

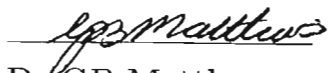
A CLASSICAL APPROACH FOR
THE ANALYSIS OF
GENERALIZED LINEAR MIXED MODELS

by

MJ Hammujuddy

Submitted in fulfilment of the academic requirements
for the degree of
Master of Science
in the
School of Mathematical & Statistical Sciences
Faculty of Science
University of KwaZulu-Natal
Howard College Campus
Durban 4041
South Africa

As the candidate's supervisor, I have approved this dissertation for
submission.



Dr GB Matthews

December 2004

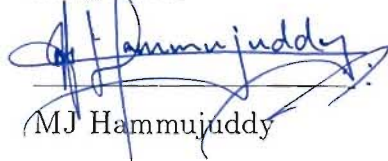
Abstract

Generalized linear mixed models (GLMMs) accommodate the study of overdispersion and correlation inherent in hierarchically structured data. These models are an extension of generalized linear models (GLMs) and linear mixed models (LMMs). The linear predictor of a GLM is extended to include an unobserved, albeit realized, vector of Gaussian distributed random effects. Conditional on these random effects, responses are assumed to be independent. The objective function for parameter estimation is an integrated quasi-likelihood (IQL) function which is often intractable since it may consist of high-dimensional integrals. Therefore, an exact maximum likelihood analysis is not feasible. The penalized quasi-likelihood (PQL) function, derived from a first-order Laplace expansion to the IQL about the optimum value of the random effects and under the assumption of slowly varying weights, is an approximate technique for statistical inference in GLMMs. Replacing the conditional weighted quasi-deviance function in the Laplace-approximated IQL by the generalized chi-squared statistic leads to a corrected profile quasi-likelihood function for the restricted maximum likelihood (REML) estimation of dispersion components by Fisher scoring. Evaluation of mean parameters, for fixed dispersion components, by iterative weighted least squares (IWLS) yields joint estimates of fixed effects and random effects. Thus, the PQL criterion involves repeated fitting of a Gaussian LMM with a linked response vector and a conditional iterated weight matrix. In some instances, PQL estimates fail to converge to a neighbourhood of their true values. Bias-corrected PQL estimators (CPQL) have hence been proposed, using asymptotic analysis and simulation. The pseudo-likelihood algorithm is an alternative estimation procedure for GLMMs. Global score statistics for hypothesis testing of overdispersion, correlation and heterogeneity in GLMMs has been developed as well as individual score statistics for testing null dispersion components separately. A conditional mean squared error of prediction (CMSEP) has also been considered as a general measure of predictive uncertainty. Local influence measures for testing the robustness of parameter estimates, by inducing minor perturbations into GLMMs, are recent advances in the study of these models. Commercial statistical software is available for the analysis of GLMMs.

Preface and Declaration

The study described in this dissertation was carried out in the School of Mathematical & Statistical Sciences at the University of KwaZulu-Natal, Howard College Campus, and was completed under the supervision of Dr GB Matthews.

The material contained in this dissertation has not, to my knowledge, been published elsewhere, nor has this dissertation been submitted in any form for the award of any degree or diploma to another university or tertiary institution. Where use has been made of the work of others, it is duly referenced in the text.



MJ Hammujuddy

December 2004

Acknowledgements

I wish to express my sincere gratitude to:

- Dr GB Matthews, my supervisor, for introducing me to Generalized Linear Mixed Models, a class of models that has instilled in me a deeper understanding of data analytic techniques, and for laying down the foundations for the study of statistics, coupled with an introduction to the SAS System, during my undergraduate years.
- Professor JH Swart, my undergraduate mentor, for inspiring me to take up higher studies in statistics, in spite of the mathematical rigour involved, and for his many words of advice.
- Faculty members in the School of Mathematical & Statistical Sciences, especially Professor AI Dale and Dr WH Moolman, for the indispensable knowledge imparted to me.
- Dr S Moopanar and Dr DE North for their trust in me as a graduate assistant.
- Members of staff of the EG Malherbe Library, especially those in the inter-library loan section, for their kind help.
- Mr S Greenwood for the computing facilities made available to me in the science postgraduate local area network.
- My parents and sisters for their concern, motivation and support, and their tolerance of my absence in their lives.
- Mauritian students at the former University of Natal for their kindness and lively company.
- My friends, especially the FD Patel and DE Moolla families of Durban and Stanger respectively, for their hospitality during my stay in South Africa.
- Mrs J Sylaides for typesetting this dissertation with \LaTeX and for her administrative assistance.
- The Student Funding Centre for the award of the university's Special Master's and Graduate Scholarships.

Dedicated to my father, Mr F Hammujuddy

Contents

Chapter 1 Introduction	1
Chapter 2 An overview of generalized linear models	4
2.1 Introduction	4
2.2 Classical linear models	5
2.3 Generalized linear models	6
2.3.1 Mean and variance	7
2.4 Fisher scoring for maximum likelihood estimation of fixed effects	10
2.4.1 Iterative weighted least squares	14
2.4.2 Algorithm	15
2.5 Measures of discrepancy	16
2.6 Quasi-likelihood functions	17
2.6.1 Quasi-deviance function	18
2.7 Some remarks	19
Chapter 3 An outline of linear mixed models	20
3.1 Introduction	20
3.2 Linear mixed models	21
3.3 Restricted maximum likelihood	25
3.4 Mixed model equations	27
3.5 Best linear unbiased prediction	29
3.5.1 Algorithm	30
3.6 Penalized likelihood	31
3.7 Some remarks	33

Chapter 4	Some aspects of generalized linear mixed models	34
4.1	Introduction	34
4.2	Generalized linear mixed models	34
4.3	Penalized quasi-likelihood function	37
4.4	Fisher scoring for maximum PQL estimation of fixed effects and prediction of random effects	41
4.5	Corrected profile quasi-likelihood function for the estimation of components of dispersion	45
4.6	Computing PQL estimators	46
4.6.1	Algorithm	46
4.6.2	Restricted pseudo-likelihood procedure for fitting GLMMs	47
4.7	Bias in parameter estimators under PQL	50
4.8	Generalized global score test for components of dispersion	52
4.8.1	Individual dispersion component score test	58
4.9	Conditional mean squared error of prediction	59
4.9.1	Conditional standard errors of prediction	59
4.10	Some remarks	64
Chapter 5	Analysis of correlated binomial data using a GLMM: The Logistic-Gaussian model	66
5.1	Introduction	66
5.2	The Logistic-Gaussian model	67
5.3	Modelling heterogeneity in binomial clinical trials data	68
5.3.1	Analyzing the risk of respiratory tract infections by selective decontamination of the digestive tract	69
5.3.2	Results	71
5.4	Modelling overdispersion in binomial data	72
5.4.1	Analyzing the mortality of cancer cells under radiation	73
5.4.2	Results	74
5.5	Some remarks	75
5.5.1	Further remarks	76

Chapter 6 Local influence analysis for GLMMs	80
6.1 Introduction	80
6.2 Measuring local influence in GLMMs	81
6.3 Cluster weights and random effects perturbation	82
6.3.1 Notation for a two-level clustered design	82
6.3.2 Proposed scheme	83
6.4 Some remarks	86
Chapter 7 Conclusion	87
References	90
Appendix A Some results	96
A.1 Cumulants	96
A.2 Standard cumulants	97
A.3 A note on quasi-distributions	98
Appendix B SAS codes for the analyses of the respiratory tract infections and the mortality of cancer cells datasets	100

Chapter 1

Introduction

The early 1990s has seen the emergence of a new tool for data analysis in the mathematical and applied statistics literature. The generalized linear models (GLMs) (Nelder & Wedderburn 1972) and the linear mixed models (LMMs), (see, for example, McCulloch & Searle 2001, Chapter 6, pp.156–186), have been fused into a hybrid body of statistical theory and methodology known as the generalized linear mixed models (GLMMs) (Breslow 2003). This new class of models is useful for analyzing overdispersed and correlated discrete data. In these models, a vector of unobserved, albeit realized, Gaussian distributed random effects, with mean zero and dispersion matrix with unknown components of dispersion (to be estimated from the data) is introduced into the linear predictor of a GLM. Responses from a hierarchical model are assumed to be conditionally independent given the random effects. The conditional means of the observations are related to the extended linear predictor through a specified link function and their conditional variances are specified by a variance function, known prior weights and a scale factor (see Breslow & Clayton 1993; Clayton 1992).

The objective function for parameter estimation in GLMMs is an integrated¹ quasi-likelihood (IQL) function (Breslow 2003). However, for complicated problems, this function involves irreducibly high-dimensional integrals, often intractable (Breslow & Clayton 1993). Several researchers, for example, Engel & Keen (1994), McGilchrist (1994), Schall (1991) and Wolfinger & O'Connell (1993), have proposed approximate estimation procedures to circumvent this cumbersome integral. We shall describe the approach of Breslow & Clayton (1993) and use their method in an application. Furthermore, we outline and employ the technique of Wolfinger & O'Connell (1993) to illustrate overdispersion in a dataset where responses are discrete and correlated.

¹We shall use the abbreviation IQL for the integrated quasi-likelihood function.

The marginal likelihood for the observed data is obtained by taking a first-order Laplace expansion to the IQL about the maximum value of the random effects. Under certain assumptions, this approximation yields the penalized quasi-likelihood (PQL) criterion. Application of Fisher scoring to PQL-based estimating equations, for fixed values of the components of dispersion, yields joint estimates of fixed effects and random effects (by IWLS). Further adjustments give standard REML estimating equations for dispersion components (Breslow 2003; Breslow & Clayton 1993; Lin & Breslow 1996a). Thus, the PQL procedure involves repeated fitting of a Gaussian LMM with a working response vector and a (conditional) iterated weight matrix (Breslow 2003). Although the PQL estimation procedure is appealing, estimates of the components of dispersion, and hence of the regression coefficients, are biased, in some instances. Correction procedures using asymptotic analysis for bias reduction in PQL estimators have been proposed by Lin & Breslow (1996a) for GLMMs with multiple components of dispersion, thereby extending the work of Breslow & Lin (1995) in the case of a single dispersion component. Simulation-based bias reduction in parameter estimates has been considered by Pawitan (2001).

In the context of GLMMs, statistical tests for overdispersion, correlation and heterogeneity have been proposed by Lin (1997). Using Laplace expansions to the IQL, she has proposed a global score statistic for testing the hypothesis that all components of dispersion are null. Furthermore, Lin (1997) provides a Laplace-approximated individual score test for testing dispersion components separately. Bias-corrected versions of these test statistics are also available which account for the loss of degrees-of-freedom incurred due to the estimation of fixed effect parameters. Booth & Hobert (1998) have proposed a conditional mean squared error of prediction (CMSEP) as a general measure of prediction variance. The application of local influence analysis (Cook 1986) is the most recent research interest in GLMMs. Some perturbation schemes have been considered by Xiang *et al* (2003) and Zhu & Lee (2003).

This dissertation is organised as follows:

Chapters 2 and 3 provide a review of GLMs and LMMs, respectively. Also, Chapter 2 briefly outlines the concept of quasi-likelihood (Wedderburn 1974). Some of the aspects of GLMMs, as mentioned above, are considered in Chapter 4, wherein we consider the CMSEP of Booth & Hobert (1998) for the hierarchical model of Breslow & Clayton (1993). In Chapter 5, we analyze two datasets in binomial form. We compare estimates obtained with SAS GLIMMIX macro and the supposedly ‘true’ ML estimates (Breslow 2003) generated by PROC NLMIXED (Wolfinger 1999) using Version 8.2 of the SAS System for SAS/STAT (SAS Institute 2001). Note that Breslow & Clayton (1993) constrain the scale factor at unity in their PQL approach. In illustrating overdispersion in one of the datasets, this extra-dispersion parameter is estimated using the restricted pseudo-likelihood² (REPL)³ procedure of Wolfinger & O’Connell (1993). We stress the fact that none of the authors referenced herein have used the SAS System to analyze these datasets. In Chapter 6, we propose an alternative formulation of a perturbation scheme considered by Zhu & Lee (2003). Some remarks conclude Chapters 2 to 6. Finally, some directions for further research are pointed out in Chapter 7.

²The pseudo-likelihood procedure estimates the scale factor and is an extension of the PQL approach (Kuss 2002).

³REPL generates an REML-like estimate for the scale factor (Littell *et al* 1996, p.436).

Chapter 2

An overview of generalized linear models

2.1 Introduction

The theory of classical linear models (LMs) has been studied extensively in the literature. In LMs, responses are continuous and, are assumed to be independently Gaussian distributed. Moreover, the assumption of homogeneity among observations holds in LMs. Parameters are estimated by the method of maximum likelihood. Searle (1971) is an excellent reference for LMs.

Discrete data have been analyzed using various transformation techniques that depend on the nature of the data under study. The data is transformed so that linearity and normality can be achieved, and parameters are obtained by maximum likelihood procedures. In an attempt to unify the seemingly different approaches for the analysis of discrete data, Nelder & Wedderburn (1972) introduced a class of models known as generalized linear models (GLMs).

In GLMs, the assumption that the responses are independently distributed is maintained while that of normality is relaxed. The distribution of the observations is assumed to be a member of the exponential family. Fixed effects are estimated by Fisher scoring and the scale factor is given by a moment estimator. Moreover, observations may be heterogeneous. The deviance is a measure of the adequacy of a specific model. When the parametric form of the responses is unknown, the concept of quasi-likelihood (QL), in the sense of Wedderburn (1974), is useful for statistical inference. QL is based by positing a mean-variance relationship for the responses. It is also assumed that the mean of each observation is some known function of fixed effect parameters.

In this chapter, we provide an overview of generalized linear models and quasi-likelihood functions.

2.2 Classical linear models

Let the vector $\mathbf{y} = (y_1, \dots, y_n)^T$ denote realizations of a continuous random variable Y .¹ Each datum y_i can be expressed as

$$y_i = \mu_i + \epsilon_i \quad (2.2.1)$$

where μ_i is the mean of y_i and ϵ_i is a random disturbance term for $i = 1, \dots, n$. Equation (2.2.1) is referred to as a *classical linear model* (LM) equation (McCulloch & Searle 2001, p.16). Furthermore, each μ_i can be expressed as a linear combination of explanatory variables x_{ij} associated with fixed effects β_j , $j = 1, \dots, p$. Equation (2.2.1) can be rewritten as

$$\begin{aligned} y_i &= \sum_{j=1}^p x_{ij} \beta_j + \epsilon_i \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \end{aligned} \quad (2.2.2)$$

or, in matrix notational form, as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (2.2.3)$$

where \mathbf{y} is an $n \times 1$ response vector, \mathbf{X} is a model matrix of order $n \times p$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects and $\boldsymbol{\epsilon}$ is of length n . Following Wedderburn (1976), we assume that \mathbf{X} is of full rank p .

In LMs, elements of \mathbf{y} are assumed to be independently Gaussian distributed, and are homoscedastic with variance φ_0 . Thus, the ϵ_i 's are uncorrelated and it is traditional to assume that $\boldsymbol{\epsilon}$ has mean zero. Hence, we have

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Gamma_0) \quad (2.2.4)$$

where $\Gamma_0 \equiv \Gamma(\varphi_0) = \varphi_0 \mathbf{I}_n$, with \mathbf{I}_n being an identity matrix of order n .

¹We shall use \mathbf{y} to denote a random vector and Y , a random variable.

Therefore,

$$E(\mathbf{y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

that is,

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{2.2.5a}$$

where $\boldsymbol{\mu}$, of length n , is the mean vector of \mathbf{y} and the variance-covariance matrix of \mathbf{y} is given by

$$\text{Var}(\mathbf{y}) = \text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Gamma}(\varphi_0) \tag{2.2.5b}$$

(The notation $\text{Var}(\mathbf{y})$ will be used throughout to denote the variance-covariance matrix of \mathbf{y}).

That is, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Gamma}_0)$. Let the parameter vector of interest be denoted by $\boldsymbol{\psi}^T = (\boldsymbol{\beta}^T, \varphi_0)$. The likelihood function for the observed data has the form

$$\begin{aligned} L(\boldsymbol{\psi}) &= L(\boldsymbol{\beta}, \varphi_0) \\ &= |2\pi\boldsymbol{\Gamma}(\varphi_0)|^{-n/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Gamma}^{-1}(\varphi_0)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \end{aligned} \tag{2.2.6}$$

It is straightforward to show that the *maximum likelihood* (ML) estimators of $\boldsymbol{\beta}$ and φ_0 are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{2.2.7a}$$

and

$$\hat{\varphi}_0 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{2.2.7b}$$

respectively. Stanner & Duffy (1989, p.290) hold φ_0 at 1 and state that $\hat{\boldsymbol{\beta}}$ is the unique maximum of the log-likelihood function $l \equiv l(\boldsymbol{\beta}, \varphi_0) = \ln L(\boldsymbol{\psi})$, and hence of $L(\boldsymbol{\psi})$, since l is strictly concave and where $(\mathbf{X}^T \mathbf{X})^{-1}$ exists. When no distributional assumption is made about \mathbf{y} , the method of *ordinary least squares* (OLS) or *generalized least squares* (GLS) may be employed to estimate $\hat{\boldsymbol{\beta}}$ (McCulloch & Searle 2001, p.116).

2.3 Generalized linear models

Suppose now that the vector $\mathbf{y} = (y_1, \dots, y_n)^T$ denotes realizations of a random variable Y . The assumption that the y_i 's are independently distributed

is maintained, as in Section 2.2, while that of normality is relaxed. The probability function (Dobson 1990, p.27) of y_i is assumed to be a member of the exponential family of distributions which, in its canonical form (McCulloch & Searle 2001, p.138), is given by

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}u_i[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\} \quad (2.3.1)$$

where u_i is a known prior weight attached to y_i , θ_i is the *natural parameter* and ϕ , the *scale factor* ($i = 1, \dots, n$). θ_i is specific to y_i whereas ϕ is common to all y_i 's (Green & Silverman 1994, p.92).

Discrete data exhibit a non-linear relationship between $\boldsymbol{\mu}$ and $\mathbf{X}\boldsymbol{\beta}$ (McCullagh & Nelder 1989, Sections 1.2.3–1.2.9, pp.10–17; McCulloch & Searle 2001, pp.135–136). Linearity can be achieved by using the *generalized linear models* (GLMs) methodology, originally propounded by Nelder & Wedderburn (1972). In GLMs, the *random component* refers to Y and $\mathbf{X}\boldsymbol{\beta}$ is called the *systematic component*. A monotonic² and twice differentiable function g (Wedderburn 1976) applied to $\boldsymbol{\mu}$ results in the following linear relationship:

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (2.3.2)$$

or, more compactly, as

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \quad (2.3.3)$$

where $g(\boldsymbol{\mu}) = (g(\mu_1), \dots, g(\mu_n))^T$, with $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ (Lin & Breslow 1996a). $g(\cdot)$ is called the *link function* and $\boldsymbol{\eta}$, the *linear predictor*. When there exists a sufficient statistic $\mathbf{X}^T \mathbf{y}$ whose dimension equals that of $\boldsymbol{\beta}$ in $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, $g(\cdot)$ is referred to as the *canonical link function*, for then $\boldsymbol{\theta} = \boldsymbol{\eta}$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ with θ_i as defined under equation (2.3.1) (McCullagh & Nelder 1989, p.32).

2.3.1 Mean and variance

Certain regularity conditions are required to derive the mean and variance of y_i , $i = 1, \dots, n$ (McCulloch & Searle 2001, p.140). Some of these conditions are quoted without proof by Casella & Berger (2002, p.516). We use the results from Dobson (1990, Appendix A, pp.142–144) to find $E(y_i)$ and $\text{Var}(y_i)$.

²The monotonicity of a function is discussed by Casella & Berger (2002, pp.50–51).

Let $l_i \equiv l(\theta_i, \phi; y_i)$ denote the log-likelihood function for y_i . Then, from equation (2.3.1), we have

$$l_i = \phi^{-1}u_i[y_i\theta_i - b(\theta_i)] + c(y_i, \phi) \quad (2.3.4)$$

The expected value of the score $\frac{\partial l_i}{\partial \theta_i}$ is given by

$$\begin{aligned} 0 &= E\left(\frac{\partial l_i}{\partial \theta_i}\right) \\ &= \phi^{-1}u_i\left[y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}\right] \end{aligned}$$

that is, $\mu_i = b'(\theta_i)$ (2.3.5)

where $b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}$. The second-order partial derivative³ of l_i with respect to θ_i is

$$\begin{aligned} \frac{\partial^2 l_i}{\partial \theta_i^2} &= -\phi^{-1}u_i \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \\ &= -\phi^{-1}u_i b''(\theta_i) \end{aligned} \quad (2.3.6)$$

Now,

$$E\left(\frac{\partial l_i}{\partial \theta_i}\right)^2 = -E\left(\frac{\partial^2 l_i}{\partial \theta_i^2}\right)$$

$$(\phi^{-1}u_i)^2 E(y_i - \mu_i)^2 = \phi^{-1}u_i b''(\theta_i)$$

Hence

$$\text{Var}(y_i) = \phi u_i^{-1} V(\mu_i) \quad (2.3.7)$$

where $V(\mu_i) = b''(\theta_i)$ is called the *variance function* that depends on μ_i . Thus, GLMs accommodate unequal variances among responses (McCullagh & Nelder 1989, p.14 and p.29). Such a dependence is not observed for the

³Primes shall denote differentiation of a function with respect to the appropriate parameter.

Gaussian distribution, a member of the exponential family (McCulloch & Searle 2001, p.141).

For canonical link functions, where $\eta_i = \theta_i$ and $\frac{\partial \eta_i}{\partial \theta_i} = 1$, we have

$$\begin{aligned}
 \text{(i)} \quad V(\mu_i) &= \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \\
 &= \frac{\partial b'(\theta_i)}{\partial \theta_i} \\
 &= \frac{\partial \mu_i}{\partial \theta_i} \\
 &= \frac{\partial \mu_i}{\partial \eta_i} \left[\frac{\partial \theta_i}{\partial \eta_i} \right]^{-1} \\
 &= \left[\frac{\partial \eta_i}{\partial \mu_i} \right]^{-1} \\
 &= \left[\frac{\partial g(\mu_i)}{\partial \mu_i} \right]^{-1} \\
 &= [g'(\mu_i)]^{-1}
 \end{aligned} \tag{2.3.8}$$

It follows, from equation (2.3.8), that

$$V(\mu_i)g'(\mu_i) = 1 \tag{2.3.9}$$

(ii) From equation (2.3.5), we have $\theta_i = a(\mu_i)$, where $a(\cdot)$ defines the inverse of $b'(\cdot)$. If $g(\cdot) = a(\cdot)$, then from equation (2.3.2), $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ (Green & Silverman 1994, p.93).

2.4 Fisher scoring for maximum likelihood estimation of fixed effects

Most regularity conditions are satisfied by the exponential family of distributions such that there exists a unique global maximum of the log-likelihood function $l(\boldsymbol{\beta}; \mathbf{y})$. This maximum is given by solutions of the equations

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

which are non-linear (and closed form solutions are unavailable). The *Fisher scoring*, a numerical optimization technique and an amended form of the Newton-Raphson method for solving non-linear equations, is used for the estimation of fixed effects (Dobson 1990, p.40). We restrict our attention to GLMs with canonical links whereby the Fisher scoring and the Newton-Raphson method reduce to the same algorithm (McCullagh & Nelder 1989, p.43).

Fisher scoring for $\boldsymbol{\beta}$ has the form

$$\boldsymbol{\beta}^{(\tau+1)} = \boldsymbol{\beta}^{(\tau)} + [\mathcal{I}(\boldsymbol{\beta}^{(\tau)})]^{-1} \nabla(\boldsymbol{\beta}^{(\tau)}) \quad (2.4.1)$$

where $\mathcal{I}(\boldsymbol{\beta}^{(\tau)})$, the *information matrix*, and $\nabla(\boldsymbol{\beta}^{(\tau)})$, the gradient vector, are evaluated at the τ th iteration. Note that $\mathcal{I}(\boldsymbol{\beta}^{(\tau)}) = -E(\mathbf{H}^{(\tau)})$, where \mathbf{H} is the Hessian, and $\nabla(\boldsymbol{\beta}^{(\tau)}) = \left. \frac{\partial l}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(\tau)}}$ (McCulloch & Searle 2001, p.143; Searle *et al* 1992, pp.292–295). It can be shown that the score function for β_j is given by

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{u_i(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} \quad (2.4.2)$$

For the elements of \mathbf{H} , we have

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_k \partial \beta_j} &= \frac{\partial}{\partial \beta_k} \left(\frac{\partial l}{\partial \beta_j} \right) \\ &= \frac{\partial}{\partial \beta_k} \left(\sum_{i=1}^n \frac{u_i(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} \right) \\ &= \sum_{i=1}^n \frac{\partial}{\partial \mu_i} \left(\frac{u_i(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} \right) \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n \left\{ \frac{u_i x_{ij}}{\phi V(\mu_i) g'(\mu_i)} \frac{\partial(y_i - \mu_i)}{\partial \mu_i} + u_i x_{ij} (y_i - \mu_i) \frac{\partial}{\partial \mu_i} \left(\frac{1}{\phi V(\mu_i) g'(\mu_i)} \right) \right\} \\
&\quad \times \frac{1}{g'(\mu_i)} x_{ik} \\
&= \sum_{i=1}^n \left\{ \frac{-u_i x_{ij} x_{ik}}{\phi V(\mu_i) [g'(\mu_i)]^2} + u_i (y_i - \mu_i) x_{ij} x_{ik} \left[\frac{V(\mu_i) g''(\mu_i) + g'(\mu_i) V'(\mu_i)}{\phi [V(\mu_i)]^2 [g'(\mu_i)]^3} \right] \right\} \tag{2.4.3}
\end{aligned}$$

for $i = 1, \dots, n$; $j, k = 1, \dots, p$. For canonical links, it has been shown in Section (2.3.1) that $V(\mu_i) = [g(\mu_i)]^{-1}$. Therefore,

$$V'(\mu_i) = -[g'(\mu_i)]^{-2} g''(\mu_i) \tag{2.4.4}$$

Substituting equation (2.4.4) in the numerator of the second term on the right-hand side of the curly brackets in equation (2.4.3) yields

$$[g'(\mu_i)]^{-1} g''(\mu_i) - g'(\mu_i) [g'(\mu_i)]^{-2} g''(\mu_i) = 0$$

(Lin 1997). Thus,

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = - \sum_{i=1}^n \frac{u_i x_{ij} x_{ik}}{\phi V(\mu_i) [g'(\mu_i)]^2} \tag{2.4.5}$$

In matrix form, equation (2.4.5) is written as

$$\frac{\partial^2 l}{\partial \beta \partial \beta^T} = \mathbf{H} = -\phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X}$$

where \mathbf{W} is an $n \times n$ diagonal matrix with elements

$$w_i = [u_i^{-1} V(\mu_i) [g'(\mu_i)]^2]^{-1} \tag{2.4.6}$$

on the diagonal. Hence,

$$\mathcal{I}(\beta) = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} \tag{2.4.7}$$

Now, consider the score equations given by

$$\begin{aligned}\frac{\partial l_i}{\partial \eta_i} &= \frac{\partial l_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \\ &= \frac{u_i(y_i - \mu_i)}{\phi V(\mu_i)} \frac{1}{g'(\mu_i)}\end{aligned}\quad (2.4.8)$$

From equation (2.4.6) we have

$$V(\mu_i) = w_i^{-1} u_i [g'(\mu_i)]^{-2} \quad (2.4.9)$$

Substituting equation (2.4.9) into equation (2.4.8) gives, after some simplification,

$$\frac{\partial l_i}{\partial \eta_i} = \phi^{-1} w_i g'(\mu_i) (y_i - \mu_i) \quad (2.4.10)$$

In matrix form, equation (2.4.10) becomes

$$\frac{\partial l}{\partial \boldsymbol{\eta}} = \phi^{-1} \mathbf{W} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) \quad (2.4.11)$$

where $\boldsymbol{\Delta}$ is a diagonal matrix of order n with elements $g'(\mu_i)$ on the diagonal.

Therefore,

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \frac{\partial l}{\partial \boldsymbol{\eta}} \\ &= \mathbf{X}^T \frac{\partial l}{\partial \boldsymbol{\eta}} \\ &= \phi^{-1} \mathbf{X}^T \mathbf{W} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu})\end{aligned}\quad (2.4.12)$$

Also,

$$\begin{aligned}
-E \left(\frac{\partial^2 l_i}{\partial \eta_i^2} \right) &= E \left(\frac{\partial l_i}{\partial \eta_i} \right)^2 \\
&= \phi^{-2} w_i^2 [g'(\mu_i)]^2 E(y_i - \mu_i)^2 \\
&= \phi^{-2} w_i^2 [g'(\mu_i)]^2 \phi u_i^{-1} V(\mu_i) \\
&= \phi^{-1} w_i^2 u_i^{-1} V(\mu_i) [g'(\mu_i)]^2 \\
&= \phi^{-1} w_i
\end{aligned} \tag{2.4.13}$$

In matrix form, we have equation (2.4.13) as

$$-E \left(\frac{\partial^2 l}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \right) = \phi^{-1} \mathbf{W} \tag{2.4.14}$$

Thus,

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\beta}) &= -E \left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) \\
&= -E \left(\mathbf{X}^T \frac{\partial^2 l}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \mathbf{X} \right) \\
&= \mathbf{X}^T E \left(-\frac{\partial^2 l}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \right) \mathbf{X} \\
&= \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X}
\end{aligned} \tag{2.4.15}$$

where $\boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{W}$ and $\boldsymbol{\Delta}$ are evaluated at the τ th iteration (Green & Silverman 1994, p.95; McCulloch & Searle 2001, pp.141–142). The *asymptotic dispersion matrix* is the inverse of the information matrix and, for $\hat{\boldsymbol{\beta}}$, is given by

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \phi(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \tag{2.4.16}$$

where \mathbf{W} is evaluated at $\hat{\boldsymbol{\beta}}$ (Green & Silverman 1994, p.97) and ϕ is given by a moment estimator (see Section 2.4.2).

2.4.1 Iterative weighted least squares

Fisher scoring for β , after substituting the appropriate identities for $\mathcal{I}(\beta^{(\tau)})$ and $\nabla(\beta^{(\tau)})$ into equation (2.4.1) becomes

$$\beta^{(\tau+1)} = \beta^{(\tau)} + (\mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{(\tau)}) \quad (2.4.17)$$

Multiplying the above equation by $\mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X}$ throughout yields

$$\begin{aligned} (\mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X}) \beta^{(\tau+1)} &= (\mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X}) \beta^{(\tau)} + \mathbf{X}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{(\tau)}) \\ &= \mathbf{X}^T \mathbf{W}^{(\tau)} [\mathbf{X} \beta^{(\tau)} + \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{(\tau)})] \\ &= \mathbf{X}^T \mathbf{W}^{(\tau)} \zeta^{(\tau)} \end{aligned} \quad (2.4.18)$$

where we define

$$\zeta^{(\tau)} = \mathbf{X} \beta^{(\tau)} + \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{(\tau)}) \quad (2.4.19)$$

Equation (2.4.18) has the normal⁴ equations form for using weighted least squares and equation (2.4.19) is the *adjusted dependent response vector* or *linked response vector* $g(\mathbf{y})$ (Wolfinger & O'Connell 1993).

Moreover,

$$E(\zeta_i^{(\tau)}) = \mathbf{x}_i^T \beta^{(\tau)} \quad (2.4.20)$$

and

$$\begin{aligned} \text{Var}(\zeta_i^{(\tau)}) &= \text{Var}(y_i - \mu_i^{(\tau)}) [g'(\mu_i^{(\tau)})]^2 \\ &= \phi u_i^{-1} V(\mu_i^{(\tau)}) [g'(\mu_i^{(\tau)})]^2 \\ &= \phi w_i^{-1(\tau)} \end{aligned} \quad (2.4.21)$$

From equations (2.4.20) and (2.4.21), it follows that $\zeta^{(\tau)} \sim N(\mathbf{X} \beta^{(\tau)}, \phi \mathbf{W}^{-1(\tau)})$. That is, equation (2.4.19) is an LM with $\epsilon^{(\tau)} = \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{(\tau)})$. The ML estimator of β is thus given by

$$\begin{aligned} \hat{\beta} &= \beta^{(\tau+1)} \\ &= (\mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(\tau)} \zeta^{(\tau)} \end{aligned} \quad (2.4.22)$$

⁴The word 'normal' in normal equations has no relation to the Normal distribution but is related to the use of orthogonal projections in the development of the theory of least squares (Jones 1993, p.33).

We have illustrated that the Fisher scoring for evaluating the estimator of β is a *weighted least squares* (WLS) regression of the adjusted or *working* response vector $\zeta^{(\tau)}$ on \mathbf{X} with a *working weight matrix* $\mathbf{W}^{(\tau)}$. In general, $\zeta^{(\tau)}$ and $\mathbf{W}^{(\tau)}$ depend on $\beta^{(\tau)}$ (that is, ζ and \mathbf{W} are updated at each iteration). The procedure for finding $\hat{\beta}$ is referred to as *iterative weighted least squares* (IWLS). See Dobson (1990, pp.40–41), Green & Silverman (1994, pp.94–95) and McCullagh & Nelder (1989, pp.40–43).

2.4.2 Algorithm

Given \mathbf{y} and \mathbf{X} :

Step 1: Initialize iteration by setting $\beta^{(\tau)} = \beta^{(0)}$, obtaining $\eta^{(0)} = \mathbf{X}\beta^{(0)}$.

Step 2: Evaluate $\zeta^{(0)}$ and $\mathbf{W}^{(0)}$. Substitute these values into equation (2.4.22) to obtain $\beta^{(1)}$.

Step 3: Test for convergence: If $|\beta^{(\tau+1)} - \beta^{(\tau)}| \rightarrow \mathbf{0}$, terminate iteration. Set $\beta^{(\tau+1)} = \hat{\beta}$. Otherwise, increment τ by 1, and repeat steps 1 and 2.

The term ϕ vanishes in equation (2.4.1). However, ϕ can be estimated by using a moment estimator (McCulloch & Searle 2001, p.154), which is given by

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{u_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \quad (2.4.23)$$

$$= \frac{1}{n-p} X^2 \quad (2.4.24)$$

where X^2 is the (*weighted*) *generalized Pearson chi-squared statistic* (McCullagh & Nelder 1989, p.34). If the model fitted is adequate, then equation (2.4.24) is approximately unbiased for ϕ . This assertion holds provided p is small relative to n (McCullagh & Nelder 1989, p.127). In the above algorithm, to avoid singularities in the initial iteration, some modifications to the data are required (Littell *et al* 1996, p.507).

2.5 Measures of discrepancy

A crucial aspect of statistical analyses is to assess the goodness- (or badness-) ⁵-of-fit of the models under study. The *generalized likelihood ratio statistic* (Dobson 1990, p.56) is used for such an assessment and for GLMs is of the form

$$\lambda = \frac{L(\mathbf{y}, \phi; \mathbf{y})}{L(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})}$$

or

$$\ln \lambda = l(\mathbf{y}, \phi; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y}) \quad (2.5.1)$$

where $l(\mathbf{y}, \phi; \mathbf{y})$ is the maximum log-likelihood achievable under the *saturated* model and $l(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})$ denotes the log-likelihood maximized over $\boldsymbol{\beta}$ for a fixed value of ϕ . Let $\tilde{\theta}_i \equiv \theta_i(y_i)$ and $\hat{\theta}_i \equiv \theta_i(\hat{\mu}_i)$ be estimates of the natural parameter θ_i under y_i and $\hat{\mu}_i$, respectively. From equation (2.5.1), we have

$$\begin{aligned} & \sum_{i=1}^n \phi^{-1} u_i \{ [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)] - [y_i \hat{\theta}_i - b(\hat{\theta}_i)] \} \\ &= \sum_{i=1}^n \phi^{-1} u_i \{ [y_i(\tilde{\theta}_i - \hat{\theta}_i)] - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \} \\ &= \sum_{i=1}^n \phi^{-1} d_i \end{aligned} \quad (2.5.2)$$

where $d_i = u_i \{ [y_i(\tilde{\theta}_i - \hat{\theta}_i)] - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \}$.

The (*weighted*) *scaled deviance* is defined as

$$\begin{aligned} D^s &\equiv D^s(\mathbf{y}; \hat{\boldsymbol{\mu}}) \\ &= 2[l(\mathbf{y}, \phi; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})] \\ &= 2 \sum_{i=1}^n \phi^{-1} d_i \\ &= 2 \ln \lambda \end{aligned} \quad (2.5.3)$$

⁵See Nelder (2000).

and the (*weighted*) *deviance* is given by

$$\begin{aligned}
 D &\equiv D(\mathbf{y}; \hat{\boldsymbol{\mu}}) \\
 &= \phi D^s \\
 &= 2 \sum_{i=1}^n d_i
 \end{aligned}
 \tag{2.5.4}$$

where the contribution of each d_i from y_i is referred to as the *deviance increment* (Green & Silverman 1994, p.96; McCullagh & Nelder 1989, pp.33–34). By Theorem 10.3.3 of Casella & Berger (2002, p.490), the (scaled) deviance converges asymptotically to a χ^2 -distribution with $n - p$ degrees-of-freedom. Thus, the current model is rejected if and only if $-2 \ln \lambda \geq X_{df, \alpha}^2$, where $df = n - p$ and α , the *asymptotic size test*, has to be defined.

The X^2 -statistic is an alternative measure of discrepancy. Asymptotically, both D and X^2 follow approximately a χ^2 -distribution, and are exactly χ^2 -distributed for normal theory LMs (McCullagh & Nelder 1989, p.34).

2.6 Quasi-likelihood functions

Consider the following argument:

$$\begin{aligned}
 \frac{\partial l_i}{\partial \mu_i} &= \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \\
 &= \frac{u_i(y_i - \mu_i)}{\phi V(\mu_i)}
 \end{aligned}
 \tag{2.6.1}$$

Equation (2.6.1) depends solely on the first two moments of y_i , $i = 1, \dots, n$. Suppose that the distribution of y_i does not belong to the exponential family. Such an assumption is reasonable in the absence of sufficient information about y_i (McCullagh & Nelder 1989, p.324). It is then difficult to postulate a likelihood function for statistical inference. Consequently, it becomes necessary to define a likelihood-type function with properties similar to those of likelihood functions proper.

Wedderburn (1974) argued that if a mean-variance relationship can be specified for y_i , then full parametric assumptions for the observed data can be

relaxed. He defined a quasi-likelihood⁶ function $Q_i(y_i; \mu_i)$ by the relation

$$\frac{\partial Q_i(y_i; \mu_i)}{\partial \mu_i} = \frac{u_i(y_i - \mu_i)}{\phi V(\mu_i)} \quad (2.6.2)$$

where we attach weights u_i to each independent y_i . It is also assumed that there exists a function g such that $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, where g is as defined in Section 2.3. On integration, equation (2.6.2) becomes

$$Q_i(y_i; \mu_i) = \int_{y_i}^{\mu_i} \frac{u_i(y_i - s)}{\phi V(s)} ds + \text{some function of } y_i \quad (2.6.3)$$

It can be proved that quasi-likelihood (QL) functions have properties similar to those of log-likelihood functions. Furthermore, the log-likelihood function for y_i is identical to the quasi-likelihood if and only if the distribution of y_i is a member of the exponential family (Wedderburn 1974).

Maximum quasi-likelihood (MQL) estimates of $\hat{\boldsymbol{\beta}}$ are evaluated via Fisher scoring and the scale factor is calculated from the moment estimator which is given by equation (2.4.23). See McCullagh & Nelder (1989, pp.327–328).

2.6.1 Quasi-deviance function

The (*weighted*) *quasi-deviance* function is defined as

$$\begin{aligned} d_i &= -2[Q_i(y_i; \mu_i) - Q_i(y_i; y_i)] \\ &= -2 \int_{y_i}^{\mu_i} \frac{u_i(y_i - s)}{V(s)} ds \end{aligned} \quad (2.6.4)$$

where $d_i > 0$ except at $y_i = \mu_i$. The total (quasi)-deviance D is the sum of d_i over the observations. Moreover, D is a function that depends on y_i and μ_i only, and is independent of ϕ (McCullagh & Nelder 1989, p.327 ; Nelder & Pregibon 1987). Breslow (2003) refers to equation (2.6.4) as the *weighted deviance*. See also Nelder & Lee (1992).

⁶Strictly quasi-log-likelihood (Nelder 2000).

2.7 Some remarks

In GLMs, hypotheses testing of fixed effect parameters can be done by using the *Wald statistic* or the *score statistic*. These statistics are asymptotically chi-squared distributed with degrees-of-freedom equal to the length of the parameter vector. An alternative approach is to specify each hypothesis in terms of a model. Deviances are then compared for competing models having identical distribution and link function, but with unequal number of parameters. See Dobson (1990, pp.61–62).

An assessment of hypothesized link functions and testing of hypotheses of nested subsets of covariates in the linear predictor of a GLM can be performed by using the likelihood ratio or score tests. Although these tests can be applied to Wedderburn's (1974) definition of a quasi-likelihood, they cannot be used to compare different variance functions, as remarked by Nelder & Pregibon (1987). These authors introduce an *extended quasi-likelihood* (EQL) function which allows for the comparison of various forms of all the components of a GLM and where the random component is specified by its first two moments only.

Chapter 3

An outline of linear mixed models

3.1 Introduction

A set of data may be classified in terms of factors, nested or crossed, that identify the source of a data point. These factors often consist of several different levels. For statistical inference, focus is then on the effects of these levels on a response variable. Suppose there exists an infinite set of levels of a factor: Effects due to its levels are called random effects. This is because only a random sample of those levels of that factor are likely to occur in the data. Therefore, random effects are random variables, assumed to be independently distributed with mean zero. When fixed effects and random effects simultaneously occur in a dataset, in addition to the random disturbance term, the resulting statistical model is referred to as a linear mixed model (LMM) (McCulloch & Searle 2001, pp.2–4, p.9 and p.13).

In many practical situations, there are usually several random factors. It is assumed that the levels of a factor are independent of each other, the levels of other factors and the residual effects. The variance of an observation is an aggregate of the variances of the levels of the different factors together with the variance of the residual effects. These variances are termed variance components (Harville 1977). In LMMs, responses are assumed to be Gaussian distributed (Searle *et al* 1992, p.233). Also, responses with common random (effect) terms are positively correlated (Engel & Keen 1994). However, given the random effects, conditional independence among responses may be achieved (Breslow & Clayton 1993).

In this chapter, we provide an outline of LMMs.¹ We focus on the estimation of components of dispersion by the method of restricted maximum likelihood (REML) of Patterson & Thompson (1971). Estimators of fixed effects and best linear unbiased predictors (BLUP) of random effects are obtained from the (Henderson) mixed model equations (MMEs), which can be used to iteratively evaluate REML (or ML) estimators of dispersion com-

¹In some sections we rely on ideas from Breslow & Clayton (1993) and Lin & Breslow (1996a), with an identity link function, however.

ponents (Searle *et al*, pp.275–286). We briefly comment on the concept of penalized likelihood.

3.2 Linear mixed models

Let $\mathbf{y} = (y_1, \dots, y_n)^T$ denote the response vector of a continuous random variable Y . Each data point y_i represents a response variable on the i th of n units, which may be blocked as in the study of repeated measures. For statistical modelling of random effects that occur in the data, y_i may be expressed, not only in terms of $\mathbf{x}_i^T \boldsymbol{\beta}$ and ϵ_i (as in Section 2.2), but also as a linear combination of a $q \times 1$ vector of explanatory (or indicator) variables \mathbf{z}_i associated with random effects (Breslow & Clayton 1993). Thus, y_i can be written as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma} + \epsilon_i \quad (3.2.1)$$

or, more compactly, as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad (3.2.2)$$

where \mathbf{Z} is an incidence matrix of order $n \times q$ and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effects. The dimensions of \mathbf{y} , full rank model matrix \mathbf{X} , fixed effects vector $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ are as defined in Section 2.2. Equation (3.2.2) is known as a *linear mixed model* (LMM) equation – ‘mixed’ because of the simultaneous presence of fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{\gamma}$, other than the random disturbance term $\boldsymbol{\epsilon}$.

By their nature, random effects are random variables and, in certain prediction problems (see, for example, Robinson 1991), are assumed to be independently Gaussian distributed with expectation zero – that is,

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \Gamma) \quad (3.2.3)$$

where $\Gamma = \text{Var}(\boldsymbol{\gamma})$. For ML estimation, \mathbf{y} is assumed to follow a Gaussian distribution because, in a practical sense, normality leads to mathematically tractable methodology, even for unbalanced data (Searle *et al* 1992, p.233).

From the above assumptions, the mean and the variance of \mathbf{y} are given by

$$\begin{aligned} E(\mathbf{y}) &= E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} \end{aligned} \quad (3.2.4a)$$

and, with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Gamma_0)$,

$$\begin{aligned} \Psi &= \text{Var}(\mathbf{y}) \\ &= \text{Var}(\boldsymbol{\epsilon}) + \mathbf{Z} \text{Var}(\boldsymbol{\gamma}) \mathbf{Z}^T \\ &= \Gamma_0 + \mathbf{Z}\Gamma\mathbf{Z}^T \end{aligned} \quad (3.2.4b)$$

It is assumed that $\text{Cov}(\boldsymbol{\gamma}, \boldsymbol{\epsilon}^T) = \mathbf{0}$ (Harville 1977) – that is, $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are independent. Therefore, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \Psi)$.

Analyses of hierarchically structured data may involve multiple, say c , sources of random variation. For statistical inference, \mathbf{Z} is partitioned as

$$\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_c) \quad (3.2.5a)$$

where the incidence matrix \mathbf{Z}_m , of order $n \times q_m$, is associated with the m th random effect $\boldsymbol{\gamma}_m$ having q_m levels – that is, of order $q_m \times 1$ – from the corresponding partitioning of the random effects vector $\boldsymbol{\gamma}$ such that

$$\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_c^T) \quad (3.2.5b)$$

where $m = 1, \dots, c$. Then, from equations (3.2.5), equation (3.2.2) becomes

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \dots + \mathbf{Z}_c\boldsymbol{\gamma}_c + \boldsymbol{\epsilon} \\ &= \mathbf{X}\boldsymbol{\beta} + \sum_{m=1}^c \mathbf{Z}_m\boldsymbol{\gamma}_m + \boldsymbol{\epsilon} \end{aligned} \quad (3.2.6)$$

It is further assumed, following equation (3.2.3), that the random effects $\boldsymbol{\gamma}_m$ are independently Gaussian distributed with means zero – that is,

$$\text{Cov}(\boldsymbol{\gamma}_m, \boldsymbol{\gamma}_k^T) = \mathbf{0} \quad (3.2.7a)$$

for $m, k = 1, \dots, c$ ($m \neq k$), and

$$\boldsymbol{\gamma}_m \sim N(\mathbf{0}, \Gamma_m) \quad (3.2.7b)$$

where $\Gamma_m \equiv \Gamma(\varphi_m) = \varphi_m \mathbf{I}_{q_m}$, with \mathbf{I}_{q_m} being an identity matrix of order q_m .

Lin (1997) notes that q_m represents the amount of information available on estimating dispersion component φ_m associated with γ_m .

Thus,

$$\begin{aligned}\text{Var}(\boldsymbol{\gamma}) &= \Gamma(\boldsymbol{\varphi}) \\ &= \text{diag}(\Gamma_1, \dots, \Gamma_c) \\ &= \text{diag}(\varphi_1 \mathbf{I}_{q_1}, \dots, \varphi_c \mathbf{I}_{q_c})\end{aligned}\quad (3.2.7c)$$

where $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_c)^T$ (Lin & Breslow 1996a). Furthermore, the following assumption holds (Searle *et al* 1992, p.233):

$$\text{Cov}(\boldsymbol{\gamma}_m, \boldsymbol{\epsilon}^T) = \mathbf{0}$$

for all $m = 1, \dots, c$. The parameter φ_m represents the common variance of the levels of the m th random effect, whereas φ_0 is the common variance of the residual effects $\boldsymbol{\epsilon}$. The variances $\varphi_0, \varphi_1, \dots, \varphi_c$ are called *variance components* (Harville 1977) or *dispersion components* (McCullagh & Nelder 1989, p.432) because they are the components of the variance of an observation y_i (McCulloch & Searle 2001, p.161).

Let the parameter vector of interest be denoted by $\boldsymbol{\psi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\varphi}^{*T})$, where $\boldsymbol{\varphi}^* = (\varphi_0, \varphi_1, \dots, \varphi_c)^T$. The likelihood function for the observed data has the form

$$\begin{aligned}L(\boldsymbol{\psi}) &= L(\boldsymbol{\beta}, \Psi) \\ &= |2\pi|^{-n/2} |\Psi(\boldsymbol{\varphi}^*)|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Psi^{-1}(\boldsymbol{\varphi}^*) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right]\end{aligned}\quad (3.2.8)$$

where $\Psi \equiv \Psi(\boldsymbol{\varphi}^*)$.

Differentiating $l \equiv l(\boldsymbol{\psi}) = \ln L(\boldsymbol{\psi})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}^*$, and setting first-order partial derivatives to zero, leads to score equations

$$\mathbf{X}^T \Psi^{-1} \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \Psi^{-1} \mathbf{y} \quad (3.2.9a)$$

for fixed effects, and estimating equations

$$\text{tr}(\Psi^{-1} \mathbf{Z}_m \mathbf{Z}_m^T) = \mathbf{y}^T \mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T \mathbf{P} \mathbf{y} \quad (3.2.9b)$$

for components of dispersion, where

$$\mathbf{P} = \Psi^{-1} - \Psi^{-1} \mathbf{X} (\mathbf{X}^T \Psi^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Psi^{-1} \quad (3.2.9c)$$

for $m = 0, 1, \dots, c$. Equation (3.2.9b) is non-linear in the dispersion components and has to be solved numerically, by an iterative procedure.

Fisher scoring may be adopted for evaluating ψ , and has the form

$$\psi^{(\tau+1)} = \psi^{(\tau)} + [\mathcal{I}(\psi^{(\tau)})]^{-1} \left. \frac{\partial l}{\partial \psi} \right|_{\psi=\psi^{(\tau)}} \quad (3.2.10)$$

where

$$\frac{\partial l}{\partial \psi} = \left(\left(\frac{\partial l}{\partial \beta} \right)^T, \left(\frac{\partial l}{\partial \varphi^*} \right)^T \right)^T$$

with

$$\frac{\partial l}{\partial \beta} = \mathbf{X}^T \Psi^{-1} \mathbf{y} - \mathbf{X}^T \Psi^{-1} \mathbf{X} \beta \quad (3.2.11a)$$

and

$$\frac{\partial l}{\partial \varphi_m} = -\frac{1}{2} \text{tr}(\Psi^{-1} \mathbf{Z}_m \mathbf{Z}_m^T) + \frac{1}{2} (\mathbf{y} - \mathbf{X} \beta)^T \Psi^{-1} \mathbf{Z}_m \mathbf{Z}_m^T \Psi^{-1} (\mathbf{y} - \mathbf{X} \beta) \quad (3.2.11b)$$

The information matrix for ψ is given by

$$\mathcal{I}(\psi) = \begin{pmatrix} \mathbf{X}^T \Psi^{-1} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2} \text{tr}(\Psi^{-1} \mathbf{Z}_m \mathbf{Z}_m^T \Psi^{-1} \mathbf{Z}_k \mathbf{Z}_k^T) \end{pmatrix} \quad (3.2.11c)$$

for $m, k = 0, 1, \dots, c$. Equations (3.2.11) are evaluated at the τ th cycle of iteration. The iteration is initialized at $\tau = 0$ and continues until convergence is achieved. The ML estimators of β and Ψ are denoted by $\hat{\beta}$ and $\hat{\Psi} \equiv \Psi(\hat{\varphi}^*)$, respectively. Asymptotically,

$$\text{Var}(\hat{\beta}) \rightarrow (\mathbf{X}^T \hat{\Psi}^{-1} \mathbf{X})^{-1} \quad (3.2.12a)$$

$$\text{Cov}(\hat{\beta}, \hat{\varphi}^*) \rightarrow \mathbf{0} \quad (3.2.12b)$$

$$\text{Var}(\hat{\varphi}^*) \rightarrow 2 \text{tr}(\hat{\Psi}^{-1} \mathbf{Z}_m \mathbf{Z}_m^T \hat{\Psi}^{-1} \mathbf{Z}_k \mathbf{Z}_k^T) \quad (3.2.12c)$$

In general, ML estimators are not unbiased but, in the limit, they are consistent. Moreover, equations (3.2.12) are only exact as the sample size tends to infinity and may, for small-sized samples of data, result in underestimation of variances of the ML estimators (Searle *et al* 1992, pp.234–240 and p.313).

3.3 Restricted maximum likelihood

Estimators of components of dispersion are generally biased (in a downward direction) when using the method of maximum likelihood (Harville 1977). This is due to the fact that ML does not take into account the loss of degrees-of-freedom incurred due to the estimation of fixed effect parameters. Patterson & Thompson (1971) rectified this bias by a modification of the ML method, which is now known as the *restricted maximum likelihood* (REML) method. It is a procedure that makes allowance for the evaluation of unknown parameters in the mean by using null contrasts solely for the estimation of dispersion components (Lee & Nelder 2003). In other words, there exists a set of maximum linearly independent vectors \mathbf{k}_r^T such that $E(\mathbf{k}_r^T \mathbf{y}) = \mathbf{0}$. This implies that $\mathbf{k}_r^T \mathbf{X} \boldsymbol{\beta} = \mathbf{0}$ such that $\mathbf{k}_r^T \mathbf{X} = \mathbf{0}$ for all $\boldsymbol{\beta}$. Upon defining $\mathbf{K} = (\mathbf{k}_1, \dots, \mathbf{k}_r)$, we have $\mathbf{K}^T \mathbf{X} = \mathbf{0}$, with \mathbf{K}^T having full row rank $r = n - p$, where $p = \text{rank}(\mathbf{X})$ (McCulloch & Searle 2001, p.176; Searle *et al* 1992, pp.250–251).

Given $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \Psi)$, we have $E(\mathbf{K}^T \mathbf{y}) = \mathbf{0}$ and $\text{Var}(\mathbf{K}^T \mathbf{y}) = \mathbf{K}^T \Psi \mathbf{K}$. Therefore, $\mathbf{K}^T \mathbf{y} \sim N(\mathbf{0}, \mathbf{K}^T \Psi \mathbf{K})$. The restricted log-likelihood function has the form

$$l(\Psi) = -\frac{1}{2}(n - p) \log |2\pi| - \frac{1}{2} \log |\mathbf{K}^T \Psi \mathbf{K}| - \frac{1}{2} \mathbf{y}^T \mathbf{K} (\mathbf{K}^T \Psi \mathbf{K})^{-1} \mathbf{K}^T \mathbf{y} \quad (3.3.1)$$

The non-linear estimating equations for the components of dispersion are given by

$$\text{tr}(\mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T) = \mathbf{y}^T \mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T \mathbf{P} \mathbf{y} \quad (3.3.2)$$

for $m = 0, 1, \dots, c$, and \mathbf{P} is as defined in equation (3.2.9c) and equals $\mathbf{K}(\mathbf{K}^T \Psi \mathbf{K})^{-1} \mathbf{K}^T$. Fisher scoring for solving equation (3.3.2) is given by

$$\varphi^{*(\tau+1)} = \varphi^{*(\tau)} + \left[\mathcal{I}(\varphi^{*(\tau)}) \right]^{-1} \left. \frac{\partial l}{\partial \varphi^*} \right|_{\varphi^* = \varphi^{*(\tau)}} \quad (3.3.3)$$

where

$$\frac{\partial l}{\partial \varphi^*} = -\frac{1}{2} \text{tr}(\mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T) + \frac{1}{2} \mathbf{y}^T \mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T \mathbf{P} \mathbf{y} \quad (3.3.4a)$$

and

$$\mathcal{I}(\varphi^*) = \frac{1}{2} \text{tr}(\mathbf{P} \mathbf{Z}_k \mathbf{Z}_k^T \mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T) \quad (3.3.4b)$$

for $m, k = 0, 1, \dots, c$. The asymptotic dispersion matrix of φ^* is given by

$$\text{Var}(\varphi^*) = [\mathcal{I}(\varphi^*)]^{-1} \quad (3.3.4c)$$

At convergence of the Fisher scoring for evaluating the dispersion components, the REML estimator of φ^* is denoted by $\hat{\varphi}^*$. Fixed effect parameters are not evaluated by the REML method. However, with $\hat{\varphi}^*$, the GLS estimator for fixed effects is given by

$$\begin{aligned} \hat{\beta} &= \beta(\hat{\varphi}^*) \\ &= (\mathbf{X}^T \hat{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Psi}^{-1} \mathbf{y} \end{aligned} \quad (3.3.4d)$$

where $\hat{\Psi} \equiv \Psi(\hat{\varphi}^*)$ (McCulloch & Searle 2001, pp.308–310; Searle *et al* 1992, pp.252–254, p.256 and p.313).

Robinson (1987) recommends REML estimators when using dispersion components to obtain better and more efficient estimates of other effects, while Robinson (1991) states, from a classical point of view, that these estimators seem to have the best credentials for unbalanced data. See McCulloch & Searle (2001, Section 6.10, pp.177–178) and Searle *et al* (1992, Section 6.8, pp.254–255) for some of the merits of REML estimators.

3.4 Mixed model equations

In LMMs, elements of \mathbf{y} are assumed to be Gaussian distributed. But those y_i 's with common random (effect) terms are positively correlated for $i = 1, \dots, n$ (Engel & Keen 1994). However, conditional independence among responses can be achieved by conditioning \mathbf{y} on the random effects $\boldsymbol{\gamma}$. Following Breslow & Clayton (1993), but with $g(\cdot)$ as identity link function, the conditional mean and conditional variance of $\mathbf{y}|\boldsymbol{\gamma}$ are given by

$$\begin{aligned} E(\mathbf{y}|\boldsymbol{\gamma}) &= \boldsymbol{\mu}^\boldsymbol{\gamma} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \\ &= \mathbf{X}\boldsymbol{\beta} + \sum_{m=1}^c \mathbf{Z}_m \boldsymbol{\gamma}_m \end{aligned} \quad (3.4.1a)$$

and

$$\begin{aligned} \text{Var}(\mathbf{y}|\boldsymbol{\gamma}) &= \text{Var}(\boldsymbol{\epsilon}) \\ &= \boldsymbol{\Gamma}_0 \end{aligned} \quad (3.4.1b)$$

respectively. Thus, $\mathbf{y}|\boldsymbol{\gamma} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \boldsymbol{\Gamma}_0)$. The joint² density function of $\mathbf{y}|\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}$ has the form

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\gamma}) &= f(\mathbf{y}|\boldsymbol{\gamma})f(\boldsymbol{\gamma}) \\ &= |2\pi|^{-\frac{1}{2}(n+q)} |\boldsymbol{\Gamma}_0|^{-\frac{1}{2}} |\boldsymbol{\Gamma}|^{-\frac{1}{2}} \times \\ &\quad \exp \left\{ -\frac{1}{2} \begin{pmatrix} \boldsymbol{\gamma} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Gamma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_0 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\gamma} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma} \end{pmatrix} \right\} \end{aligned} \quad (3.4.2)$$

where $f(\cdot)$ and $f(\cdot|\cdot)$ denote density function and conditional density function, respectively, and $q = \sum_{m=1}^c q_m$, the total number of random effects $\boldsymbol{\gamma}_m$, $m = 1, \dots, c$. It can be shown that by maximizing $f(\mathbf{y}, \boldsymbol{\gamma})$ – that is, by equating to zero first-order partial derivatives of equation (3.4.2) with respect to $\boldsymbol{\beta}$, first, then to $\boldsymbol{\gamma}$, leads to the following equations:

$$\mathbf{X}^T \boldsymbol{\Gamma}_0^{-1} \mathbf{X} \boldsymbol{\beta} + \mathbf{X}^T \boldsymbol{\Gamma}_0^{-1} \mathbf{Z} \boldsymbol{\gamma} = \mathbf{X}^T \boldsymbol{\Gamma}_0^{-1} \mathbf{y} \quad (3.4.3a)$$

$$\mathbf{Z}^T \boldsymbol{\Gamma}_0^{-1} \mathbf{X} \boldsymbol{\beta} + (\mathbf{Z}^T \boldsymbol{\Gamma}_0^{-1} \mathbf{Z} + \boldsymbol{\Gamma}^{-1}) \boldsymbol{\gamma} = \mathbf{Z}^T \boldsymbol{\Gamma}_0^{-1} \mathbf{y} \quad (3.4.3b)$$

²Equation (3.4.2) is not a likelihood function proper as $\boldsymbol{\gamma}$ is unobserved, albeit realized (Lee & Nelder 1996).

Equations (3.4.3) are called *mixed model equations* (MMEs), and are expressed more compactly as

$$\begin{pmatrix} \mathbf{X}^T \Gamma_0^{-1} \mathbf{X} & \mathbf{X}^T \Gamma_0^{-1} \mathbf{Z} \\ \mathbf{Z}^T \Gamma_0^{-1} \mathbf{X} & \mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T \Gamma_0^{-1} \mathbf{y} \\ \mathbf{Z}^T \Gamma_0^{-1} \mathbf{y} \end{pmatrix} \quad (3.4.4)$$

(McCulloch & Searle 2001, p.258; Robinson 1991; Searle *et al* 1992, pp.275–276).

The coefficient matrix

$$\begin{pmatrix} \mathbf{X}^T \Gamma_0^{-1} \mathbf{X} & \mathbf{X}^T \Gamma_0^{-1} \mathbf{Z} \\ \mathbf{Z}^T \Gamma_0^{-1} \mathbf{X} & \mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1} \end{pmatrix} \quad (3.4.5)$$

of the linear system (3.4.4) is symmetric positive definite or semidefinite, a property that can be exploited for computational purposes (Harville 1977). Assuming that the components of dispersion are known and equal to $\hat{\boldsymbol{\varphi}}^*$, solutions $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively, to equation (3.4.4) are called *mixed model solutions*. Furthermore, the variance-covariance matrix of estimation errors

$$E \left\{ \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix}^T \right\}$$

equals the inverse of expression (3.4.5), provided that the model matrix \mathbf{X} is of full rank (Robinson 1991).

It may occur, at some τ th cycle of an iterative procedure that one of the elements of the dispersion components is zero. Having a null dispersion component makes Γ non-invertible, so that Γ^{-1} does not exist. To circumvent this drawback, the following linear system may be used:

$$\begin{pmatrix} \mathbf{X}^T \Gamma_0^{-1} \mathbf{X} & \mathbf{X}^T \Gamma_0^{-1} \mathbf{Z} \Gamma \\ \mathbf{Z}^T \Gamma_0^{-1} \mathbf{X} & \mathbf{I} + \mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} \Gamma \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma}_* \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T \Gamma_0^{-1} \mathbf{y} \\ \mathbf{Z}^T \Gamma_0^{-1} \mathbf{y} \end{pmatrix} \quad (3.4.6)$$

where $\boldsymbol{\gamma} = \Gamma \boldsymbol{\gamma}_*$. System (3.4.6) does not require inverting Γ (Harville 1977) and its coefficient matrix is not symmetric (Searle *et al* 1992, p.284).

3.5 Best linear unbiased prediction

Consider equations (3.4.3): From equation (3.4.3b), we have

$$\gamma = (\mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{Z}^T \Gamma_0^{-1} (\mathbf{y} - \mathbf{X}\beta) \quad (3.5.1)$$

Substituting equation (3.5.1) into equation (3.4.3a) gives

$$\mathbf{X}^T \Gamma_0^{-1} \mathbf{X}\beta + \mathbf{X}^T \Gamma_0^{-1} \mathbf{Z} (\mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{Z}^T \Gamma_0^{-1} (\mathbf{y} - \mathbf{X}\beta) = \mathbf{X}^T \Gamma_0^{-1} \mathbf{y}$$

So

$$\begin{aligned} & \mathbf{X}^T \left[\Gamma_0^{-1} - \Gamma_0^{-1} \mathbf{Z} (\mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{Z}^T \Gamma_0^{-1} \right] \mathbf{X}\beta \\ &= \mathbf{X}^T \left[\Gamma_0^{-1} - \Gamma_0^{-1} \mathbf{Z} (\mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{Z}^T \Gamma_0^{-1} \right] \mathbf{y} \end{aligned}$$

Thus

$$\mathbf{X}^T \Psi^{-1} \mathbf{X}\beta = \mathbf{X}^T \Psi^{-1} \mathbf{y} \quad (3.5.2a)$$

where $\Psi^{-1} = \Gamma_0^{-1} - \Gamma_0^{-1} \mathbf{Z} (\mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{Z}^T \Gamma_0^{-1}$.

Therefore, with Ψ set equal to its ML estimate $\hat{\Psi}$, we have

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T \hat{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Psi}^{-1} \mathbf{y} \\ &= \beta(\hat{\varphi}^*) \end{aligned} \quad (3.5.2b)$$

Now replacing $(\mathbf{Z}^T \Gamma_0^{-1} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{Z}^T \Gamma_0^{-1} = \Gamma \mathbf{Z}^T \Psi^{-1}$ into equation (3.5.1) yields, at $\Psi = \hat{\Psi}$,

$$\begin{aligned} \hat{\gamma} &= E(\gamma|\mathbf{y}) \\ &= \hat{\Gamma} \mathbf{Z}^T \hat{\Psi}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \gamma(\hat{\beta}) \\ &= \gamma(\beta(\hat{\varphi}^*)) \end{aligned} \quad (3.5.3)$$

$\hat{\gamma}$ in equation (3.5.3) is referred to as the *best linear unbiased predictor* (BLUP) of γ (Searle *et al* 1992, pp.276–277).

It is to be noted that $\hat{\gamma}$ is a *linear* function of \mathbf{y} (Robinson 1991). Since γ is a random variable, the term *predictor*³ is often used to predict its realized values. Furthermore, the criteria *best* and *unbiasedness* in the acronym

³See Robinson (1991) for a discussion on ‘estimating’ or ‘predicting’ realized values of a random variable.

BLUP respectively implies minimum mean squared error of prediction and $E_y(\hat{\gamma}) = E_y[E\gamma|y(\gamma|y)] = E(\gamma)$ (Searle *et al* 1992, pp.261-262).

The elements of the MMEs can be used to evaluate ML and REML estimates of components of dispersion. The REML estimating equations for $\varphi_0, \varphi_1, \dots, \varphi_c$ are given by equation (3.3.2). For $m = 0$,

$$\varphi_0^{(\tau+1)} = \frac{1}{n-p} \mathbf{y}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(\tau)} - \mathbf{Z}\boldsymbol{\gamma}^{(\tau)}) \quad (3.5.4a)$$

and for $m = 1, \dots, c$

$$\varphi_m^{(\tau+1)} = \frac{1}{[q_m - tr(\mathbf{T}_{mm}^{(\tau)})]} \boldsymbol{\gamma}^{(\tau)T} \boldsymbol{\gamma}^{(\tau)} \quad (3.5.4b)$$

where \mathbf{T}_{mm} is the (m, m) th submatrix of

$$\mathbf{T} = (\mathbf{I} + \mathbf{Z}^T \mathbf{S} \mathbf{Z} \boldsymbol{\Gamma})^{-1} \quad (3.5.5)$$

with

$$\mathbf{S} = \boldsymbol{\Gamma}_0^{-1} - \boldsymbol{\Gamma}_0^{-1} \mathbf{X} (\mathbf{X}^T \boldsymbol{\Gamma}_0^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Gamma}_0^{-1}$$

and

$$\mathbf{P} = \mathbf{S} - \mathbf{S} \mathbf{Z} (\boldsymbol{\Gamma}^{-1} + \mathbf{Z}^T \mathbf{S} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{S}$$

evaluated at the τ th cycle of iteration.

3.5.1 Algorithm

Given \mathbf{y} , \mathbf{X} and \mathbf{Z} :

Step 1: Initialize iteration by setting $\varphi_0^{(\tau)} = \varphi_0^{(0)}$ and $\varphi_m^{(\tau)} = \varphi_m^{(0)}$, $m = 1, \dots, c$.

Step 2: Solve equation (3.4.6) for $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\gamma}_*^{(0)}$, then calculate $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\Gamma}^{(0)} \boldsymbol{\gamma}_*^{(0)}$.

Step 3: Calculate $\mathbf{T}^{(0)}$ from equation (3.5.5).

Step 4: Determine $\varphi_0^{(1)}, \varphi_1^{(1)}, \dots, \varphi_c^{(1)}$ from equations (3.5.4).

Step 5: Test for convergence: If $|\varphi_0^{(\tau+1)} - \varphi_0^{(\tau)}| \rightarrow 0$ and $|\varphi_m^{(\tau+1)} - \varphi_m^{(\tau)}| \rightarrow 0$, terminate iteration. Set $\varphi_0^{(\tau+1)} = \hat{\varphi}_0$ and $\varphi_m^{(\tau+1)} = \hat{\varphi}_m$. With these REML estimates, compute $\beta^{(\tau+1)} = \hat{\beta}$, $\gamma^{(\tau+1)} = \hat{\gamma}$ and $\mathbf{T}^{(\tau+1)} = \hat{\mathbf{T}}$. Otherwise, repeat steps 2, 3 and 4, by incrementing τ by 1.

The (REML) information matrix is determined at $\hat{\varphi}_0, \hat{\varphi}_1, \dots, \hat{\varphi}_c$ and has the form

$$\mathcal{I}(\hat{\varphi}^*) = \frac{1}{2} \begin{pmatrix} \frac{n - q + \text{tr}(\hat{\mathbf{T}}^2)}{\hat{\varphi}_0^2} & \frac{\text{tr}(\hat{\mathbf{T}}_{mm}) - \sum_{l=1}^c \text{tr}(\hat{\mathbf{T}}_{ml}\hat{\mathbf{T}}_{lm})}{\hat{\varphi}_0\hat{\varphi}_m} \\ \text{symmetry} & \frac{\delta_{mk}[q_m - 2\text{tr}(\hat{\mathbf{T}}_{mm})] + \text{tr}(\hat{\mathbf{T}}_{mk}\hat{\mathbf{T}}_{km})}{\hat{\varphi}_k\hat{\varphi}_m} \end{pmatrix}$$

where $\delta_{mk} = \begin{cases} 1 & \text{for } m = k \\ 0 & \text{for } m \neq k \end{cases}$ (Searle *et al* 1992, pp.282–286).

3.6 Penalized likelihood

Suppose there exists a twice-differentiable curve \mathbb{C} defined on an interval $[a, b]$ of \mathbb{R} . The residual sum of squares

$$\sum_{i=1}^n [y_i - \mathbb{C}(t_i)]^2 \quad (3.6.1)$$

is an important criterion for assessing the goodness-of-fit of \mathbb{C} to the data y_i , $i = 1, \dots, n$. t_i is called a *knot* and satisfies $a < t_1 < \dots < t_n < b$. On that interval, \mathbb{C} may be fluctuating rapidly. Therefore, some smoothness conditions need to be placed on \mathbb{C} to study the more slowly varying ‘trend’ in the data, if there is any such trend. To compromise between goodness-of-fit and the ‘roughness’ of \mathbb{C} on $[a, b]$, there is a need to define a measure to quantify roughness. This quantity is given by

$$\int_a^b [\mathbb{C}''(t)]^2 dt \quad (3.6.2)$$

and is referred to as the *integrated squared second derivative*. The penalized sum of squares is defined as

$$S(\mathbb{C}) = \sum_{i=1}^n [y_i - \mathbb{C}(t_i)]^2 + \alpha \int_a^b [\mathbb{C}''(t)]^2 dt \quad (3.6.3)$$

where $\alpha (> 0)$ is called a *smoothing parameter* and

$$\alpha \int_a^b [\mathbb{C}''(t)]^2 dt \quad (3.6.4)$$

is the *roughness penalty* term. $\hat{\mathbb{C}}$, the penalized least squares estimator, minimizes the functional $S(\mathbb{C})$ over the class of all twice-differentiable functions \mathbb{C} . In the context of simple linear regression, any curve \mathbb{C} interpolating the data maximizes the likelihood function. The *penalized log-likelihood function* is given by

$$l_p(\mathbb{C}) = -\frac{1}{2}\varphi_0^{-1} \sum_{i=1}^n [y_i - \mathbb{C}(t_i)]^2 - \frac{1}{2}\lambda \int_a^b [\mathbb{C}''(t)]^2 dt \quad (3.6.5)$$

where $\lambda = \alpha\varphi_0^{-1}$, with φ_0 being the constant variance of the y_i 's. Maximizing $l_p(\mathbb{C})$ is equivalent to minimizing $S(\mathbb{C})$. $\hat{\mathbb{C}}$ is then called the *maximum penalized likelihood estimator* (MPLE) (Green & Silverman 1994, pp.4-5, p.10, p.50 and p.98).

By analogy to Theorem 2.1 and equation (4.4) of Green & Silverman (1994, p.13 and p.65), but with weights equal to unity for all y_i 's, it can be deduced that, for LMMs, the term $\gamma^T \Gamma^{-1} \gamma$ in the 'joint' log-likelihood function (ignoring constant terms)

$$(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma)^T \Gamma_0^{-1} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma) + \gamma^T \Gamma^{-1} \gamma$$

(Robinson 1991) is a *penalty function*. This function prevents arbitrary values of γ from being selected and forces them to be near zero: This constraint is called a *shrinkage effect* (McCulloch & Searle 2001, p.283).

3.7 Some remarks

When fitting models with at least two dispersion components, predictors of random effects are the natural generalization of the concept of residuals. Outlying values of certain random effects will indicate that groups of data points may not fit a model. Thus, observing predictors can be regarded as a more sensitive test for detecting outliers than simply observing residuals. Moreover, Robinson (1991) assumes that the components of dispersion are known when estimating fixed effects and predicting random effects. The unbiasedness criterion holds when the dispersion components have to be estimated, under the condition that these components are translation-invariant and are even functions of \mathbf{y} . In other words, $\hat{\beta}$ and $\hat{\gamma}$ remain unbiased. In general, therefore, predictors of γ need not be changed. In contrast, their estimated precisions have to be modified. However, in practical situations, Robinson (1991) states that this modification is at times ignored or calculations based on the best point estimate of the dispersion components are interpreted conservatively.

See Searle (1995) and Searle *et al* (1992, Section 6.4, pp.242–243) for comments on the caveats pertaining numerical optimization when estimating components of dispersion.

Chapter 4

Some aspects of generalized linear mixed models

4.1 Introduction

In Chapters 2 and 3, we provided a review of GLMs and LMMs, respectively. These methodologies are now extended to form a new class of models known as generalized linear mixed models (GLMMs). These models are useful for the analysis of overdispersed and correlated discrete data. In a GLMM, responses are assumed to be conditionally independent given a vector of Gaussian distributed random effects. The assumption of moment conditions suffice for the observations.

In this chapter, we provide some aspects of GLMMs, as briefly described in Chapter 1.

4.2 Generalized linear mixed models

Let $\mathbf{y} = (y_1, \dots, y_n)^T$ denote a vector of correlated discrete responses. This supposition immediately relaxes the assumption of independence among responses in GLMs and that of normality in LMMs. Furthermore, it is assumed that the (conditional) probability function of each y_i need not belong to the exponential family for $i = 1, \dots, n$. Thus, Wedderburn's (1974) concept of quasi-likelihood, coupled with GLM and LMM methodologies, can be adopted for the statistical modelling of the random vector \mathbf{y} . Given a vector of Gaussian distributed random effects $\boldsymbol{\gamma}$, responses are assumed to be (conditionally) independent with conditional means

$$E(y_i|\boldsymbol{\gamma}) = \mu_i^{\boldsymbol{\gamma}} = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma}) \quad (4.2.1a)$$

and conditional variances

$$\text{Var}(y_i|\boldsymbol{\gamma}) = \phi u_i^{-1} V(\mu_i^{\boldsymbol{\gamma}}) \quad (4.2.1b)$$

where $V(\mu_i^{\boldsymbol{\gamma}})$ is referred to as the (*conditional*) *variance function*. The linear predictor of a GLM is extended to include $\boldsymbol{\gamma}$ and is expressed as

$$\eta_i^{\boldsymbol{\gamma}} = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma} \quad (4.2.2a)$$

or, more compactly, as

$$\begin{aligned}\boldsymbol{\eta}^\gamma &= g(\boldsymbol{\mu}^\gamma) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}\end{aligned}\tag{4.2.2b}$$

Following the rationale behind equations (3.2.5), equation (4.2.2b) can be rewritten as

$$\boldsymbol{\eta}^\gamma = \mathbf{X}\boldsymbol{\beta} + \sum_{m=1}^c \mathbf{Z}_m \boldsymbol{\gamma}_m\tag{4.2.2c}$$

In equations (4.2.1) and (4.2.2), the link function $g(\cdot)$, scale factor ϕ , known prior weights u_i , dimensions of model matrices \mathbf{X} (of full rank p) and \mathbf{Z} , lengths of fixed effects vector $\boldsymbol{\beta}$ and random effects vector $\boldsymbol{\gamma}$ (hence of $\boldsymbol{\gamma}_m$) are as appropriately defined in Chapters 2 and 3. See Section 3.2 for the distributional and statistical properties of $\boldsymbol{\gamma}_m$, $m = 1, \dots, c$. Clearly, $\boldsymbol{\eta}^\gamma$ is of order $n \times 1$, in conformity with that of \mathbf{y} . The class of models described herein is an extension of GLMs and LMMs, and is referred to as *generalized linear mixed models* (GLMMs) (Breslow 2003; Breslow & Clayton 1993; Clayton 1992; Lin & Breslow 1996a).

Let $\boldsymbol{\psi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\varphi}^T)$, with ϕ constrained at unity, be the parameter vector of interest. The objective function for estimating $\boldsymbol{\psi}$ has the form

$$\begin{aligned}L(\boldsymbol{\psi}) &= \exp[l(\boldsymbol{\beta}, \boldsymbol{\varphi})] \propto |\Gamma(\boldsymbol{\varphi})|^{-1/2} \int_{\mathbb{R}^q} \exp \left[\sum_{i=1}^n Q_i(y_i; \mu_i^\gamma) - \frac{1}{2} \boldsymbol{\gamma}^T \Gamma^{-1}(\boldsymbol{\varphi}) \boldsymbol{\gamma} \right] d\boldsymbol{\gamma} \\ &\propto |\Gamma(\boldsymbol{\varphi})|^{-1/2} \int_{\mathbb{R}^q} \exp \left[-\frac{1}{2} \sum_{i=1}^n \left[-2 \int_{y_i}^{\mu_i^\gamma} \frac{u_i(y_i - s)}{\phi V(s)} ds \right] - \frac{1}{2} \boldsymbol{\gamma}^T \Gamma^{-1}(\boldsymbol{\varphi}) \boldsymbol{\gamma} \right] d\boldsymbol{\gamma} \\ &= |2\pi|^{-q/2} |\Gamma(\boldsymbol{\varphi})|^{-1/2} \int_{\mathbb{R}^q} \exp \left[-\frac{1}{2\phi} \sum_{i=1}^n d_i(y_i; \mu_i^\gamma) - \frac{1}{2} \boldsymbol{\gamma}^T \Gamma^{-1}(\boldsymbol{\varphi}) \boldsymbol{\gamma} \right] d\boldsymbol{\gamma}\end{aligned}\tag{4.2.3}$$

where

$$d_i(y_i; \mu_i^\gamma) = -2 \int_{y_i}^{\mu_i^\gamma} \frac{u_i(y_i - s)}{V(s)} ds$$

is the (*conditional weighted quasi*-)deviance function. Equation (4.2.3) is referred to as the *integrated quasi-likelihood function* (IQL) (Breslow 2003;

Breslow & Clayton 1993). An alternative formulation to equation (4.2.3) is given by Lin & Breslow (1996a) and has the form

$$L(\boldsymbol{\psi}) = \exp[l(\boldsymbol{\beta}, \boldsymbol{\varphi})]$$

$$\propto |\Gamma(\boldsymbol{\varphi})|^{-1/2} \int_{\mathbb{R}^q} \exp \left[\sum_{i=1}^n l_i(\boldsymbol{\beta}; \boldsymbol{\gamma}) - \frac{1}{2} \boldsymbol{\gamma}^T \Gamma^{-1}(\boldsymbol{\varphi}) \boldsymbol{\gamma} \right] d\boldsymbol{\gamma}$$

where

$$l_i(\boldsymbol{\beta}; \boldsymbol{\gamma}) \propto \int_{y_i}^{\mu_i^{\boldsymbol{\gamma}}} \frac{u_i(y_i - s)}{\phi V(s)} ds \quad (4.2.4)$$

is the *conditional log-quasi-likelihood* of $\boldsymbol{\beta}|\boldsymbol{\gamma}$.

$L(\boldsymbol{\psi})$ in equation (4.2.3) involves a q -dimensional integral. A full likelihood analysis when $c = 1$ has been performed by Breslow & Lin (1995) in a study of a series of matched pairs of binary outcomes, where a one-dimensional Gaussian quadrature is feasible (cf. Lin & Breslow 1996a). In contrast, in complicated situations, the IQL becomes intractable (Breslow & Clayton 1993) – that is, an analytic solution to $L(\boldsymbol{\psi})$ does not often exist (Zhu & Lee 2003). However, the Laplace expansion, employed by Breslow & Clayton (1993), can be used to approximate $L(\boldsymbol{\psi})$. Under some adjustments, this approximation leads to estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$.

4.3 Penalized quasi-likelihood function

The Laplace approximation to the IQL has the form

$$\begin{aligned}
 l &\equiv l(\psi) \\
 &\approx -\frac{q}{2} \log |2\pi| - \frac{1}{2} \log |\Gamma| \\
 &\quad + \log \int_{\mathbb{R}^q} \exp \left[\sum_{i=1}^n Q_i(y_i; \mu_i^\gamma) - \frac{1}{2} \gamma^T \Gamma^{-1}(\varphi) \gamma \right] d\gamma \\
 &= -\frac{q}{2} \log |2\pi| - \frac{1}{2} \log |\Gamma| + \log \int_{\mathbb{R}^q} \exp[h(\gamma)] d\gamma
 \end{aligned} \tag{4.3.1}$$

where

$$h(\gamma) = \sum_{i=1}^n Q_i(y_i; \mu_i^\gamma) - \frac{1}{2} \gamma^T \Gamma^{-1}(\varphi) \gamma \tag{4.3.2}$$

Now,

$$\log \int_{\mathbb{R}^q} \exp[h(\gamma)] d\gamma \approx \frac{q}{2} \log |2\pi| + h(\gamma) + \frac{\partial h(\gamma)}{\partial \gamma} - \frac{1}{2} \log \left| \frac{-\partial^2 h(\gamma)}{\partial \gamma \partial \gamma^T} \right| \tag{4.3.3}$$

Substituting equation (4.3.3) into equation (4.3.1) yields

$$l \approx -\frac{1}{2} \log |\Gamma| + h(\gamma) - \frac{1}{2} \log \left| \frac{-\partial^2 h(\gamma)}{\partial \gamma \partial \gamma^T} \right| \tag{4.3.4}$$

where

$$\begin{aligned}
\left. \frac{\partial h(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \right|_{\boldsymbol{\gamma}=\tilde{\boldsymbol{\gamma}}} &= \sum_{i=1}^n \frac{\partial Q_i(y_i; \mu_i^\gamma)}{\partial \gamma_m} - \frac{1}{2} \frac{\partial(\boldsymbol{\gamma}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \\
&= \sum_{i=1}^n \frac{\partial Q_i(y_i; \mu_i^\gamma)}{\partial \mu_i^\gamma} \frac{\partial \mu_i^\gamma}{\partial \eta_i^\gamma} \frac{\partial \eta_i^\gamma}{\partial \gamma_m} - \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma} \\
&= \sum_{i=1}^n \frac{u_i(y_i - \mu_i^\gamma)}{\phi V(\mu_i^\gamma)} \frac{1}{g'(\mu_i^\gamma)} z_{im} - \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma} \\
&= \sum_{i=1}^n \frac{(y_i - \mu_i^\gamma) g'(\mu_i^\gamma)}{\phi u_i^{-1} V(\mu_i^\gamma) [g'(\mu_i^\gamma)]^2} z_{im} - \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma} \\
&= \mathbf{Z}^T \mathbf{W} \Delta (\mathbf{y} - \boldsymbol{\mu}^\gamma) - \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma} \\
&= 0
\end{aligned} \tag{4.3.5}$$

with $\tilde{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\varphi})$ being the solution to equation (4.3.5) that minimizes $h(\boldsymbol{\gamma})$. By contrast to equation (2.4.6), we define $w_i^{-1} = \phi u_i^{-1} V(\mu_i^\gamma) [g'(\mu_i^\gamma)]^2$ to be the diagonal elements of \mathbf{W} . Δ is as defined in Section 2.4, but with μ_i replaced by μ_i^γ ($i = 1, \dots, n$; $m = 1, \dots, c$).

The matrix (of order q) of second-order partial derivatives of $h(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ is given by

$$\left. \frac{\partial^2 h(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \right|_{\boldsymbol{\gamma}=\tilde{\boldsymbol{\gamma}}} = \frac{\partial}{\partial \gamma_k} \left(\sum_{i=1}^n \frac{u_i(y_i - \mu_i^\gamma)}{\phi V(\mu_i^\gamma) g'(\mu_i^\gamma)} z_{im} \right) - \frac{\partial(\boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$$

$$\begin{aligned}
&= \frac{\partial}{\partial \mu_i^\gamma} \left(\sum_{i=1}^n \frac{u_i(y_i - \mu_i^\gamma)}{\phi V(\mu_i^\gamma) g'(\mu_i^\gamma)} \right) \frac{\partial \mu_i^\gamma}{\partial \eta_i^\gamma} \frac{\partial \eta_i^\gamma}{\partial \gamma_k} z_{im} - \Gamma^{-1} \\
&= \left[\frac{1}{\phi V(\mu_i^\gamma) g'(\mu_i^\gamma)} \sum_{i=1}^n \frac{\partial(y_i - \mu_i^\gamma) u_i}{\partial \mu_i^\gamma} \right. \\
&\quad \left. + \sum_{i=1}^n u_i(y_i - \mu_i^\gamma) \frac{\partial}{\partial \mu_i^\gamma} \left(\frac{1}{\phi V(\mu_i^\gamma) g'(\mu_i^\gamma)} \right) \right] \frac{1}{g'(\mu_i^\gamma)} z_{im} z_{ik} - \Gamma^{-1} \quad (4.3.6a)
\end{aligned}$$

$$= - \sum_{i=1}^n \frac{z_{ik} z_{im} u_i}{\phi V(\mu_i^\gamma) [g'(\mu_i^\gamma)]^2} - \Gamma^{-1} \quad (4.3.6b)$$

since, for canonical links, the term

$$\text{REM} = \left[\sum_{i=1}^n u_i(y_i - \mu_i^\gamma) \frac{\partial}{\partial \mu_i^\gamma} \left(\frac{1}{\phi V(\mu_i^\gamma) g'(\mu_i^\gamma)} \right) \right] \frac{1}{g'(\mu_i^\gamma)} z_{im} z_{ik}$$

reduces to zero due to the fact that

$$V(\mu_i^\gamma) g'(\mu_i^\gamma) = 1 \text{ implies } \frac{\partial (V(\mu_i^\gamma) g'(\mu_i^\gamma))}{\partial \mu_i^\gamma} = 0$$

Furthermore, $E(\text{REM}) = 0$, and consequently REM is of lower order¹ in probability as a function of n than the terms in equation (4.3.6b), thus negligible. We then have

$$\left. \frac{\partial^2 h(\gamma)}{\partial \gamma \partial \gamma^T} \right|_{\gamma=\hat{\gamma}} \approx -(\mathbf{Z}^T \mathbf{W} \mathbf{Z} + \Gamma^{-1}) \quad (4.3.7a)$$

$$= -\Gamma^{-1} (\mathbf{I} + \mathbf{Z}^T \mathbf{W} \mathbf{Z} \Gamma) \quad (4.3.7b)$$

¹Without loss of generality, REM is of stochastic order $O_p(n^{1/2})$ (Booth & Hobert 1998).

Substitution of equation (4.3.7b) into equation (4.3.4) gives

$$\begin{aligned}
l_{\gamma=\tilde{\gamma}} &\approx -\frac{1}{2} \log |\Gamma| + \sum_{i=1}^n Q_i(y_i; \mu_i^{\tilde{\gamma}}) - \frac{1}{2} \gamma^T \Gamma^{-1}(\varphi) \gamma \\
&\quad - \frac{1}{2} \log |\Gamma^{-1}(\mathbf{I} + \mathbf{Z}^T \mathbf{W} \mathbf{Z} \Gamma)| \\
&= \sum_{i=1}^n Q_i(y_i; \mu_i^{\tilde{\gamma}}) - \frac{1}{2} \gamma^T \Gamma^{-1}(\varphi) \gamma \\
&\quad - \frac{1}{2} \log |\Gamma| - \frac{1}{2} \log |\mathbf{I} + \mathbf{Z}^T \mathbf{W} \mathbf{Z} \Gamma| + \frac{1}{2} \log |\Gamma| \\
&= \sum_{i=1}^n Q_i(y_i; \mu_i^{\tilde{\gamma}}) - \frac{1}{2} \gamma^T \Gamma^{-1}(\varphi) \gamma - \frac{1}{2} \log |\mathbf{I} + \mathbf{Z}^T \mathbf{W} \mathbf{Z} \Gamma| \\
&= -\frac{1}{2\phi} \sum_{i=1}^n d_i(y_i; \mu_i^{\tilde{\gamma}}) - \frac{1}{2} \gamma^T \Gamma^{-1}(\varphi) \gamma - \frac{1}{2} \log |\mathbf{I} + \mathbf{Z}^T \mathbf{W} \mathbf{Z} \Gamma| \quad (4.3.8)
\end{aligned}$$

where $\tilde{\gamma}$ maximizes the sums of the first two terms in equation (4.3.8). Under the assumption that the weights w_i (at $\tilde{\gamma}$) vary slowly (or not at all) as a function of the mean, equation (4.3.8) becomes

$$l_{\gamma=\tilde{\gamma}} = -\frac{1}{2\phi} \sum_{i=1}^n d_i(y_i; \mu_i^{\tilde{\gamma}}) - \frac{1}{2} \gamma^T \Gamma^{-1}(\varphi) \gamma \quad (4.3.9)$$

where β is chosen to maximize the first term in equation (4.3.9) such that l is maximized by the joint solutions $(\hat{\beta}, \hat{\gamma}) = (\hat{\beta}(\varphi), \hat{\gamma}(\varphi))$, with $\hat{\gamma}(\varphi) = \tilde{\gamma}(\hat{\beta}(\varphi))$ and φ being held fixed. Equation (4.3.9) is referred to as the (log-)penalized quasi-likelihood function (PQL) (Breslow & Clayton 1993; Lin & Breslow 1996a; McCulloch & Searle 2001, pp.281–283). The assumption of slowly varying weights require that the components of dispersion are relatively small (Engel & Keen 1994).

4.4 Fisher scoring for maximum PQL estimation of fixed effects and prediction of random effects

The quasi-score equations for β and γ are given by

$$\left. \frac{\partial l}{\partial \beta} \right|_{\gamma=\tilde{\gamma}} = \mathbf{X}^T \mathbf{W} \Delta(\mathbf{y} - \boldsymbol{\mu} \gamma) \quad (4.4.1a)$$

and

$$\left. \frac{\partial l}{\partial \gamma} \right|_{\gamma=\tilde{\gamma}} = \mathbf{Z}^T \mathbf{W} \Delta(\mathbf{y} - \boldsymbol{\mu} \gamma) - \Gamma^{-1} \gamma \quad (4.4.1b)$$

respectively. Fisher scoring for estimating β and predicting γ has the form

$$\begin{aligned} \begin{pmatrix} \beta^{(\tau+1)} \\ \gamma^{(\tau+1)} \end{pmatrix} &= \begin{pmatrix} \beta^{(\tau)} \\ \gamma^{(\tau)} \end{pmatrix} + \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} & \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} & \Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z} \end{pmatrix}^{-1} \\ &\times \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)}(\mathbf{y} - \boldsymbol{\mu} \gamma^{(\tau)}) \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)}(\mathbf{y} - \boldsymbol{\mu} \gamma^{(\tau)}) - \Gamma^{-1} \gamma^{(\tau)} \end{pmatrix} \end{aligned} \quad (4.4.2)$$

where

$$\begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} & \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} & \Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z} \end{pmatrix} \quad (4.4.3)$$

is expression (3.4.5) with Γ_0^{-1} replaced by $\mathbf{W}^{(\tau)}$. Equations (4.4.2) and (4.4.3) hold at $\gamma = \tilde{\gamma}$.

Multiplying equation (4.4.2) by the matrix of order $p+q$ in expression (4.4.3) gives:

$$\begin{aligned}
& \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} & \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} & \Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(\tau+1)} \\ \boldsymbol{\gamma}^{(\tau+1)} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} & \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} & \Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(\tau)} \\ \boldsymbol{\gamma}^{(\tau)} \end{pmatrix} \\
&+ \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu} \boldsymbol{\gamma}^{(\tau)}) \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu} \boldsymbol{\gamma}^{(\tau)}) - \Gamma^{-1} \boldsymbol{\gamma}^{(\tau)} \end{pmatrix}
\end{aligned}$$

We then have

$$\begin{aligned}
& \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} \boldsymbol{\beta}^{(\tau+1)} + \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \boldsymbol{\gamma}^{(\tau+1)} \\
&= \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} \boldsymbol{\beta}^{(\tau)} + \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \boldsymbol{\gamma}^{(\tau)} + \mathbf{X}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu} \boldsymbol{\gamma}^{(\tau)}) \quad (4.4.4a)
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} \boldsymbol{\beta}^{(\tau+1)} + (\Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z}) \boldsymbol{\gamma}^{(\tau+1)} \\
&= \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} \boldsymbol{\beta}^{(\tau)} + (\Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z}) \boldsymbol{\gamma}^{(\tau)} \\
&+ \mathbf{Z}^T \mathbf{W}^{(\tau)} \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu} \boldsymbol{\gamma}^{(\tau)}) - \Gamma^{-1} \boldsymbol{\gamma}^{(\tau)} \quad (4.4.4b)
\end{aligned}$$

Equations (4.4.4) can be simplified to

$$\begin{aligned}
& \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} \boldsymbol{\beta}^{(\tau+1)} + \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \boldsymbol{\gamma}^{(\tau+1)} \\
&= \mathbf{X}^T \mathbf{W}^{(\tau)} [\mathbf{X} \boldsymbol{\beta}^{(\tau)} + \mathbf{Z} \boldsymbol{\gamma}^{(\tau)} + \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu} \boldsymbol{\gamma}^{(\tau)})] \\
&= \mathbf{X}^T \mathbf{W}^{(\tau)} \boldsymbol{\zeta}^{(\tau)} \quad (4.4.5a)
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} \boldsymbol{\beta}^{(\tau+1)} + (\Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z}) \boldsymbol{\gamma}^{(\tau+1)} \\
&= \mathbf{Z}^T \mathbf{W}^{(\tau)} [\mathbf{X} \boldsymbol{\beta}^{(\tau)} + \mathbf{Z} \boldsymbol{\gamma}^{(\tau)} + \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{\boldsymbol{\gamma}^{(\tau)}})] \\
&= \mathbf{Z}^T \mathbf{W}^{(\tau)} \boldsymbol{\zeta}^{(\tau)} \tag{4.4.5b}
\end{aligned}$$

where

$$\boldsymbol{\zeta}^{(\tau)} = \mathbf{X} \boldsymbol{\beta}^{(\tau)} + \mathbf{Z} \boldsymbol{\gamma}^{(\tau)} + \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{\boldsymbol{\gamma}^{(\tau)}}) \tag{4.4.6a}$$

exists at $\tilde{\boldsymbol{\gamma}}$ and $\boldsymbol{\epsilon}^{(\tau)} = \Delta^{(\tau)} (\mathbf{y} - \boldsymbol{\mu}^{\boldsymbol{\gamma}^{(\tau)}})$. Equations (4.4.5) can be rewritten as

$$\begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} & \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} & \Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(\tau+1)} \\ \boldsymbol{\gamma}^{(\tau+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \boldsymbol{\zeta}^{(\tau)} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \boldsymbol{\zeta}^{(\tau)} \end{pmatrix} \tag{4.4.6b}$$

Therefore,

$$\begin{pmatrix} \boldsymbol{\beta}^{(\tau+1)} \\ \boldsymbol{\gamma}^{(\tau+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{X} & \mathbf{X}^T \mathbf{W}^{(\tau)} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{X} & \Gamma^{-1} + \mathbf{Z}^T \mathbf{W}^{(\tau)} \mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}^T \mathbf{W}^{(\tau)} \boldsymbol{\zeta}^{(\tau)} \\ \mathbf{Z}^T \mathbf{W}^{(\tau)} \boldsymbol{\zeta}^{(\tau)} \end{pmatrix} \tag{4.4.7}$$

The linked response $\boldsymbol{\zeta}^{(\tau)}$ in equation (4.4.6a) has the form of a Gaussian LMM equation. Thus, at $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$,

$$E(\zeta_i^{(\tau)}) = \mathbf{x}_i^T \boldsymbol{\beta}^{(\tau)} \tag{4.4.8a}$$

and

$$\begin{aligned}
\text{Var}(\zeta_i^{(\tau)}) &= \text{Var}(\epsilon_i^{(\tau)}) + \mathbf{z}_i^T \text{Var}(\boldsymbol{\gamma}) \mathbf{z}_i \\
&= \text{Var}(y_i - \mu_i^{\boldsymbol{\gamma}^{(\tau)}}) [g'(\mu_i^{\boldsymbol{\gamma}^{(\tau)}})]^2 + \mathbf{z}_i^T \text{Var}(\boldsymbol{\gamma}) \mathbf{z}_i \\
&= \phi u_i^{-1} V(\mu_i^{\boldsymbol{\gamma}^{(\tau)}}) [g'(\mu_i^{\boldsymbol{\gamma}^{(\tau)}})]^2 + \mathbf{z}_i^T \text{Var}(\boldsymbol{\gamma}) \mathbf{z}_i \\
&= w_i^{-1(\tau)} + \mathbf{z}_i^T \text{Var}(\boldsymbol{\gamma}) \mathbf{z}_i \tag{4.4.8b}
\end{aligned}$$

for $i = 1, \dots, n$. Equations (4.4.8) can be rewritten more compactly as

$$E(\zeta^{(\tau)}) = \mathbf{X}\beta^{(\tau)} \quad (4.4.9a)$$

and

$$\text{Var}(\zeta^{(\tau)}) = \mathbf{W}^{-1(\tau)} + \mathbf{Z}\Gamma\mathbf{Z}^T = \Psi \quad (4.4.9b)$$

Thus, $\zeta^{(\tau)} \sim N(\mathbf{X}\beta^{(\tau)}, \Psi)$ and $\epsilon \sim N(\mathbf{0}, \mathbf{W}^{-1(\tau)})$. Since $\zeta^{(\tau)}$ is approximately Gaussian distributed and equation (4.4.6b) corresponds to equation (3.4.4), with Γ_0^{-1} replaced by $\mathbf{W}^{(\tau)}$, we have

$$\begin{aligned} \hat{\beta} &= \beta^{(\tau+1)} \\ &= (\mathbf{X}^T\Psi^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Psi^{-1}\zeta^{(\tau)} \end{aligned} \quad (4.4.10a)$$

for the maximum PQL estimator for fixed effects, with ϕ and φ held fixed, and

$$\hat{\gamma} = \tilde{\gamma}^{(\tau+1)} = \Gamma\mathbf{Z}^T\Psi^{-1}(\zeta^{(\tau)} - \mathbf{X}\hat{\beta}) \quad (4.4.10b)$$

for the maximum PQL predictor for random effects. Equations (4.4.10) correspond to equations (3.5.2b) and (3.5.3), respectively (with \mathbf{y} replaced by $\zeta^{(\tau)}$).

We have illustrated that the Fisher scoring described in Section 2.4.1 for the evaluation of the estimator of fixed effect parameters can be extended to obtain predictors of random effects via IWLS, with ϕ and φ held fixed throughout the iterative procedure. Alternatively to equation (4.4.7), the following system of equations may be used to iteratively evaluate $\hat{\beta}$ and $\hat{\gamma}$, as cautioned in Section 3.4:

$$\begin{pmatrix} \mathbf{X}^T\mathbf{W}^{(\tau)}\mathbf{X} & \mathbf{X}^T\mathbf{W}^{(\tau)}\mathbf{Z}\Gamma \\ \mathbf{Z}^T\mathbf{W}^{(\tau)}\mathbf{X} & \mathbf{I} + \mathbf{Z}^T\mathbf{W}^{(\tau)}\mathbf{Z}\Gamma \end{pmatrix} \begin{pmatrix} \beta^{(\tau+1)} \\ \gamma_*^{(\tau+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{W}^{(\tau)}\zeta^{(\tau)} \\ \mathbf{Z}^T\mathbf{W}^{(\tau)}\zeta^{(\tau)} \end{pmatrix} \quad (4.4.11)$$

where $\gamma = \Gamma\gamma_*$, and thus with $\hat{\beta}, \hat{\gamma} = \Gamma\hat{\gamma}_*$. See also Breslow & Clayton (1993) and Green & Silverman (1994, pp.106–107). Littell *et al* (1996, p.435) refer to equation (4.4.6b) as the *generalized mixed model equations*.

4.5 Corrected profile quasi-likelihood function for the estimation of components of dispersion

Breslow & Clayton (1993) motivate an approximate profile log-quasi-likelihood function for the estimation of dispersion components by substituting the maximized PQL value from equation (4.3.9) into the Laplace-approximated IQL given by equation (4.3.8), where $\mathbf{W} = \mathbf{W}(\hat{\boldsymbol{\beta}}(\varphi), \hat{\boldsymbol{\gamma}}(\varphi))$. To draw a close correspondence with the normal theory LM profile likelihood for $\boldsymbol{\zeta}^{(\tau)}$, they replace the (conditional weighted quasi-) deviance increment $\sum_{i=1}^n d_i(y_i; \mu_i^{\boldsymbol{\gamma}})$

by the Pearson's X^2 -statistic $\sum_{i=1}^n u_i(y_i - \mu_i^{\boldsymbol{\gamma}})^2 [V(\mu_i^{\boldsymbol{\gamma}})]^{-1}$ and ignore the dependence of $\mathbf{W}_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}}$ on φ . Furthermore, to take into account the loss of degrees-of-freedom incurred due to the estimation of $\boldsymbol{\beta}$, Breslow & Clayton (1993) parallel the profile likelihood correction of Cox & Reid (1987) for normal theory LM and quote the REML formulation of Patterson & Thompson (1971) for estimating φ . However, the orthogonality property of $(\boldsymbol{\beta}, \varphi)$ and the information matrix $\mathbf{X}^T \boldsymbol{\Psi}^{-1} \mathbf{X}$ of $\hat{\boldsymbol{\beta}}(\varphi)$ are not exact for GLMMs. Moreover, equation (13) of Breslow & Clayton (1993) cannot be employed as an objective function for solving their equations (14) and (15) since the dependence of $\mathbf{W}_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}}$ on φ is ignored in the evaluation of first-order partial derivatives of $\text{Var}(\boldsymbol{\zeta}^{(\tau)})$ with respect to φ_m , $m = 1, \dots, c$.

We adopt the REML formulation for estimating φ as summarized in Section 3.3. The restricted profile log-quasi-likelihood function for $\boldsymbol{\zeta}^{(\tau)}$ has the form (ignoring constant terms)

$$l(\boldsymbol{\Psi}) = -\frac{1}{2} \log |\mathbf{K}^T \boldsymbol{\Psi} \mathbf{K}| - \frac{1}{2} \boldsymbol{\zeta}^{(\tau)T} \mathbf{K} (\mathbf{K}^T \boldsymbol{\Psi} \mathbf{K})^{-1} \mathbf{K}^T \boldsymbol{\zeta}^{(\tau)} \quad (4.5.1)$$

where $\boldsymbol{\Psi} = \mathbf{W}_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}}^{-1} + \mathbf{Z} \boldsymbol{\Gamma} \mathbf{Z}^T$. The profile (quasi-)score equations for φ are given by

$$\frac{\partial l}{\partial \varphi} = -\frac{1}{2} \text{tr}(\mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T) + \frac{1}{2} \boldsymbol{\zeta}^{(\tau)T} \mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T \mathbf{P} \boldsymbol{\zeta}^{(\tau)} \quad (4.5.2)$$

and the information matrix by

$$\mathcal{I}(\varphi) = \frac{1}{2} \text{tr}(\mathbf{P} \mathbf{Z}_k \mathbf{Z}_k^T \mathbf{P} \mathbf{Z}_m \mathbf{Z}_m^T) \quad (4.5.3)$$

where

$$\mathbf{P} = \Psi^{-1} - \Psi^{-1} \mathbf{X} (\mathbf{X}^T \Psi^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Psi^{-1} \quad (4.5.4)$$

Fisher scoring for solutions of φ then has the form

$$\varphi^{(\tau+1)} = \varphi^{(\tau)} + [\mathcal{I}(\varphi^{(\tau)})]^{-1} \left. \frac{\partial l}{\partial \varphi} \right|_{\varphi = \varphi^{(\tau)}} \quad (4.5.5)$$

where ϕ is held fixed at unity. The scale factor may be estimated if so desired (see Wolfinger & O'Connell 1993).

4.6 Computing PQL estimators

We present an iterative procedure for the estimation of fixed effect parameters and components of dispersion, and the prediction of random effects, by following Breslow & Clayton (1993) and Green & Silverman (1994, p.107).

4.6.1 Algorithm

Step 0: Generating initial values

Under the assumption that the y_i 's are independent – that is, $\Gamma = \mathbf{0}$, a preliminary estimate of $\beta (= \beta^{(0)})$ is obtained via the GLM iterative method of Section 2.4. Thereafter, residuals are used to calculate initial values of $\varphi (= \varphi^{(0)})$. ϕ is fixed at 1. $\gamma (= \gamma^{(0)})$ is calculated from equation (4.4.10b).

Given $\phi = 1$, \mathbf{y} , \mathbf{X} and \mathbf{Z} :

Loop I: Evaluation of fixed effect estimators and predictors of random effects

Step 1(I): Initialize iteration with starting values $\beta^{(\tau)} = \beta^{(0)}$, $\gamma^{(\tau)} = \gamma^{(0)}$, and $\varphi^{(\tau)} = \varphi^{(0)}$ (held fixed). Determine $\zeta^{(\tau)} = \zeta^{(0)}$ (equation 4.4.6a) and $\mathbf{W}^{(\tau)} = \mathbf{W}^{(0)}$.

Step 2(I): Calculate new estimates $\beta^{(1)}$, then $\gamma^{(1)}$, via equations (4.4.10).

Step 3(I): Test for convergence: If $|\boldsymbol{\beta}^{(\tau+1)} - \boldsymbol{\beta}^{(\tau)}| \rightarrow \mathbf{0}$
 $(\Rightarrow |\boldsymbol{\gamma}^{(\tau+1)} - \boldsymbol{\gamma}^{(\tau)}| \rightarrow \mathbf{0})$, terminate iteration. Otherwise, increase τ by 1 and, repeat steps 1(I) and 2(I).

Loop II: Estimation of components of dispersion

With $\boldsymbol{\zeta}^{(\tau+1)}$ and $\mathbf{W}^{(\tau+1)}$ (both held fixed) from Loop I, and $\boldsymbol{\varphi}^{(0)}$ as starting value, a new estimate $\boldsymbol{\varphi}^{(1)}$ of $\boldsymbol{\varphi}$ is calculated via a one-step Fisher scoring (equation (4.5.5)). The one-step procedure is halved if $\Gamma(\boldsymbol{\varphi}^{(1)})$ is not positive definite.

Loop I is repeated with $\boldsymbol{\varphi}^{(1)}$ held fixed and, $\boldsymbol{\beta}^{(\tau+1)}$ and $\boldsymbol{\gamma}^{(\tau+1)}$ as initial values. Cycling between loops I and II, and updating the appropriate parameters within their corresponding loops, will eventually lead to estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\varphi}}$ of $\boldsymbol{\varphi}$, and predictors $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$.

4.6.2 Restricted pseudo-likelihood procedure for fitting GLMMs

The PQL approach, described in the previous section, is a *subject-specific* (SS) approach where focus is on predictions of $\boldsymbol{\gamma}$ for individuals (subjects) and their relation to the population parameters $\boldsymbol{\beta}$. This is in contrast to the *population-averaged* (PA) approach where emphasis is on $\boldsymbol{\beta}$ and variability due to $\boldsymbol{\gamma}$ is treated as a nuisance parameter. See Wolfinger & O'Connell (1993). In this section, we briefly outline the *pseudo-likelihood* (PL) procedure, developed by these authors, for estimation of model parameters in GLMMs. They motivate their method via two analytic and one probabilistic approximations.

The GLMM equation is written as

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

such that $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$ and,

$$E(\boldsymbol{\epsilon}|\boldsymbol{\mu}) = \mathbf{0} \text{ and } \text{Cov}(\boldsymbol{\epsilon}|\boldsymbol{\mu}) = \Gamma_{0\boldsymbol{\mu}}^{1/2}\Gamma_0\Gamma_{0\boldsymbol{\mu}}^{1/2}$$

with $\Gamma_{0\boldsymbol{\mu}}$ being a diagonal matrix, evaluated at $\boldsymbol{\mu}$, of a known GLM variance function and Γ_0 is unknown. Variance modelling in $\boldsymbol{\gamma}$ and Γ is related to the

SS approach, whereas modelling in Γ_0 corresponds to the PA method. The first analytic approximation is based on the assumption that $\hat{\beta}$ and $\hat{\gamma}$ are known, with $\hat{\mu} = g^{-1}(\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma})$, such that

$$\tilde{\epsilon} = \mathbf{y} - \hat{\mu} - (g^{-1})'(\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma})(\mathbf{X}\beta - \mathbf{X}\hat{\beta} + \mathbf{Z}\gamma - \mathbf{Z}\hat{\gamma}) \quad (4.6.1)$$

where $(g^{-1})'(\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma})$ is an $n \times n$ diagonal matrix of the first derivative of g^{-1} evaluated at $\hat{\beta}$ and $\hat{\gamma}$.

$\tilde{\epsilon}$ in equation (4.6.1) is a first-order Taylor approximation to ϵ , where the expansion is about $\hat{\beta}$ and $\hat{\gamma}$. In the probabilistic approximation, it is assumed that

$$\tilde{\epsilon}|\beta, \gamma \sim N(\mathbf{0}, \Gamma_{0\mu}^{1/2} \Gamma_0 \Gamma_{0\mu}^{1/2}) \quad (4.6.2)$$

The last (analytic) approximation involves substituting $\hat{\mu}$ for μ in the covariance matrix. From the Gaussian approximation in (4.6.2), it follows that

$$g'(\hat{\mu})(\mathbf{y} - \hat{\mu})|\beta, \gamma \sim N[\mathbf{X}\beta - \mathbf{X}\hat{\beta} + \mathbf{Z}\gamma - \mathbf{Z}\hat{\gamma}, g'(\hat{\mu})\Gamma_{0\hat{\mu}}^{1/2} \Gamma_0 \Gamma_{0\hat{\mu}}^{1/2} g'(\hat{\mu})] \quad (4.6.3)$$

where

$$(g^{-1})'(\mathbf{x}_i^T \hat{\beta} + \mathbf{z}_i^T \hat{\gamma}) = \frac{1}{g'(\hat{\mu}_i)}$$

for each $i = 1, \dots, n$ ($g'(\hat{\mu})$ is an $n \times n$ diagonal matrix). Upon defining

$$\zeta = g(\hat{\mu}) + g'(\hat{\mu})(\mathbf{y} - \hat{\mu})$$

it follows from (4.6.3) that

$$\zeta|\beta, \gamma \sim N[\mathbf{X}\beta + \mathbf{Z}\gamma, g'(\hat{\mu})\Gamma_{0\hat{\mu}}^{1/2} \Gamma_0 \Gamma_{0\hat{\mu}}^{1/2} g'(\hat{\mu})] \quad (4.6.4)$$

The above approximations yield a weighted Gaussian LMM with $\widehat{\mathbf{W}} = \Gamma_{0\hat{\mu}}^{-1} [g'(\hat{\mu})]^{-2}$, which reduces to $\Gamma_{0\hat{\mu}}$ for canonical link functions.

The *restricted pseudo-likelihood* (REPL) function, which takes into account the loss of degrees-of-freedom due to the estimation of β , has the form

$$\begin{aligned}
 l &\equiv l(\Gamma_0^*, \Gamma^*; \zeta) \\
 &= -\frac{1}{2} \log |\Psi| - \frac{(n-p)}{2} \log |\rho^T \Psi^{-1} \rho| \\
 &\quad - \frac{1}{2} \log |\mathbf{X}^T \Psi^{-1} \mathbf{X}| - \frac{(n-p)}{2} \left[1 + \log \left(\frac{2\pi}{n-p} \right) \right] \quad (4.6.5)
 \end{aligned}$$

where Γ_0^* and Γ^* are reparameterized forms of Γ_0 and Γ , respectively. These reparameterizations are in terms of ratios of the scale factor ϕ .

In equation (4.6.5),

$$\Psi = \mathbf{W}^{-1/2} \Gamma_0^* \mathbf{W}^{-1/2} + \mathbf{Z} \Gamma^* \mathbf{Z}^T$$

and

$$\rho = \zeta - \mathbf{X}(\mathbf{X}^T \Psi^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Psi^{-1} \zeta$$

In general, estimates $\hat{\Gamma}^*$ and $\hat{\Gamma}_0^*$ are obtained by maximizing l numerically. Thereafter, estimates of β , γ and ϕ are evaluated from

$$\hat{\beta} = (\mathbf{X}^T \hat{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Psi}^{-1} \zeta \quad (4.6.6a)$$

$$\hat{\gamma} = \hat{\Gamma}^* \mathbf{Z}^T \hat{\Psi}^{-1} \hat{\rho} \quad (4.6.6b)$$

$$\hat{\phi} = \frac{1}{n-p} \hat{\rho}^T \hat{\Psi}^{-1} \hat{\rho} \quad (4.6.6c)$$

Updated values of Γ and Γ_0 are obtained by maximizing l again using estimates from equations (4.6.6). Iterating between equations (4.6.5) and (4.6.6) yields the REPL method. See Wolfinger & O'Connell (1993, pp.238–239) for their algorithm for fitting the above procedure. Furthermore, these authors state that their PL/REPL approach allows non-trivial covariance structures for both Γ and Γ_0 to be specified.

4.7 Bias in parameter estimators under PQL

The dependence of the weight matrix $\mathbf{W}_{\gamma=\tilde{\gamma}}^{(\tau)}$ (and hence of the linked response $\zeta_{\gamma=\tilde{\gamma}}^{(\tau)}$) on the components of dispersion φ is ignored in the heuristic argument leading to the corrected profile log-quasi-likelihood function for estimating φ (Breslow & Clayton 1993). Ignoring this dependence may have statistical consequences (Pawitan 2001). Analyses of discrete data, in particular binary data, by PQL estimators are known to produce biased estimates of β and φ (Lin & Breslow 1996b). The bias in φ stems from the fact that $\zeta_{\gamma=\tilde{\gamma}}^{(\tau)}$ varies with respect to φ through $\hat{\beta}$ and $\hat{\gamma}$, where $\hat{\gamma} = \tilde{\gamma}(\hat{\beta}(\varphi))$. If φ is far from the true parameter value φ_0 , then information about φ_0 is biased when evaluating $\zeta_{\gamma=\tilde{\gamma}}^{(\tau)}$ at φ . Also, the marginal variance Ψ of $\zeta_{\gamma=\tilde{\gamma}}^{(\tau)}$ should theoretically be independent of γ . However, this uncorrelatedness property does not hold for GLMMs. Thus, the likelihood function for φ given $\zeta_{\gamma=\tilde{\gamma}}^{(\tau)}$, and hence the estimating equations for φ , contain biased information (see Pawitan 2001).

For GLMMs with canonical link functions and a single component of dispersion, Breslow & Lin (1995) analytically derive a correction factor and a first-order corrected estimator to reduce bias in, respectively, φ_m ($1 = m = c$) and β for estimation under PQL. Their asymptotic analysis involve Laplace expansions to the IQL for small values of φ . For larger values, their correction factor performs reasonably well asymptotically, thus improving the estimate of fixed effects. The correction procedures of Breslow & Lin (1995) have been extended by Lin & Breslow (1996a) for the case $1 < m \leq c$. The latter authors provide corrected PQL matrices² for φ , first- and second-order corrected PQL estimators for fixed effect parameters, and propose a four-step algorithm to estimate these parameters. See Lin & Breslow (1996a) for mathematical derivations of their correction procedures and comments on the relative merits of the corrected PQL estimators. Lin & Breslow (1996b) summarize the results of Lin & Breslow (1996a). Engel & Keen (1994) claim that I-MINQUE, which is equivalent to REML, may reduce bias in φ when $\gamma \sim N(\mathbf{0}, \Gamma)$ in contrast to ML.

²Under regularity condition II of Lin & Breslow (1996a, p.1011), the correction matrices are not applicable to fully crossed designs.

Bias correction of parameter estimators in GLMMs can also be performed by simulation. This is done by Pawitan (2001), who exploits the quasi-likelihood concept of Wedderburn (1974). He considers a *pilot* working vector and proposes a two-stage estimation procedure for φ_m ($1 = m = c$). He suggests holding $\hat{\zeta} \equiv \zeta_{(\beta, \gamma) = (\hat{\beta}, \hat{\gamma})}$ and $\hat{\beta}$ fixed in the second stage. The first stage for estimating φ is based on existing methods (see Section 4.6). Then, re-estimate φ_m by maximizing $L(\varphi)$ ³ as a function of φ_m , where φ_m enters $L(\varphi)$ through the marginal variance Ψ only (in the second stage). The maximization of $L(\varphi)$ is achieved by employing numerical derivatives or derivative-free techniques (see Kennedy & Gentle 1980, pp.469–475). A new estimate $\hat{\varphi}_m$ is then obtained. Thereafter, $\hat{\beta}$ is updated via its corresponding estimating equation.

Pawitan (2001) states that all less-than-full likelihood methods, including analytical correction procedures, have an upper bound which is computationally increased by his proposition for estimating dispersion components based on discrete data. In a simulation study of binomial data, the bias-corrected $L(\varphi)$ estimate is comparable to a full likelihood estimate when φ is small to moderate. This is not the case for large values of φ , however.

³In bias correction procedures, either asymptotically or by simulation, it is sufficient to employ ML estimating equations for φ (Lin & Breslow 1996a). See also Pawitan's (2001) equation (11).

4.8 Generalized global score test for components of dispersion

The degree of overdispersion, correlation and heteroscedasticity, as they occur in GLMMs, is gauged by the magnitude of the components of dispersion φ . However, hypothesis testing of φ is cumbersome in GLMMs since the IQL may involve high-dimensional integrals. A global score test for the hypothesis $H_0 : \varphi = \mathbf{0}$, for all φ_m , $m = 1, \dots, c$, has been proposed by Lin (1997), for the hierarchical model of Section 4.2. This test⁴ exists in closed form and is appealing since it requires fitting ordinary GLMs. Furthermore, the parametric assumptions for the random effects γ are relaxed. Thus, the global score test is, in a sense, robust. Relaxing the normality assumption for γ requires that the kurtosis is nearly null (Engel & Keen 1994). The IQL has the form

$$\begin{aligned} L(\psi) &\equiv L(\beta, \varphi) \\ &= \exp[l(\beta, \varphi)] \\ &= \int_{\mathbb{R}^c} \exp \left[\sum_{i=1}^n l_i(\beta; \gamma) \right] dF(\gamma; \varphi) \end{aligned} \quad (4.8.1)$$

where $l_i(\beta; \gamma)$ is as defined in equation (4.2.4), with γ following some *zero mean kernel distribution* $F(\gamma; \varphi)$ whose variance, in the notation of Hall & Praestgaard (2001), is given by $\text{Var}_{\varphi}(\gamma) = \Gamma(\varphi)$ and ϕ is assumed known.

In deriving the global score test, the score $\left. \frac{\partial l(\beta, \varphi)}{\partial \varphi} \right|_{\varphi=0}$ is required. Obtaining this quantity from equation (4.8.1) is difficult since the integral may be intractable. This difficulty is circumvented by taking an expansion of $l(\beta, \varphi)$ about $\varphi = \mathbf{0}$ by using the Laplace approximation.

Lin (1997) expands equation (4.8.1), for small φ , by taking a quadratic expansion of $\sum_{i=1}^n l_i(\beta; \gamma)$ about $\gamma = \mathbf{0}$, the true means of γ , before integration.

⁴The global score test is applicable to both crossed and clustered designs, but not to *fully* crossed designs (Lin 1997, p.315).

A second-order Taylor expansion yields

$$\begin{aligned} \exp \left[\sum_{i=1}^n l_i(\beta; \gamma) \right] &= \exp \left[\sum_{i=1}^n l_i(\beta; \mathbf{0}) \right] \times \left\{ 1 + \sum_{i=1}^n \frac{\partial l_i(\beta; \mathbf{0})}{\partial \eta_i} \mathbf{z}_i^T \gamma \right. \\ &\quad + \frac{1}{2} \gamma^T \left[\left[\sum_{i=1}^n \frac{\partial l_i(\beta; \mathbf{0})}{\partial \eta_i} \mathbf{z}_i \right] \left[\sum_{i=1}^n \frac{\partial l_i(\beta; \mathbf{0})}{\partial \eta_i} \mathbf{z}_i^T \right] \right. \\ &\quad \left. \left. + \sum_{i=1}^n \frac{\partial^2 l_i(\beta; \mathbf{0})}{\partial \eta_i^2} \mathbf{z}_i \mathbf{z}_i^T \right] \gamma + o(\|\gamma\|) \right\} \end{aligned}$$

Equation (4.8.1) can then be written as

$$\begin{aligned} L(\beta, \varphi) &= E \left[\exp \left[\sum_{i=1}^n l_i(\beta; \gamma) \right] \right] \\ &= \exp \left[\sum_{i=1}^n l_i(\beta; \mathbf{0}) \right] \left\{ 1 + \sum_{i=1}^n \frac{\partial l_i(\beta; \mathbf{0})}{\partial \eta_i} \mathbf{z}_i^T E(\gamma) \right. \\ &\quad + \frac{1}{2} \text{tr} \left[\left[\sum_{i=1}^n \frac{\partial l_i(\beta; \mathbf{0})}{\partial \eta_i} \mathbf{z}_i \right] \left[\sum_{i=1}^n \frac{\partial l_i(\beta; \mathbf{0})}{\partial \eta_i} \mathbf{z}_i^T \right] \right. \\ &\quad \left. \left. + \sum_{i=1}^n \frac{\partial^2 l_i(\beta; \mathbf{0})}{\partial \eta_i^2} \mathbf{z}_i \mathbf{z}_i^T \right] E(\gamma \gamma^T) + o(\|\varphi\|) \right\} \\ &= \exp \left[\sum_{i=1}^n l_i(\beta; \mathbf{0}) \right] \left\{ 1 + \frac{1}{2} \text{tr} \left[\mathbf{Z}^T \left(\frac{\partial l(\beta; \mathbf{0})}{\partial \boldsymbol{\eta}} \frac{\partial l(\beta; \mathbf{0})}{\partial \boldsymbol{\eta}^T} \right. \right. \right. \\ &\quad \left. \left. \left. + \frac{\partial^2 l(\beta; \mathbf{0})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \right) \mathbf{Z} \Gamma(\varphi) \right] + o(\|\varphi\|) \right\} \end{aligned} \tag{4.8.2}$$

where $\frac{\partial l(\boldsymbol{\beta}; \mathbf{0})}{\partial \boldsymbol{\eta}}$ is an $n \times 1$ vector with elements $\frac{\partial l_i(\boldsymbol{\beta}; \mathbf{0})}{\partial \eta_i}$ and $\frac{\partial^2 l(\boldsymbol{\beta}; \mathbf{0})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} = \text{diag} \left(\frac{\partial^2 l_i(\boldsymbol{\beta}; \mathbf{0})}{\partial \eta_i^2} \right)$, for $i = 1, \dots, n$. The marginal log-quasi-likelihood function is then given by

$$\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\varphi}) &= \sum_{i=1}^n l_i(\boldsymbol{\beta}; \mathbf{0}) \\
&\quad + \frac{1}{2} \text{tr} \left[\mathbf{Z}^T \left(\frac{\partial l(\boldsymbol{\beta}; \mathbf{0})}{\partial \boldsymbol{\eta}} \frac{\partial l(\boldsymbol{\beta}; \mathbf{0})}{\partial \boldsymbol{\eta}^T} + \frac{\partial^2 l(\boldsymbol{\beta}; \mathbf{0})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \right) \mathbf{Z} \Gamma(\boldsymbol{\varphi}) \right] + o(\|\boldsymbol{\varphi}\|) \\
&= \sum_{i=1}^n l_i(\boldsymbol{\beta}; \mathbf{0}) + \frac{1}{2} \text{tr} \left[(\mathbf{W} \Delta^{-1} (\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T \Delta^{-1} \mathbf{W} - \mathbf{W}_0) \right. \\
&\quad \left. \times \mathbf{Z} \Gamma(\boldsymbol{\varphi}) \mathbf{Z}^T \right] + o(\|\boldsymbol{\varphi}\|)
\end{aligned} \tag{4.8.3}$$

where Lin (1997) defines

$$\begin{aligned}
\mathbf{W}_0 &= -\frac{\partial^2 l(\boldsymbol{\beta}; \mathbf{0})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \\
&= \text{diag}(w_{0i}) \\
&= \text{diag}(w_i + e_i(y_i - \mu_i))
\end{aligned}$$

with e_i given by the second term in square brackets in equation (2.4.3). Furthermore, Lin (1997) defines $\Delta = \text{diag}(\delta_i)$, with $\delta_i = [g'(\mu_i)]^{-1}$ and $\mathbf{W} = E(\mathbf{W}_0)$.

The $c \times 1$ efficient score vector $U_\varphi(\hat{\beta}_0)$ has elements

$$\begin{aligned}
U_{\varphi_m}(\hat{\beta}_0) &= \left. \frac{\partial l(\beta, \varphi)}{\partial \varphi_m} \right|_{\varphi=0, \beta=\hat{\beta}_0} \\
&= \frac{1}{2} \text{tr} \left\{ \left[\mathbf{W} \Delta^{-1} (\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T \Delta^{-1} \mathbf{W} - \mathbf{W}_0 \right] \mathbf{Z} \frac{\partial \Gamma}{\partial \varphi_m} \mathbf{Z}^T \right\} \\
&= \frac{1}{2} \left\{ \left[(\mathbf{y} - \boldsymbol{\mu})^T \Delta^{-1} \mathbf{W} \mathbf{Z} \frac{\partial \Gamma}{\partial \varphi_m} \mathbf{Z}^T \mathbf{W} \Delta^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right] \right. \\
&\quad \left. - \text{tr} \left(\mathbf{W}_0 \mathbf{Z} \frac{\partial \Gamma}{\partial \varphi_m} \mathbf{Z}^T \right) \right\}
\end{aligned} \tag{4.8.4}$$

where $\hat{\beta}_0$ is the ML estimator of β under $\varphi = 0$. $U_{\varphi_m}(\hat{\beta}_0)$ in equation (4.8.4) gives a comparison of the weighted actual and nominal observed covariance of $\mathbf{y} = (y_1, \dots, y_n)^T$; $m = 1, \dots, c$.

Let the information matrix be partitioned as

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\varphi} \\ \mathcal{I}_{\beta\varphi}^T & \mathcal{I}_{\varphi\varphi} \end{pmatrix} \tag{4.8.5a}$$

where $\mathcal{I}_{\varphi\varphi} = E \left(\frac{\partial l}{\partial \varphi} \frac{\partial l}{\partial \varphi^T} \right) \Big|_{\varphi=0}$, $\mathcal{I}_{\beta\varphi} = E \left(\frac{\partial l}{\partial \beta} \frac{\partial l}{\partial \varphi^T} \right) \Big|_{\varphi=0}$ and $\mathcal{I}_{\beta\beta} = E \left(\frac{\partial l}{\partial \beta} \frac{\partial l}{\partial \beta^T} \right) \Big|_{\varphi=0}$. Furthermore,

$$U_{\varphi_m} = \frac{\partial l}{\partial \varphi} = \sum_{i=1}^n b_{ii}^m [w_i^2 \delta_i^{-2} (y_i - \mu_i)^2 - w_i - e_i (y_i - \mu_i)]$$

and

$$U_\beta = \frac{\partial l}{\partial \beta} = \sum_{i=1}^n x_i w_i \delta_i^{-1} (y_i - \mu_i)$$

Lin (1997).

It can be shown that

$$\begin{aligned}
\mathcal{I}_{\varphi_m \varphi_k} &= E(U_{\varphi_m} U_{\varphi_k}) \\
&= \frac{1}{4} \left(\sum_{i=1}^n b_{ii}^m b_{ii}^k r_{ii} + 2 \sum_{i < i'} b_{ii}^m b_{ii'}^k r_{ii'} \right) \\
&= \frac{1}{4} \mathbf{J}^T (\mathbf{B}_m \cdot \mathbf{R} \cdot \mathbf{B}_k) \mathbf{J}
\end{aligned} \tag{4.8.5b}$$

where $\mathbf{B}_m = \mathbf{Z} \frac{\partial \Gamma}{\partial \varphi_m} \mathbf{Z}^T = (b_{ii}^m)$, with \mathbf{b}^m being an $n \times 1$ vector with elements b_{ii}^m , \mathbf{R} is an $n \times n$ matrix with diagonal elements $r_{ii} = w_i^4 \delta_i^{-4} \kappa_{4i} + 2w_i^2 + e_i^2 \kappa_{2i} - 2w_i^2 \delta_i^{-2} e_i \kappa_{3i}$ and off-diagonal elements $r_{ii'} = 2w_i w_{i'}$, for $i \neq i'$. \mathbf{J} is a vector of ones and $\mathbf{G} \cdot \mathbf{H}$ is the component-wise multiplication of conformable matrices \mathbf{G} and \mathbf{H} . Now,

$$\begin{aligned}
\mathcal{I}_{\beta \varphi_m} &= E(U_{\beta} U_{\varphi_m}) \\
&= \frac{1}{2} E \left\{ \left[\sum_{i=1}^n \mathbf{x}_i w_i \delta_i^{-1} (y_i - \mu_i) \right] \right. \\
&\quad \times \left. \left[\sum_{i=1}^n b_{ii}^m \left[w_i^2 \delta_i^{-2} (y_i - \mu_i)^2 - w_i - e_i (y_i - \mu_i) \right] \right] \right\} \\
&= \frac{1}{2} \left\{ \sum_{i=1}^n \mathbf{x}_i b_{ii}^m \left[w_i^3 \delta_i^{-3} E(y_i - \mu_i)^3 - w_i \delta_i^{-1} e_i E(y_i - \mu_i)^2 \right] \right\} \\
&= \frac{1}{2} \sum_{i=1}^n \mathbf{x}_i b_{ii}^m (w_i^3 \delta_i^{-3} \kappa_{3i} - w_i \delta_i^{-1} e_i \kappa_{2i}) \\
&= \frac{1}{2} \sum_{i=1}^n \mathbf{x}_i c_i b_{ii}^m \\
&= \frac{1}{2} \mathbf{X}^T \mathbf{C} \mathbf{b}^m
\end{aligned} \tag{4.8.5c}$$

where $\mathbf{C} = \text{diag}(c_i) = \text{diag}(w_i^3 \delta_i^{-3} \kappa_{3i} - w_i \delta_i^{-1} e_i \kappa_{2i})$. It is straightforward to

show that

$$\mathcal{I}_{\beta\beta} = \mathbf{X}^T \mathbf{W} \mathbf{X} \quad (4.8.5d)$$

Lin (1997) defines the *efficient information matrix* as

$$\mathcal{I}^{eff} = \mathcal{I}_{\varphi\varphi} - \mathcal{I}_{\beta\varphi}^T \mathcal{I}_{\beta\beta}^{-1} \mathcal{I}_{\beta\varphi}$$

The *generalized global score statistic* is then defined as

$$\chi_G^2 = (U_\varphi(\hat{\beta}_0))^T (\mathcal{I}^{eff}(\hat{\beta}_0))^{-1} (U_\varphi(\hat{\beta}_0)) \quad (4.8.6)$$

for testing $H_0 : \varphi = \mathbf{0}$. Under certain regularity conditions, and using Slutsky's Theorem (Casella & Berger 2002, p.239), Lin (1997) proves that

- (i) χ_G^2 is asymptotically chi-squared distributed with c degrees-of-freedom under $\varphi = \mathbf{0}$;
- (ii) χ_G^2 is a *locally asymptotically most powerful test* (LAMPT) if $c = 1$, and is a *locally asymptotically most stringent test* (LAMST)⁵ if $c > 1$.

Therefore, the null hypothesis $H_0 : \varphi = \mathbf{0}$ is rejected if and only if $\chi_G^2 \geq \chi_{c;\alpha}^2$, where α is the asymptotic size test. A bias, when n is small, is incurred in χ_G^2 due to the estimation of β . Lin (1997) provides a bias-corrected efficient score statistic $U_{\#\varphi_m}(\hat{\beta}_0)$, which is of the same form as in equation (4.8.4) except that \mathbf{W}_0 is replaced by $\mathbf{W}_{\#0}$, where

$$\mathbf{W}_{\#0} = \text{diag}((1 - h_i)w_i + e_i(y_i - \mu_i))$$

with h_i being the i th diagonal element of the hat matrix

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2}$$

⁵See Bhat & Nagnur (1965).

The corresponding bias-corrected global score statistic is then given by

$$\chi_{\#G}^2 = (U_{\#\varphi}(\hat{\beta}_0))^T (\mathcal{I}^{eff}(\hat{\beta}_0))^{-1} (U_{\#\varphi}(\hat{\beta}_0))$$

where $\chi_{\#G}^2$ is the bias-corrected form of χ_G^2 .

According to a simulation study of binary data by Lin (1997), $\chi_{\#G}^2$ performs slightly better than χ_G^2 in terms of size and power. As φ increases, the powers of the tests converge to 1 rapidly. Furthermore, the performance of χ_G^2 and $\chi_{\#G}^2$ is high when the number of levels q_m of each random effect is moderate or large. For small q_m , for example when $q_m < 15$, critical values of χ_G^2 and $\chi_{\#G}^2$ are suspected to be less precise. Hall & Praestgaard (2001) state that Lin's (1997) global score test has power in all directions from $\varphi = \mathbf{0}$ and can thus be regarded as an omnibus test.

4.8.1 Individual dispersion component score test

A drawback of the global score test statistic χ_G^2 is that it does not single out particular φ_m 's, $m = 1, \dots, c$, that do not conform with the hypothesis $H_0 : \varphi = \mathbf{0}$. Thus, a statistic for testing the null hypothesis $H_0 : \varphi_m = 0$ against a one-sided alternative hypothesis $H_a : \varphi_m > 0$ becomes necessary. For such a test, parameter estimates of β and φ , without φ_m , are required. These estimates can be obtained via Fisher scoring by dropping moment assumptions for the random effects and introducing the stronger normality assumptions. Efficient score and efficient information matrix for φ_m do not have closed-form structures, and have to be approximated by Laplace expansions. Moreover, it is well-known that, for small n , estimation of β results in a loss of degrees-of-freedom. It is therefore recommended to use Laplace-approximated efficient score and efficient information matrix for φ_m under REML to construct an individual score statistic for testing $H_0 : \varphi_m = 0$ against $H_a : \varphi_m > 0$. Under certain regularity conditions, the Laplace-based score statistic for testing $\varphi_m = 0$ follows a standard normal distribution, asymptotically (under $H_0 : \varphi_m = 0$), and is a LAMPT. See Lin (1997) for the mathematical derivation of the individual dispersion component test.

Simulation-based analysis of binary data by Lin (1997) indicates that the bias-corrected individual score statistic for φ_m does not perform well. As the binomial denominator increases, an improvement, in terms of size and power, of this Laplace-approximated test becomes apparent.

4.9 Conditional mean squared error of prediction

In the analysis of Gaussian LMMs, standard errors of prediction are traditionally calculated as the square root of an estimate of the *unconditional mean squared error of prediction* (UMSEP). However, the UMSEP cannot be adapted to GLMMs, where the conditional variance of the random effects γ depends on the data, as stated by Booth & Hobert (1998). These authors have proposed a general measure of prediction variance, the *conditional mean squared error of prediction* (CMSEP), in GLMMs under full ‘joint’ distributional assumptions for the data and the random effects.

Booth & Hobert (1998) motivate their CMSEP for a two-level GLMM where the distribution of \mathbf{y} , given the random effects γ , belongs to the exponential family and where random effects are Gaussian distributed with mean zero. We formulate the CMSEP for the hierarchical model of Section 4.2 (without the use of rigorous asymptotic analysis). In our attempt, we follow the notation of Booth & Hobert (1998) by dropping the superscript γ in η_i and μ_i , $i = 1, \dots, n$. We assume that the scale factor ϕ is known. Furthermore, when the conditional distribution of the responses belong to the exponential family, the canonical parameter is given by

$$\theta_i^\gamma = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \gamma$$

(Sinha 2004).

4.9.1 Conditional standard errors of prediction

We focus on standard errors of prediction for η_i of equation (4.2.2a). For known $\boldsymbol{\psi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\varphi}^T)$, a point predictor for η_i is

$$\begin{aligned} \eta_i &\equiv \eta_i(\boldsymbol{\psi}; y_i) \\ &= E_{\boldsymbol{\psi}}(\eta_i | y_i) \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T E_{\boldsymbol{\psi}}(\gamma | y_i) \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma}(\boldsymbol{\psi}; y_i) \end{aligned} \tag{4.9.1a}$$

where a point predictor for γ is its conditional mean $\boldsymbol{\gamma}(\boldsymbol{\psi}; y_i) \equiv E_{\boldsymbol{\psi}}(\boldsymbol{\gamma} | y_i)$,

$i = 1, \dots, n$. The prediction variance for η_i is given by

$$\begin{aligned}
\nu_i &\equiv \nu_i(\boldsymbol{\psi}; y_i) \\
&= \text{Var}_{\boldsymbol{\psi}}(\eta_i | y_i) \\
&= \text{Var}_{\boldsymbol{\psi}}[\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma}(\boldsymbol{\psi}; y_i)] \\
&= \mathbf{z}_i^T \text{Var}_{\boldsymbol{\psi}}(\boldsymbol{\gamma} | y_i) \mathbf{z}_i
\end{aligned} \tag{4.9.1b}$$

When $\boldsymbol{\psi}$ is unknown, and if $\widehat{\boldsymbol{\psi}}$ is a consistent (REML)⁶ estimate of $\boldsymbol{\psi}$, equation (4.9.1b) becomes

$$\widehat{\nu}_i \equiv \nu_i(\widehat{\boldsymbol{\psi}}; y_i) \tag{4.9.2a}$$

Clearly, a point predictor for η_i is then

$$\begin{aligned}
\widehat{\eta}_i &\equiv \eta_i(\widehat{\boldsymbol{\psi}}; y_i) \\
&= \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} + \mathbf{z}_i^T \boldsymbol{\gamma}(\widehat{\boldsymbol{\psi}}; y_i)
\end{aligned} \tag{4.9.2b}$$

with associated standard error of prediction $\sqrt{\widehat{\nu}_i}$. There exists a caveat, however; $\widehat{\nu}_i$ fails to account for the sampling variability associated with $\widehat{\boldsymbol{\psi}}$, and is thus termed *naive*. Also, a bias is incurred by the substitution of $\widehat{\boldsymbol{\psi}}$ into $\widehat{\nu}_i$. It is assumed that as $n, q_m \rightarrow \infty$ ($m = 1, \dots, c$), the number of observations at any level of any random effect is bounded for all n . Therefore, $q_m = O(n)$ (Lin 1997, Condition 2, Appendix 2, p.323).

The CMSEP is defined as

$$\begin{aligned}
\text{CMSEP}(\boldsymbol{\psi}; y_i) &= E_{\boldsymbol{\psi}}[(\eta_i - \widehat{\eta}_i)^2 | y_i] \\
&= E_{\boldsymbol{\psi}} \{ [(\eta_i - \eta_i(\boldsymbol{\psi}; y_i)) + (\eta_i(\boldsymbol{\psi}; y_i) - \widehat{\eta}_i)]^2 | y_i \}
\end{aligned} \tag{4.9.3a}$$

In equation (4.9.3a), the predictor $\eta_i(\boldsymbol{\psi}; y_i)$ is added and subtracted under the assumption that $\boldsymbol{\psi}$ is known. Equation (4.9.3a) simplifies to

$$\begin{aligned}
\text{CMSEP}(\boldsymbol{\psi}; y_i) &= E_{\boldsymbol{\psi}} \{ [(\eta_i - \eta_i(\boldsymbol{\psi}; y_i))^2 | y_i] \\
&\quad + 2[(\eta_i - \eta_i(\boldsymbol{\psi}; y_i))(\eta_i(\boldsymbol{\psi}; y_i) - \widehat{\eta}_i) | y_i] \\
&\quad + [(\eta_i(\boldsymbol{\psi}; y_i) - \widehat{\eta}_i)^2 | y_i] \}
\end{aligned} \tag{4.9.3b}$$

$$\begin{aligned}
&= \text{Var}_{\boldsymbol{\psi}}(\eta_i | y_i) + E_{\boldsymbol{\psi}}[(\eta_i(\boldsymbol{\psi}; y_i) - \widehat{\eta}_i)^2 | y_i] \\
&= \nu_i(\boldsymbol{\psi}; y_i) + \text{corr}_i(\boldsymbol{\psi}; y_i)
\end{aligned} \tag{4.9.3c}$$

⁶Booth & Hobert (1998) use ML.

where, in equation (4.9.3b), the second term reduces to zero since

$$[(\eta_i - \eta_i(\boldsymbol{\psi}; y_i))|y_i] \text{ and } [(\eta_i(\boldsymbol{\psi}; y_i) - \hat{\eta}_i)|y_i]$$

are conditionally independent (given y_i). The term $corr_i(\boldsymbol{\psi}; y_i)$ is a non-negative correction term that accounts for the sampling variability of $\hat{\boldsymbol{\psi}}$.

Now,

$$\begin{aligned} corr_i(\boldsymbol{\psi}; y_i) &= E_{\boldsymbol{\psi}}[(\eta_i - \hat{\eta}_i)^2|y_i] \\ &= E_{\boldsymbol{\psi}} \left\{ [(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \gamma(\boldsymbol{\psi}; y_i)) - (\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \mathbf{z}_i^T \gamma(\hat{\boldsymbol{\psi}}; y_i))]^2 | y_i \right\} \\ &= E_{\boldsymbol{\psi}} \left\{ [\mathbf{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{z}_i^T (\gamma(\boldsymbol{\psi}; y_i) - \gamma(\hat{\boldsymbol{\psi}}; y_i))]^2 | y_i \right\} \end{aligned} \quad (4.9.4a)$$

A second-order Taylor expansion of $\gamma(\boldsymbol{\psi}; y_i)$ about $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$ yields:

$$\gamma(\boldsymbol{\psi}; y_i) = \gamma(\hat{\boldsymbol{\psi}}; y_i) + \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\psi}} (\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) + O_p(|\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}|^2) \quad (4.9.4b)$$

Substituting equation (4.9.4b) into equation (4.9.4a) gives, after some rearrangement:

$$\begin{aligned} &corr_i(\boldsymbol{\psi}; y_i)|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \\ &= E_{\boldsymbol{\psi}} \left\{ \left[\mathbf{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{z}_i^T \left(\frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\beta}}, \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\varphi}} \right) ((\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T, (\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}})^T)^T \right]^2 | y_i \right\} \\ &= E_{\boldsymbol{\psi}} \left\{ \left[\left(\mathbf{x}_i^T + \mathbf{z}_i^T \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\beta}} \right) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \mathbf{z}_i^T \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\varphi}} (\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}) \right]^2 | y_i \right\} \\ &= \left(\mathbf{x}_i^T + \mathbf{z}_i^T \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\beta}}, \mathbf{z}_i^T \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\varphi}} \right) \times [\mathcal{I}(\boldsymbol{\psi})]^{-1} \\ &\quad \times \left(\mathbf{x}_i^T + \mathbf{z}_i^T \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\beta}}, \mathbf{z}_i^T \frac{\partial \gamma(\boldsymbol{\psi}; y_i)}{\partial \boldsymbol{\varphi}} \right)^T \end{aligned} \quad (4.9.5)$$

where

$$\mathcal{I}(\boldsymbol{\psi})|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = E_{\boldsymbol{\psi}}[(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})^T]$$

is the information matrix for ψ based on the assumed GLMM, and is given by equation (4.8.5a), with corresponding elements that are given by equations (4.8.5b) – (4.8.5d).

In equation (4.9.5), we require expressions for $\frac{\partial \gamma(\psi; y_i)}{\partial \beta}$ and $\frac{\partial \gamma(\psi; y_i)}{\partial \varphi}$ to be evaluated at $\psi = \hat{\psi}$. Booth & Hobert (1998) define the conditional expectation $\gamma(\psi; y_i)$ by a function M such that

$$M(\gamma; \psi) = \frac{\partial l(\gamma; \psi)}{\partial \gamma} \approx \mathbf{0} \quad (4.9.6a)$$

where $l(\gamma; \psi)$ is defined by equation (4.3.9). From equation (4.4.1b), we have

$$\frac{\partial l(\gamma; \psi)}{\partial \gamma} = \sum_{i=1}^n \mathbf{z}_i w_i (y_i - \mu_i) g'(\mu_i) - \Gamma^{-1} \gamma \quad (4.9.6b)$$

Now, $\frac{\partial \gamma(\psi; y_i)}{\partial \beta} \approx - \left[\frac{\partial M}{\partial \gamma} \right]^{-1} \frac{\partial M}{\partial \beta}$, where $\frac{\partial M}{\partial \gamma} \left(= \frac{\partial^2 l(\gamma; \psi)}{\partial \gamma \partial \gamma^T} \right)$ is given by equation (4.3.7a). Therefore,

$$\begin{aligned} \frac{\partial M}{\partial \beta} &= \frac{\partial}{\partial \beta} \left(\frac{\partial l(\gamma; \psi)}{\partial \gamma} \right) \\ &= \frac{\partial}{\partial \beta} \left(\sum_{i=1}^n \mathbf{z}_i w_i (y_i - \mu_i) g'(\mu_i) \right) \\ &= \frac{\partial}{\partial \mu_i \gamma} \left(\sum_{i=1}^n \mathbf{z}_i w_i (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} \right) \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} \\ &= \sum_{i=1}^n \mathbf{z}_i w_i \frac{\partial (y_i - \mu_i)}{\partial \mu_i \gamma} \mathbf{x}_i^T \\ &= - \sum_{i=1}^n \mathbf{z}_i w_i \mathbf{x}_i^T \\ &= -\mathbf{Z}^T \mathbf{W} \mathbf{X} \end{aligned} \quad (4.9.6c)$$

We then have

$$\frac{\partial \gamma(\psi; y_i)}{\partial \beta} = -(\mathbf{Z}^T \mathbf{W} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{X} \quad (4.9.6d)$$

Also, $\frac{\partial \gamma(\psi; y_i)}{\partial \varphi} \approx - \left[\frac{\partial M}{\partial \gamma} \right]^{-1} \frac{\partial M}{\partial \varphi}$, where we use equation (4.9.6b) to obtain

$$\begin{aligned} \frac{\partial M}{\partial \varphi_m} &= \frac{\partial}{\partial \varphi_m} \left(\frac{\partial l(\gamma; \psi)}{\partial \gamma} \right) \\ &= -\frac{\partial \Gamma^{-1}}{\partial \varphi_m} \gamma \\ &= \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \varphi_m} \Gamma^{-1} \right) \gamma \end{aligned}$$

Thus,

$$\frac{\partial \gamma(\psi; y_i)}{\partial \varphi_m} = (\mathbf{Z}^T \mathbf{W} \mathbf{Z} + \Gamma^{-1})^{-1} \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \varphi_m} \Gamma^{-1} \right) \gamma \quad (4.9.6e)$$

for $m = 1, \dots, c$. Furthermore, from equation (4.9.1b), we have the following expression for ν_i :

$$\begin{aligned} \nu_i &\approx \mathbf{z}_i^T \left[-E \left(\frac{\partial^2 l(\gamma; \psi)}{\partial \gamma \partial \gamma^T} \right) \right]^{-1} \mathbf{z}_i \\ &= \mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z} + \Gamma^{-1})^{-1} \mathbf{z}_i \end{aligned} \quad (4.9.6f)$$

An expression for the CMSEP can be obtained by taking a second-order Taylor expansion of $\hat{\nu}_i$ about ψ (see Booth & Hobert 1998, equation (26)). We conclude that the conditional standard error of prediction for $\hat{\eta}_i$ at $\psi = \hat{\psi}$ is given by

$$\sqrt{\hat{\nu}_i - \text{bias}_i(\hat{\psi}; y_i) + \text{corr}_i(\hat{\psi}; y_i)} \quad (4.9.7)$$

where $\text{bias}_i(\hat{\psi}; y_i) (= E_{\psi}[\hat{\nu}_i - \nu_i(\psi; y_i)|y_i])$ is the parametric *bootstrap* (see Urban Hjorth 1994) estimate of the conditional bias incurred in $\hat{\nu}_i$ through $\hat{\psi}$. Moreover, the bias term should not be ignored if the sampling variability of $\hat{\psi}$ is of importance (Booth & Hobert 1998).

4.10 Some remarks

Hall & Praestgaard (2001) have proposed an improvement of Lin's (1997) global score test by concentrating this test into directions that are given by constraining φ to form a positive semidefinite covariance matrix. They argue, by simulations and theoretical justifications, that their *projected* or *restricted* score test has more power than Lin's (1997) unrestricted test. However, their restricted test is a mixture of chi-squared distributions with associated degrees-of-freedom less than or equal to c . In their endeavour, Hall & Praestgaard (2001) specify a true density for the conditional distribution of $\mathbf{y}|\gamma$. They recommend that the score statistic and information matrix, both bias-corrected, be employed in a restricted, bias-corrected global score test for homogeneity to achieve better performance in terms of size and power. This particular test adjusts for the loss of degrees-of-freedom due to the estimation of β .

Lee & Nelder (1996) have proposed a broader class of models known as the *hierarchical generalized linear models* (HGLMs), of which GLMMs are a special case. In an HGLM, the distribution of the random effects is conjugate to that of the observed data \mathbf{y} . For inference in HGLMs, Lee & Nelder (1996) define a *hierarchical* or *h-likelihood*. By its definition, the *h-likelihood* does not require obtaining the marginal likelihood, and therefore, intractable high-dimensional integrals, usually associated with GLMMs, are circumvented. Maximizing the *h-likelihood* leads, asymptotically, to efficient fixed effect estimators that are equivalent to those obtained from the marginal likelihood approach and to random effect estimates that are best unbiased predictors. These estimators (predictors) are called *maximum h-likelihood estimators* (MHLEs). Components of dispersion are estimated by an adjusted profile *h-likelihood* (APHL) procedure that takes into account the loss of degrees-of-freedom due to the estimation of β . A scaled deviance test has also been proposed by Lee & Nelder (1996) to assess model adequacy. This goodness-of-fit test, however, uses the conditional distribution of \mathbf{y} given γ only, and thus, cannot be used for testing components of dispersion. Moreover, they propose a test criterion to detect the absence of random effects γ_m that is based on testing the null hypothesis $H_0 : \varphi_m = 0$, $m = 1, \dots, c$. This statistic can also be used to test for the equality of random effects against their independence. Recently, Lee & Nelder (2003) have considered

extended-REML estimators, based on *double* EQL functions. They show that this extension results in efficient estimation of dispersion components.

Lee & Nelder (1996) suggest the Wald⁷ test and the likelihood ratio test for inference on fixed effects. The former test statistic is also available for testing dispersion components. However, since null hypothesis testing of dispersion components places φ on the boundary of the parameter space, these statistics follow a mixed chi-squared distribution, under some regularity conditions, instead of a chi-squared distribution (Lin 1997).

The *method of simulated moments* (MSM) has been proposed in the literature for the estimation of fixed effect parameters and dispersion components. In brief, estimating equations based on the *method of moments* (MM) are obtained by equating sample moments of sufficient statistics to their expectations. However, the high-dimensional integrals involved in these expectations hampers statistical inference in GLMMs. The method of simulated moments provides an approximation to these integrals. Simulation results indicate that MSM estimators of fixed effects and dispersion components, though consistent, are quite inefficient (see Jiang 1998).

⁷Engel *et al* (1995) also suggest the Wald statistic for hypothesis testing of fixed effects.

Chapter 5

Analysis of correlated binomial data using a GLMM: The Logistic-Gaussian model

5.1 Introduction

We mentioned in Chapter 4 that correlated discrete data exhibit overdispersion and heterogeneity. For the analysis of GLMMs, the SAS Institute has implemented the necessary codes in the GLIMMIX macro to provide solutions to the PQL algorithm of Breslow & Clayton (1993) and to the pseudo-likelihood algorithm of Wolfinger & O'Connell (1993). However, it is known that PQL estimators for components of dispersion, and hence for regression coefficients, are biased, especially with binary outcomes (see Section 4.7). 'True' ML estimation for GLMMs is now feasible, a result of recent developments in numerical integration (Breslow 2003).

The adaptive Gaussian quadrature, which is one of the best techniques to approximate integrated likelihood functions, is now implemented in SAS PROC NLMIXED, where the approximation is maximized by a dual quasi-Newton algorithm which is the default algorithm (Wolfinger 1999). However, it is to be noted that this quadrature requires that the dimensionality of the integrations be in the low single digits for the analysis of clustered data (Breslow 2003).

Schall (1991) presents a binomial dataset concerning an experiment that measures the mortality of cancer cells under radiation and Smith *et al* (1995) report a dataset, in binomial form, regarding an assessment of the effects of antibiotics on the risk of respiratory tract infections in patients in intensive care units. The latter dataset has been analyzed (in terms of log-odds ratios) by Smith *et al* (1995) from a Bayesian perspective, and by Turner *et al* (2000) from a classical point of view within the framework of multilevel modelling. We note that Turner *et al* (2000) use the MLn (Woodhouse *et al* 1996) and MLwiN (Goldstein *et al* 1998) software.

In this chapter, we analyze the above-mentioned datasets and compare results to the parameter vector of interest $\boldsymbol{\psi}$, with the GLIMMIX macro and PROC NLMIXED. We model heterogeneity in the respiratory tract infections data by constraining the scale factor at unity (Breslow & Clayton 1993) where $\boldsymbol{\psi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\varphi}^T)$, whereas in the modelling of overdispersion in the cancer cells data, it is estimated¹ (Wolfinger & O'Connell 1993) where $\boldsymbol{\psi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\varphi}^{*T})$, with $\boldsymbol{\varphi}^{*T} = (\phi, \boldsymbol{\varphi}^T)$. Note that, for the former dataset, we modify the SAS statements of Brown & Prescott (1999, p.196) to parallel results generated by PROC NLMIXED (see Wolfinger 1999).

5.2 The Logistic-Gaussian model

Let the vector $\mathbf{y}^* = (y_1^*, \dots, y_n^*)^T$ denote n binomial responses and let $\mathbf{y} = (y_1, \dots, y_n)^T$ denote the corresponding vector of observed proportions, where $y_i = u_i^{-1}y_i^*$ ($i = 1, \dots, n$) with u_i being the binomial denominators. Conditional on a vector of Gaussian distributed random effects (see equation (3.2.3)), the data points y_i are assumed to be independent with (conditional) means that are given by equation (4.2.1a) and (conditional) variances

$$\text{Var}(y_i|\boldsymbol{\gamma}) = \phi u_i^{-1} \mu_i^\gamma (1 - \mu_i^\gamma) \quad (5.2.1)$$

The resulting GLMM (equation (4.2.2a) is then written as

$$\eta_i^\gamma = g(\mu_i^\gamma) = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma} \quad (5.2.2a)$$

where $g(\mu_i^\gamma)$ is the (*conditional*) logit function (McCullagh & Nelder 1989, p.31) such that

$$\text{logit}(\mu_i^\gamma) = \ln \left(\frac{\mu_i^\gamma}{1 - \mu_i^\gamma} \right) \quad (5.2.2b)$$

Therefore, we have

$$\mu_i^\gamma = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma})} \quad (5.2.2c)$$

In matrix notational form, equation (5.2.2a) is written as

$$\text{logit}(\boldsymbol{\mu}^\gamma) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \quad (5.2.2d)$$

¹The GLIMMIX macro was written to fit the pseudo-likelihood algorithm of Wolfinger & O'Connell (1993) which is an extension of the PQL algorithm of Breslow & Clayton (1993) where the scale factor is estimated (see Kuss 2002).

In the presence of multiple sources of random variation, equations (4.2.2c) and (5.2.2d) become

$$\text{logit}(\boldsymbol{\mu}^\gamma) = \mathbf{X}\boldsymbol{\beta} + \sum_{m=1}^c \mathbf{Z}_m \boldsymbol{\gamma}_m \quad (5.2.3)$$

and is referred to as the *Logistic-Gaussian* model. Ignoring a constant term, the term $l_i(\boldsymbol{\beta}; \boldsymbol{\gamma})$ in the IQL, viz. equation (4.2.4), has the form

$$\phi^{-1} u_i \left[y_i \ln \left(\frac{\mu_i^\gamma}{1 - \mu_i^\gamma} \right) + \ln(1 - \mu_i^\gamma) \right] \quad (5.2.4)$$

(Lin & Breslow 1996b). Moreover, the y_i 's may be blocked to form clusters (see Section 6.3.1 for the notation of a two-level GLMM).

5.3 Modelling heterogeneity in binomial clinical trials data

Meta-analysis is increasingly used to assess the same treatments from a combination of results from multiple clinical trials so that a more precise overall estimate of treatment effects can be achieved. It is not plausible for a treatment to vary across trials. However, if treatment estimates are related to the circumstances and locations sampled by the trials, effects due to trial and trial-treatment interaction should be fitted as random. This approach increases the standard errors of treatment estimates to reflect heterogeneity across trials. Furthermore, treating trial effects and trial-treatment effects as random, respectively (i) increases the precision of treatment estimates from the aggregation of information from both trial error and residual strata, and (ii) allows modelling of variability at the observation level by the trial-treatment dispersion component, which is retained in the model provided that it is positive (although not 'proven' to be statistically significant), while the scale factor is held fixed at unity.

See Brown & Prescott (1999, Chapter 5, pp.171–198) for a comprehensive account of randomized clinical trials.

5.3.1 Analyzing the risk of respiratory tract infections by selective decontamination of the digestive tract

A suggested strategy in the prevention of infections acquired in intensive care units (ICU) is the selective decontamination of the digestive tract to prevent carriage of potentially pathogenic micro-organisms from the oropharynx, stomach and gut. A meta-analysis of 22 randomized trials was performed to investigate the clinical benefits of such a strategy. In each trial, patients in ICU were randomized to either a treatment group, where they received different combinations of oral non-absorbable antibiotics, or to a control group, where no treatment was received. Table 5.1 below (Table 1 in Smith *et al* (1995)) reports the number of patients who were diagnosed with respiratory tract infections in these groups.

Table 5.1 Respiratory tract infections in treated and control groups of 22 trials

Trial	Infections ^a /Total	
	Treated	Control
1	7/47	25/54
2	4/38	24/41
3	20/96	37/95
4	1/14	11/17
5	10/48	26/49
6	2/101	13/84
7	12/161	38/170
8	1/28	29/60
9	1/19	9/20
10	22/49	44/47
11	26/162	30/160
12	31/200	40/185
13	9/39	10/41
14	22/193	40/185
15	0/45	4/46
16	31/131	60/140
17	4/75	12/75
18	31/220	42/225
19	7/55	26/57
20	3/91	17/92
21	14/25	23/23
22	3/65	6/68

^a Our analyses are coded in terms of patients who were *not* infected.

Let y_{kt}^* ($k = 1, \dots, 22$; $t = \text{drug/control}$) denote the number of patients who responded favourably to the k th trial-treatment combination – clearly, those who did not, is given by $u_{kt} - y_{kt}^*$, where u_{kt} is the number of patients assigned to the t th treatment at the k th trial. Moreover, let the proportion corresponding to favourable responses be denoted by $y_{kt} = y_{kt}^*/u_{kt}$ (Littell *et al* 1996, p.437). The y_{kt}^* 's are assumed to be conditionally independent (given γ) – that is,

$$y_{kt}^* | \gamma_k \sim \text{binomial}(u_{kt}, y_{kt})$$

(Booth & Hobert 1998). Also $\gamma^T = (\gamma_1^T, \gamma_2^T)$, where

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Gamma(\varphi_1) & \mathbf{0} \\ \mathbf{0} & \Gamma(\varphi_2) \end{pmatrix} \right)$$

with γ_1 and γ_2 being random effects vectors corresponding to trial and trial-treatment interaction, respectively. The Logistic-Gaussian model is then given by

$$\ln \left(\frac{\mu_{kt}^\gamma}{1 - \mu_{kt}^\gamma} \right) = \beta_0 + \beta_1 x_{kt} + \gamma_{1k} + \gamma_{2kt} \quad (5.3.1)$$

where x_{kt} is an indicator variable such that

$$x_{kt} = \begin{cases} 1, & \text{for drug} \\ 0, & \text{for control} \end{cases}$$

5.3.2 Results

Table 5.2 compares parameter estimates for the regression coefficients and dispersion components obtained by fitting the respiratory tract infections data with the GLIMMIX macro (PQL) and PROC NLMIXED (ML).

Table 5.2 Parameter estimates \pm standard errors for the respiratory tract infections data (p -values in parentheses)

Parameter	GLIMMIX ^a (PQL) ^b	NLMIXED ^c (ML)
β_0	0.6170 \pm 0.2447 (0.0188)	0.7397 \pm 0.2182 (0.0029)
β_1	1.3426 \pm 0.1877 (< 0.0001)	1.0535 \pm 0.08898 (< 0.0001)
φ_1	0.9644 \pm 0.3737 (0.0049)	0.9370 \pm 539.80 (0.9986)
φ_2	0.2501 \pm 0.1205 (0.0190)	0.07349 \pm 14.3641 (0.9960)

^aScale factor constrained at unity; deviance = 15.9021.

^bPQL estimation under REML.

^c Scale factor and deviance not computed.

The GLIMMIX macro produces a deviance of 15.9021. Relative to $\chi_{0.95,42}^2 = 58.12$, there is no evidence of a lack-of-fit for the above Logistic-Gaussian model (Littell *et al* 1996, p.446). PQL parameter estimates are greater than those obtained under ML, except for the intercept term. Standard errors for $\beta = (\beta_0, \beta_1)^T$ under PQL are larger than those under ML. However, standard errors for $\varphi = (\varphi_1, \varphi_2)^T$ are very large under ML. We note that the p -values, generated by the GLIMMIX macro, that are associated with β and φ are derived from the t -statistic and z -statistic respectively, whereas those obtained from PROC NLMIXED are calculated from the t -statistic for all parameters. Furthermore, the computed p -values for the respiratory tract infections data indicate that estimates for PQL and ML regression coefficients, as well as PQL dispersion components, are statistically significant at the 5% level of significance. Caution should be exercised when interpreting the p -values that are associated with ML components of dispersion estimates since the sampling distributions of these components tend to be skewed (Wolfinger 1999). We retain ML estimates for φ_1 and φ_2 following the comments in Section 5.3. Furthermore, the positive value of β_1 implies that the treatment (drug) significantly increases the chance of a favourable cure (Wolfinger 1999).

PROC NLMIXED presumptively generates accurate results (Breslow 2003). Therefore, we conclude that, for the respiratory tract infections data, the GLIMMIX macro overestimates the model parameters, except for β_0 .

5.4 Modelling overdispersion in binomial data

Schall (1991, Table 2, p.726) presents a dataset, in binomial form, that summarizes the number of cancer cells that survived under radiation. We describe the experiment below. The data was found to be seriously overdispersed. The term *overdispersion* is used when the variance of the data under study exceeds the theoretical binomial scale factor which equals one.

5.4.1 Analyzing the mortality of cancer cells under radiation

The experiment involves measuring the mortality of cancer cells placed in a radiation chamber. Three dishes, each containing 400 cells, were irradiated at a time (or occasion). Thereafter, the number of surviving cells were counted. Since cells die naturally, dishes with cells were placed in the radiation chamber without being irradiated. The natural mortality of the cells could therefore be established. Twenty-seven binomial responses on nine occasions were available for analysis. Only the zero-dose effect was of interest. The data for this experiment are shown in Table 5.3.

Table 5.3 Cell irradiation data

Occasion	Dish	Number of cells survived/400
1	1	178
1	2	193
1	3	217
2	4	109
2	5	112
2	6	115
3	7	66
3	8	75
3	9	80
4	10	118
4	11	125
4	12	137
5	13	123
5	14	146
5	15	170
6	16	115
6	17	130
6	18	133
7	19	200
7	20	189
7	21	173
8	22	88
8	23	76
8	24	90
9	25	121
9	26	124
9	27	136

5.4.2 Results

A preliminary analysis of the cancer cells data reveals a Pearson X^2 -statistic of 492.9681, with $df = 27 - 1 = 26$ degrees-of-freedom. The estimate $\hat{\phi} = 18.9603$ for the scale factor is well above the binomial scale factor of 1. This is an indication that the data is highly overdispersed. To adjust for the extra-binomial variation, a vector of Gaussian distributed random effects γ_1 , which measures the variability across occasions, is introduced into the linear predictor on the logit scale such that

$$\ln \left(\frac{\mu_{ki}^\gamma}{1 - \mu_{ki}^\gamma} \right) = \beta_0 + \gamma_{1k} \quad (5.4.1)$$

where $i = 1, \dots, 3$ and $k = 1, \dots, 9$. Overdispersion persists since $\hat{\phi} = 1.8167$. A second vector of random effects $\gamma_2 \sim N(\mathbf{0}, \Gamma(\varphi_2))$, which measures the variability between dishes, is introduced into equation (5.4.1). The resulting model equation is

$$\ln \left(\frac{\mu_{ki}^\gamma}{1 - \mu_{ki}^\gamma} \right) = \beta_0 + \gamma_{1k} + \gamma_{2ki} \quad (5.4.2)$$

REPL parameter estimates generated by the GLIMMIX macro and ML estimates obtained from PROC NLMIXED are shown in Table 5.4.

Table 5.4 Parameter estimates \pm standard errors for the cancer cells data (p -values in parentheses)

Parameter	GLIMMIX ^a (REPL)	NLMIXED ^b (ML)
β_0	-0.7522 ± 0.1662 (0.0014)	-0.7527 ± 0.08989 (< 0.0001)
φ_1	0.2253 ± 0.1148 (0.0248)	0.2004 ± 66.7178 (0.9976)
φ_2	0.006431 ± 0.007741 (0.2030)	0.01218 ± 1.1419 (0.9916)

^aScale factor = 0.9986 ± 0.0136 (< 0.0001).

^bScale factor not computed.

The computed estimate for the scale factor by the GLIMMIX macro is 0.9986. This is clear evidence that overdispersion in the cancer cells data has been adjusted for by the incorporation of γ_1 and γ_2 into the linear predictor on the logit scale. The value of β_0 under the two procedures is almost identical with a small difference between the standard errors. REPL estimates for the dispersion components are very close to those obtained under ML. However, the REPL estimate for φ_1 is larger and that for φ_2 is slightly smaller than the corresponding ML dispersion components estimates. Standard errors for φ_1 and φ_2 under ML are greater than those under REPL. The p -values associated with β_0 , under both REPL and ML, and that associated with φ_1 , under REPL, indicate that these parameters are statistically significant at an α size test of 0.05. The p -value associated with the REPL estimate for φ_2 can be misleading since the decision not to reject the null hypothesis $H_0 : \varphi_2 = 0$ results in $\hat{\phi}$ being equal to 1.8167 – an indication of overdispersion in the data. Moreover, following the cautionary note about the p -values generated by PROC NL MIXED in Section 5.3.2, ML estimates for φ_1 and φ_2 are retained in spite of their associated p -values being nearly 1. Not rejecting the null hypothesis $H_0 : \varphi = \mathbf{0}$ under ML implies fitting an ordinary logistic regression and not adjusting for overdispersion.

We analyzed the cancer cells data by coding the GLIMMIX macro in terms of 0's and 1's. In terms of the original model, the computed deviance, with 26 degrees-of-freedom is 21.4731, which is less than $\chi_{0.95;26}^2 = 38.89$. Thus, there is no evidence of a lack-of-fit for the conditional model given γ_1 and γ_2 (Littell *et al* 1996, p.446).

5.5 Some remarks

For PQL to perform adequately, a rule of thumb is that the expected numbers of ‘successes’ and ‘failures’ for each response should be at least 5. For response probabilities in the mid-range, the binomial denominator should therefore be greater than 10. Larger denominators are required if many of the probabilities are near the boundaries in the interval $[0, 1]$. See Breslow (2003).

SAS codes for the analyses of the above datasets are to be found in Appendix B. Codes for Lin's (1997) χ_G^2 -test and those for the CMSEP of Booth

& Hobert (1998) are not implemented in the GLIMMIX macro. However, regarding the datasets analyzed in this chapter, the hypothesis $H_0 : \varphi = \mathbf{0}$ would be rejected at the 5% level of significance since the χ_G^2 -statistic would then be greater than or equal to $\chi_{0.95;c}^2 = 5.99$, where $c = 2$. Moreover, we would expect the CMSEP to be slightly larger than the (unconditional) standard errors of prediction ‘StdErrPred’ printed by using the option ‘out =_pred’ in the GLIMMIX macro (see also Booth & Hobert 1998, Table 6, p.270). We note that values for the standard errors of prediction generated by PROC NLMIXED are smaller than those computed by the GLIMMIX macro for both datasets.

5.5.1 Further remarks

For binomial data, PQL performs well when the denominators are large, for then the distributions of the responses are approximately Gaussian (Breslow 2003). Moreover, predicted values for responses in GLMMs depend on estimates $\hat{\beta}$, for fixed effects, and predictors $\hat{\gamma}$, for random effects. Residuals for individual outcomes are then given by $y_i - \hat{\mu}_i^{\hat{\gamma}}$, where $\hat{\mu}_i^{\hat{\gamma}} = g^{-1}(\mathbf{x}_i^T \hat{\beta} + \mathbf{z}_i^T \hat{\gamma})$ and $i = 1, \dots, n$. However, these residuals are heteroscedastic and should therefore be standardized (see Brown & Prescott 1999, p.134), and are thus given by

$$res_i = \frac{y_i - \hat{\mu}_i^{\hat{\gamma}}}{\sqrt{V(\hat{\mu}_i^{\hat{\gamma}})}} \quad (5.5.1)$$

where $V(\hat{\mu}_i^{\hat{\gamma}}) = \hat{\mu}_i^{\hat{\gamma}}(1 - \hat{\mu}_i^{\hat{\gamma}})$. We shall refer to equation (5.5.1)² as the (*conditional*) *Pearson residuals* – ‘conditional’ in the sense that they explicitly depend on $\hat{\gamma}$.

For LMMs, Brown & Prescott (1999, p.78) use Normal probability plots to check the assumption of normality of residuals. Pawitan (2001) uses these plots to illustrate the distribution of the linked response $g(\mathbf{y})$ for a Logistic-Gaussian model for different values of the dispersion components. Similarly, we employ such plots in an attempt to verify the approximate normality of binomial responses (as mentioned in the previous paragraph) by an examination of the ordered (conditional) Pearson residuals for the respiratory tract

²Zhu & Lee (2003, p.301) define res_i similarly.

infections data and the cancer cells data using the GLIMMIX macro and PROC NL MIXED. The plots are shown in Figures 5.1-5.4.

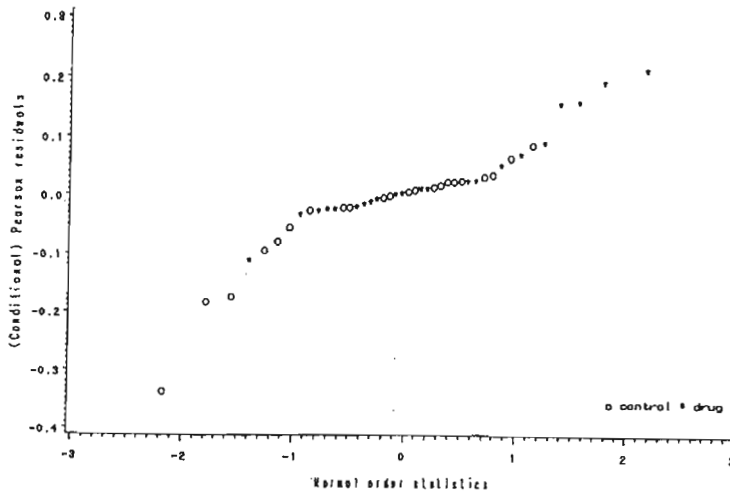


Figure 5.1 Normal probability plot for respiratory tract infections data under PQL

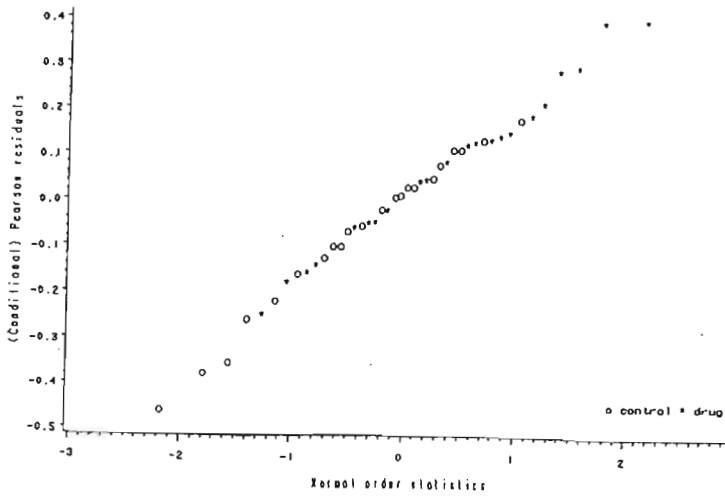


Figure 5.2 Normal probability plot for respiratory tract infections data under NL

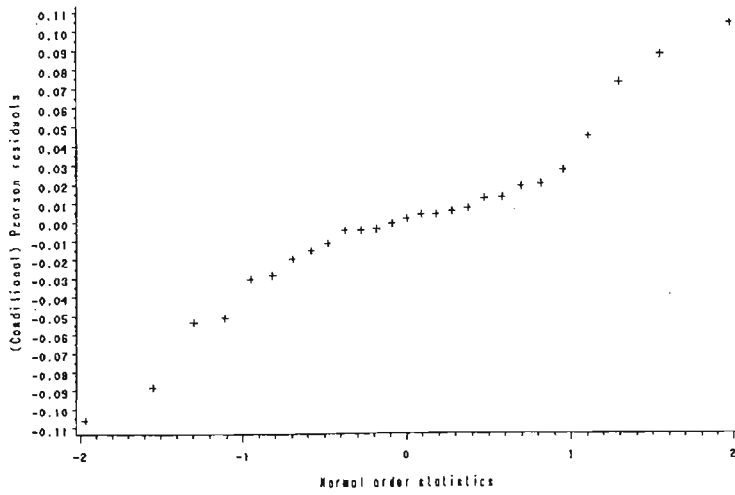


Figure 5.3 Normal probability plot for mortality of cancer cells data under REPL

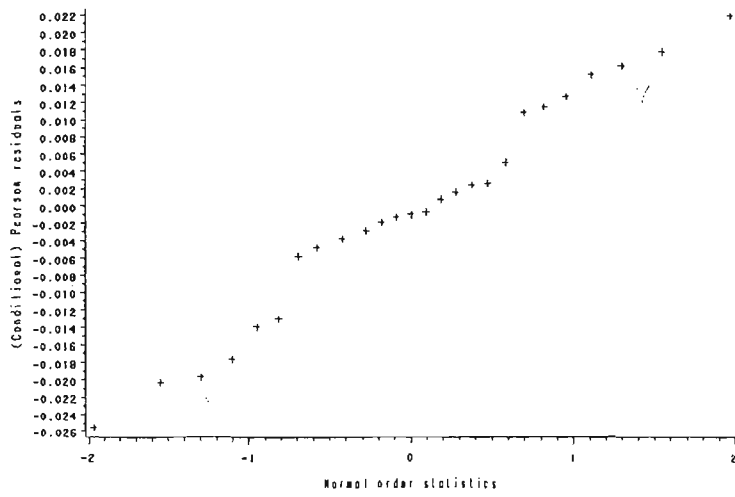


Figure 5.4 Normal probability plot for mortality of cancer cells data under ML

The Normal plot is markedly straight for the respiratory tract infections data, and somewhat straighter for the cancer cells data, under ML than those under PQL and REPL. Thus, we believe that PROC NLMIXED produces a better fit for both datasets. Furthermore, we point out that the former dataset contains many response probabilities that are in close proximity to 1, with one probability of 1, one near zero and one zero probability. We suspect that these near-boundary values are the reason why PQL has not performed well as opposed to the latter dataset where response probabilities are in the interval $[0.165, 0.5425]$. We suggest that an in-depth analysis be performed to identify the effect of influential responses, and/or trials, on parameter estimates for the respiratory tract infections data (see Chapter 6 for a theoretical proposition of local influence analysis in GLMMs).

The (conditional) Pearson residuals are defined as

$$res'_i = \frac{y_i - \hat{\mu}_i^{\hat{\gamma}}}{\sqrt{\hat{\phi} V(\hat{\mu}_i^{\hat{\gamma}})}} \quad (5.5.2)$$

in the GLIMMIX macro. For the cancer cells data, $\hat{\phi}^{-1/2} = 1.00070 \approx 1$ and, therefore, the difference between equation (5.5.1) and equation (5.5.2) is infinitesimal. Thus, the Normal plots in Figures 5.3 and 5.4 can be used to assess the fit produced by the GLIMMIX macro and PROC NLMIXED.

We recommend that PROC NLMIXED be used in conjunction with the GLIMMIX macro, especially for binomial data. Moreover, the SAS statements should be coded in terms of 0's and 1's. In that way, the computed estimate for the scale factor will indicate overdispersion (or underdispersion) in the data. Thus, misleading results (see Section 5.4.2) can be avoided.

Chapter 6

Local influence analysis for GLMMs

6.1 Introduction

Statistical models are often approximate descriptions of more complicated (stochastic) processes that generate a set of data. As a consequence, these models are, in many situations, not exact. Because of this inaccuracy, a study of the variation in key results of an analysis, under minor perturbations of the hypothesized models, is essential. If no influence on the results is detected, then the sample is robust with respect to the induced perturbations. Otherwise, there is cause for concern. See Cook (1986).

In the context of GLMMs, local influence measures to assess cluster influence on parameter estimates under several perturbation schemes have been proposed by Xiang *et al* (2003). They follow the approach of Cook (1986) and the GLMM formulation of McGilchrist (1994). In contrast, Zhu & Lee (2003) consider local influence analysis for GLMMs based on a Q -function associated with the conditional expectation of the complete-data likelihood function in the EM algorithm (Dempster *et al* 1977; McLachlan & Krishnan 1997), and where random effects are treated as missing data. Furthermore, Zhu & Lee (2003) state that the Q -displacement function and the likelihood displacement function of Cook (1986) have similar behaviours and statistical properties.

Following Cook's (1986) methodology for local influence analysis and the GLMM approach of Breslow & Clayton (1993), a perturbation scheme to assess cluster-specific¹ local influence on parameter estimates in GLMMs (Xiang *et al* 2003) with canonical link functions is proposed in this chapter. It is assumed that the conditional distribution of the responses, given Gaussian distributed random effects, belongs to the exponential family.

¹Simultaneous changes in weights of all clusters and random effects (Zhu & Lee 2003).

6.2 Measuring local influence in GLMMs

The parameter vector of interest is denoted by $\boldsymbol{\psi}^T = (\phi, \boldsymbol{\beta}^T, \boldsymbol{\varphi}^T)$. Let the (maximum) penalized quasi-likelihood (PQL) estimate be $\hat{\boldsymbol{\psi}}$ and let $l(\boldsymbol{\psi})$ denote the penalized (log-) quasi-likelihood function. Consider an r -dimensional perturbation vector $\boldsymbol{\omega} = (\omega_1, \dots, \omega_r)^T$, where $\boldsymbol{\omega}$ is restricted to vary in an open region $\Omega \subset \mathbb{R}^r$. Let $\hat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$ be the (maximum) PQL estimate given $\boldsymbol{\omega}$. The corresponding perturbed PQL function is denoted by $l(\boldsymbol{\psi}|\boldsymbol{\omega})$ for a given $\boldsymbol{\omega} \in \Omega$. The following assumption holds: There exists a point $\boldsymbol{\omega}_0 \in \Omega$ such that $l(\boldsymbol{\psi}) = l(\boldsymbol{\psi}|\boldsymbol{\omega}_0)$ for all $\boldsymbol{\psi}$ ($\boldsymbol{\omega}$ is chosen so that the application is meaningful). Moreover, $l(\boldsymbol{\psi}|\boldsymbol{\omega})$ is assumed to be continuous and twice-differentiable in $(\boldsymbol{\psi}^T, \boldsymbol{\omega}^T)$. The likelihood displacement function is defined as

$$LD(\boldsymbol{\omega}) = 2[l(\hat{\boldsymbol{\psi}}) - l(\hat{\boldsymbol{\psi}}_{\boldsymbol{\omega}})] \quad (6.2.1)$$

where $LD(\boldsymbol{\omega})$ measures the amount of the displacement of $\hat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$ from $\hat{\boldsymbol{\psi}}$ with respect to the contours of the PQL function. Equation (6.2.1) can be regarded in terms of the asymptotic uncertainty band for $\boldsymbol{\psi}$ – that is,

$$\{\boldsymbol{\psi} | 2[l(\hat{\boldsymbol{\psi}}) - l(\boldsymbol{\psi})] < \chi_{\alpha, df}^2\}$$

where α is the level of significance of a chi-squared distribution with $df = 1 + p + c$ degrees-of-freedom. A graph $G(\boldsymbol{\omega})$ of $LD(\boldsymbol{\omega})$ against $\boldsymbol{\omega}$ reveals the influence of a particular perturbation scheme. $G(\boldsymbol{\omega})$ is called an *influence graph* and is a geometric surface generated by values of the $(r + 1) \times 1$ vector $(\boldsymbol{\omega}^T, LD(\boldsymbol{\omega}))^T$ for varying $\boldsymbol{\omega} \in \Omega$. The behaviour of $G(\boldsymbol{\omega})$ is analyzed using geometric normal curvatures around $\boldsymbol{\omega}_0$ because it is not feasible to evaluate $LD(\boldsymbol{\omega})$ for every $\boldsymbol{\omega} \in \Omega$. The normal curvature at $\boldsymbol{\omega}_0$ is defined as

$$C(\boldsymbol{\lambda}) = 2 \left| \boldsymbol{\lambda}^T \frac{\partial^2 LD(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T} \boldsymbol{\lambda} \right|$$

where $\boldsymbol{\lambda}$ is a non-null unit vector in \mathbb{R}^r . It can be shown that

$$C(\boldsymbol{\lambda}) = 2 \left| \boldsymbol{\lambda}^T \left(\frac{\partial^2 l(\boldsymbol{\psi}|\boldsymbol{\omega})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\omega}^T} \right)^T \left(\frac{\partial^2 l(\boldsymbol{\psi}|\boldsymbol{\omega})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \right)^{-1} \left(\frac{\partial^2 l(\boldsymbol{\psi}|\boldsymbol{\omega})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\omega}^T} \right) \boldsymbol{\lambda} \right| \quad (6.2.2)$$

$\forall \boldsymbol{\lambda} \in \Omega$ and $\|\boldsymbol{\lambda}\| = 1$.

In equation (6.2.2), $\frac{\partial^2 l(\boldsymbol{\psi}|\boldsymbol{\omega})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\omega}^T}$ is an $r \times (1 + p + c)$ matrix and $\frac{\partial^2 l(\boldsymbol{\psi}|\boldsymbol{\omega})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}$ is of order $1 + p + c$. The local influence of the perturbation of $\boldsymbol{\omega}$ on $\hat{\boldsymbol{\psi}}$, in the direction of $\boldsymbol{\lambda}$, is given by $C(\boldsymbol{\lambda})$. The largest eigenvector $\boldsymbol{\lambda}_{\max}$ associated with C_{\max} , the largest eigenvalue, is the most important diagnostic. It indicates the direction in which the greatest local change of $LD(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_0$ can be achieved by a specific perturbation scheme. $\boldsymbol{\lambda}_{\max}$ reveals data points influencing the sensitivity of $LD(\boldsymbol{\omega})$. Therefore, plots of $\boldsymbol{\lambda}_{\max}$ suffice, in some cases, for gauging influence irrespective of $|C_{\max}|$. See Cook (1986) and Xi-ang *et al* (2003).

6.3 Cluster weights and random effects perturbation

6.3.1 Notation for a two-level clustered design

For clustered designs, the data are assumed to be arranged in a series of K independent clusters (Lin 1997). Let y_{ki} denote the i th response ($i = 1, \dots, n_k$) from the k th cluster ($k = 1, \dots, K$), and let $\boldsymbol{\gamma}_k$ be a q -dimensional vector of Gaussian distributed random effects such that $\boldsymbol{\gamma}_k \sim N(\mathbf{0}, \Gamma(\boldsymbol{\varphi}_k))$ that is associated with the k th cluster. Given the random effects, the conditional distribution of the responses is assumed to be a member of the exponential family with linear predictor $\eta_{ki}^{\boldsymbol{\gamma}_k} = g(\boldsymbol{\mu}_{ki}^{\boldsymbol{\gamma}_k}) = \mathbf{x}_{ki}^T \boldsymbol{\beta} + \mathbf{z}_{ki}^T \boldsymbol{\gamma}_k$ where $\boldsymbol{\mu}_{ki}^{\boldsymbol{\gamma}_k} = E(y_{ki}|\boldsymbol{\gamma}_k)$ and, \mathbf{x}_{ki} and \mathbf{z}_{ki} are p - and q -vectors of covariates associated with y_{ki} , respectively (Booth & Hobert 1998). Moreover, let \mathbf{y}_k denote the $n_k \times 1$ response vector from the k th cluster with $n_k \times p$ model matrix \mathbf{X}_k associated with fixed effects and $n_k \times q$ incidence matrix \mathbf{Z}_k associated with random effects. Then, the linear predictor becomes $\boldsymbol{\eta}_k^{\boldsymbol{\gamma}_k} = g(\boldsymbol{\mu}_k^{\boldsymbol{\gamma}_k}) = \mathbf{X}_k \boldsymbol{\beta} + \mathbf{Z}_k \boldsymbol{\gamma}_k$ (Lin 1997). The objective function for $\boldsymbol{\psi}$ (ignoring constant terms) has the form

$$L(\boldsymbol{\psi}) = \sum_{k=1}^K \log \int_{\mathbb{R}^q} \prod_{i=1}^{n_k} f(y_{ki}|\boldsymbol{\gamma}_k) |\Gamma(\boldsymbol{\varphi})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{\gamma}_k^T \Gamma^{-1}(\boldsymbol{\varphi}) \boldsymbol{\gamma}_k\right) d\boldsymbol{\gamma}_k$$

where

$$f(y_{ki}|\boldsymbol{\gamma}_k) = \exp\left\{\phi^{-1} u_{ki} [y_{ki} \theta_{ki} - b(\theta_{ki})] + c(y_{ki}, \phi)\right\}$$

and the PQL function $l(\psi)$ has the form

$$\sum_{k=1}^K \left[\sum_{i=1}^{n_k} [\phi^{-1} u_{ki} (y_{ki} \theta_{ki} - b(\theta_{ki})) + c(y_{ki}, \phi)] - \frac{1}{2} \gamma_k^T \Gamma^{-1}(\varphi) \gamma_k - \frac{1}{2} \log |\Gamma(\varphi)| \right] \quad (6.3.1)$$

(Zhu & Lee 2003), where $\mu_{ki}^{\gamma_k} = \frac{\partial b(\theta_{ki})}{\partial \theta_{ki}}$. In the following proposed perturbation scheme, $\omega_0^T = \mathbf{1}_n$, where $n = \sum_{k=1}^K n_k$, so that $l(\psi|\omega_0) = l(\psi)$ (Xiang *et al* 2003). θ_{ki} should strictly be written as $\theta_{ki}^{\gamma_k}$. However, for convenience, we avoid the latter notation.

6.3.2 Proposed scheme

A perturbation scheme for simultaneous changes in the weights of all clusters and random effects is proposed, where the perturbation vector is $\omega = (\omega_1, \dots, \omega_K)^T$. The perturbed PQL function $l(\psi|\omega)$ has the form

$$\sum_{k=1}^K \omega_k \left\{ \sum_{i=1}^{n_k} [\phi^{-1} u_{ki} (y_{ki} \theta_{ki} - b(\theta_{ki})) + c(y_{ki}, \phi)] - \frac{1}{2} \gamma_k^T \Gamma^{-1} \gamma_k - \frac{1}{2} \log |\Gamma| \right\} \quad (6.3.2)$$

Now,

$$\frac{\partial^2 l(\psi|\omega)}{\partial \omega_k \partial \phi} = \sum_{i=1}^{n_k} \left[-\phi^{-2} u_{ki} (y_{ki} \theta_{ki} - b(\theta_{ki})) + \frac{dc(y_{ki}, \phi)}{d\phi} \right] \quad (6.3.3a)$$

$$\frac{\partial^2 l(\psi|\omega)}{\partial \omega_k \partial \beta^T} = \sum_{i=1}^{n_k} \frac{\phi^{-1} u_{ki} (y_{ki} - \mu_{ki}^{\gamma_k}) \mathbf{x}_{ki}^T}{V(\mu_{ki}^{\gamma_k}) [g'(\mu_{ki}^{\gamma_k})]^2} \quad (6.3.3b)$$

and

$$\begin{aligned} \frac{\partial^2 l(\psi|\omega)}{\partial \omega_k \partial \varphi_k} &= \frac{\partial}{\partial \varphi_k} \left(\frac{\partial l(\psi|\omega)}{\partial \omega_k} \right) \\ &= \frac{1}{2} \gamma_k^T \Gamma^{-1} \frac{\partial \Gamma}{\partial \varphi_k} \Gamma^{-1} \gamma_k - \frac{1}{2} \text{tr} \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \varphi_k} \right) \\ &= \frac{1}{2} \text{tr} \left[\Gamma^{-1} \frac{\partial \Gamma}{\partial \varphi_k} \Gamma^{-1} (\gamma_k \gamma_k^T - \Gamma) \right] \end{aligned} \quad (6.3.3c)$$

Equations (6.3.3) together with the following second-order derivatives are necessary for assessing local influence:

$$\frac{\partial^2 l(\psi|\omega)}{\partial\phi\partial\phi} = \sum_{k=1}^K \sum_{i=1}^{n_k} \left\{ [2\phi^{-3}u_{ki}(y_{ki}\theta_{ki} - b(\theta_{ki}))] + \frac{d^2c(y_{ki}, \phi)}{d\phi^2} \right\} \quad (6.3.4a)$$

$$\frac{\partial^2 l(\psi|\omega)}{\partial\beta\partial\beta^T} = -\mathbf{X}^T \mathbf{W} \mathbf{X} \quad (6.3.4b)$$

where $\mathbf{W} = \text{diag}(\mathbf{W}_1, \dots, \mathbf{W}_K)$, with $\mathbf{W}_k = [\phi u_{ki}^{-1} V(\mu_{ki}^{\gamma_k}) [g'(\mu_{ki}^{\gamma_k})]^2]^{-1}$ and $\mathbf{X}^T = (\mathbf{X}_1^T, \dots, \mathbf{X}_K^T)$ (Zhu & Lee 2003),

$$\begin{aligned} \frac{\partial^2 l(\psi|\omega)}{\partial\phi\partial\beta^T} &= \frac{\partial}{\partial\phi} \left(\sum_{i=1}^n \frac{\phi^{-1} u_{ki}(y_{ki} - \mu_{ki}^{\gamma_k}) \mathbf{x}_{ki}^T}{V(\mu_{ki}^{\gamma_k}) g'(\mu_{ki}^{\gamma_k})} \right) \\ &= - \sum_{k=1}^K \sum_{i=1}^{n_k} \frac{\phi^{-2} u_{ki}(y_{ki} - \mu_{ki}^{\gamma_k}) \mathbf{x}_{ki}^T}{V(\mu_{ki}^{\gamma_k}) g'(\mu_{ki}^{\gamma_k})} \end{aligned} \quad (6.3.4c)$$

$$\begin{aligned}
\frac{\partial^2 l(\varphi|\omega)}{\partial\varphi\partial\varphi^T} &= \frac{1}{2}tr \left\{ \frac{\partial}{\partial\varphi_m} \left(\gamma^T \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \Gamma^{-1} \gamma_k \right) - \frac{\partial}{\partial\varphi_m} \left(\Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \right) \right\} \\
&= \frac{1}{2}tr \left\{ \gamma_k^T \frac{\partial}{\partial\varphi_m} \left(\Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \Gamma^{-1} \right) \gamma_k - \frac{\partial}{\partial\varphi_m} \left(\Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \right) \right\} \\
&= \frac{1}{2}tr \left\{ \gamma_k^T \left[\frac{\partial\Gamma^{-1}}{\partial\varphi_m} \frac{\partial\Gamma}{\partial\varphi_k} \Gamma^{-1} + \Gamma^{-1} \frac{\partial}{\partial\varphi_m} \left(\frac{\partial\Gamma}{\partial\varphi_k} \right) \Gamma^{-1} \right. \right. \\
&\quad \left. \left. + \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \frac{\partial\Gamma^{-1}}{\partial\varphi_m} \right] \gamma_k - \left[\frac{\partial\Gamma^{-1}}{\partial\varphi_m} \frac{\partial\Gamma}{\partial\varphi_k} + \Gamma^{-1} \frac{\partial}{\partial\varphi_m} \left(\frac{\partial\Gamma}{\partial\varphi_k} \right) \right] \right\} \\
&= \frac{1}{2}tr \left\{ \gamma_k^T \left[-\Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_m} \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \Gamma^{-1} + \Gamma^{-1} \frac{\partial^2\Gamma}{\partial\varphi_m\partial\varphi_k} \Gamma^{-1} \right. \right. \\
&\quad \left. \left. - \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_m} \Gamma^{-1} \right] \gamma_k \right. \\
&\quad \left. - \left[-\Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_m} \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} + \Gamma^{-1} \frac{\partial^2\Gamma}{\partial\varphi_m\partial\varphi_k} \right] \right\} \\
&= \frac{1}{2}tr \left\{ \gamma_k^T \left(-2\Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_m} \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \Gamma^{-1} \right) \gamma_k + \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_m} \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \right. \\
&\quad \left. + \gamma_k^T \Gamma^{-1} \frac{\partial^2\Gamma}{\partial\varphi_m\partial\varphi_k} \Gamma^{-1} \gamma_k - \Gamma^{-1} \frac{\partial^2\Gamma}{\partial\varphi_m\partial\varphi_k} \right\} \\
&= \frac{1}{2}tr \left\{ \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_m} \Gamma^{-1} \frac{\partial\Gamma}{\partial\varphi_k} \Gamma^{-1} (-2\gamma_k \gamma_k^T + \Gamma) \right. \\
&\quad \left. + \Gamma^{-1} \frac{\partial^2\Gamma}{\partial\varphi_m\partial\varphi_k} \Gamma^{-1} (\gamma_k \gamma_k^T - \Gamma) \right\}
\end{aligned} \tag{6.3.4d}$$

for $m, k = 1, \dots, K$.

Equations (6.3.3) and (6.3.4) are used to obtain measures for local influence. The proposed perturbation scheme is similar to Scheme 4 of Zhu & Lee (2003). Furthermore, $\hat{\psi}$ may be obtained as outlined in Section 4.6. (In deriving equations (6.3.3c) and (6.3.4d), we employed differential calculus identities from McCulloch & Searle (2001, pp.297–299).

6.4 Some remarks

Xiang *et al* (2002) identify influential clusters by considering cluster-wise deletion. Random effects are assumed to be independent. Therefore, deletion of a particular cluster, say m , has no effect on $\hat{\gamma}_k$, $m \neq k$. Thus, γ_k, φ_k and ϕ can be regarded as nuisance parameters. The focus is then on the regression coefficients estimator. They propose a first-order approximation to $\hat{\beta}_{(m)}$ to measure the influence of a deleted m th cluster on $\hat{\beta}$. This approximation is used to derive a generalized Cook's distance. Furthermore, they investigate *masking effects* in GLMMs via two procedures:

- (i) Joint influence of paired cluster-wise deletion on $\hat{\beta}$ is assessed by a generalized joint Cook's distance.
- (ii) Conditional influence of cluster k , say, after deletion of cluster m , say, on $\hat{\beta}$ is gauged by a generalized conditional Cook's distance.

A stochastic *robust Monte Carlo Newton-Raphson* (RMCNR) procedure, which avoids the computational difficulties involving high-dimensional integrals often encountered in GLMMs, has recently been proposed in the literature for bounded influence² *robust maximum likelihood* (RML) estimation of the parameter vector $\psi^T = (\beta^T, \varphi^T)$ with $\phi = 1$. Simulation results indicate that the RML method is useful in downweighting influential data points, which originate when a small proportion of the data deviate from their 'true' underlying distribution and may come from an arbitrary distribution, when evaluating these parameters. Furthermore, it has been shown that the RML estimator for β is consistent and is asymptotically Gaussian distributed under certain regularity conditions. Robust analysis is appealing in the sense that no information on the data is completely lost since all the observations are included in the study. See Sinha (2004) for details.

²The study of robust statistics is mathematically rigorous. See Hampel *et al* (1986) for a treatment of the theory of robustness based on the notion of *influence functions*.

Chapter 7

Conclusion

We indicate some directions for further research in the context of GLMMs.

The bias correction procedure of Lin & Breslow (1996a) for the PQL regression coefficients in their equation (18) is a partial second-order corrected estimator. Evaluating the difference between this equation and an exact second-order corrected estimator should be investigated. Furthermore, for binary data, corrected PQL estimators of regression coefficients may not be necessary since uncorrected ones give satisfactory results under ‘small dispersion asymptotics’ (Breslow & Clayton 1993) – that is, when the binomial denominator increases (Lin & Breslow 1996a). Bias correction via simulation can be performed for multiple components of dispersion (Pawitan 2001).

In a fully crossed design, the number of observations at a particular level of a specific factor is proportional to the number of levels of another factor. In such a case, the global score test may not be applicable as the number of responses and the number of levels tend to infinity. Thus, a study of the asymptotic distribution of the χ_G^2 -statistic for a fully crossed design is required. The performance of the individual dispersion component test based on Laplace approximation is unsatisfactory for binary data. As the binomial denominator increases, the Laplace approximation performs better, and so does the approximate individual score test. It would therefore be worthwhile to study the asymptotic bias of the Laplace approximation. See Lin (1997).

Properties of the tests proposed by Lin (1997), as well as the CMSEP of Booth & Hobert (1998), when applied to the HGLMs of Lee & Nelder (1996) should be studied. Lahiri & Rao (1995) have shown that an estimate of the UMSEP derived by Prasad & Rao (1990) is at times valid when the assumption of normality for the random effects is relaxed (cf. Booth & Hobert 1998). It would be of interest to investigate the CMSEP based solely on moment assumptions for the responses and the random effects.

Uncertainty bands for predicted random effects may be constructed using bootstrap techniques. However, there exists controversies regarding these methods: They do not fully reflect the relevant uncertainties. These drawbacks can be remedied by an examination of posterior distributions within the Bayesian framework (Breslow & Clayton 1993). Order-restricted inference, although theoretically and computationally complicated, increases efficiency (Hall & Praestgaard 2001). It would be of interest to investigate such inferential procedure when moment assumptions for the responses are relaxed.

PROC NLMIXED has been developed for fitting non-linear mixed models, where random effects enter the models non-linearly. Because of this non-linear relationship, the procedure has no direct analog to the REML method and, therefore, only the ML method is available (Wolfinger 1999). Moreover, when a very small number of correlated random effects is observed in each cluster in the analysis of clustered data, PROC NLMIXED is still restricted (Breslow 2003).

Our proposition for local influence analysis for GLMMs is motivated on the basis that the conditional distribution of the observations, given Gaussian distributed random effects, belongs to the exponential family. Cook's (1986) idea for local influence analysis relies on a well-behaved likelihood and makes use of log-likelihood contours for assessing local influence. Since quasi-likelihood functions (Wedderburn 1974) share similar properties to log-likelihood functions proper, it is expected that when parametric assumptions are relaxed for the responses and the random effects, local influence measures can be derived. In this context, $LD(\omega)$ may be termed *quasi-likelihood displacement function*. It would also be of interest to program the SAS GLIMMIX macro for local influence analysis in GLMMs.

Sinha's (2004) proposition for robust estimation of model parameters may well be extended by assuming that moment conditions suffice for the responses. It would be worth investigating robust analysis for REML estimation in GLMMs. The estimators may be called *robust penalized quasi-likelihood* (RPQL) estimators. We note that Sinha (2004) does not place any parametric form on the random effects.

Finally, we conclude that the key paper by Breslow & Clayton (1993) was a major breakthrough for the analysis of correlated discrete data. We believe that further research should be envisaged within the framework of GLMMs when the uncorrelatedness assumption for the random effects is relaxed, as in the analysis of longitudinal data where random effects are often time-dependent (Sinha 2004).

References

- Bhat, B.R. & Nagnur, B.N. (1965). Locally asymptotically most stringent tests and Lagrangian multiplier tests of linear hypotheses. *Biometrika* 52, pp. 459-468.
- Booth, J.G. & Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* 93, pp. 262-272.
- Breslow, N. (2003). *Whither PQL?* University of Washington Biostatistics Working Paper Series, Working Paper 192.
<http://www.bepress.com/uwbiostat/paper192>
- Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88, pp. 9-25.
- Breslow, N.E. & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 82, pp. 81-91.
- Brown, H. & Prescott, R. (1999). *Applied Mixed Models in Medicine*. John Wiley & Sons Ltd, Chichester, England.
- Casella, G. & Berger, R.L. (2002). *Statistical Inference*. (Second Edition). Duxbury, Wadsworth Group, California.
- Clayton, D. (1992). Generalized linear mixed models in biostatistics. *The Statistician* 41, pp. 327-328.
- Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, New Jersey.
<http://www.ats.ucla.edu/stat/sas/examples/vizdata/chapter1.htm>
- Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society* B48, pp. 133-169.
- Cox, D.R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society* B49, pp. 1-39.

- Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B39*, pp. 1-38.
- Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*. Chapman & Hall, London.
- Engel, B., Buist, W. & Visscher, A. (1995). Inference for threshold models with variance components from the generalized linear mixed model perspective. *Genetics Selection Evolution 27*, pp. 15-32.
- Engel, B. & Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica 48*, pp. 1-22.
- Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M.J.R. (1998). *A User's Guide to MLwiN*. Institute of Education, London.
- Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London.
- Hall, D.B. & Praestgaard, J.T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika 88*, pp. 739-751.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, Inc., New York.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems (with comment). *Journal of the American Statistical Association 72*, pp. 320-340.
- Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association 93*, pp. 720-729.
- Kennedy, W.J. & Gentle, J.E. (1980). *Statistical Computing*. Marcel Dekker, Inc., New York.

- Kuss, O. (2002). How to use SAS for logistic regression with correlated data. *Proceedings of the 27th Annual SAS Users Group International Conference (SUGI 27)*, Paper 261-27.
- Lahiri, P. & Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association* **90**, pp. 758-766.
- Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society* **B58**, pp. 619-678.
- Lee, Y. & Nelder, J.A. (1999). The robustness of the quasiliikelihood estimator. *The Canadian Journal of Statistics* **27**, pp. 321-327.
- Lee, Y. & Nelder, J.A. (2003). Extended-REML estimators. *Journal of Applied Statistics* **30**, pp. 845-856.
- Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84**, pp. 309-326.
- Lin, X. & Breslow, N.E. (1996a). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, pp. 1007-1016.
- Lin, X. & Breslow, N.E. (1996b). Analysis of correlated binomial data in logistic-normal models. *Journal of Statistical Computation and Simulation* **55**, pp. 133-146.
- Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute Inc., Cary, North Carolina.
- McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. (Second Edition). Chapman & Hall, London.
- McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Inc., New York.
- McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society* **B56**, pp. 61-69.

- McLachlan, G.J. & Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc., New York.
- Nelder, J.A. (2000). Quasi-likelihood and pseudo-likelihood are not the same thing. *Journal of Applied Statistics* **27**, pp. 1007-1011.
- Nelder, J.A. & Lee, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: Some Comparisons. *Journal of the Royal Statistical Society B54*, pp. 273-284.
- Nelder, J.A. & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, pp. 221-232.
- Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A135*, pp. 370-384.
- Patterson, H.D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, pp. 545-554.
- Pawitan, Y. (2001). Two-staged estimation of variance components in generalized linear mixed models. *Journal of Statistical Computation and Simulation* **69**, pp. 1-17.
- Prasad, N.G.N. & Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* **85**, pp. 163-171.
- Robinson, D.L. (1987). Estimation and use of variance components. *The Statistician* **36**, pp. 3-14.
- Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects (with comment). *Statistical Science* **6**, pp. 15-51.
- SAS Institute Inc. (2001). *The SAS System for Windows (Version 8.2)*. SAS Institute Inc., Cary, North Carolina.
- SAS Institute Inc. (2002). *GLIMMIX: A SAS macro for fitting generalized linear mixed models using PROC MIXED and the Output Delivery System (ODS)*. Requires SAS/STAT Version 8. SAS Institute Inc., Cary, North Carolina.
<http://ftp.sas.com/techsup/download/stat/glmm800.html>

- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, pp. 719-727.
- Schabenberger, O. (1998). *An Introduction to Publication Quality Graphics In SAS for Windows*. SAS Institute Inc., Cary, North Carolina. <http://home.nc.rr.com/schabenb/SASGraph.html#Adding%20a%20legend>
- Searle, S.R. (1971). *Linear Models*. John Wiley & Sons, Inc., New York.
- Searle, S.R. (1995). An overview of variance component estimation. *Metrika* **42**, pp. 215-230.
- Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. John Wiley & Sons, Inc., New York.
- Sinha, S.K. (2004). Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association* **99**, pp. 451-460.
- Smith, T.C., Spiegelhalter, D.J. & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* **14**, pp. 2685-2699.
- Stanner, T.J. & Duffy, D.E. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.
- Turner, R.M., Omar, R.Z., Yang, M., Goldstein, H. & Thompson, S.G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **19**, pp. 3417-3432.
- Urban Hjorth, J.S. (1994). *Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap*. Chapman & Hall, London.
- Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, pp. 439-447.
- Wedderburn, R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, pp. 27-32.

- Wolfinger, R.D. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. *Proceedings of the 24th Annual SAS Users Group International Conference (SUGI 24)*, pp. 287-294.
- Wolfinger, R. & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, pp. 233-243.
- Woodhouse, G., Rasbash, J., Goldstein, H., Yang, M. & Plewis, I. (1996). *Multilevel Modelling Applications: A Guide for Users of MLn*. Institute of Education, London.
- Xiang, L., Lee, A.H. & Tse, S.-K. (2003). Assessing local cluster influence in generalized linear mixed models. *Journal of Applied Statistics* **30**, pp. 349-359.
- Xiang, L., Tse, S.-K. & Lee, A.H. (2002). Influence diagnostics for generalized linear mixed models: Applications to clustered data. *Computational Statistics & Data Analysis* **40**, pp. 759-774.
- Zhu, H.-T. & Lee, S.-Y. (2003). Local influence for generalized linear mixed models. *The Canadian Journal of Statistics* **31**, pp. 293-309.

Appendix A

Some results

A.1 Cumulants

Let κ_{ri} denote the r th cumulant of y_i ($i = 1, \dots, n$) which is defined as

$$\kappa_{ri} = b^{(r)}(\theta_i) a^{r-1}(\phi) \quad (\text{A.1.1})$$

where $b^{(r)}(\theta_i)$ denotes the r th derivative of $b(\theta_i)$ with respect to θ_i . When $r = 2$,

$$\begin{aligned} \kappa_{2i} &= b^{(2)}(\theta_i) a(\phi) \\ &= \phi u_i^{-1} V(\mu_i) \end{aligned} \quad (\text{A.1.2})$$

where $b^{(2)}(\theta_i) = V(\mu_i)$ (and $a(\phi) = \phi u_i^{-1}$ (McCullagh & Nelder 1989, p.29)).

The third and fourth cumulants of y_i are related to κ_{2i} through the equation

$$\kappa_{(r+1)i} = \kappa_{2i} \frac{\partial \kappa_{ri}}{\partial \mu_i} \quad (\text{A.1.3})$$

for $r = 2, 3$ (Lin 1997; McCullagh & Nelder 1989, p.45). Thus,

$$\begin{aligned} \kappa_{3i} &= \kappa_{2i} \frac{\partial \kappa_{2i}}{\partial \mu_i} = \phi u_i^{-1} V(\mu_i) \frac{\partial \phi u_i^{-1} V(\mu_i)}{\partial \mu_i} \\ &= (\phi u_i^{-1})^2 V(\mu_i) V'(\mu_i) \end{aligned} \quad (\text{A.1.4})$$

and

$$\begin{aligned} \kappa_{4i} &= \kappa_{2i} \frac{\partial \kappa_{3i}}{\partial \mu_i} = (\phi u_i^{-1}) V(\mu_i) \frac{\partial (\phi u_i^{-1})^2 V(\mu_i) V'(\mu_i)}{\partial \mu_i} \\ &= (\phi u_i^{-1})^3 V(\mu_i) [V(\mu_i) V''(\mu_i) + V'(\mu_i) V'(\mu_i)] \\ &= (\phi u_i^{-1})^3 V(\mu_i) \{V(\mu_i) V''(\mu_i) + [V'(\mu_i)]^2\} \end{aligned} \quad (\text{A.1.5})$$

A.2 Standard cumulants

The *standard cumulants* (McCullagh & Nelder 1989, p.351) are given by

$$\begin{aligned}\rho_{3i}^2 &= \frac{\kappa_{3i}^2}{\kappa_{2i}^3} \\ &= \phi u_i^{-1} \frac{[V(\mu_i)]^2 [V'(\mu_i)]^2}{[V(\mu_i)]^3} \\ &= \phi u_i^{-1} V(\mu_i) \left[\frac{V'(\mu_i)}{V(\mu_i)} \right]^2\end{aligned}$$

Thus,

$$\rho_{3i} = \sqrt{\phi u_i^{-1} V(\mu_i)} \left[\frac{\partial \ln V(\mu_i)}{\partial \mu_i} \right] \quad (\text{A.2.1})$$

and

$$\begin{aligned}\rho_{4i} &= \frac{\kappa_{4i}}{\kappa_{2i}^2} \\ &= \phi u_i^{-1} V(\mu_i) \frac{\{V(\mu_i)V''(\mu_i) + [V'(\mu_i)]^2\}}{[V(\mu_i)]^2} \\ &= \phi u_i^{-1} \left\{ \frac{\partial^2 V(\mu_i)}{\partial \mu_i^2} + V(\mu_i) \left[\frac{\partial \ln V(\mu_i)}{\partial \mu_i} \right]^2 \right\}\end{aligned} \quad (\text{A.2.2})$$

Equations (A.2.1) and (A.2.2) can be expressed as

$$\rho_{3i} = \sqrt{\frac{\phi u_i^{-1}}{V(\mu_i)}} \frac{\partial V(\mu_i)}{\partial \mu_i}$$

and

$$\rho_{4i} = \phi u_i^{-1} \frac{\partial^2 V(\mu_i)}{\partial \mu_i^2} + \rho_{3i}^2$$

respectively (McCullagh & Nelder 1989, p.361).

A.3 A note on quasi-distributions

By using a suitable normalization, a quasi-likelihood Q can be transformed into a *quasi-distribution* with frequency function

$$f_Q = \frac{\exp(Q)}{\int \exp(Q)dy} \quad (\text{A.3.1})$$

and likelihood

$$l_Q = Q - \log \left(\int \exp(Q)dy \right) \quad (\text{A.3.2})$$

The ML equations are given by

$$\frac{\partial l_Q}{\partial \beta} = \frac{\partial Q}{\partial \beta} - \frac{\partial}{\partial \beta} \log \left(\int \exp(Q)dy \right) \quad (\text{A.3.3})$$

The difference between $\frac{\partial l_Q}{\partial \beta} (= 0)$ and $\frac{\partial Q}{\partial \beta} (= 0)$ is given by

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta} \log \left(\int \exp(Q)dy \right) \\ &= \frac{\partial}{\partial \mu} \log \left(\int \exp(Q)dy \right) \frac{\partial \mu}{\partial \beta} \\ &= \frac{1}{\int \exp(Q)dy} \left(\frac{\partial}{\partial \mu} \int \exp(Q)dy \right) \frac{\partial \mu}{\partial \beta} \\ &= \frac{\partial \mu}{\partial \beta} \frac{1}{\int \exp(Q)dy} \int \frac{\partial(\exp(Q))}{\partial \mu} dy \\ &= \frac{\partial \mu}{\partial \beta} \frac{1}{\int \exp(Q)dy} \int \exp(Q) \frac{\partial Q}{\partial \mu} dy \\ &= \frac{\partial \mu}{\partial \beta} \int \frac{\exp(Q)}{\int \exp(Q)dy} \frac{\partial Q}{\partial \mu} dy \\ &= \frac{\partial \mu}{\partial \beta} \int f_Q \frac{y - \mu}{\phi V(\mu)} dy \\ &= \frac{\partial \mu}{\partial \beta} \frac{\mu^* - \mu}{\phi V(\mu)} \end{aligned} \quad (\text{A.3.4})$$

where $\mu^* = \int y f_Q dy$ is called the *quasi-mean* (Nelder & Lee 1992) and $\int f_Q dy = 1$.

The normalizing factor of such a quasi-distribution depends on the mean. This dependence does not exist for GLMs. Therefore, MQL estimators and those derived from a quasi-distribution do not coincide. If the normalizing factor changes slowly with the mean, as is the case in general, then these two estimators differ by a small amount. Thus, MQL estimates are expected to be good approximations to ML estimates obtained from the quasi-distribution. See Nelder (2000).

(In the derivation of equation (A.3.4), we have, for notational convenience, suppressed subscript i in y_i , ($i = 1, \dots, n$) and j in β_j ($j = 1, \dots, p$), and the integral is over (y_i, μ_i) . We have also omitted the weights u_i).

In GLMs, cumulants are obtained from derivatives of $b(\theta)$. The third and fourth cumulants of a quasi(-likelihood) distribution are usually well approximated by formulae derived from $b(\theta)$ that would hold if an exact GLM exists. The choice of a non-Gaussian quasi-likelihood implies assuming that the distribution of the data errors is skew (Nelder 2000). Given a variance function, Lee & Nelder (1999) recommend the use of the MQL estimator for fixed effects on the basis of its robustness and conservatism.¹

¹Conservatism implies minimizing a maximum risk (Lee & Nelder 1999, p.321).

Appendix B

SAS codes for the analyses of the respiratory tract infections and the mortality of cancer cells datasets

```
/*PQL ANALYSIS OF RESPIRATORY TRACT INFECTIONS DATA (SMITH ET AL 1995) UNDER REML*/
/*Include GLIMMIX macro*/
data resp_infec;
  input trial trt $ fav unfav;
  nki=fav+unfav;
  if fav=0 then fav=0.1/nki;
  datalines;
1 drug 40 7
1 cntl 29 25
2 drug 34 4
2 cntl 17 24
3 drug 76 20
3 cntl 58 37
4 drug 13 1
4 cntl 6 11
5 drug 38 10
5 cntl 23 26
6 drug 99 2
6 cntl 71 13
7 drug 149 12
7 cntl 132 38
8 drug 27 1
8 cntl 31 29
9 drug 18 1
9 cntl 11 9
10 drug 27 22
10 cntl 3 44
11 drug 137 25
11 cntl 130 30
12 drug 169 31
12 cntl 145 40
13 drug 30 9
13 cntl 31 10
14 drug 171 22
14 cntl 145 40
15 drug 45 0
15 cntl 42 4
16 drug 100 31
16 cntl 80 60
17 drug 71 4
17 cntl 63 12
18 drug 189 31
18 cntl 183 42
19 drug 48 7
19 cntl 31 26
20 drug 88 3
20 cntl 75 17
```

```

21 drug 11 14
21 cntl 0 23
22 drug 62 3
22 cntl 62 6
;
%glimmix(data=resp_infec,
  procopt=covtest,
  stmts=%str(
    class trial trt;
    model fav/nki=trt/ddfm=satterth;
    random trial trial*trt;
    estimate 'beta0' intercept 1 trt 1 0;
    estimate 'beta1' trt -1 1;
    parms (0) (0) (1)/eqcons=3;),
  error=binomial,link=logit,out=_pred)
run;
proc print data=_pred(keep=trial trt mu reschi stderpred);run;
legend across=1 frame position=(bottom right inside) mode=share value=('o control * drug');
axis1 label=(angle=90 f=simulate '(Conditional) Pearson residuals');
axis2 label=(f=simulate 'Normal order statistics');
symbol1 color=black font=swiss value='o' repeat=1;run;
symbol2 color=black font=swiss value='*' repeat=1;run;
footnote 'Figure 5.1 Normal probability plot for respiratory tract infections data under PQL';
proc rank out=reschi normal=tukey;var reschi;ranks rank;
proc gplot data=reschi;plot reschi*rank=trt/legend=legend vaxis=axis1 haxis=axis2;run;
/*-----*/
/*-----*/

/*ML ANALYSIS OF RESPIRATORY TRACT INFECTIONS DATA (SMITH ET AL 1995)
  WITH PROC NLMIXED (WOLFINGER 1999)*/
data resp_infec_NL;
  input trial trt y n;
  datalines;
1 1 40 47
1 0 29 54
2 1 34 38
2 0 17 41
3 1 76 96
3 0 58 95
4 1 13 14
4 0 6 17
5 1 38 48
5 0 23 49
6 1 99 101
6 0 71 84
7 1 149 161
7 0 132 170
8 1 27 28
8 0 31 60
9 1 18 19
9 0 11 20
10 1 27 49
10 0 3 47
11 1 137 162
11 0 130 160

```

```

12 1 169 200
12 0 145 185
13 1 30 39
13 0 31 41
14 1 171 193
14 0 145 185
15 1 45 45
15 0 42 46
16 1 100 131
16 0 80 140
17 1 71 75
17 0 63 75
18 1 189 220
18 0 183 225
19 1 48 55
19 0 31 57
20 1 88 91
20 0 75 92
21 1 11 25
21 0 0 23
22 1 62 65
22 0 62 68
run;
proc nlmixed data=resp_infec_NL;
  parms beta0=0.6 beta1=1.2 s2u1=0.95 cb12=0 s2u2=0.05;
  eta=beta0+beta1*trt+u1+u2;
  expeta=exp(eta);
  p=expeta/(1+expeta);
  predict p out=p;
  reschi=((y/n)-p)/(sqrt((p*(1-p)))) /*Pearson residuals*/;
  predict reschi out=reschi;
  model y ~ binomial(n,p);
  random u1 u2 ~ normal([0,0],[s2u1,cb12,s2u2])
  subject=trial;id reschi;
run;
proc print data=p(keep=trial trt pred reschi stderrpred);run;
legend across=1 frame position=(bottom right inside) mode=share value=('o control * drug');
axis1 label=(angle=90 f=simulate '(Conditional) Pearson residuals');
axis2 label=(f=simulate 'Normal order statistics');
symbol1 color=black font=swiss value='o' repeat=1;run;
symbol2 color=black font=swiss value='*' repeat=1;run;
footnote 'Figure 5.2 Normal probability plot for respiratory tract infections data under ML';
proc rank out=reschi normal=tukey;var reschi;ranks rank;
proc gplot data=reschi;plot reschi*rank=trt/legend=legend vaxis=axis1 haxis=axis2;run;
/*-----*/
/*-----*/
/*REPL ANALYSIS OF CANCER CELLS DATA UNDER RADIATION (SCHALL 1991)*/
/*Include GLIMMIX macro*/
data cells;
  input occasion dish $ survived not_survived;
  nki=survived+not_survived;
  if survived=0 then survived=0.1/nki;
  datalines;
1 1 178 222
1 2 193 207

```

```

1 3 217 183
2 1 109 291
2 2 112 288
2 3 115 285
3 1 66 334
3 2 75 325
3 3 80 320
4 1 118 282
4 2 125 275
4 3 137 263
5 1 123 277
5 2 146 254
5 3 170 230
6 1 115 285
6 2 130 270
6 3 133 267
7 1 200 200
7 2 189 211
7 3 173 227
8 1 88 312
8 2 76 324
8 3 90 310
9 1 121 279
9 2 124 276
9 3 136 264
;
/*Ordinary GLM*/
%glimmix(data=cells,
  procopt=covtest,
  stmts=%str(
    class occasion dish;
    model survived/nki=;
    parms (1);)
)
run;
/*GLMM with one random effect due to occasion*/
%glimmix(data=cells,
  procopt=covtest,
  stmts=%str(
    class occasion dish;
    model survived/nki=;
    random occasion;
    parms (0) (1);)
)
run;

/*GLMM with two random effects (due to occasion and dish) computed in terms of 0's and 1's*/
data new;
set cells;
do i=1 to survived;
  y=1;
output;
end;
do i=1 to not_survived;
  y=0;

```

```

output;
end;
%glimmix(data=new,
  procopt=covtest,
  stmts=%str(
    class occasion dish;
model y=/ddfm=satterth ;
random occasion dish;),
error=binomial,
link=logit);
run;
/*Computing deviance with dispersion components obtained from 0-1 data(=new)*/
%glimmix(data=cells,
  procopt=covtest,
  stmts=%str(
    class occasion dish;
    model survived/nki=/ddfm=satterth;
    random occasion dish;
    parms (0.2253) (0.006431) (0.9986)/eqcons=3;),
  error=binomial,link=logit,out=_pred)
run;
proc print data=_pred(keep=occasion dish mu reschi stderrpred);run;
axis1 label=(f=simulate angle=90 '(Conditional) Pearson residuals');
axis2 label=(f=simulate 'Normal order statistics');
footnote 'Figure 5.3 Normal probability plot for mortality of cancer cells data under REPL';
proc rank out=reschi normal=tukey;var reschi;ranks rank;
proc gplot data=reschi;plot reschi*rank/vaxis=axis1 haxis=axis2;run;
/*-----*/
/*-----*/
/*ML ANALYSIS OF CANCER CELLS DATA UNDER RADIATION (SCHALL 1991) WITH
PROC NLMIXED (WOLFINGER 1999)*/
data cells;
input occasion dish y n;
datalines;
1 1 178 400
1 2 193 400
1 3 217 400
2 1 109 400
2 2 112 400
2 3 115 400
3 1 66 400
3 2 75 400
3 3 80 400
4 1 118 400
4 2 125 400
4 3 137 400
5 1 123 400
5 2 146 400
5 3 170 400
6 1 115 400
6 2 130 400
6 3 133 400
7 1 200 400
7 2 189 400
7 3 173 400

```



```

8 1 88 400
8 2 76 400
8 3 90 400
9 1 121 400
9 2 124 400
9 3 136 400
run;
/*GLMM with two dispersion components (due to occasion and dish)*/
proc nlmixed data=cells;
  parms beta0=-0.7 s2u1=0.2 cb12=0 s2u2=0.05;
  eta=beta0+u1+u2;
  expeta=exp(eta);
  p=expeta/(1+expeta);
  predict p out=p;
  reschi=((y/n)-p)/(sqrt((p*(1-p))))/*Pearson residuals*/;
  predict reschi out=reschi;
  model y ~ binomial(n,p);
  random u1 u2 ~ normal([0,0],[s2u1,cb12,s2u2])
  subject=dish;id reschi;
run;
proc print data=p(keep=occasion dish pred reschi stderrpred);run;
axis1 label=(angle=90 f=simulate '(Conditional) Pearson residuals');
axis2 label=(f=simulate 'Normal order statistics');
footnote ' Figure 5.4 Normal probability plot for mortality of cancer cells data under ML';
proc rank out=reschi normal=tukey;var reschi;ranks rank;
proc gplot data=reschi;plot reschi*rank/vaxis=axis1 haxis=axis2;run;
/*-----*/
/*-----*/

```

/* NOTE

The GLIMMIX macro is available at

<http://ftp.sas.com/techsup/download/stat/glmm800.html>

SAS GLIMMIX codes used to compute parameter estimates are available in:

[1] Brown & Prescott (1999, p.196)

[2] Kuss (2002)

[3] Littell et al (1996, pp.439-440, p.446 and p.600)

SAS codes used to generate the Normal probability plots are available in:

[1] Brown & Prescott (1999, p.147)

[2] Cleveland (1993)

Download

<http://home.nc.rr.com/schabenb/SASGraph.html#Adding%20a%20legend>

[3] Schabenberger (1998)

Download

<http://www.ats.ucla.edu/stat/sas/examples/vizdata/chapter1.htm> */