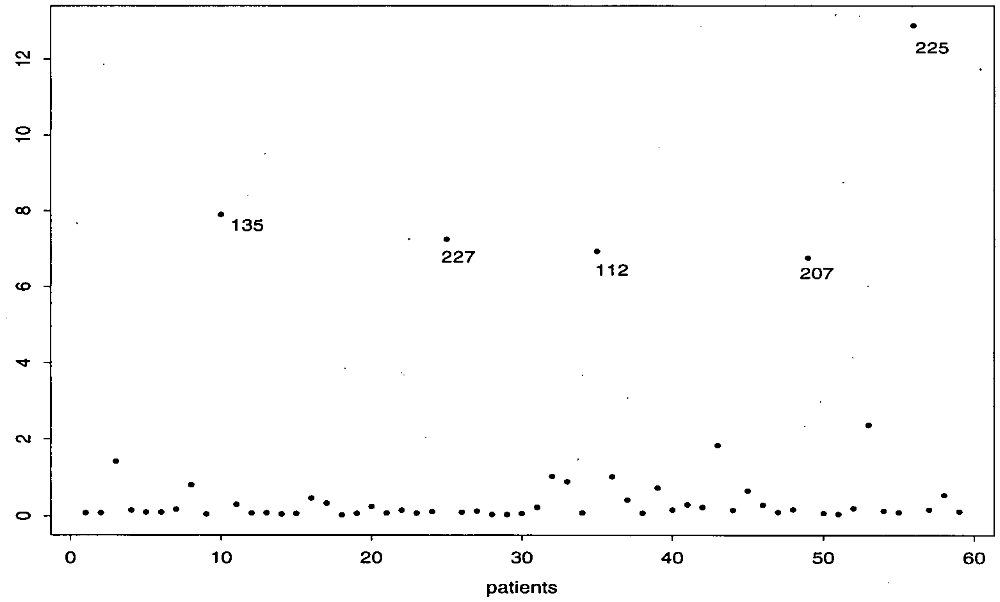


Figure 8.5: Heterogeneity plot for epilepsy data.

8.1.5 Computational aspects

With regards to computing, we ran Splus on a Sparc ultra machine. All parameter estimates for the Model(ω^2) updated in monotone directions after five iterations. Those included regression and dispersion parameters as well as standard errors for regression parameters. All estimates at the fifth iteration are close to those in Table 8.2. The same two decimal point precision for all parameter estimates was achieved at the eighteenth iteration. At this iteration, the absolute value of the standardized components of $\boldsymbol{\psi}(\boldsymbol{\beta})$ were less than 0.001. The absolute distance between regression parameter estimates and their updates were less than 0.00003. We stopped the program after sixty iterations. The maximum of the absolute values of the standardized components of $\boldsymbol{\psi}(\boldsymbol{\beta})$ and the absolute distance between regression parameter estimates and their updates were less than $2E - 12$ and $3E - 13$, respectively. Each iteration took about 10 seconds of user time and 0.01 seconds of system time on the Sparc ultra machine. Twenty iterations took about four minutes.

As to the other two level random effects models, the computing time for Model(ω_j^2) was slightly longer than that for Model(ω^2). For Model(ω_i^2), all parameter estimates stabilized after eight iterations, but they converged much more slowly although still in a monotone direction. Those estimates approached the values reported in Table 8.2 after about fifty iterations. Each iteration took about 15 seconds on the same machine.

8.2 Binomial data

We illustrate the orthodox BLUP approach to binomial data with analysis of the seed germination data described in Section 2.4.2.

8.2.1 Paired Poisson-Tweedie models

We analyze the seed germination data based on the paired Poisson-Tweedie models. The model notation is similar to that for the epilepsy data. The estimates and corresponding standard errors are displayed in Table 8.3.

Table 8.3: Parameter estimates for the seed germination data based on paired Poisson mixed models.

	Paired GLM		Model($\sigma^2 = 0$)	Model(ω^2)	Model(ω_j^2)
Parameter	est.*	s.e.*	s.e.	s.e.	s.e.
Constant	-0.558	0.126	0.349	0.181	0.179
Seed (2)	0.146	0.223	0.512	0.289	0.287
Extract (2)	1.318	0.177	0.475	0.250	0.248
Interaction	-0.778	0.306	0.708	0.398	0.395
			est.	est.	est.
σ^2				0.229	0.226
ω_1^2			0.265	0.042	0.055
ω_2^2					0.026

$*$ est. and s.e. represent estimates and standard errors respectively.

As indicated in Chapter 7, the asymmetric binomial mixed models can be considered based on the paired two level Poisson-Tweedie models with the second level random effects having unequal dispersion parameters ω_j^2s between the two groups $j = 1, 2$. We started our modeling with this asymmetric model, Model(ω_j^2), but there is little evidence for this asymmetry as the difference between $\hat{\omega}_1^2$ and $\hat{\omega}_2^2$ is small relative to $\hat{\sigma}^2$. Thus we proceed to fit the symmetric paired Poisson mixed model with one and two levels of random effects, namely Model($\sigma^2 = 0$) and Model(ω^2).

The estimates and standard errors presented in the first and second columns were obtained from the standard paired Poisson generalized linear model, namely, the paired Poisson generalized linear models without random effects. This standard paired Poisson generalized linear model corresponds to the standard binomial generalized linear models. The regression parameter estimates and their corresponding standard errors obtained from these two models are exactly the same.

The regression parameter estimates obtained from all above paired Poisson-Tweedie models with various distributional assumptions for random effects are almost identical to those obtained from the standard binomial generalized linear models. Breslow and Clayton (1993) also reported similar results from their simulation study based on some simulated data sets generated from a balanced binomial mixed model design. Their marginal quasi-likelihood (MQL) approach to binomial mixed models led to regression coefficient estimates identical to those obtained from standard logistic regression; however neither the MQL nor PQL regression parameter estimates for this seed germination data were identical to those obtained from standard logistic regression.

On the other hand, the standard error estimates obtained from the paired Poisson mixed models are much larger than those obtained from the standard paired Poisson regression models. Furthermore, the standard error estimates clearly vary from one random effects model to another. That is, the standard error estimates depend heavily on the random effects model assumptions. This is different from the results for the epilepsy data.

8.2.2 Model checking

We performed residual analysis based on the paired Poisson-Tweedie models. Since the procedure is exactly the same as those presented in last section, we present the residual plots for Model(ω^2) only. The normal plots of the level 1, 2 and 3 residuals for Model(ω^2) are displayed in Figure 8.6 and the scatter plots of level 2 and 3 residuals are displayed in Figure 8.7. Normal plots at all three levels exhibit moderate curvature, whereas the scatter plots of level 2 and 3 residuals against log-fitted values show moderate upward trends. Some patterns usually exist for a small data set with discrete responses.

8.2.3 Comparison of different approaches

Breslow and Clayton (1993) and Lee and Nelder (1996) also analyzed the seed germination data. The estimates and the corresponding standard errors are presented in Table 8.4.

Table 8.4: Parameter estimates for the seed data based on binomial models.

	GLM		PQL		HGLM	
Parameter	est.*	s.e.*	est.	s.e.	est.	s.e.
Constant	−0.558	0.126	−0.542	0.190	−0.543	0.187
Seed (2)	0.146	0.223	0.077	0.308	0.080	0.303
Extract (2)	1.318	0.177	1.339	0.270	1.337	0.265
Interaction	−0.778	0.306	−0.825	0.430	−0.822	0.423
σ^2			0.098		0.045	

$*$ est. and s.e. represent estimates and standard errors respectively.

The standard errors obtained from the paired Poisson-Tweedie models with one

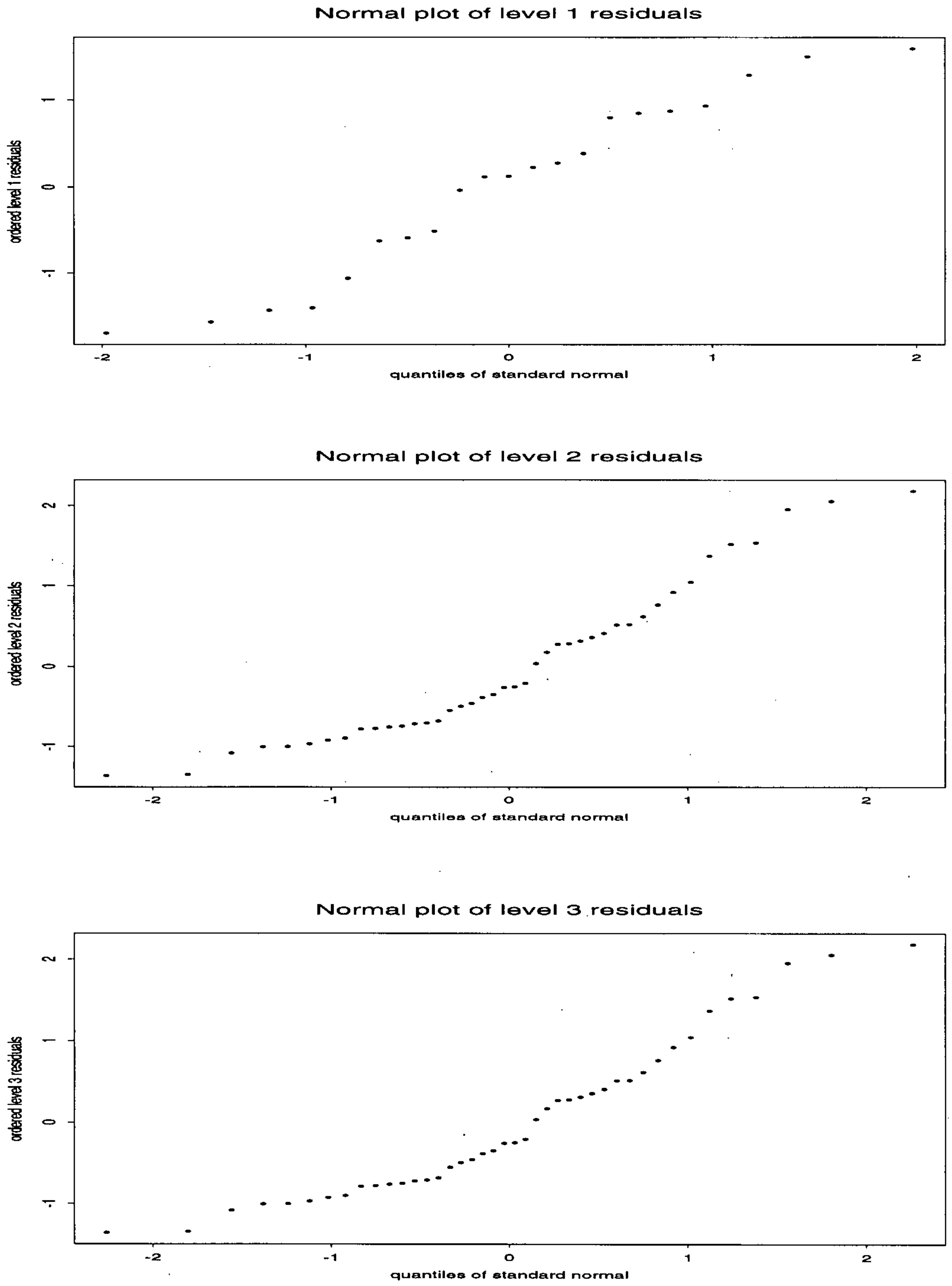


Figure 8.6: Normal plots of residuals for seed germination data.

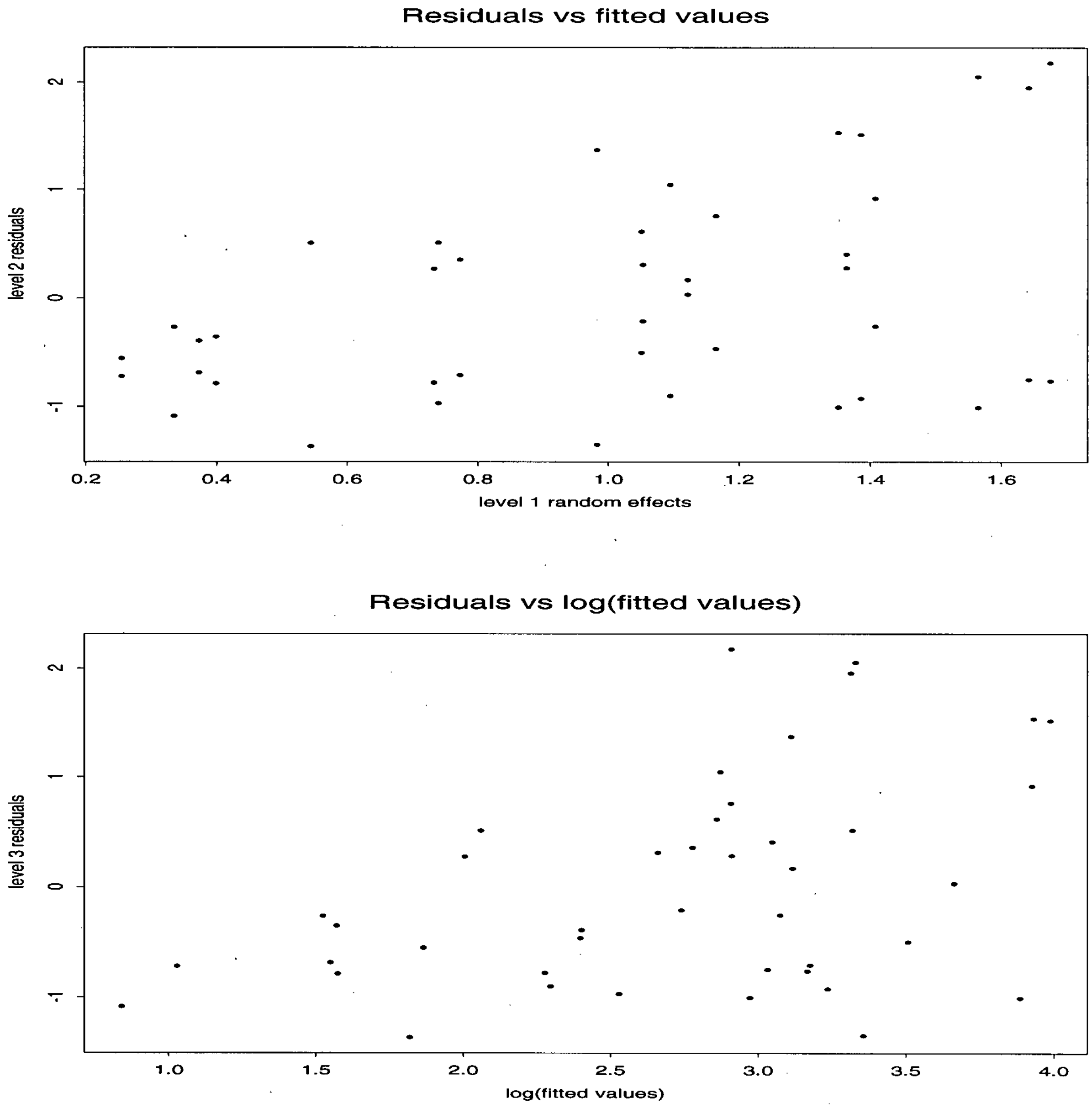


Figure 8.7: Scatter plots of residuals for seed germination data.

level of random effects are almost twice those obtained from penalized quasi-likelihood and maximum hierarchical likelihood approaches, whereas the symmetric paired Poisson with two levels of random effects model gives similar, but slightly smaller standard error estimates than those obtained from penalized quasi-likelihood and maximum hierarchical likelihood approaches. The Model($\sigma^2 = 0$) corresponds to binomial-beta model (Lee and Nelder 1996); however the former gives quite different standard error estimates than the latter. This difference is not easy to explain. Further investigation is needed.

8.3 Continuous data

We illustrate our approach to clustered continuous data with an analysis of the cake baking data described in Section 2.4.3. Cochran and Cox (1957) analyzed the cake baking data using analysis of variance. The normal plot of residuals for the ANOVA model in Figure 8.8 shows a curvature. Firth and Harris (1991) re-analyzed the cake baking data using multiplicative models and they found strong support for the hypothesis of a constant coefficient of variation. Thus we analyze the cake baking data based on the conventional gamma mixed models, Model(ω^2), where the two levels of random effects are at the batch and observation levels. Temperature and recipe are taken as factors. We screen these factors based on the Wald test. We found that the recipe effect and any interaction between temperature and recipe appear negligible, but the temperature effect is highly statistically significant. In fact, Figure 8.9 shows that there is an increasing trend in breaking angles as the temperature increases.

Cochran and Cox (1957) and Firth and Harris (1991) reached similar results to ours except they found the recipe effect is significant at the statistical significance level of 0.05. The normal plots of level 1, 2 and 3 residuals in Figure 8.10 do not reveal serious patterns except some curvature for the level 1 residuals. The scatter plot for level 2 residuals in Figure 8.11 does not exhibit any serious pattern, whereas the scatter plot for level 3 residuals shows an upward trend. Noting that the dispersion parameter ρ^2 is nearly zero, one would expect that the effect of level 3 residuals on the model be small. To have a systematic comparison, we display estimates and corresponding standard errors for ANOVA, gamma and gamma mixed models without interaction in Table 8.5. The parameter estimates from ANOVA model are quite different from those obtained using standard generalized linear model and gamma mixed model; however the regression parameters for the latter two models are again almost identical for this balanced design. After accounting for random effects, the recipe factor becomes less statistically significant, whereas the temperature factor becomes more statistically significant.

Table 8.5: Parameter estimates for cake baking data.

	ANOVA		GLM		Model(ω^2)	
Parameters	est.	s.e.	est.	s.e.	est.	s.e.
Constan	32.122	0.474	3.466	0.015	3.466	0.030
R1	0.989	0.821	−0.023	0.018	−0.023	0.037
R2	0.819	0.474	−0.008	0.010	−0.008	0.021
T1	0.598	0.335	0.034	0.026	0.034	0.014
T2	1.092	0.260	0.028	0.015	0.028	0.008
T3	0.647	0.212	0.020	0.010	0.020	0.006
T4	−0.739	0.581	0.033	0.008	0.033	0.005
T5	−0.261	0.335	0.020	0.007	0.020	0.004
σ^2					0.038	
ω^2					0.019	
ρ^2					0.00025	

R and T represent recipe and temperature factors.

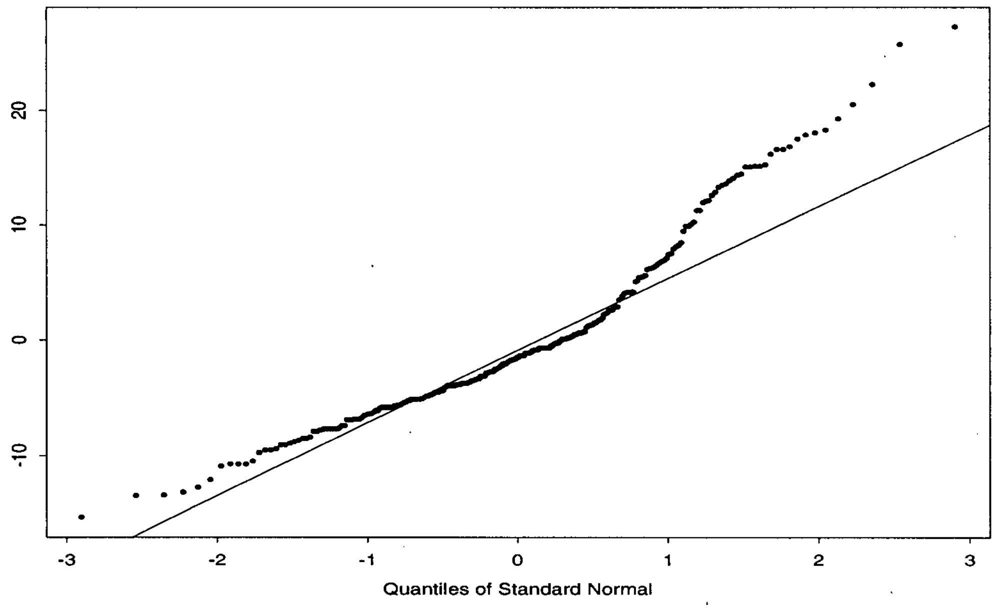


Figure 8.8: Normal plot of ANOVA model residuals for cake baking data.

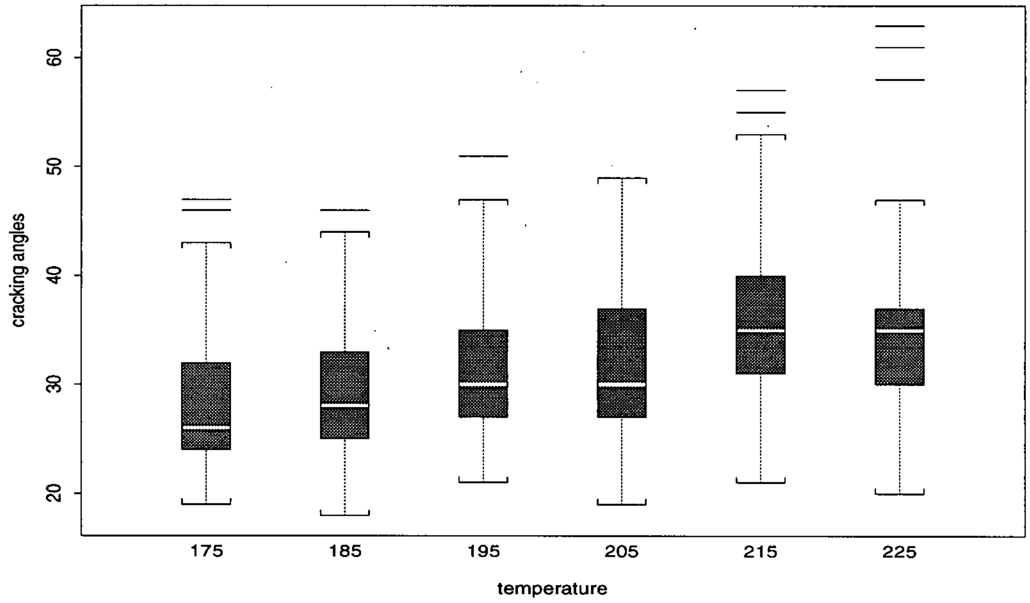
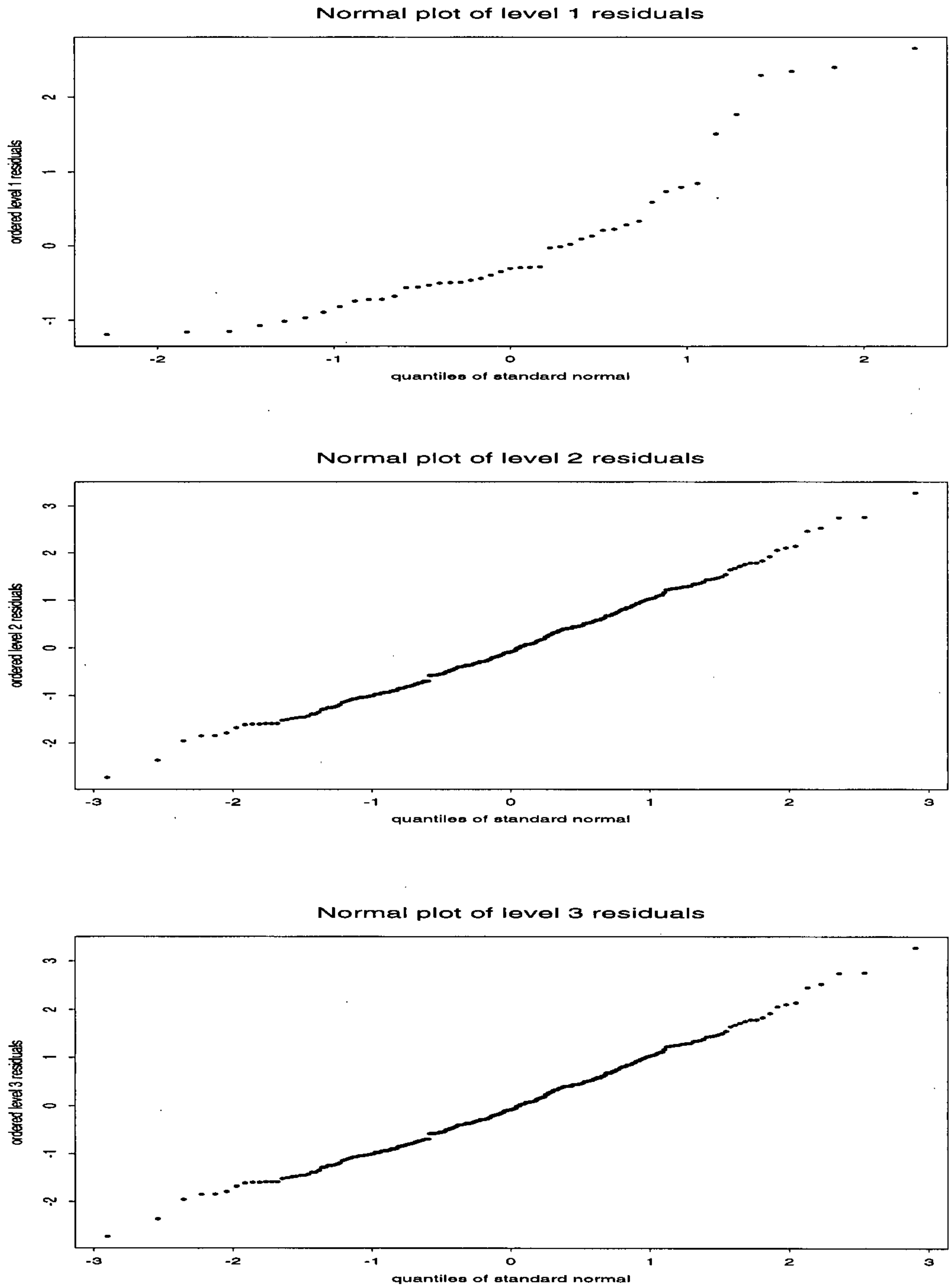


Figure 8.9: Boxplots of cake baking data by temperature.


Normal plot of level 1 residuals

Normal plot of level 2 residuals

Normal plot of level 3 residuals

Figure 8.10: Normal plots of residuals for cake baking data.

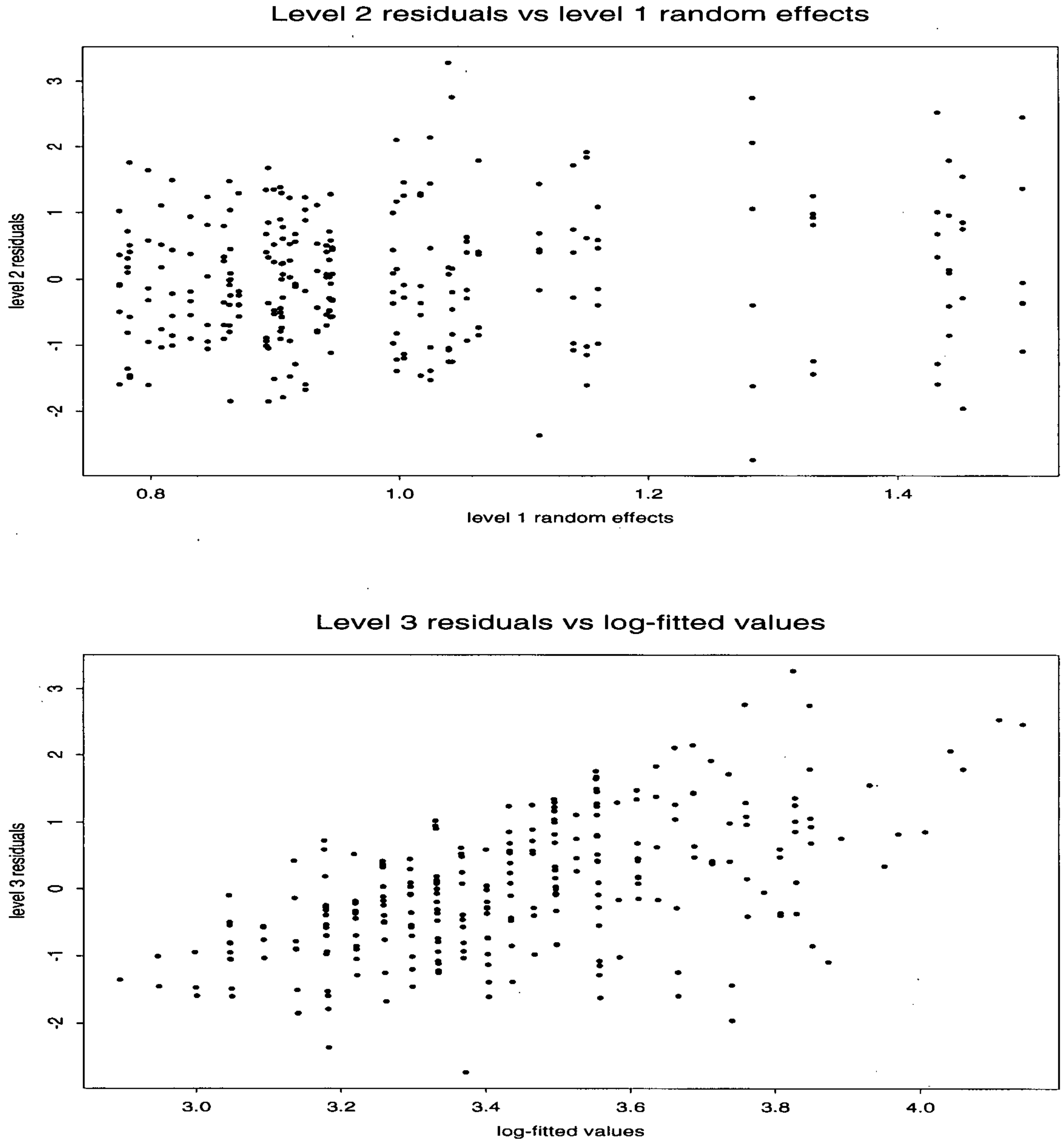


Figure 8.11: Scatter plots of residuals for cake baking data.

Chapter 9

Simulation

In this chapter, we evaluate the performance of the orthodox BLUP approach based on a simulation study. We will focus on Poisson mixed models since these models play an important role in the analysis of discrete data. Illustrative examples of applications of Poisson mixed models to analyze count data and binomial data were presented in the previous chapter. A possible connection between Poisson mixed models and the random effects Cox proportional hazard models for censored data will be discussed in next chapter.

9.1 Results over 100 simulations

To investigate the performance of the orthodox BLUP approach under realistic conditions, we 'replicate' the epilepsy data set 100 times via simulation using Poisson-gamma model. To generate the random effects and responses via simulation from this model, we take the covariates Constant, Base, Trt, Age, Visit and Base.Trt of epilepsy data as the covariates for simulation. The corresponding regression and dispersion parameter estimates for the epilepsy data are taken as true model parameters. There are listed in Table 9.1 as 'true values' β_0, β_1, $\ldots$, β_5, σ^2 and ω^2, respectively. Each of these 100 data sets is then simulated through the following three steps:

- Step 1:

  Generate 59 samples from Gamma$(1, \sigma^2)$, denoted by $u_1^{(k)}, \ldots, u_{59}^{(k)}$;

- Step 2:

  Generate 4 samples from Gamma$(u_i^{(k)}, \omega^2 u_i^{(k)})$ for each i, denoted by $u_{ij}^{(k)}$, $j = 1, 2, 3, 4$ and $i = 1, \ldots, 59$;

- Step 3:

  Generate a sample from Poisson $\left(u_{ij} \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta})\right)$ for each (i, j), denoted by $y_{ij}^{(k)}$, $j = 1, 2, 3, 4$ and $i = 1, \ldots, 59$, where the design matrix $\mathbf{X}$ is taken exactly the same as that of epilepsy data.

The regression and dispersion parameter were fixed at the corresponding estimates for epilepsy data, that is, $\boldsymbol{\beta} = (1.30, 0.88, -0.88, 0.50, -0.23, 0.34)$, $\sigma^2 = 0.24$ and $\omega^2 = 0.44$, respectively. The 100 data sets are obtained by repeating this procedure for $k = 1, \ldots, 100$.

We analyzed each of these 100 data sets based on the standard Poisson regression model (GLM) and Poisson-Tweedie models with and without degree of freedom correction for small samples, denoted by BLUP.c and BLUP, respectively. The summaries are presented in the next subsection.

9.1.1 Summary statistics

The averages of regression and dispersion parameter estimates over 100 simulations are displayed in Table 9.1.

Table 9.1: Averages of parameter estimates over 100 simulations.

	β_0	β_1	β_2	β_3	β_4	β_5	σ^2	ω^2
true value	-1.30	0.88	-0.88	0.50	-0.23	0.34	0.24	0.44
GLM	-1.37	0.88	-0.78	0.51	-0.26	0.29		
BLUP	-1.36	0.87	-0.89	0.51	-0.25	0.34	0.17	0.50
BLUP.c	-1.31	0.87	-0.88	0.50	-0.25	0.34	0.23	0.61

The regression parameters are reasonably estimated by both GLM and orthodox BLUP approaches with or without small-sample correction. On the other hand, the dispersion parameters σ^2 and ω^2 are underestimated and overestimated respectively. This phenomenon is different from simulation studies reported by Breslow and Clayton (1993) where the dispersion parameters were always underestimated. However the possibility of overestimation was previously predicted by Engel and Keen (1996). The small-sample correction alleviates the negative bias of the estimate of σ^2, but worsens the positive bias of the estimate of ω^2.

The standard errors of the estimates over 100 simulations and the averages of the 100 estimated standard errors are termed as simulated and estimated standard errors, respectively (Lin and Breslow 1996b). Table 9.2 displays the simulated and estimated standard errors. The expressions of the simulated and estimated standard errors for β_1 are as follows:

$$\text{simulated s.e.}(\hat{\beta}_1) = \sqrt{\frac{1}{100}\sum_{k=1}^{100}(\hat{\beta}_1^{(k)} - \frac{1}{100}\sum_{k=1}^{100}\hat{\beta}_1^{(k)})^2},$$

and

$$\text{estimated s.e.}(\hat{\beta}_1) = \frac{1}{100}\sum_{k=1}^{100} \text{s.e.}(\hat{\beta}_1^{(k)}).$$

Table 9.2: Simulated and estimated standard errors.

	s.e.	β_0	β_1	β_2	β_3	β_4	β_5	σ^2	ω^2
GLM	Sim.	1.65	0.17	0.60	0.48	0.28	0.31		
	Est.	0.41	0.04	0.15	0.12	0.10	0.06		
BLUP	Sim.	1.42	0.15	0.43	0.41	0.22	0.23	0.07	0.07
	Est.	1.25	0.14	0.42	0.36	0.24	0.22		
BLUP.c	Sim.	1.42	0.15	0.43	0.41	0.22	0.23	0.07	0.09
	Est.	1.39	0.16	0.47	0.41	0.26	0.24		

∗ Simulated s.e.: s.e. of estimates over 100 simulations.

∗∗ Estimated s.e.: average of 100 estimated s.e.s.

Clearly the GLM standard error estimates are seriously negatively biased. In contrast, the orthodox BLUP approach with small-sample correction slightly overestimates standard errors for regression parameters, except for the intercept β_0. On the other hand, the orthodox BLUP approach without small-sample correction slightly underestimates the standard errors for the regression parameters, severely so for the intercept.

Table 9.3: Mean squared errors over 100 simulations.

	β_0	β_1	β_2	β_3	β_4	β_5	σ^2	ω^2
GLM	2.73	0.03	0.36	0.23	0.08	0.10		
BLUP	2.01	0.02	0.18	0.17	0.05	0.05	0.01	0.01
BLUP.c	2.01	0.02	0.19	0.17	0.05	0.05	0.01	0.04

To evaluate the overall performance of the estimates over 100 simulations, we display the mean squared errors (MSEs) in Table 9.3. As expected, the mean squared errors of regression parameter estimates based on standard Poisson model are larger than those based on the orthodox BLUP approaches. According to mean squared errors, the orthodox BLUP approach without small-sample correction performed slightly better than the orthodox BLUP approach with small-sample correction. All three approaches gave exceptionally large mean squared errors of the intercept β_0.

9.1.2 Confidence and prediction intervals

Based on the simulated and estimated standard errors, we constructed simulated and estimated 95% confidence intervals for the parameters and 95% prediction intervals for the random effects (u_1 and u_{11}) assuming normality of the estimators and predictors. We present in Table 9.4 the counts of the number of times the true values are covered by 95% confidence intervals or prediction intervals. The kth confidence intervals for β_1 are defined as follows:

$$\text{simulated C.I. for } \beta_1 = \left(\hat{\beta}_1^{(k)} - 1.96 \text{ simulated s.e.}(\hat{\beta}_1),\ \hat{\beta}_1^{(k)} + 1.96 \text{ simulated s.e.}(\hat{\beta}_1)\right),$$

and

$$\text{estimated C.I. for } \beta_1 = \left(\hat{\beta}_1^{(k)} - 1.96 \text{ estimated s.e.}(\hat{\beta}_1^{(k)}),\ \hat{\beta}_1^{(k)} - 1.96 \text{ estimated s.e.}(\hat{\beta}_1^{(k)})\right).$$

As expected, the GLM estimated 95% confidence intervals performed very poorly. On the other hand, the estimated 95% confidence intervals based on the orthodox BLUP approaches performed reasonably well, especially those with small-sample correction. The orthodox BLUP approaches also gave reasonable 95% prediction intervals, but with lower coverage counts for u_1 and higher coverage counts for u_{11}.

Table 9.4: Coverage counts of 95% confidence and prediction intervals.

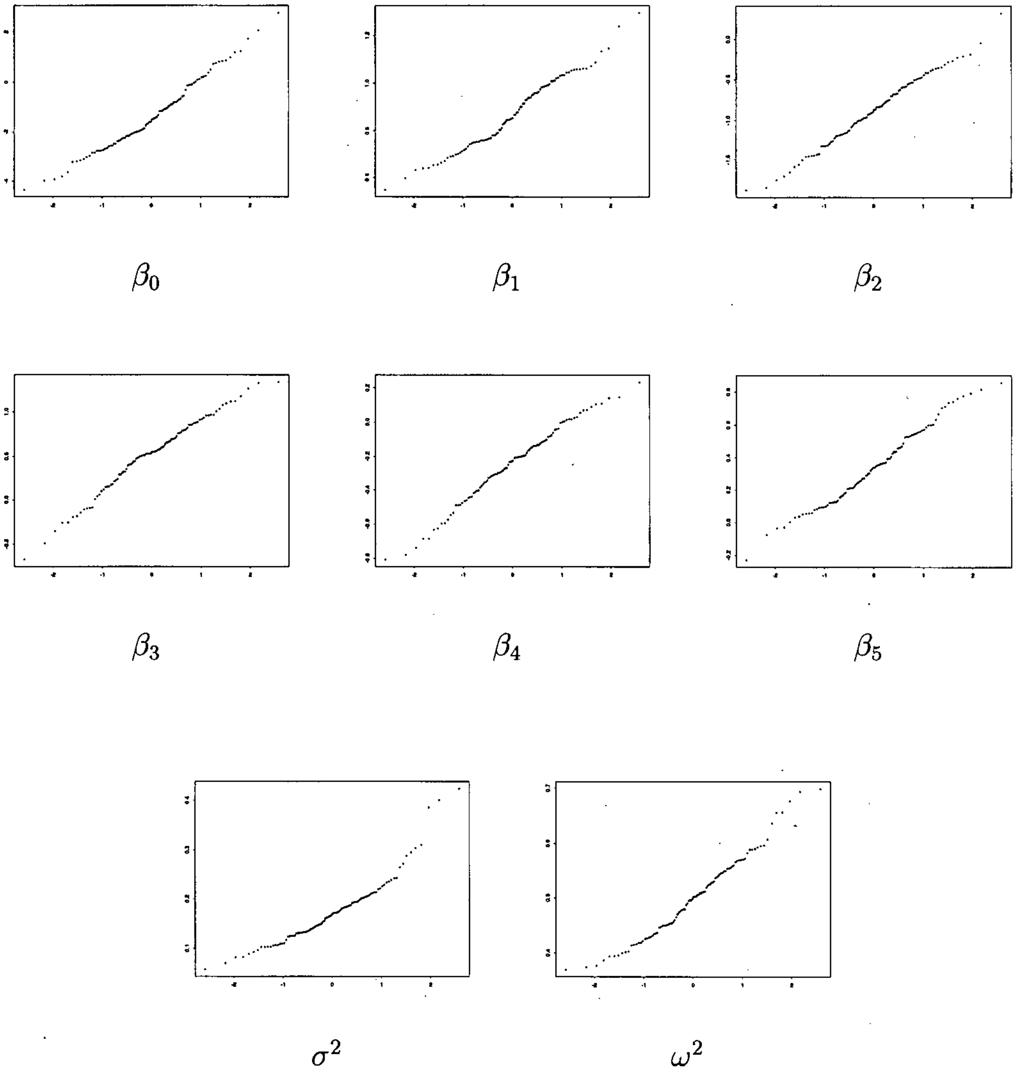
		β_0	β_1	β_2	β_3	β_4	β_5	u_1	u_{11}
GLM	Sim.	95	96	94	94	94	93		
	Est.	35	28	38	35	44	31		
BLUP	Sim.	96	97	96	95	94	96	88	100
	Est.	90	96	94	91	97	92	84	98
BLUP.c	Sim.	96	97	94	95	94	97	92	100
	Est.	95	96	95	94	98	97	91	100

9.1.3 Normality of parameter estimates and random effects

To check the normality of parameter estimators, we did a normal plot for each of regression and dispersion parameter estimates over the 100 simulations. The results are displayed in Figure 9.1. The departures from normality do not seem serious even for the dispersion parameters.

To assess the performance of the orthodox BLUP predictors, we plotted the simulated random effects versus their orthodox BLUP predictors over the 100 simulations. In particular, we plotted $u_1^{(k)}$ versus $\hat{u}_1^{(k)}$ and $u_{11}^{(k)}$ versus $\hat{u}_{11}^{(k)}$ over $k = 1, \ldots, 100$ in the upper rows of Figure 9.2. Perfect prediction would lead to diagonal lines in these two plots. The prediction for the second level of random effects seems to be better than that for the first level random effects. Normal plots of the orthodox BLUP predictors of random effects u_1 and u_{11} over 100 simulations are displayed in the lower row of Figure 9.2. The clear curvatures in these normal plots imply that random effects predictors are not well approximated by normal distributions.

The numerical results and plots from this simulation study show that the param-



β_0

β_1

β_2

β_3

β_4

β_5

σ^2

ω^2

Figure 9.1: Normal plots of parameters over 100 simulations: x and y axes are quantiles of standard normal distribution and ordered parameter estimates over 100 simulations, respectively.

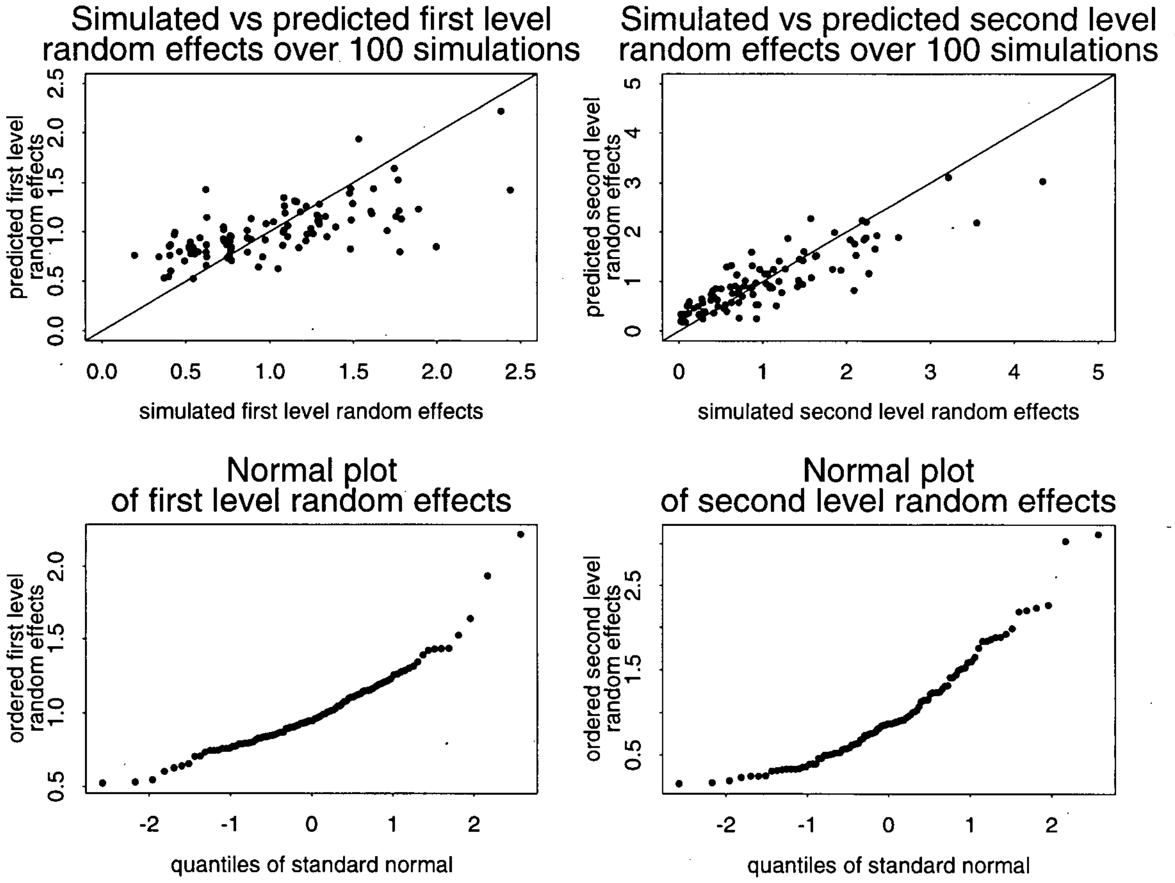


Figure 9.2: Plots for random effects predictors $\hat{u}_1$ and $\hat{u}_{11}$ over 100 simulations.

eters are reasonably well estimated through orthodox BLUP approach.

9.2 Residual analysis

In Chapter 8, we did model checking through residual plots. Serious irregularities exhibited in residual plots may indicate serious violation of model assumptions. To help to set up standards for judging irregularity exhibited by residual plots, we will study the behavior of residuals through simulation in this section.

In this section, we make comparison and contrast of residual plots for the epilepsy data and simulated data. The normal plots for the level 1, 2 and 3 residuals for the epilepsy data and simulated data are displayed at row 1, 2, and 3 of Figure 9.3 where the normal plots for the epilepsy data and simulated data are on the left and right hand side, respectively. The histograms and scatter plots for the epilepsy data and simulated data are displayed in Figure 9.4 and 9.5, respectively, in the same fashion as the normal plots. The simulation shows that the residual plots may exhibit some patterns even if the data were generated from a Poisson-gamma model. That is, moderate patterns in residual plots may not indicate violation of the model assumptions.

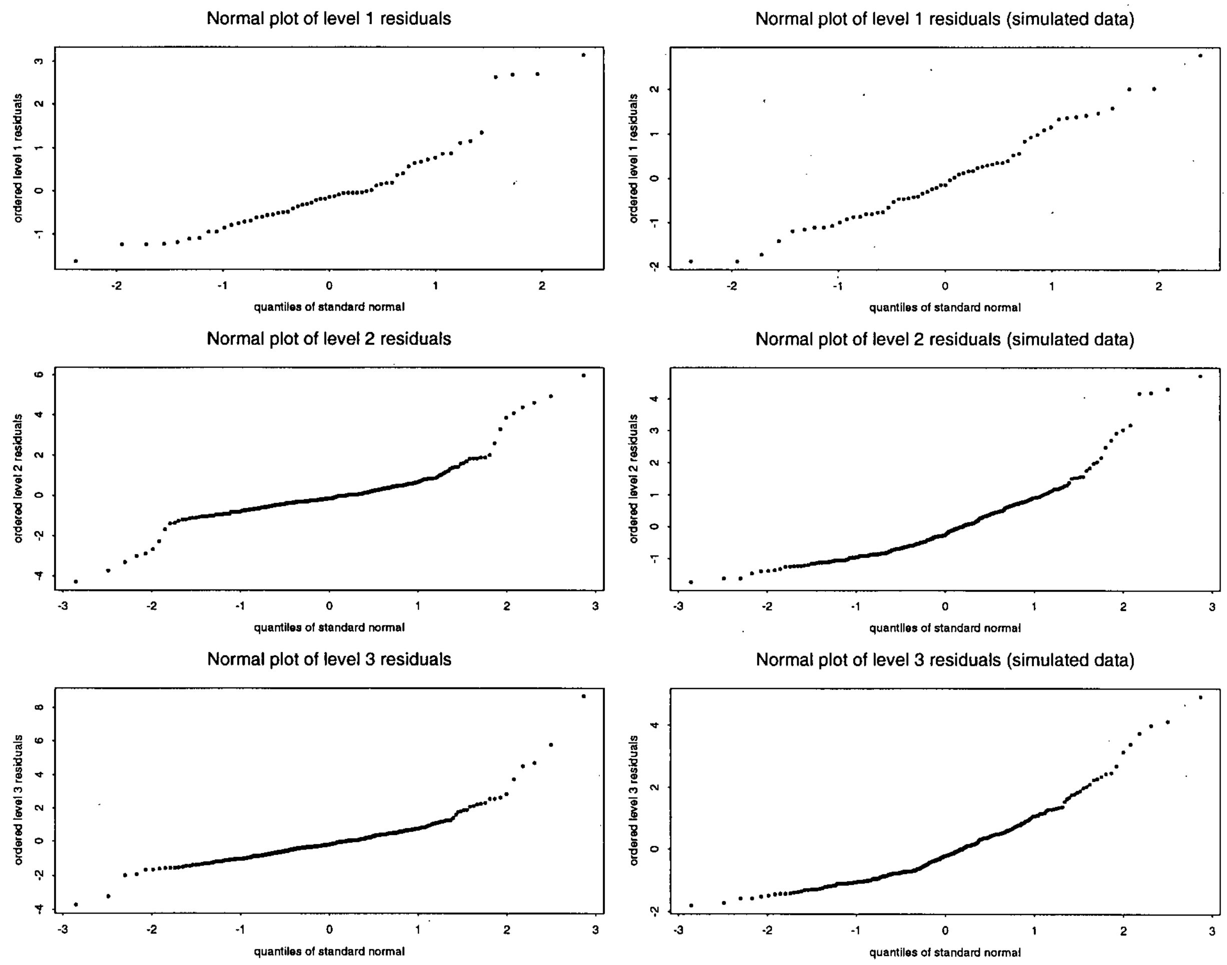
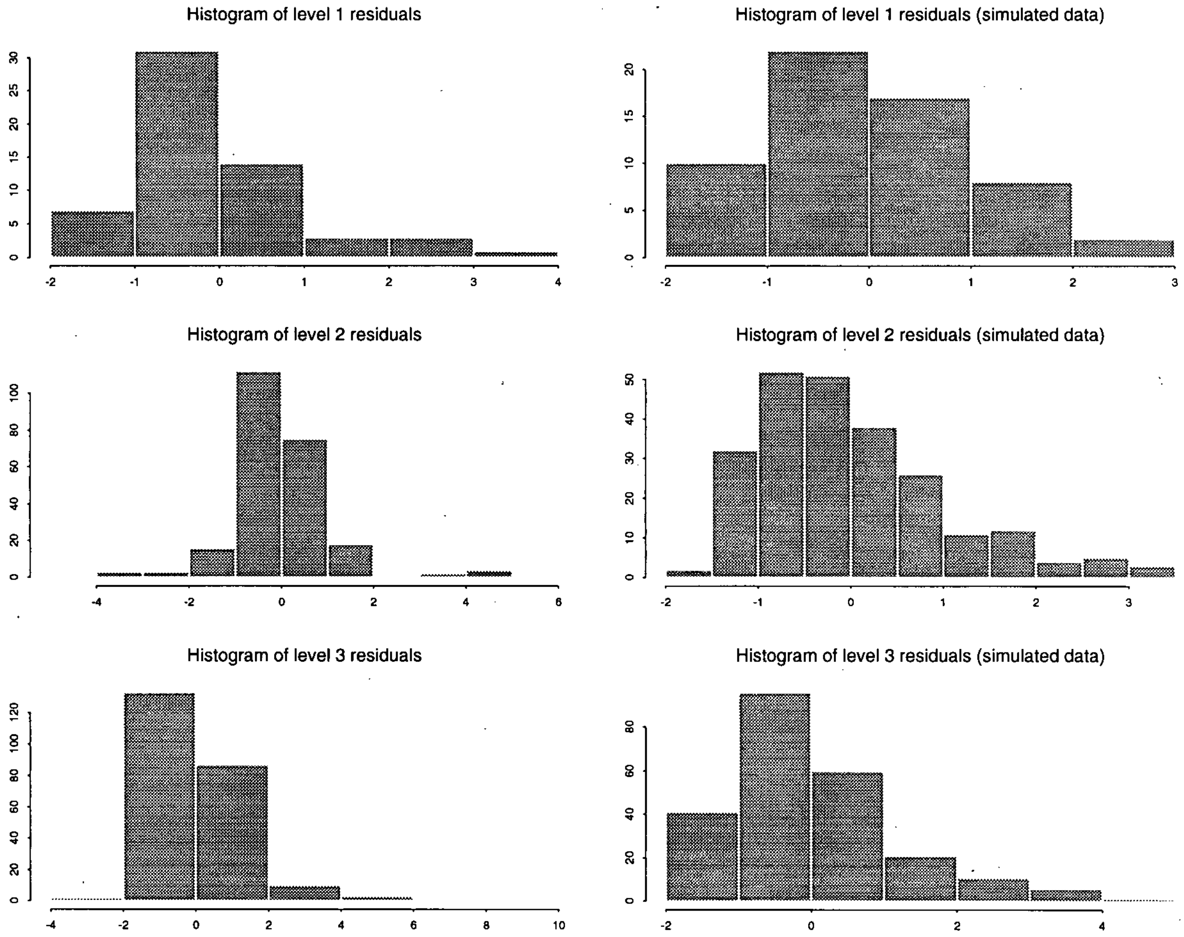


Figure 9.3: Comparison of normal plots of level 1, 2 and 3 residuals for epilepsy data and simulated data.


Histogram of level 1 residuals

Histogram of level 1 residuals (simulated data)

Histogram of level 2 residuals

Histogram of level 2 residuals (simulated data)

Histogram of level 3 residuals

Histogram of level 3 residuals (simulated data)

Figure 9.4: Comparison of histograms of level 1, 2 and 3 residuals for epilepsy data and simulated data.

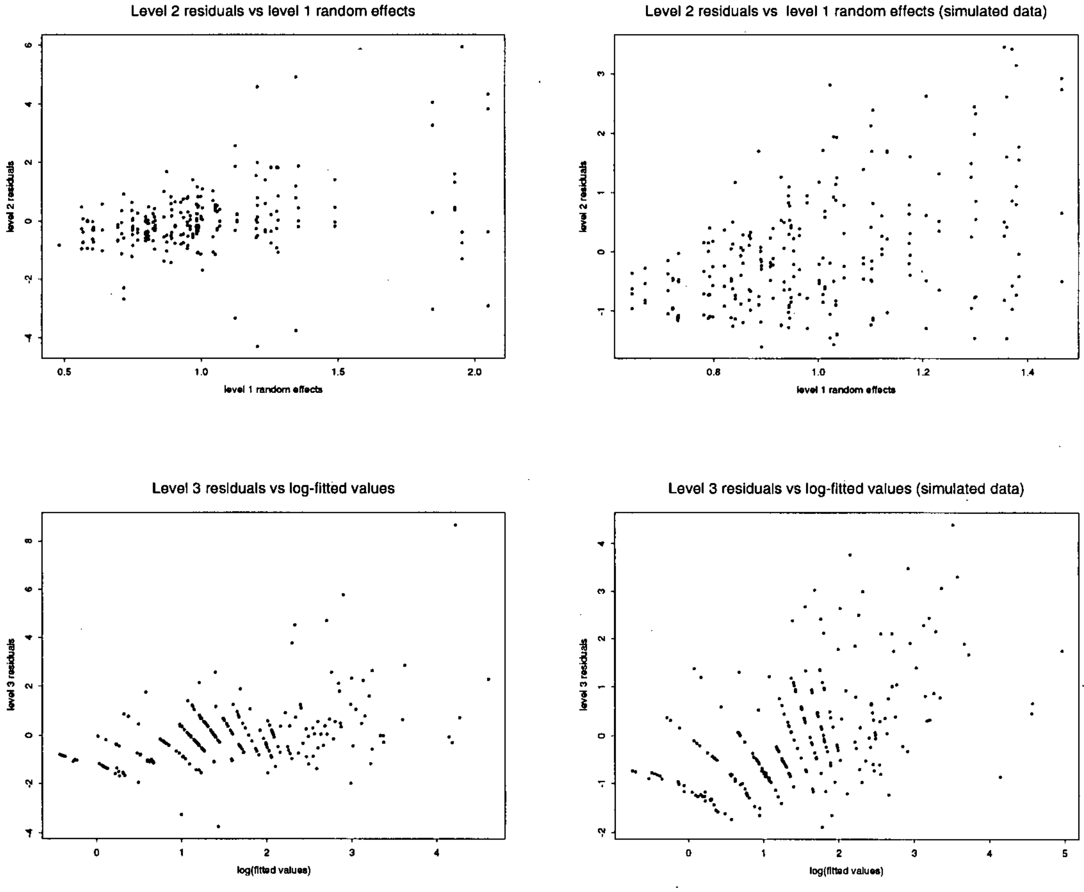


Figure 9.5: Comparison of scatter plots of level 2 and 3 residuals for epilepsy data and simulated data.

Chapter 10

Discussion

10.1 Conclusion

In this thesis we have introduced a new class of generalized linear mixed models, Tweedie mixed models, and adopted the orthodox BLUP in the fitting algorithm. This approach has several advantages.

1. Tweedie mixed models take both the distributional shape and intra-dependence of clustered data into account. The resulting variance component decomposition of the structure of the covariance matrix of responses is not only very interpretable, but also makes the model fitting much simpler than modal predictor approaches; all expressions for the estimating equations were explicitly derived based on (3.13) so there is no need to invert large matrices.

2. The orthodox approach provides us a common method for computing parameter estimates and random effects predictors for all Tweedie mixed models; therefore we can study Tweedie mixed models as a single class, rather than as a collection of

unrelated different models. Furthermore, the conventional Tweedie mixed models are shown to have close connection with well-known Poisson-gamma, binomial-beta and Poisson-lognormal models. The semiparametric interpretation of the conventional Tweedie mixed models allows specification only the first and second moments of the unobserved random effects.

3. The orthodox BLUP for random effects in (4.1), estimated score function for regression parameter in Theorem (5.1) and adjusted Pearson estimators for dispersion parameters in (5.10), (5.15) and (5.17) show that our orthodox BLUP approach relies on only the first and second moment structure of $(\mathbf{U}, \mathbf{Y})$. It is important to note that the orthodox BLUP approach requires specification only of the first and second moments of the random effects model. Tweedie mixed models play an interpretable role within this larger family of models just as exponential dispersion models do within generalized linear models.

4. The asymptotic justification of orthodox BLUP approach does not rely on any kind of normality of random effects. In contrast, the asymptotic justifications of penalized quasi-likelihood and maximum hierarchical likelihood approaches largely rely on the approximate normality for random effects or 'the right transformed normality for random effects on the right scale' , respectively. It is important to note that our asymptotic variance of regression parameter estimator is not affected by the variability of estimating dispersion parameters.

Our orthodox BLUP approach leads to the same estimating equation for the regression parameters as that obtained by the generalized estimating equation approach. This estimating equation was also reached via the quasi-likelihood function

(McCullagh 1983) and multilevel modelling (Goldstein 1995) approaches. All these approaches except the orthodox BLUP approach are marginal modelling instead of conditional modelling approaches, that is, they are useful to make the 'population-averaged' instead of 'subject-specific' inferences on the mean (Schabenberger 1996; Burton et al. 1998). On the other hand, the orthodox BLUP approach explicitly incorporates random effects and mergers the 'subject-specific' and the 'population-averaged' inferences using the same model. In addition, the marginal covariance structures for our models are generally different from those considered by the other three approaches.

In the development of random effects modelling methodologies, the model checking has generally been ignored (see e.g. Zeger and Karim 1991; Breslow and Clayton 1993) or done via normality checking of residuals (Lee and Nelder 1996); however the justification of the normality of residuals has not been found in the literature. It appears from our simulation study that moderate departures from normality may not necessarily indicate departures from the model assumptions.

We illustrated the orthodox BLUP approach to analyses of clustered count, binomial and continuous data using Tweedie mixed models. In addition, our compound Poisson-Tweedie models may be applied to positive continuous data with a positive probability component at zero. One data example is car insurance data, for example, where the zero component corresponds to the case of no claims for a given insurance policy and the positive part of the distribution is the claim-size distribution (Jørgensen and Souze 1994). Another data example is the customers' expenditures in a certain shopping center.

In this thesis, we take dispersion parameters as nuisance parameters and their standard errors are not estimated; however it would be of interest to develop tests for the presence of random effects as well as issues such as over-dispersion and heterogeneity. Testing for such hypotheses is complicated because the null values lie on the boundary of the dispersion parameter space. Bootstrap methods may be used to construct such confidence intervals (Lele 1991), but these are controversial as regards their ability to fully reflect the relevant uncertainties.

10.2 Further study

In this section, we discuss some further research of the orthodox BLUP approach to generalized linear mixed models.

10.2.1 More than two levels of random effects

We have presented the Tweedie mixed model with two levels of random effects in detail. This model is useful to handle three-level hierarchical structure; however, hierarchies involving more levels also occur frequently in practice (Goldstein 1995). One such example is multi-center longitudinal clinical trials with centers further nested within geographical regions such as cities or countries.

The extension of the Tweedie mixed models with two levels of random effects is straightforward. The estimation procedure for more than two levels of nested random effects models is much the same as that for two-level models. The estimated score function and sensitivity matrix have the same global matrix expression as in Theorem 5.1. The key derivation of the orthodox BLUP approach is the calculation of the orthodox BLUP prediction of random effects which involves the inverse of the covariance matrix of the response. However this inverse can be obtained by repeatedly applying (3.13) to the covariance matrix of the response, twice for the models with two levels of random effects, three times for three levels of random effects, and so on.

We have derived explicit expressions for all quantities of interest. These expressions are useful for theoretical study. In practice, those explicit expressions are not necessary for computing purposes. Calculation of the covariance matrix of the response can be programmed based on (3.13), and all quantities of interest can be evaluated in terms of this inverse. This remark will be especially useful to resolve the computing issue for these extended models.

10.2.2 Crossed designs

In our analyses of the epilepsy data, we considered the nested random effects only. In clinical trials, the nature and degree of the variability of such epileptic seizure counts over time may be as important as its average behavior (Thall and Vail 1990). This aspect of the epilepsy data may be analyzed through the following simple crossed factor Tweedie mixed model with $p = 1$:

$$Y_{ijk}|U = u, V = v \sim Tw_p \left\{ \mu_{ijk}(u_i + v_j), \rho^2 (u_i + v_j)^{1-p} \right\},$$

where $U_1, \ldots, U_m, V_1, \ldots, V_n$ are mutually independent with $\mathrm{E}(U_i) = \mathrm{E}(V_j) = \frac{1}{2}$.

Replacing the random effects in the partially observed score function by their orthodox BLUP predictors would give us an estimated score function as follows:

$$\begin{aligned}\boldsymbol{\psi}(\boldsymbol{\beta}) &= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{N_{ij}} \frac{1}{\rho^2}\mu_{ijk}^{1-p}[Y_{ijk} - \mu_{ijk}(\widehat{U}_i + \widehat{V}_j)] \\ &= \mathbf{X}^{\top}\mathrm{diag}\,(\mathrm{E}(\mathbf{Y}))\,\mathrm{Var}^{-1}(\mathbf{Y})\,(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))\,.\end{aligned}$$

This estimating equation is clearly unbiased, but its asymptotic properties have yet to be studied. Ways of handling crossed designs are still rather limited in the literature of generalized linear mixed models. Further work on crossed factor random effects models would be of interest. The development of models for complicated crossed factor designs such as salamander mating data (McCullagh and Nelder 1989) is under way.

10.2.3 Survival data analysis

Incorporating random effects into Cox proportional hazard models has gained increasing attention in analyses of epidemiological or event history data. However, these models pose considerable theoretical difficulties in the development of estimation and inference procedures (Clayton 1991). Following Whitehead's (1980) idea of fitting the traditional Cox model using Poisson modelling techniques, we showed elsewhere that the random effects Cox model can also be fitted using random effects Poisson modelling techniques. To be more specific, we consider a Cox model with two levels of random effects as follows:

Let the kth individual in the jth sub-cluster of the ith cluster be indexed by (i, j, k). Let the hazard function for individual (i, j, k) at time t be denoted by $h_{ijk}(t)$. We assume that, given random effects $\mathbf{U} = \mathbf{u}$, the hazard functions for individuals are conditionally independent with

$$h_{ijk}(t) = h_0(t) U_{ij} \exp(\mathbf{x}_{ijk}^\top \boldsymbol{\beta}), \tag{10.1}$$

where $h_0(t)$ is the baseline hazard function and U_{ij}s are positive random effects, or 'frailties', shared by all individuals of the same cluster.

Suppose further that

- $\tau_1, \ldots, \tau_q$ are distinct death times;
- m_h is the multiplicity of deaths at time τ_h;
- the risk set at τ_h is $\mathcal{R}(\tau_h) = \{(i,j,k) : t_{ijk} \geq \tau_h\}$, where t_{ijk} is the observed survival time for individual (i,j,k).

Let $Y_{ijk,h}$ be 1 if individual (i,j,k) dies at τ_h and 0 otherwise. The fitting of random effects Cox model is equivalent to fitting the following auxiliary random effects Poisson model:

$$Y_{ijk,h} | \mathbf{U} = \mathbf{u} \quad \sim \quad \text{Poisson}\left(u_{ij} \exp(\alpha_h + \mathbf{x}_{ijk}^\top \boldsymbol{\beta}) \right). \tag{10.2}$$

Given random effects, Peto's version of the conditional partial likelihood (Cox and Oakes 1984) and conditional Poisson likelihood are

$$p\ell(\boldsymbol{\beta}; \mathbf{Y} | \mathbf{U} = \mathbf{u}) = \prod_{h=1}^{q} \frac{\prod_{(i,j,k) \in \mathcal{R}(\tau_h)} u_{ij}^{Y_{ijk,h}} [\exp(\mathbf{x}_{ijk}^\top \boldsymbol{\beta}]^{Y_{ijk,h}} (m_h!)}{[\sum_{(i,j,k) \in \mathcal{R}(\tau_h)} u_{ij} \exp(\mathbf{x}_{ijk}^\top \boldsymbol{\beta}]^{m_h}}, \tag{10.3}$$

and

$$\ell(\alpha, \boldsymbol{\beta}; \mathbf{Y} | \mathbf{U} = \mathbf{u}) = \prod_{h=1}^{q} \frac{\prod_{(i,j,k) \in \mathcal{R}(\tau_h)} u_{ij}^{Y_{ijk,h}} [\exp(\alpha_h + \mathbf{x}_{ijk}^\top \boldsymbol{\beta})]^{Y_{ijk,h}}}{\exp[\sum_{(i,j,k) \in \mathcal{R}(\tau_h)} u_{ij} \exp(\alpha_h + {}_{ijk}^\top \boldsymbol{\beta})]} \tag{10.4}$$

It can be shown that, given random effects, Peto's version of the conditional partial likelihood and conditional Poisson likelihood lead to the same maximum likelihood estimates for regression parameters. Hence the orthodox BLUP approach may be adopted to fit the random effects Cox models via Poisson mixed models.

Appendix A

Data sets

Table A.1: Seed germination data (a)seeds: O. Aegyptiaca 75 and 73, (b)root extracts: bean and cucumber

O. Aegyptiaca 75						O. Aegyptiaca 73					
Bean			Cucumber			Bean			Cucumber		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	0.26	5	6	0.83	8	16	0.50	3	12	0.25
23	62	0.37	53	74	0.72	10	30	0.33	22	41	0.54
23	81	0.28	55	72	0.76	8	28	0.29	15	30	0.50
26	51	0.51	32	51	0.63	23	45	0.51	32	51	0.63
17	39	0.44	46	79	0.58	0	4	0	3	7	0.43
			10	13	0.77						

Table A.2: Epilepsy data: successive two-week seizure counts for 59 epileptics. (a) Trt: treatment 0=placebo, 1=progabide); (b) Base: eight-week baseline seizure counts; (c)Age (in years).

ID	Y_1	Y_2	Y_3	Y_4	Base	Age	Trt
104	5	3	3	3	11	31	0
106	3	5	3	3	11	30	0
107	2	4	0	5	6	25	0
114	4	4	1	4	8	36	0
116	7	18	9	21	66	22	0
118	5	2	8	7	27	29	0
123	6	4	0	2	12	31	0
126	40	20	23	12	52	42	0
130	5	6	6	5	23	37	0
135	14	13	6	0	10	28	0
141	26	12	6	22	52	36	0
145	12	6	8	4	33	24	0
201	4	4	6	2	18	23	0
202	7	9	12	14	42	36	0
205	16	24	10	9	87	26	0
206	11	0	0	5	50	26	0
210	0	0	3	3	18	28	0
213	37	29	28	29	111	31	0
215	3	5	2	5	18	32	0
217	3	0	6	7	20	21	0
219	3	4	3	4	12	29	0
220	3	4	3	4	9	21	0
222	2	3	3	5	17	32	0
226	8	12	2	8	28	25	0
227	18	24	76	25	55	30	0
230	2	1	2	1	9	40	0
234	3	1	4	2	10	19	0
238	13	15	13	12	47	22	0
101	11	14	9	8	76	18	1
102	8	7	9	4	38	32	1

Table A.2: continued

ID	Y_1	Y_2	Y_3	Y_4	Base	Age	Trt
103	0	4	3	0	19	20	1
108	3	6	1	3	10	30	1
110	2	6	7	4	19	18	1
111	4	3	1	3	24	24	1
112	22	17	19	16	31	30	1
113	5	4	7	4	14	35	1
117	2	4	0	4	11	27	1
121	3	7	7	7	67	20	1
122	4	18	2	5	41	22	1
124	2	1	1	0	7	28	1
128	0	2	4	0	22	23	1
129	5	4	0	3	13	40	1
137	11	14	25	15	46	33	1
139	10	5	3	8	36	21	1
143	19	7	6	7	38	35	1
147	1	1	2	3	7	25	1
203	6	10	8	8	36	26	1
204	2	1	0	0	11	25	1
207	102	65	72	63	151	22	1
208	4	3	2	4	22	32	1
209	8	6	5	7	41	25	1
211	1	3	1	5	32	35	1
214	18	11	28	13	56	21	1
218	6	3	4	0	24	41	1
221	3	5	4	3	16	32	1
225	1	23	19	8	22	26	1
228	2	3	0	1	25	21	1
232	0	0	0	0	13	36	1
236	1	4	3	2	12	37	1

Table A.3: Cake baking data: breaking angles (degrees)

		Temperature					
	Rep.	175°	185°	195°	205°	215°	225°
	1	42	46	47	39	53	42
	2	47	29	35	47	57	45
	3	32	32	37	43	45	45
	4	26	32	35	24	39	26
	5	28	30	31	37	41	47
	6	24	22	22	29	35	26
	7	26	23	25	27	33	35
Recipe I	8	24	33	23	32	31	34
	9	24	27	28	33	34	23
	10	24	33	27	31	30	33
	11	33	39	33	28	33	30
	12	28	31	27	39	35	43
	13	29	28	31	29	37	33
	14	24	40	29	40	40	31
	15	26	28	32	25	37	33
	1	39	46	51	49	55	42
	2	35	46	47	39	52	61
	3	34	30	42	35	42	35
	4	25	26	28	46	37	37
	5	31	30	29	35	40	36
	6	24	29	29	29	24	35
	7	22	25	26	26	29	36
Recipe II	8	26	23	24	31	27	37
	9	27	26	32	28	32	33
	10	21	24	24	27	37	30
	11	20	27	33	31	28	33
	12	23	28	31	34	31	29
	13	32	35	30	27	35	30
	14	23	25	22	19	21	35
	15	21	21	28	26	27	20
	1	46	44	45	46	48	63
	2	43	43	43	46	47	58
	3	33	24	40	37	41	38
	4	38	41	38	30	36	35
	5	21	25	31	35	33	23
	6	24	33	30	30	37	35
	7	20	21	31	24	30	33
Recipe III	8	24	23	21	24	21	35
	9	24	18	21	26	28	28
	10	26	28	27	27	35	35
	11	28	25	26	25	38	28
	12	24	30	28	35	33	28
	13	28	29	43	28	33	37
	14	19	22	27	25	25	35
	15	21	28	25	25	31	25

Bibliography

[1] Aichison, J. and Ho, C.H. (1989). The multivariate Poisson log-normal distribution. *Biometrika* **76**, 643-653.

[2] Anderson, D.A. and Aitkin, M. (1985) Variance component models with binary responses: interviewer variability. *J. Roy. Statist. Soc. Ser. B* **47**, 203-210.

[3] Artes, R. and Jørgensen, B. (1998). Longitudinal data estimating equations for dispersion models. Technical report RT-MAE 9802, Institute of Mathematical Statistics, University of São Paulo.

[4] Atkinson, A.C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13-20.

[5] Breslow, N.E. (1984) Extra-Poisson variation in log-linear models. *Applied Statist.* **33**, 38-44.

[6] Breslow, N.E. (1996). Generalized linear models: Checking assumptions and strengthening conclusions. *Statistica Applicata* **8**, 23-41.

[7] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed model. *J. Amer. Statist. Assoc.* **88**, 9-25.

[8] Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**, 81-91.

[9] Brillinger, D. and Preisler, H.K. (1986) Two examples of quantal data analysis. *Proceedings of the* 13*th International Biometrics Conference* Seattle: The Biometrics Society. 95-113.

[10] Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods* 2nd ed. New York: Springer-Verlag.

[11] Burton, P., Gurrin, L., and Sly, P. (1998). Tutorial in biastatistics. Extending the simple linear regression model to account for correlated responese: an introduction to generalized estimating equations and multi-level mixed modelling. *Statist. Med.* **17**, 1261-1291.

[12] Clayton, D.G. (1991) A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467-485.

[13] Clayton, D.G. (1996). Discussion of the paper by Lee Y., and Nelder J.A., Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58**, 667.

[14] Cochran, W.G. and Cox, D.M. (1957). *Experimental Designs.* 2ed. New York: Wiley.

[15] Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data.* London: Chapman & Hall.

[16] Crowder, M.J. (1978). Beta-binomial Anova for proportions. *Appl. Statist.* **27**, 34-37.

[17] Crowder, M.J. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory* **3**, 305-330.

[18] Crowder, M.J. (1987). On linear and quadratic estimating function. *Biometrika* **74**, 591-597.

[19] Dean, C.B. and Lawless, J.F. (1989). A mixed Poisson-inverse Gaussian regression model. *Canad. J. Statist.* **17**, 171-181.

[20] Dempster, A.P., Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

[21] Draper, D. (1996). Discussion of the paper by Lee Y., and Nelder J.A., Hierarchical generalized linear models J. Roy. Statist. Soc., Ser. B **58**, 662.

[22] Engel, B. and Keen, A. (1996). Discussion of the paper by Lee Y., and Nelder J. A., Hierarchical generalized linear models, J. Roy. Statist. Soc., Ser. B **58**, 656.

[23] Firth, D. and Harris, I.R. (1991). Quasi-likelihood for multiplicative random effects. *Biometrika* **78**, 545-555.

[24] Fitzmaurice, G.M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**, 309-317.

[25] George, E.I., Makov, U.E. and Smith, I.G. (1987). Conjugate likelihood distributions. *Scand. J. Statist.* **20**, 147-156.

[26] Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likehihood for dependent data. *J. Roy. Statist. Soc., Ser. B* **54**, 657-699.

[27] Gilks, W.R., Wang, C.C., Yvonnet, B., and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biomatrics* **49**, 441-453.

[28] Gilmour, A.R., Anderson, R. and Rae, A.L. (1985) The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593-599.

[29] Glifford, P. (1993). Discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **55**, 53-54.

[30] Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277-284.

[31] Goldstein, H. (1995). *Multilevel Statistical Models.* New York: Wiley.

[32] Gray, J., Jesson, D., Goldstein, H., Hedger, J. and Rasbash, J. (1995). A multilevel analysis of school improvement: changes in schools' performance over time. *School Effectiveness and School Improvement,* **6**, 97-114.

[33] Harvey, A.C. (1981). *Time Series Models.* Oxford: Allan.

[34] Harville, D.A. and Mee, R.W. (1984) A mixed model procedure for analyzing ordered categorical data. *Biometrics* **40**, 393-408.

[35] Harville, D. (1977). Maximum likelihood approaches to variance component estimation and related problems. *J. Amer. Statist. Assoc.* **72**, 320-340.

[36] Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933-944.

[37] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biomatrics* **31**, 423-447.

[38] Hinde, J. (1982) Compound Poisson regression models. In GLIM 82: *Proceedings of International Conference on Generalized Models.* ed. R. Gilchrist, Berlin: Springer. 109-121.

[39] Hobert, J. and Casella, G. (1996) The effect of improper priors on Gibbs Sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91**, 1461-1473.

[40] Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc. Ser. B* **49**, 127-162.

[41] Jørgensen, B. (1997). *The Theory of Dispersion Models.* London: Chapman & Hall.

[42] Jørgensen, B. and Labouriau, R.S. (1995). *Exponential Family and Theoretical Inference.* Lecture notes. Department of Statistics, University of British Columbia, Vancouver, Canada.

[43] Jørgensen, B., Labouriau, R. and Lundbye-Christensen, S. (1996a). Linear growth curve analysis based on exponential dispersion model. *J. Roy. Statist. Soc. Ser. B* **58**, 573-592.

[44] Jørgensen, B., Lundbye-Christensen, S., Song, X.-K. and Sun, L. (1996b). A longitudinal study of emergency room visits and air pollution for Prince George, British Columbia. *Statist. Med.* **15**, 823-836.

[45] Jørgensen, B., Lundbye-Christensen, S., Song, X.-K. and Sun, L. (1996c). State-space models for multivariate longitudinal data of mixed types. *Canad. J. Statist.* **24**, 385-402.

[46] Jørgensen, B., Martínez, J.R. and Tsao, M. (1994). Asymptotic behaviour of the variance function. *Scand. J. Statist.* **21**, 223-243.

[47] Jørgensen, B. and Souza, M.C.P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuarial J.* **1**, 69-93.

[48] Karim, M.R. and Zeger, S.L. (1992). Generalized linear models with random effects; Salamander Mating Revisited. *Biometrics* **48**, 631-644.

[49] Knudsen, S.J. (1998). Estimating functions and Separate Inference. Cand. Scient. thesis, Department of Theoretical Statistics, Aarhus University.

[50] Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, 581-590.

[51] Lambert, D. and Roeder, K. (1995). Overdispersion diagnostics for generalized linear models. *J. Amer. Statist. Assoc.* **90**, 1225-1236.

[52] Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *Canad. J. of Stat.* **15**, 209-225.

[53] Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58**, 619-678.

[54] Lele, S. (1991). Resampling using estimating equations. *Theory of Estimating Equations.*(ed. Godambe, V.P.) Oxford: Clarendon Press.

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

[56] Lin, X. and Breslow, N.E. (1996a). Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. *J. Amer. Statist. Assoc.* **91**, 1007-1016.

[57] Lin, X. and Breslow, N.E. (1996b). Analysis of correlated binomial data in logistic-normal models. *J. Statist. Comput. Simul.* **55**, 130-146.

[58] Lipsitz, S.R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50**, 270-278.

[59] McCullagh, P. (1983). Quasi-likelihood functions *Annals of statistics* **11**, 59-67.

[60] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* 2nd ed. London: Chapman & Hall.

[61] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**, 162-170.

[62] McGilchrist, C.A. (1994). Estimation in generalized mixed models. *J. Roy. Statist. Soc. Ser. B* **56**, 61-69.

[63] McGilchrist, C.A.and Yau, K.K.W. (1995). The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. *Commun. Statist., Theory and Methods* **24**, 2963-2980.

[64] McLeish, D.L. and Small, C.G. (1988). *The Theory and Applications of Statistical Inference Functions.* Lecture Notes in Statistics **44**, New York: Springer-Verlag.

[65] Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. *Biometrics* **74**, 247-257.

[66] Natarajan, R. and McCullogh, C.E. (1995). A note on existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**, 639-643.

[67] Paik, M.C., Tsai, W.Y. and Ottman, R. (1994). Multivariate survival analysis using piecewise gamma frailty. *Biometrics* **50**, 975-988.

[68] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* 2nd ed. New York: Wiley.

[69] Ripley, B.D. and Venables, W.N. (1994). *Modern Applied Statistics with Splus.* New York: Springer-Verlag.

[70] Robinson, C.K. (1991). That BLUP Is a Good Thing: The Estimation of Random Effects. *Statist. Sci.* **6**, 15-32.

[71] Schabenberger, O. (1996). Population-averaged and subject-specific approaches for clustered categorical data. *J. Statist. Simul.* **54**, 231-253.

[72] Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727.

[73] Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **55**, 3-23.

[74] Stiratelli, R., Laird, N. and Ware, J. (1984) Random effects models for serial observations with binary responses. *Biometrics* **40**, 961-971.

[75] Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657-671.

[76] Waclawiw, M.A. and Liang, K.Y. (1993). Prediction of random effects in the generalized linear model. *J. Amer. Statist. Assoc.* **88**, 171-178.

[77] Whitehead, J. (1980) Fitting Cox's regression model to survival data using GLIM. *Applied Statistics* **29**, 268-275.

[78] Williams, D.A. (1982) Extra-binomial variation in logistic linear models. *Applied Statist.* **31**, 144-148.

[79] Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *J. Statist. Comput. Simul.* **48**, 233-243.

[80] Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects; A Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**, 79-86.

[81] Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.