ON APPROXIMATE LIKELIHOOD INFERENCE IN THE POISSON MIXED MODEL

ZHEN-DE QU

On Approximate Likelihood Inference in the Poisson Mixed Model

by

Zhen-De Qu

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirement for the degree of

Master of Science in Statistics

Department of Mathematics and Statistics

Memorial University of Newfoundland

January 1995

St. John's Newfoundland

ISBN 0-612-01905-5

Abstract

The application of the Poisson mixed model has been hampered by the difficulty of computation in evaluating the marginal likelihood of the parameters involved. Many approximate approaches have recently been proposed for inference about the generalized linear mixed model which refers to the Poisson mixed model as a special case, for example, the penalized quasi-likelihood (PQL) approach of Breslow and Clayton (1993), and the generalized estimating function (GEF) approach of Waclawiw and Liang (1993). We show in the thesis that both the PQL and GEF produce inconsistent inference for the variance component in the Poisson mixed model. The thesis then proposes a two-step approximate likelihood approach (AL) for the estimation of three types of parameters (fixed effect parameters, random effects and their variance component) in the Poisson mixed model. In the first step, an approximate likelihood function of count data is constructed to estimate the fixed effect parameters and the variance component by applying a conjugate Bayesian theorem. In the second step, the random effects are estimated by minimizing their approximate posterior mean square error. Our estimates are always consistent for both the fixed effect parameters and the variance component. When the actual variance component is near zero, our estimates are almost efficient for both the fixed effect parameters and the variance component, and are almost optimal for the random effects. When the actual variance component is away from zero, our estimates are always asymptotically unbiased for the fixed effect parameters, whereas our estimate is asymptotically negative biased for the variance component. Another desirable merit is that, unlike the existing approaches mentioned above, our estimates for both the

fixed effect parameters and the variance component only depend on the distribution of random effects rather than the estimates of random effects. An important finding is that the asymptotic covariance of our estimates for the fixed effect parameters will become smaller in general as the variance component, an index of the intra-cluster association, increases, and can be noticeably reduced by assigning the values of the fixed effect covariates as different as possible among different observations in any cluster. However, if the fixed effect covariate has the same or almost equal values among different observations in any cluster, the asymptotic variance of the estimate for the corresponding fixed effect parameter may increase as the variance component gets larger. This feature may be useful in designing a valid experiment or sampling for the Poisson mixed model. Unless the variance component is small, the fixed effect covariates should be designed to have values as different as possible among different observations in any cluster. It is further shown, through simulation, the proposed approach performs better than the PQL and GEF approaches.

Acknowledgements

First of all, I would like to thank my supervisory committee: Professors Brajendra Sutradhar, Uditha Balasooriya and Roy Bartelett for their guidance and support in conducting this thesis.

I take great pleasure in acknowledging the instructions and encouragement I received from Professors Chu-In Charles Lee, Rajendra Jain, Ning-Zhong Shi , Hong Wang, Mark Reimers, Nadine Hoenig, David Chu and Louise Dionne. I also want to thank Professors Edgar Goodaire, Bruce Watson, Ms Judy Lee, the Maple Leaf and many other professors and graduate students at the Department of Mathematics and Statistics for their hospitality and friendship.

As well, I am grateful to the School of Graduate Studies and the Department of Mathematics and Statistics for providing me with financial support in the form of a Graduate Student Scholarship and Graduate Assistantship to make my stay at the Memorial University of Newfoundland possible.

Finally, I would like to dedicate this thesis to Mr. Yu-Chu Zhu, my dearest teacher at high school.

Contents

List of Tables

Chapter 1

Introduction

Define y_{ij} as the jth ($j = 1, \ldots, n_i$) count observation associated with a $p \times 1$ observed vector x_{ij} of fixed effect covariates for the ith ($i = 1, \ldots, k$) cluster. Here n_i is the size of the i-th cluster and k is the total number of clusters. Let β denote a $p \times 1$ vector of unknown fixed effect parameters associated with the observed vectors $x_{i1}, \ldots, x_{in_i}$ of the fixed effect covariates, and γ_i denote univariate random effects. Given γ_i, the n_i observations y_{ij} ($j = 1, \ldots, n_i$) within the ith cluster are assumed to be independent, and to follow the Poisson distribution, yielding

$$f(y_i \mid \gamma_i) = \prod_{j=1}^{n_i} \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \exp(-\mu_{ij}) \tag{1.1}$$

where $f(y_i|\gamma_i)$ denotes the conditional probability density of $y_i = (y_{i1}, \ldots, y_{in_i})^T$ for a given γ_i, and

$$\mu_{ij} = E(y_{ij} \mid \gamma_i) = Var(y_{ij} \mid \gamma_i), \tag{1.2}$$

in which $E(y_{ij} \mid \gamma_i)$ and $Var(y_{ij} \mid \gamma_i)$ are the conditional mean and variance of y_{ij} for a given γ_i respectively. The conditional covariance of y_i given γ_i is

$$Cov(y_{ij}, y_{ij'} \mid \gamma_i) = \begin{cases} Var(y_{ij} \mid \gamma_i) = \mu_{ij} & \text{if } j = j' \\ 0 & \text{if } j \neq j'. \end{cases} \tag{1.3}$$

Let

$$\eta_{ij} = \log(\mu_{ij}),$$

then the conditional density in 1.1 can be reparametrized into the natural exponential family form:

$$f(y_i \mid \gamma_i) = \frac{1}{\prod_{j=1}^{n_i} y_{ij}!} \exp\{\sum_{j=1}^{n_i} y_{ij}\eta_{ij} - \sum_{j=1}^{n_i} \exp(\eta_{ij})\}. \tag{1.4}$$

Moreover, we model

$$\eta_{ij} = \log(\mu_{ij}) = x_{ij}^{\top}\beta + \gamma_i, \tag{1.5}$$

and assume that random effects γ_i $(i = 1, \ldots, n_i)$ are identically, independently and normally distributed with mean zero and variance σ^2, that is.

$$\gamma_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \tag{1.6}$$

where σ^2 is usually unknown, and is called the variance component. Now because

$$\begin{aligned} E(\exp(\gamma_i)) &= \exp(\frac{\sigma^2}{2}) \\ Var(\exp(\gamma_i)) &= \exp(2\sigma^2) - \exp(\sigma^2) \end{aligned} \tag{1.7}$$

it then follows that

$$E(y_{ij}) = E_{\gamma_i}[E_{y_{ij}|\gamma_i}(y_{ij} \mid \gamma_i)]$$

$$= E_{\gamma_i}[\exp(x_{ij}^{\top}\beta + \gamma_i)]$$

$$= \exp(x_{ij}^{\top}\beta + \frac{\sigma^2}{2}), \quad j = 1, \ldots, n_i. \tag{1.8}$$

$$\begin{aligned} Var(y_{ij}) &= E_{\gamma_i}[Var_{y_{ij}|\gamma_i}(y_{ij} \mid \gamma_i)] + Var_{\gamma_i}[E_{y_{ij}|\gamma_i}(y_{ij} \mid \gamma_i)] \\ &= E_{\gamma_i}[\exp(x_{ij}^{\top}\beta + \gamma_i)] + Var_{\gamma_i}[\exp(x_{ij}^{\top}\beta + \gamma_i)] \\ &= \exp(x_{ij}^{\top}\beta + \frac{\sigma^2}{2}) + \exp(2x_{ij}^{\top}\beta)[\exp(2\sigma^2) - exp(\sigma^2)], \\ &\qquad j = 1, \ldots, n_i. \end{aligned} \tag{1.9}$$

and

$$\begin{aligned} Cov(y_{ij}, y_{ij'}) &= E(y_{ij}y_{ij'}) - E(y_{ij})E(y_{ij'}) \\ &= E_{\gamma_i}[E_{y_{ij}|\gamma_i}(y_{ij}y_{ij'} \mid \gamma_i)] - \exp(x_{ij}^{\top}\beta + \frac{\sigma^2}{2})\exp(x_{ij'}^{\top}\beta + \frac{\sigma^2}{2}) \\ &= E_{\gamma_i}[\exp(x_{ij}^{\top}\beta + x_{ij'}^{\top}\beta + 2\gamma_i)] - \exp(x_{ij}^{\top}\beta + x_{ij'}^{\top}\beta + \sigma^2) \\ &= \exp(x_{ij}^{\top}\beta + x_{ij'}^{\top}\beta + 2\sigma^2) - \exp(x_{ij}^{\top}\beta + x_{ij'}^{\top}\beta + \sigma^2) \\ &= \exp(x_{ij}^{\top}\beta + x_{ij'}^{\top}\beta)[exp(2\sigma^2) - \exp(\sigma^2)], \quad \mathrm{j} \neq \mathrm{j}'. \end{aligned} \tag{1.10}$$

The correlation of y_{ij} and $y_{ij'}$ $(j \neq j')$ increases when σ^2 becomes larger as follows:

$$\begin{aligned} Corr(y_{ij}, y_{ij'}) &= \frac{Cov(y_{ij}, y_{ij'})}{\sqrt{Var(y_{ij})Var(y_{ij'})}} \\ &= \begin{cases} 0 & \text{if } \sigma^2 = 0 \\ 1 & \text{if } \sigma^2 \to \infty. \end{cases} \end{aligned} \tag{1.11}$$

Therefore, σ^2 may be considered as the index of the intra-cluster association parameter of the observations in a cluster.

The above model 1.4 along with 1.5 and 1.6 is the so called Poisson mixed model. The thesis deals with the improved estimation methods for this Poisson mixed model parameters including the random effect components.

The unified approach to fitting the Poisson mixed model, based on the maximum likelihood estimation of fixed effects and variance components, and the empirical Bayesian estimation of random effects, is statistically desired. However, it would involve an integral which does not possess an analytic solution. Many approaches have been proposed in order to avoid this difficulty, as described in detail in the next chapter.

Recently, Waclawiw and Liang (1993) used the Poisson mixed model to analyze a count data set of acquired immune deficiency syndrome (AIDS) cases. More specifically, they simultaneously estimated the AIDS incidence growth rate across 42 strata indexed by seven risk groups and six geographic regions, based on the number of AIDS cases collected over 5 ($n_i = 5$) consecutive quarterly time intervals during the period from January 1982 to March 1983. They also estimated γ_i ($i = 1, \ldots, k$) which represented the stratum-specific AIDS growth rates over and above the average growth rates, which is decided by the fixed effects. In fact, Waclawiw and Liang (1993) developed a three-step iterative estimation procedure for the estimation of three types of parameters β, γ_i and σ^2 in the generalized linear mixed model, which accommodated the Poisson mixed model as a special case. This three-step iterative procedure can be described as follows:

1. Assuming an initial fixed value for σ^2, the fixed effect parameters β are updated by using the generalized estimating equation approach of Zeger et al. (1988). Note that

this procedure does not presume specific values of the random effects but only the knowledge that random effects have a Gaussian distribution, as in 1.6.

2. Assuming that σ^2 and β are fixed, the Stein-type estimators for the random effects γ_i $(i = 1, \ldots, k)$ are developed with the introduction of estimating functions.

3. Assuming that β and γ_i $(i = 1, \ldots, n_i)$ are fixed, the variance of the random effects σ^2 is updated by using a moment method under the assumption that the effects of a cross-product term are negligible. Details about the validity of such assumptions are, however, not known. As shown in Chapter 4, their estimate of σ^2 is not consistent.

The above three steps of the iterative procedure describe a complete cycle. Note that with a new updated value for σ^2 from the third step, another full cycle is prompted, and the procedure continues in a circular fashion until convergence in σ^2 or one of the other parameters is achieved. But whether such a convergence would be achieved is unknown.

The generalized linear mixed model, similar to those of Waclawiw and Liang (1993), was also analyzed by Breslow and Clayton (1993). However, unlike Waclawiw and Liang (1993), Breslow and Clayton (1993) did not assume specific density functions for y_i given γ_i, where γ_i is a vector of multivariate normal distributed random effects. Instead, they assumed that for a given γ_i, the first and second conditional moments of y_i existed, and the second conditional moment was a specified function of the first conditional moment. Breslow and Clayton (1993) first used the penalized quasi-likelihood estimation approach to estimate β and γ_i. They then generated a modified profile quasi-likelihood function for inference on σ^2. Six practical problems were discussed to illustrate the wide range of applications of

their approach. For example, Breslow and Clayton (1993) used the Poisson mixed model to analyze a count data set of seizures from 59 epileptics who were randomized to a new drug or a placebo as an adjuvant to the standard chemotherapy during the two weeks before each of four clinic visits. In another example, they applied the Poisson mixed model to analyze another count data set of breast cancer rates in Iceland according to year of birth in 11 cohorts from $1800-1849$ to $1940-1949$ and age in 13 groups from $20-24$ years to $80-84$ years. However, their "derivation" of the penalized quasi-likelihood and the modified profile quasi-likelihood involved several ad hoc adjustments and approximations for which no formal justification was given. As shown in Chapter 4, this estimate is also not consistent for σ^2.

In summary, the application of the Poisson mixed model has been hampered by the lack of the analytic form for the integral of the joint density function of clustered correlated count data and random effects with respect to the random effects in evaluating the marginal likelihood. Many approximate methods have recently been proposed, for example, the penalized quasi-likelihood approach of Breslow and Clayton (1993), and the generalized estimating function approach of Waclawiw and Liang (1993). But these methods are found to produce inconsistent inference for the variance of random effects. This inconsistent estimate of the variance component may further degrade the estimation of other parameters such as fixed effect parameters and random effects. On the other hand, both the penalized quasi-likelihood and generalized estimating function methods need the iteration among the three types of parameters, and thus usually involve a large load of computation.

In the thesis, we propose a two-step approximate likelihood approach to estimate the fixed

effect parameters, random effects and their variance component in the Poisson mixed model, based on a well-grounded fact that the logarithm of a gamma random variable is nearly normally distributed when its variance is near zero, and is more peaked around its center than the density of a normal curve with the same mean and variance when its variance is away from zero (Bartlett and Kendall 1946). In the first step, the conjugate Bayesian theorem is applied to construct an approximate likelihood function of clustered correlated count data y_i $(i = 1, \ldots, k)$ in order to estimate the β and σ^2. The resulting approximate score functions for the fixed effect parameters are surprisingly the same as the marginal estimating functions used in the GEF. As a result, if σ^2 were known, this approach would yield the same estimates for the fixed effect parameters as the GEF. When σ^2 itself needs to be estimated as in usual cases, this approach produces the approximate likelihood based consistent estimates for both the fixed effect parameters and the variance component, and any accuracy of the estimates can be achieved by increasing the number of randomly selected clusters in principle. For small σ^2, our estimates are almost efficient (in the sense that they tend to be efficient as the σ^2 goes to zero) for both the fixed effect parameters and the variance component. For large σ^2, our estimates are asymptotically unbiased for the fixed effect parameters, whereas our estimate is asymptotically negative biased for the variance component. In the second step, using the estimates of β and σ^2 from the first step, we estimate γ_i $(i = 1, \ldots, k)$ by minimizing their approximate posterior mean square error based on the empirical Bayesian procedure. The resulting estimates are almost optimal (in the sense that they tend to be optimal as the σ^2 goes to zero) for γ_i $(i = 1, \ldots, k)$ when σ^2 is small. Another desirable merit

is that, unlike the previous approaches, our estimates for both the fixed effect parameters and the variance component only depend on the distribution of the random effects rather than the estimates of the random effects. Furthermore, the proposed approach is demonstrated that the asymptotic covariance of the estimates for β will become smaller in general as σ^2, an index of the intra-cluster association, gets larger , and can be significantly reduced by using the values of the fixed effect covariates x_{ij} as different as possible among different observations ($j = 1, \ldots, n_i$) in any cluster i. However, if the fixed effect covariate has the same or almost equal values among different observations in any cluster, the asymptotic variance of the estimate for the corresponding fixed effect parameter may increase as σ^2 gets larger. This feature may be useful in designing a valid experiment or sampling for the Poisson mixed model. Unless the actual σ^2 is small, the fixed effect covariates should be designed to have values as different as possible among different observations for any cluster.

The above results for the proposed approach are presented in detail in Chapter 3. Chapter 2 introduces the historical background of the Poisson mixed model. In Chapter 4, we spell out the estimation formulae of the GEF and the PQL, for the Poisson mixed model, and also show that these two methods produce inconsistent estimation for the variance component. The performance of the proposed two-step procedure is further compared with the GEF and the PQL through a simulation study, in Chapter 5. The proposed approach appears to perform better in estimating all three types of parameters than the GEF and the PQL, for small as well as large σ^2. Our approach can be used in the clustered count data studies, which usually have a large number of clusters but relatively a small number of cluster

sizes, provided the assumptions of the Poisson mixed model are valid. Chapter 6 gives the conclusion and some suggestions for further research.

Chapter 2

Historical Background of the Poisson Mixed Model

2.1 Poisson Process

Consider a Bernoulli process defined over an interval of time (or space) so that p is the probability that an event may occur during the time interval. If the time interval is allowed to become shorter and shorter so that the probability, p, of an event occurring in the interval gets smaller and the number of trials, n, increases in such a fashion that np remains constant, then the expected number of occurrences in any total time interval remains the same. It can be shown that as n gets large and p gets small so that np remains a constant, μ, the binomial distribution approaches the Poisson distribution given by

$$f(y;\mu) = \frac{\mu^y}{y!}\exp(-\mu) \quad x = 0, 1, \ldots; \mu > 0. \tag{2.1}$$

The mean and variance of the Poisson distribution are both μ.

The Poisson distribution possesses the additive property that the sum of two independent Poisson random variables with parameters μ_1 and μ_2 is a Poisson random variable with parameter $\mu = \mu_1 + \mu_2$.

A Poisson process for a continuous time scale can be defined analogous to a Bernoulli process on a discrete time scale. The Poisson process refers to the occurrence of events along a continuous time (or location) scale. For an empirical background take random events such as disintegrations of particles, incoming telephone calls, and chromosome breakages under harmful irradiation. All occurrences are assumed to be of the same kind, and we are concerned with the total number of occurrences in an arbitrary time interval of length t. Each occurrence is represented by a point on the time axis, and hence we are really concerned with certain random placements of points on a line. The underlying physical assumption is that the forces and influences governing the process remain constant so that the probability of any particular event is the same for all time intervals of duration t, and is independent of the past development of the process. In mathematical terms this means that the process is a time-homogeneous Markov process, that is,

1. The probability of an event in any short interval t to $t + \Delta t$ is $\mu \Delta t$ (proportional to the length of the interval) for all values of t. This property is known as stationarity.

2. The probability of more than one event in any short interval t to $t + \Delta t$ is negligible in comparison to $\mu \Delta t$.

3. The number of events in any interval of time is independent of the number of events

in any other non-overlapping interval of time.

The probability mass function of the number of events y in time t for a Poisson process is given by

$$f(y;\mu t) = \frac{(\mu t)^y}{y!}\exp(-\mu t) \quad y = 0, 1, \ldots; t > 0; \mu > 0. \tag{2.2}$$

where $f(y;\mu t)$ is the probability of y events in time t.

2.2 Models for Clustered Count Data

Under idealized experimental conditions when successive events occur independently and at the same rate, then the inference for Poisson count data is relatively easy, and the traditional log linear model and maximum likelihood estimation can be used for this purpose. However, even in well-conducted laboratory experiments, departures from the idealized Poisson model are to be expected for several reasons. For example, in behavioural studies involving primates or other animals, incidents usually occur in spurts or clusters. The net effect is that the number of recorded events is more variable than the simple Poisson model would suggest. Here, unless there is strong evidence to the contrary, we avoid the assumption of Poisson variation and assume the appearance of overdispersion in Poisson count data. In biomedical applications it is also rarely the case that $Var(y) = E(y)$ as is implied by the Poisson assumption. Typically, the variance exceeds the mean (Breslow, 1984). This over-dispersion can be explained by assuming that there is natural heterogeneity among the expected responses across observations. If the means are assumed to follow a gamma distribution, the

marginal distribution of the counts is the negative binomial distribution. Specifically, this distribution arises from the assumptions that

1. conditional on μ_i, the response variable y_{ij} has a Poisson distribution with mean μ_i,

2. the μ_i are independent gamma random variables with mean μ and variance $\phi\mu^2$.

Then, the marginal distribution of y_{ij} is negative binomial with

$$E(y_{ij}) = \mu \quad and \quad Var(y_{ij}) = \mu + \phi\mu^2.$$

The use of the negative binomial model dates back at least to the work of Greenwood and Yule (1920) who modelled over-dispersed accident counts. Breslow (1984), Brillinger (1986), Lawless (1987a,b) and McCullagh and Nelder (1989, Sec. 6.2) discuss the analysis of count data when extra- Poisson variation is present. It is desirable to use a model that allows for the possibility of extra-Poisson variation if we are interested primarily in inference concerning regression parameters and if the situation is one in which overdispersion routinely occurs. Recently, Dean and Lawless (1989) develop tests for detecting extra-Poisson variation in analyzing count data. Dean (1992) further develops a unifying method for obtaining tests for overdispersion with respect to a natural exponential family which refers to the extra-Poisson variation as a special case.

The simplest extension of the negative binomial model is to assume that the μ_i depend on covariates x_i through some parametric function. The most common is the log-linear model for which

$$\log(\mu_i) = x_i'\beta. \tag{2.3}$$

Using the log-linear model to analyse independent count data with overdispersion is also discussed by Clayton and Kaldor (1987) as well as by McCullagh and Nelder (1989). Actually, Clayton and Kaldor (1987) used the log-linear model to analyze observed and expected numbers of lip cancer cases in the 56 counties of Scotland with a view toward producing a map that would display regional variations in cancer incidence yet avoid the presentation of unstable rates for the smaller counties.

One important limitation of this log-linear model for application to clustered data is that the explanatory variables in the regression above do not vary within clusters. It is unlikely that the clustered count data are independent. The responses within a cluster are generally correlated. When regression is the main focus, this dependency is a nuisance, for example when testing the overall efficiency of a new drug. In other studies, the dependency is the main focus, for example, whether a disease runs in families or how a disease tends to progress. The traditional regression assumptions that the responses are statistically independent with constant variability about their expected values are not satisfied. As a result, the classical standard regression methods such as the log linear model may give inconsistent and invalid inferences. Extensions of the log linear model which account for dependence are necessary in order to obtain valid inferences.

In general, the analysis of discrete correlated data is difficult partly because their joint distribution is hardly specified well. It is usually reasonable to assume the clustered responses from distinct clusters are independent, but within a certain cluster, the clustered response data are correlated. This distinguishes the clustered data from other types of more

complicated correlated data, and thus simplifies the inference.

2.3 Mixed Effects Model for Clustered Count Data

Although the use of mixed models has a long history dating back to the last century, it is only in recent years that these models have attracted much attention in the statistical research literature. The simplest and well developed mixed models with assumed continuous Gaussian responses are the linear mixed model, in which the response is assumed to be a linear function of explanatory variables with regression coefficients that vary from one individual to the next (see references, for example, Prasad and Rao (1990)). This variability reflects natural heterogeneity due to unmeasured factors. An example is a simple linear regression for infant growth where the coefficients represent birth weight and growth rate. Children obviously are born at different weights and have different growth rates due to genetic and environmental factors which are difficult or impossible to quantify. A mixed effects model is a reasonable description if the set of coefficients from a population of children can be thought of as a sample from a distribution. Given the actual coefficients for a family, the linear mixed effects model further assumes that the observations on children for that family are independent. The correlation among different observations arises because we cannot observe the underlaying family effect, that is, the true regression coefficients, but have only imperfect measurements of weight on each infant.

A unified approach to fitting the linear mixed model, based on a combination of the empirical Bayesian and the maximum likelihood estimation of model parameters and using

the EM algorithm was discussed by Laird and Ware (1982). Searle, Casella and McCulloch (1992) presented a broad coverage of the linear mixed model.

This idea extends naturally to regression models for discrete and non-Gaussian continuous responses. It is assumed that the data for a subject are independent observations following a generalized linear model, but that the regression coefficients can vary from person to person according to a distribution, F. To illustrate, consider a log linear model studied by Waclawiw and Liang (1993) for the probability of the number of the AIDS incidence across several geographic regions. We might assume that the AIDS incidence growth rate varies across geographic regions, reflecting their different cultures, living habits and unmeasured influences of environmental factors. This simplest model would assume that every geographic region has its own AIDS incidence growth rate but the effect of the average annual income on this probability is the same for every geographic region. This model takes the form

$$log(E(y_{ij} \mid \gamma_i)) = \beta_0 + \beta_1 x_{ij} + \gamma_i \tag{2.4}$$

where y_{ij} and x_{ij} represent the number of AIDS cases and the average annual income in the ith geographic region at the jth year, and γ_i represents the geographic region-specific random effect. Although not very reasonable, Waclawiw and Liang (1993) assume that given γ_i, the repeated observations y_{ij} $(j = 1, \ldots, n_i)$ for the ith geographic region are independent of one another. Finally the model requires an assumption about the distribution of the γ_i across geographic region in the population. Typically, a parametric model such as the Gaussian with mean zero and unknown variance, σ^2, is used. This variance represents the degree of heterogeneity across geographic regions in the AIDS incidence growth rate, not attributable

to x_{ij}.

The general specification of the generalized linear mixed model is as follows:

1. Given γ_i, the responses $y_{i1}, \ldots, y_{in_i}$ are mutually independent and follow a generalized linear model with density $f(y_{ij} \mid \gamma_i) = \exp\{[y_{ij}\theta_{ij} - \psi(\theta_{ij})]/\phi + c(y_{ij}, \phi)\}$, where θ_{ij} and ϕ are unknown parameters, and ψ and c are known functions. The conditional moments, $\mu_{ij} = E(y_{ij} \mid \gamma_i) = \psi'(\theta_{ij})$ and $v_{ij} = Var(y_{ij} \mid \gamma_i) = \psi''(\theta_{ij})\phi$, satisfy $h(\mu_{ij}) = x'_{ij}\beta + d'_{ij}\gamma_i$ and $v_{ij} = v(\mu_{ij})\phi$ where h and v are known link and variance functions, respectively, β is an unknown parameter vector, and d_{ij} is a subset of x_{ij}.

2. The random effects, γ_i, $i = 1, \ldots, k$, are mutually independent with a common underlying multivariate distribution, F.

The model that is the focus of the remainder of this thesis is the Poisson mixed model with univariate random effects as follows:

1. $\log E(y_{ij} \mid \gamma_i) = x'_{ij}\beta + \gamma_i$;

2. Given γ_i, the responses $y_{i1}, \ldots, y_{in_i}$ are independent Poisson variables with mean $E(y_{ij} \mid \gamma_i)$;

3. the γ_i are independent realizations from a normal distribution with mean zero and variance σ^2.

The basic idea underlying a mixed effects model is that there is natural heterogeneity across individuals in their regression coefficients and that this heterogeneity can be repre-

sented by a cluster effect which has a probability distribution. Correlation among observations for one cluster arises from their sharing unobservable variables, γ_i. In the mixed model, the conditional probability distributions of the responses at given different subjects belong to a single family, but the random effects vary across subjects, with a common distribution or the first two common moments, specified at the second stage. Therefore, they apparently reflect heterogeneity across groups in the regression coefficients, and association within the same group in the observations. Such mixed models have several desirable features. There is no requirement for balanced data in different groups. They allow explicit modelling and analysis of between- and within- group responses. The random effects parameters have a natural interpretation which is frequently relevant to the goals of studies, and their estimates can be used for exploratory analysis. These models also facilitate the study of fixed effects on response variables.

The mixed effects model is most useful when the objective is to make inference about individuals rather than the population average. In the above AIDS incidence growth rate example, the mixed effects model would permit inference about the AIDS incidence growth rate for a particular geographic region. The regression coefficients, β, represent the effects of the explanatory variables on an individual child's chance of infection. This is in contrast to the marginal model coefficients which describe the effect of explanatory variables on the population average.

2.4 Methods for Estimating the Poisson Mixed Model

The generalized linear mixed model was proposed by Stiratelli, Laird and Ware (1984) for the analysis of serial dichotomous responses provided by a panel of study participants. Each subject's serial responses were assumed to arise from a logistic linear model, but with regression coefficients that vary between subjects. The logistic regression parameters were assumed to be normally distributed in the population. A unified approach to fitting the mixed model, based on the maximum likelihood estimation of fixed effects and variance components, and empirical Bayesian estimation of random effects was used. They found that exact solutions were analytically and computationally infeasible, and thus proposed an approximation based on the mode of the posterior distribution of the random parameters, implemented by means of the EM algorithm. The main difficulty here encountered with either maximum likelihood or empirical Bayesian approaches is that the closed-form expressions for necessary integrals do not exist. This computational difficulty appears in other generalized mixed models such as the Poisson mixed model, and it has become a current statistical research topic with a high level of interest.

Zeger and Karim (1991) cast the generalized linear mixed models in a fully Bayesian framework and used the Gibbs sampling technique to overcome the lack of closed-form expressions for necessary integrals. Compared with early used numerical integration methods that has a long history (for example, Goodwin 1949; Crouch and Spiegelman 1990), the sampling-based approaches are conceptually simple and easy to implement for users with available computing resources but without numerical analytic integration expertise. Poten-

tial drawbacks include the intensive computations and questions about when the sampling process has achieved equilibrium (Ripley and Kirkland 1990), and the requirement that conditional or joint density distributions for all fixed and random effects, as well as variance components should be subjectively (but maybe not properly) assumed (Gelfand and Smith 1990). Zeger and Karim (1991) assumed a noninformative prior for variance components, and a flat prior for fixed effects. The validity of such assumptions is in need of careful justification for each certain case. Moreover, both numerical and sampling based approaches may produce different results. Therefore, strictly speaking, both of them are approximate inference approaches.

Breslow and Clayton (1993) and Waclawiw and Liang (1993) proposed two different but related approximate approaches to estimate the generalized mixed model in order to avoid the computational difficulties. However, as it is shown in Chapter 4, both methods produce inconsistent estimate for the variance component in the Poisson mixed model with univariate random effects.

Chapter 3

The Proposed Two-Step Approach

This chapter presents an approximate likelihood approach for the Poisson mixed model, based on the fact that the logarithm of a gamma random variable is nearly normally distributed when its variance is small, and more peaked around its center than the density of a normal curve with the same mean and variance when its variance is large (Bartlett and Kendall 1946). This approximate likelihood approach consists of two steps. In the first step, the conjugate Bayesian theorem is applied to yield the approximate likelihood for the fixed effect parameters and the variance component. In the second stage, we deduce the approximate empirical Bayesian estimation for the random effects by minimizing the approximate posterior mean square error of the random effects.

3.1 Likelihood Approximation

Although Bayes theorem can be applied to combine any prior distribution with any likelihood, it is convenient to use conjugate priors for the unknown parameters because these lead to simple answers. For example, a Poisson likelihood and gamma conjugate prior can be combined to produce a marginal likelihood with the closed-form expression for the necessary integral, whereas Poisson likelihood and normal prior can not. But the application of such a conjugate prior needs careful justification in each case.

Conjugate priors have been widely used in time series and regression problems (for example, West, Harrison and Migon 1985). Harvey and Fernandes (1989) applied this approach to structural count data models which describe only one correlated series of count data. Clayton and Kaldor (1987) applied it to analyse independent count data with overdispersion.

For the Poisson mixed model with univariate random effects, the gamma distribution for the exponential function of random effects would be a conjugate prior distribution. On the other hand, the log function of a gamma random variable is found to be nearly normally distributed for the small variance, and to be more peaked around its center than the density of normal distribution for the large variance. These interesting properties are used here to construct the approximate likelihood for clustered count data in the Poisson mixed model.

For the present model, the likelihood function for β and σ^2 has the form

$$L(\beta, \sigma^2, y) \propto \Pi_{i=1}^{k} \int f(y_i|\gamma_i)\sigma^{-1} \exp\left(-\frac{\gamma_i^2}{2\sigma^2}\right) d\gamma_i, \tag{3.1}$$

where $f(y_i|\gamma_i)$ is the conditional Poisson density as in 1.4. It is well known that the integral above does not have an analytic solution. Hence the likelihood inference requires numerical

evaluation, which is not only difficult to use , but also yields approximate inference. As a remedy, we now propose to construct an approximate likelihood function as follows.

Rewrite the conditional Poisson model 1.4 in the form

$$f(y_i|w_i) = \frac{1}{\Pi_{j=1}^{n_i} y_{ij}!} \exp\left\{\sum_{j=1}^{n_i} y_{ij}(x_{ij}^T\beta + \log w_i) - w_i \sum_{j=1}^{n_i} \exp(x_{ij}^T\beta)\right\}, \tag{3.2}$$

where $w_i = \exp(\gamma_i)$. Then the likelihood function in 3.1 is equivalent to

$$L(\beta, \sigma^2, y) \propto \Pi_{i=1}^{k} \int f(y_i|w_i)g(w_i)dw_i, \tag{3.3}$$

where $g(w_i)$ is the probability density of $w_i = \exp(\gamma_i)$. In general, $g(w_i)$ is not known, because the distribution of γ_i is not known. If γ_i is assumed to be normal, which is the case in our Poisson mixed model 3.1, then $g(w_i)$ may be computed, which by 3.3 yields the exact likelihood function. But, as it was mentioned earlier, the integral in 3.3 does not have the analytic solution. To overcome this integral problem, we suggest a gamma 'working' distribution for w_i. More specifically, we use

$$g(w_i) = \frac{\lambda^\alpha}{\Gamma(\alpha)} w_i^{\alpha-1} \exp(-\lambda w_i) \tag{3.4}$$

as the 'working' probability density of w_i, where the parameters α and λ are evaluated by equating the first two moments of this 'working' distribution to the respective moments of the correct distribution of $w_i = \exp(\gamma_i)$. That is,

$$\frac{\alpha}{\lambda} = \exp(\frac{\sigma^2}{2}), \tag{3.5}$$

and

$$\frac{\alpha}{\lambda^2} = \exp(2\sigma^2) - \exp(\sigma^2). \tag{3.6}$$

After simple algebra, one obtains:

$$\alpha = \frac{1}{\exp(\sigma^2) - 1}, \quad \lambda = \frac{1}{\exp(\frac{\sigma^2}{2})[\exp(\sigma^2) - 1]}. \tag{3.7}$$

Note that corresponding to the 'working' density 3.4 of w_i, $\gamma_i = \log w_i$ has the probability density function given by

$$h(\gamma_i) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \exp[\alpha\gamma_i - \lambda \exp(\gamma_i)], \tag{3.8}$$

yielding the moment generating function

$$m_{\gamma_i}(t) = \frac{\Gamma(\alpha + t)}{\lambda^t \Gamma(\alpha)},$$

and the cumulant generating function

$$k_{\gamma_i}(t) = \log m_{\gamma_i}(t) = \log \Gamma(\alpha + t) - t \log \lambda - \log \Gamma(\alpha).$$

Thus the first four indexes of the shape of the distribution of γ_i—mean, variance, skewness and kurtosis have the following formulae:

$$\begin{aligned} E(\gamma_i) &= \psi(\alpha) - \log \lambda, \\ Var(\gamma_i) &= \psi'(\alpha), \\ \alpha_3(\gamma_i) &= \frac{\psi''(\alpha)}{\psi'(\alpha)^{3/2}}, \end{aligned}$$

and

$$\alpha_4(\gamma_i) = 3 + \frac{\psi'''(\alpha)}{\psi'(\alpha)^2}, \tag{3.9}$$

where $\psi(\alpha)$ is the digamma function

$$\psi(\alpha) = \frac{\partial \log \Gamma(\alpha)}{\partial \alpha} = -\xi - \frac{1}{\alpha} + \sum_{j=1}^{\infty} \frac{\alpha}{j(\alpha + j)} = \psi(\alpha + 1) - \frac{1}{\alpha}, \tag{3.10}$$

in which $\xi = 0.57721\ldots$. Euler's constant. Let $\psi'(\alpha)$, $\psi''(\alpha)$ and $\psi'''(\alpha)$ represent the first, second and third order derivatives of $\psi(\alpha)$ with respect to α respectively, then

$$\psi'(\alpha) = \frac{\partial \psi(\alpha)}{\partial \alpha} = \sum_{j=0}^{\infty} \frac{1}{(\alpha + j)^2} = \psi'(\alpha + 1) + \frac{1}{\alpha^2}, \tag{3.11}$$

$$\psi''(\alpha) = \frac{\partial^3 \log \Gamma(\alpha)}{\partial \alpha^3} = -2 \sum_{j=0}^{\infty} \frac{1}{(\alpha + j)^3} = \psi''(\alpha + 1) - \frac{2}{\alpha^3}, \tag{3.12}$$

and

$$\psi'''(\alpha) = \frac{\partial^4 \log \Gamma(\alpha)}{\partial \alpha^4} = 6 \sum_{j=0}^{\infty} \frac{1}{(\alpha + j)^4} = \psi'''(\alpha + 1) + \frac{6}{\alpha^4}. \tag{3.13}$$

These results can be found in any standard textbook such as Johnson and Kotz (1970, pages 196-198) and Van der Laan and Temme (1984, pages 117-120). They show that the 'working' density function 3.8 of γ_i usually follows different distribution from the normal one of real random effects. However, it is interesting to observe that when the actual variance σ^2 is near zero, one may use Taylor's series expansion to approximate the probability density of 3.8 by a normal density with the mean zero and variance σ^2, which is the same as the original density function 1.6 of γ_i. In fact, in such a case, the density function 3.8 of γ_i can be approximated by

$$h(\gamma_i) \approx \frac{\lambda^\alpha}{\Gamma(\alpha)} \exp[\alpha\gamma_i - \lambda(1 + \gamma_i + \frac{\gamma_i^2}{2})].$$

From 3.5, we have

$$\frac{\alpha}{\lambda} \approx \exp(0) = 1,$$

that is,

$$\alpha \approx \lambda.$$

Therefore,

$$h(\gamma_i) \approx \frac{\lambda^\alpha \exp(-\lambda)}{\Gamma(\alpha)} \exp[-\lambda \frac{\gamma_i^2}{2}].$$

From 3.7, one gets,

$$\lambda \approx \frac{1}{exp(0)(1+\sigma^2-1)} = \frac{1}{\sigma^2}.$$

Finally, we have

$$\begin{aligned} h(\gamma_i) &\approx \frac{\lambda^\alpha \exp(-\lambda)}{\Gamma(\alpha)} \exp[-\frac{\gamma_i^2}{2\sigma^2}] \\ &\propto \exp[-\frac{\gamma_i^2}{2\sigma^2}] \end{aligned}$$

which is the normal density with mean zero and variance σ^2, and the same as the density of γ_i in 1.6. Thus, for small σ^2, the gamma 'working' density 3.8 for γ_i reduces almost to the true distribution of γ_i, and we can expect that our likelihood inference based on the 'working' density 3.4 of w_i would be almost efficient (in the sense that they tend to be efficient as the real σ^2 goes to zero). On the other hand, it follows from 3.9 that, when the actual σ^2 is not small, the kurtosis of the 'working' density 3.8 of γ_i is larger than 3, the kurtosis of normal random effects, because $\psi'''(\alpha) > 0$, as shown in 3.13. Therefore, the 'working' density of γ_i in 3.8 would be more peaked around its center than the density of the normal curve with the same mean and variance. These properties sufficiently justify the use of the 'working' density 3.4 of w_i in computing the approximate likelihood function for the Poisson mixed model.

Now, by using $g(w_i)$ from 3.4 in 3.3 and integrating out w_i, we obtain the approximate

log likelihood function for β and σ^2 as

$$\begin{aligned}
\ell(\beta,\sigma^2) &= \sum_{i=1}^{k}\{\alpha\log\lambda-\log\Gamma(\alpha)-\sum_{j=1}^{n_i}\log y_{ij}!+\sum_{j=1}^{n_i}y_{ij}x_{ij}^{\top}\beta \\
&\quad +\log\int w_i^{(\sum_{j=1}^{n_i}y_{ij}+\alpha-1)}\exp[-w_i(\lambda+\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta))]dw_i\} \\
&= \sum_{i=1}^{k}\left[\log\left\{\frac{\Gamma\left(\alpha+\sum_{j=1}^{n_i}y_{ij}\right)}{\Gamma(\alpha)}\right\}\right]-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log y_{ij}! \\
&\quad +\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}x_{ij}^{T}\beta-\sum_{i=1}^{k}\left[\left\{\alpha+\sum_{j=1}^{n_i}y_{ij}\right\}\log\left(\lambda+\sum_{j=1}^{n_i}\exp(x_{ij}^{T}\beta)\right)\right] \\
&\quad +k\alpha\log\lambda. \qquad (3.14)
\end{aligned}$$

The above likelihood function is exploited in the next subsection to obtain the estimates of the fixed effect parameters β and the variance component σ^2, the variance of random effects.

3.2 Two-Step Approach

1. The first step

   Let β^* and σ^{*2} be the likelihood estimates for β and σ^2 respectively, based on the approximate likelihood function given in 3.14. In step 1, these estimates are obtained by solving the score equations

$$\begin{aligned}
U_1(\beta) &= \frac{\partial\ell(\beta,\sigma^2)}{\partial\beta} \\
&= \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}y_{ij}x_{ij}-(\frac{y_i^*}{\mu_i^*})\sum_{j=1}^{n_i}x_{ij}\exp(x_{ij}^{T}\beta)\right]=0 \qquad (3.15)
\end{aligned}$$

   and

$$U_2(\sigma^2) = \frac{\partial\ell(\beta,\sigma^2)}{\partial\sigma^2}$$

$$= \alpha'(\sigma^2)\sum_{i=1}^{k}[\psi(y_i^*) - \psi(\alpha) + \log(\frac{\lambda}{\mu_i^*})]$$

$$+\lambda'(\sigma^2)\sum_{i=1}^{k}[\frac{\alpha}{\lambda} - \frac{y_i^*}{\mu_i^*}] = 0 \qquad (3.16)$$

where

$$y_i^* = \alpha + \sum_{j=1}^{n_i} y_{ij}, \quad \mu_i^* = \lambda + \sum_{j=1}^{n_i} \exp(x_{ij}^T\beta),$$

$$\alpha'(\sigma^2) = \frac{\partial\alpha}{\partial\sigma^2} = -\frac{\exp(\sigma^2)}{[\exp(\sigma^2) - 1]^2},$$

and

$$\lambda'(\sigma^2) = \frac{\partial\lambda}{\partial\sigma^2} = -\frac{3\exp(\sigma^2) - 1}{2\exp(\frac{\sigma^2}{2})[\exp(\sigma^2) - 1]^2}.$$

Note that the score function 3.15 for β is the same as the estimating function for β in Waclawiw and Liang (1993) which will be further discussed in the next chapter. Thus for known σ^2, both the present approach and the estimating function approach yield the same inference for β. But, in practice, σ^2 is rarely known. For unknown σ^2, the two approaches yield different inferences for β and σ^2. Unlike Waclawiw and Liang (1993), the present approach provides the joint approximate likelihood estimates for β and σ^2.

2. The second step.

In step 2, we deal with the prediction of the random effects γ_i $(i = 1, \ldots, k)$. Let γ_i^* be the minimum mean square error prediction of γ_i. It then follows that $\gamma_i^* = E(\gamma_i|y_i)$. Now, by exploiting the conditional density 1.4 of y_i for a given γ_i, and the 'working'

probability density 3.8 of γ_i, one obtains

$$\begin{aligned} f^*(\gamma_i \mid y_i) &= \frac{f(y_i \mid \gamma_i)h(\gamma_i)}{\int f(y_i \mid \gamma_i)h(\gamma_i)d\gamma_i} \\ &= \frac{\exp\{(\alpha^* + \sum_{j=1}^{n_i} y_{ij})\gamma_i - [\lambda^* + \sum_{j=1}^{n_i} \exp(x_{ij}\beta^*)]\exp(\gamma_i)\}}{\int \exp\{(\alpha^* + \sum_{j=1}^{n_i} y_{ij})\gamma_i - [\lambda^* + \sum_{j=1}^{n_i} \exp(x_{ij}\beta^*)]\exp(\gamma_i)\}d\gamma_i}. \end{aligned}$$

Therefore,

$$\begin{aligned} \gamma_i^* &= \frac{\int \gamma_i f(y_i \mid \gamma_i)h(\gamma_i)d\gamma_i}{\int f(y_i \mid \gamma_i)h(\gamma_i)d\gamma_i} \\ &= \psi\left(\alpha^* + \sum_{j=1}^{n_i} y_{ij}\right) - \log\left(\lambda^* + \sum_{j=1}^{n_i} \exp(x_{ij}^T\beta^*)\right) \end{aligned} \tag{3.17}$$

where α^* and λ^* are computed from 3.7 by replacing σ^2 with σ^{*2}. When the actual σ^2 is near zero, this estimate is almost optimal (in the sense that it tends to be optimal as the real σ^2 goes to zero) .

3.3 Computational Aspects

The traditional Newton Raphson iteration procedure may run into convergence problems in solving the score equations 3.15 and 3.16 simultaneously. Even if σ^2 is known, the score function 3.15 may lead to a local maximum or minimum. When σ^2 is unknown, as in the general case, the solution becomes more complicated because of the restriction of boundary for σ^2. More specifically, the iteration procedure may yield negative estimates of σ^2, when the true maximum occurs near the boundary of the parameter space $\sigma^2 = 0$. To avoid this convergence problem, we solve β and σ^2 by jointly using the modified Newton Raphson iteration based on the log-likelihood 3.14 and the general form of the EM algorithm as follows:

1. Decide the initial values of β, say $\beta^{(0)}$.

Assume $\sigma^2 = 0$, then the score function 3.15 becomes

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_{ij} - \exp(x_{ij}^T\beta)]x_{ij} = 0. \tag{3.18}$$

Therefore, we approximately have

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}[\log y_{ij} - x_{ij}^T\beta]x_{ij} \approx 0, \tag{3.19}$$

and

$$\beta^{(0)} = (\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}x_{ij}^T)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\log y_{ij} \tag{3.20}$$

2. Decide the initial values of σ^2, α and λ, say $\sigma^{2(0)}$, $\alpha^{(0)}$ and $\lambda^{(0)}$ respectively.

Maximizing the conditional likelihood 3.2 with respect to w_i yields

$$w_i^{(0)} = \frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{j=1}^{n_i}\exp(x_{ij}^T\beta^{(0)})}, \quad i = 1, \ldots, k. \tag{3.21}$$

Then the moment method is used to produce the initial value of σ^2

$$\sigma^{2(0)} = \frac{\sum_{i=1}^{k}[\log w_i^{(0)}]^2}{k}, \tag{3.22}$$

Using 3.7 yields:

$$\begin{aligned} \alpha^{(0)} &= \frac{1}{\exp(\sigma^{2(0)} - 1)}, \\ \lambda^{(0)} &= \frac{1}{\exp(\frac{\sigma^{2(0)}}{2})[\exp(\sigma^{2(0)}) - 1]}. \end{aligned} \tag{3.23}$$

Note that equations 3.18-3.22 are only used to yield the initial estimates of β, σ^2, α and λ. These initial estimates may be inaccurate and will be improved in the following steps.

3. The modified Newton Raphson iterative algorithm β.

We solve the score equation 3.15 for β by using the well-known modified Newton Raphson iterative algorithm (cf. Seber and Wild (1989, p. 599-600)), as follows,

(a) Solving the score function 3.15 by using the first-order Taylor expansion, we have,

$$\beta^{(1)} = \beta^{(0)} - (\frac{\partial U_1(\beta^{(0)})}{\partial \beta})^{-1} U_1(\beta^{(0)}) \tag{3.24}$$

where

$$\frac{\partial U_1(\beta^{(0)})}{\partial \beta} = \sum_{i=1}^{k} \frac{y_i^{*(0)}}{\mu_i^{*(0)2}} (l_i^{(0)} l_i^{(0)\top} - \mu_i^{*(0)} L_i^{(0)}) \tag{3.25}$$

with

$$l_i^{(0)} = \sum_{j=1}^{n_i} x_{ij} \exp(x_{ij}^T \beta^{(0)}).$$

$$l_i^{(0)\top} = \sum_{j=1}^{n_i} x_{ij}^T \exp(x_{ij}^T \beta^{(0)}),$$

$$y_i^{*(0)} = \alpha^{(0)} + \sum_{j=1}^{n_i} y_{ij},$$

$$L_i^{(0)} = \sum_{j=1}^{n_i} \exp(x_{ij}^T \beta^{(0)}) x_{ij} x_{ij}^T,$$

and

$$\mu_i^{*(0)} = \lambda^{(0)} + \sum_{j=1}^{n_i} \exp(x_{ij}^T \beta^{(0)}).$$

(b) We compute the log likelihood 3.14 for $\beta^{(0)}$ and $\beta^{(1)}$. If $\ell(\beta^{(1)}, \sigma^{2(0)}) > \ell(\beta^{(0)}, \sigma^{2(0)})$, then the $\beta^{(1)}$ are the modified Newton Raphson iterative estimates of β in this step: Otherwise, we return to 3.24 with half an original change size of $\beta^{(0)}$, that

is,

$$\beta^{(1)} = \beta^{(0)} - 0.5(\frac{\partial U_1(\beta^{(0)})}{\partial \beta})^{-1} U_1(\beta^{(0)}). \tag{3.26}$$

In this way, the change size of $\beta^{(0)}$ becomes shorter and shorter until the log likelihood $\ell(\beta^{(1)}, \sigma^{2(0)}) > \ell(\beta^{(0)}, \sigma^{2(0)})$. This can guarantee that our estimate $\beta^{(1)}$ of β must lead to the maximum likelihood ones.

4. The general form of the EM algorithm for σ^2.

At the present stage, we use the estimates $\beta^{(1)}$ for β from last stage and solve the score equation 3.16 for σ^2 by exploiting the general form of the EM algorithm of Dempster, Laird and Rubin (1977).

Wu (1983) explained that if a likelihood is unimodal within parameter bounds and has only stationary point, then the EM algorithm estimation converges to the unique maximum likelihood estimate. The solution is unique here due to the well-known convexity property of the log-likelihood for regular exponential families. As Searle, Casella and McCulloch (1992, pages 296-305) pointed out, an important advantage of the EM algorithm is that the iterations will always remain in the parameter space, since it is performing the maximum likelihood estimation for the complete data. Moreover, the EM algorithm usually simplifies the direct calculation of the maximum likelihood estimation.

Following the idea of Stiratelli, Laird and Ware (1984), we also think of the incomplete data as being the observed data y_{ij} and the complete data the unobservable random effects γ_i. But the application of the EM algorithm here is slightly different

from theirs in the sense that the general form rather than simple one of EM algorithm is used.

For the present model, the E-step of the EM algorithm involves finding the expectation of $\sum_{i=1}^{k} \log g(w_i|y_i, \sigma^{2(0)}, \beta^{(1)})$, where $g(w_i|y_i, \sigma^{2(0)}, \beta^{(1)})$ is the conditional density function of w_i with the gamma 'working' density given in 3.4, conditional on the observed data vector y_i and given the initial estimates $\sigma^{2(0)}$ and $\beta^{(1)}$. A straightforward algebra yields

$$E\left\{\sum_{i=1}^{k} \log g(w_i|y_i, \sigma^{2(0)}, \beta^{(1)})\right\}$$

$$= \quad k\alpha \log\lambda + (\alpha - 1)s_1^{(1)} - \lambda s_2^{(1)} - k \log\Gamma(\alpha) \tag{3.27}$$

where

$$\begin{aligned} s_1^{(1)} &= \sum_{i=1}^{k} E(\log w_i \mid y_i, \sigma^{2(0)}, \beta^{(1)}) \\ &= \sum_{i=1}^{k} \{\psi(\alpha^{(0)} + \sum_{j=1}^{n_i} y_{ij}) - \log[\lambda^{(0)} + \sum_{j=1}^{n_i} \exp(x_{ij}^{\mathrm{T}}\beta)^{(1)}]\}, \end{aligned} \tag{3.28}$$

and

$$\begin{aligned} s_2^{(1)} &= \sum_{i=1}^{k} E(w_i \mid y_i, \sigma^{2(0)}, \beta^{(1)}) \\ &= \sum_{i=1}^{k} \frac{\alpha^{(0)} + \sum_{j=1}^{n_i} y_{ij}}{\lambda^{(0)} + \sum_{j=1}^{n_i} \exp(x_{ij}^{\mathrm{T}}\beta^{(1)})}. \end{aligned} \tag{3.29}$$

Next, the M-step of the EM algorithm requires maximizing 3.27 for σ^2. Let $\sigma^{2(1)}$ be the solution. The above two stages of computations constitute a cycle.

The computation involves the function $\log\Gamma(\alpha)$ and its derivatives. They are not easy to be directly calculated from their formulas 3.10, 3.11, 3.12 and 3.13. Van der Laan

and Temme (1984) listed the following convenient approximate formula for $\log\Gamma(a)$:

$$\begin{aligned}\log\Gamma(a) &= (a-0.5)\log a - a + 0.5\log(2\pi) + \frac{1}{12a} \\ &\quad -\frac{1}{360a^3} + \frac{1}{1260a^5} - \frac{1}{1680a^7} + O(a^{-9}) \qquad (3.30)\end{aligned}$$

When a is not less than 2, the formula 3.30 can be used to compute $\log\Gamma(a)$ with very high accuracy without last term $O(a^{-9})$. When a is less than 2 but larger than 1, the same high accuracy can be guaranteed by the combination of the above formula 3.30 and the following recurrence formula,

$$\log\Gamma(a) = \log\Gamma(a+1) - \log(a), \qquad (3.31)$$

as $a+1$ is larger than 2. That is,

$$\begin{aligned}\log\Gamma(a) &= (a+0.5)\log(a+1) - a - 1 + 0.5\log(2\pi) + \frac{1}{12(a+1)} \\ &\quad -\frac{1}{360(a+1)^3} + \frac{1}{1260(a+1)^5} - \frac{1}{1680(a+1)^7} \\ &\quad -\log(a) + O((a+1)^{-9}). \qquad (3.32)\end{aligned}$$

Similarly, when a is less than 1 but larger than 0, then

$$\begin{aligned}\log\Gamma(a) &= \log\Gamma(a+2) - \log(a+1) - \log(a) \\ &= (a+1.5)\log(a+2) - a - 2 + 0.5\log(2\pi) + \frac{1}{12(a+2)} \\ &\quad -\frac{1}{360(a+2)^3} + \frac{1}{1260(a+2)^5} - \frac{1}{1680(a+2)^7} \\ &\quad -\log(a+1) - \log(a) + O((a+2)^{-9}). \qquad (3.33)\end{aligned}$$

$\psi(a)$, $\psi'(a)$, $\psi''(a)$ and $\psi'''(a)$ can be computed from the derivatives of the above formulas of 3.30, 3.32 and 3.33 of $\log\Gamma(x)$ for different values of a. For example, when

α is larger than 2, one obtains,

$$\begin{aligned}\psi(\alpha) &= \frac{\partial}{\partial\alpha}\log\Gamma(\alpha) \\ &= -\frac{0.5}{\alpha} + \log\alpha - \frac{1}{12\alpha^2} \\ &\quad + \frac{1}{120\alpha^4} - \frac{1}{252\alpha^6} - \frac{1}{240\alpha^8} + O(\alpha^{-10}). \end{aligned} \tag{3.34}$$

In order to obtain the improved estimates of β and σ^2, $\sigma^{2(1)}$ from the first cycle is used for σ^2 in the score equation 3.15 at the first stage to obtain improved estimate $\beta^{(2)}$ for β. This is done by using the modified Newton Raphson iteration procedures as in the first cycle. Then at the second stage of this second cycle, we maximize $E\left\{\sum_{i=1}^{k}\log g(w_i|y_i,\sigma^{2(1)},\beta^{(2)})\right\}$ to obtain the improved estimate $\sigma^{2(2)}$ for σ^2. These two-stage based cycles of computations continue until convergence is achieved. The final estimates are β^* and σ^{*2} for β and σ^2 respectively.

3.4 Remarks on Asymptotic Theory

3.4.1 When σ^2 is Known

By exploiting the score equation 3.15, we have the following result when σ^2 is known.

Theorem 1 *If $\sum_{i=1}^{k}(L_i - \frac{l_i l_i^T}{\mu_i^*})$ is positive definite and σ^2 is known, then the approximate likelihood estimates β^* of β are consistent, asymptotically unbiased, and $\sqrt{k}(\beta^*-\beta)$ is asymptotically $(k\to\infty)$ distributed as multivariate normal with mean zero and $p\times p$ covariance*

matrix given by

$$kV_{\beta^*} = k\exp(-\frac{\sigma^2}{2})[\sum_{i=1}^{k}(L_i - \frac{l_i l_i^T}{\mu_i^*})]^{-1} \tag{3.35}$$

with

$$l_i = \sum_{j=1}^{n_i} x_{ij}\exp(x_{ij}^T\beta), \quad l_i^T = \sum_{j=1}^{n_i} x_{ij}^T\exp(x_{ij}^T\beta),$$

$$L_i = \sum_{j=1}^{n_i}\exp(x_{ij}^T\beta)x_{ij}x_{ij}^T, \quad \textit{and} \quad \mu_i^* = \lambda + \sum_{j=1}^{n_i}\exp(x_{ij}^T\beta).$$

Proof:

It is apparent that the first-order derivative of $U_1(\beta)$ is continuous. Therefore, using the Taylor expansion and ignoring the high order terms, $\sqrt{k}(\beta^* - \beta)$ can be approximated by

$$\sqrt{k}(\beta^* - \beta) \approx [-\frac{\partial U_1(\beta)}{\partial\beta}/k]^{-1}[U_1(\beta)/k^{1/2}]. \tag{3.36}$$

By using 1.8, 1.9, 1.10, 3.5 and 3.6, we have

$$\begin{aligned} EU_1(\beta) &= \sum_{i=1}^{k}\{\sum_{j=1}^{n_i}(Ey_{ij}x_{ij}) - \frac{\alpha + \sum_{j=1}^{n_i} Ey_{ij}}{\mu_i^*}l_i\} \\ &= \sum_{i=1}^{k}\{\sum_{j=1}^{n_i}[\exp(x_{ij}^T\beta + \sigma^2/2)x_{ij}] - \frac{\alpha + \sum_{j=1}^{n_i}\exp(x_{ij}^T\beta + \sigma^2/2)}{\mu_i^*}l_i\} \\ &= \sum_{i=1}^{k}(l_i\exp(\sigma^2/2) - \frac{\alpha}{\lambda}l_i) \\ &= \sum_{i=1}^{k} l_i[\exp(\sigma^2/2) - \frac{\alpha}{\lambda}] = 0 \end{aligned}$$

and

$$\begin{aligned} Var U_1(\beta) &= \sum_{i=1}^{k} Var\{\sum_{j=1}^{n_i}(y_{ij}x_{ij}) - \frac{\alpha + \sum_{j=1}^{n_i} y_{ij}}{\mu_i^*}l_i\} \\ &= \sum_{i=1}^{k} Var[\sum_{j=1}^{n_i}(x_{ij} - \frac{l_i}{\mu_i^*})y_{ij}] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{k}\{\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i}(x_{ij}-\frac{l_i}{\mu_i^*})(Cov(y_{ij},y_{ij'})(x_{ij'}-\frac{l_i}{\mu_i^*})^{\top}\}\\
&= \sum_{i=1}^{k}\{\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i}(x_{ij}-\frac{l_i}{\mu_i^*})[\exp(2\sigma^2)-exp(\sigma^2)]exp(x_{ij}^{\top}\beta+x_{ij'}^{\top}\beta)(x_{ij'}-\frac{l_i}{\mu_i^*})^{\top}\\
&\quad+\sum_{j=1}^{n_i}exp(x_{ij}^{\top}\beta+\frac{\sigma^2}{2})(x_{ij}-\frac{l_i}{\mu_i^*})(x_{ij}-\frac{l_i}{\mu_i^*})^{\top}\}\\
&= \sum_{i=1}^{k}\{[\exp(2\sigma^2)-exp(\sigma^2)][1-\frac{\sum_{j=1}^{n_i}exp(x_{ij}^{\top}\beta)}{\mu_i^*}]^2l_il_i^{\top}\\
&\quad+\exp(\frac{\sigma^2}{2})[L_i-2\frac{l_il_i^{\top}}{\mu_i^*}+\frac{l_il_i^{\top}}{\mu_i^{*2}}\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta)]\}\\
&= \exp(\frac{\sigma^2}{2})\sum_{i=1}^{k}(L_i-\frac{l_il_i^{\top}}{\mu_i^*}).
\end{aligned}
$$

which is positive definite. Thus according to the central limit theorem, $\frac{U_1(\beta)}{k^{1/2}}$ has an asymptotically normal distribution with mean zero and covariance $\frac{Var(U_1(\beta))}{k}$. Differentiation of the score function 3.15 leads to

$$
\begin{aligned}
\frac{\partial U_1(\beta)}{\partial\beta} &= \frac{\partial U^2(\beta,\sigma^2)}{\partial\beta^2}\\
&= \sum_{i=1}^{k}\{(a+\sum_{j=1}^{n_i}y_{ij})[\frac{\sum_{j=1}^{n_i}[\exp(x_{ij}^{\top}\beta)x_{ij}]\sum_{j=1}^{n_i}[\exp(x_{ij}^{\top}\beta)x_{ij}^{\top}]}{[\lambda+\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta)]^2}\\
&\quad-\frac{\sum_{j=1}^{n_i}[\exp(x_{ij}^{\top}\beta)x_{ij}x_{ij}^{\top}]}{\lambda+\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta)}]\}\\
&= \sum_{i=1}^{k}(a+\sum_{j=1}^{n_i}y_{ij})(\frac{l_il_i^{\top}}{\mu_i^{*2}}-\frac{L_i}{\mu_i^*}) \qquad (3.37)
\end{aligned}
$$

and

$$
\begin{aligned}
E(-\frac{\partial U_1(\beta)}{\partial\beta}) &= -\sum_{i=1}^{k}\{[a+\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta+\frac{\sigma^2}{2})](\frac{l_il_i^{\top}}{\mu_i^{*2}}-\frac{L_i}{\mu_i^*})\}\\
&= \exp(\frac{\sigma^2}{2})\sum_{i=1}^{k}(L_i-\frac{l_il_i^{\top}}{\mu_i^*})\\
&= Var(U_1(\beta)). \qquad (3.38)
\end{aligned}
$$

According to large number theory, we have

$$\frac{-\partial U_1(\beta)}{\partial \beta}/k \xrightarrow{p} E\{-\frac{\partial U_1(\beta)}{\partial \beta}\}/k. \tag{3.39}$$

Therefore, combining 3.39 with 3.36 yields the proof of the theorem. □

Theorem 1 indicates that the asymptotic covariance of the approximate likelihood estimates β^* depends on the variance of random effects σ^2, which is the index of the intra-cluster association within the same cluster in the observations, as explained in Chapter 1. When the actual σ^2 is zero, the approximate likelihood estimates β^* are exactly the classical maximum likelihood estimates for fully independent count data, and are asymptotically efficient, with the asymptotic covariance

$$V_{\beta^*} = (\sum_{i=1}^{k} I_i)^{-1}. \tag{3.40}$$

When the actual σ^2 is away from zero, and so large that λ is very small compared with $\sum_{j=1}^{n_i} \exp(x_{ij}^T \beta)$, the asymptotic covariance 3.35 can be approximated by

$$\begin{aligned} V_{\beta^*} &\approx \exp(-\frac{\sigma^2}{2})\{\sum_{i=1}^{k}[I_i - \frac{I_i I_i^T}{\sum_{j=1}^{n_i}\exp(x_{ij}^T\beta)}]\}^{-1} \\ &= \exp(-\frac{\sigma^2}{2})\{\sum_{i=1}^{k}\frac{\sum_{j,j'=1}^{n_i}[\exp(x_{ij}^T\beta + x_{ij'}^T\beta)(x_{ij}-x_{ij'})(x_{ij}-x_{ij'})^T]}{2\sum_{j=1}^{n_i}\exp(x_{ij}^T\beta)}\}^{-1}. \end{aligned} \tag{3.41}$$

The last equality 3.41 is from the following lemma:

Lemma 1 *For any $n \times 1$ vector u_j and its positive scalar function a_j,*

$$\sum_{j=1}^{n}(a_j u_j u_j^T)\sum_{j=1}^{n}a_j - \sum_{j=1}^{n}(a_j u_j)\sum_{j=1}^{n}(a_j u_j^T) = \frac{1}{2}\sum_{j,j'=1}^{n}[a_j a_{j'}(u_j - u_{j'})(u_j - u_{j'})^T] \geq 0 \tag{3.42}$$

Proof:

$$\sum_{j,j'=1}^{n}[a_j a_{j'}(u_j - u_{j'})(u_j - u_{j'})^\top] = \sum_{j,j'=1}^{n}[a_j a_{j'} u_j u_j^\top - a_j a_{j'} u_j u_{j'}^\top - a_j a_{j'} u_{j'} u_j^\top + a_j a_{j'} u_{j'} u_{j'}^\top]$$
$$= 2\sum_{j=1}^{n}(a_j u_j u_j^\top)\sum_{j=1}^{n} a_j - 2\sum_{j=1}^{n}(a_j u_j)\sum_{j=1}^{n}(a_j u_j^\top)$$

and

$$\sum_{j,j'=1}^{n}[a_j a_{j'}(u_j - u_{j'})(u_j - u_{j'})^\top] \geq 0,$$

which yield the proof of the lemma. □

The approximate formula 3.41 of the asymptotic covariance of β^* for large σ^2 reveals the following fact. As the actual σ^2 gets larger, the asymptotic covariance of β^* will become smaller in general, unless the corresponding fixed effect covariate has the same or nearly equal values ($x_{ij} = x_{ij'}$ or $x_{ij} \approx x_{ij'}$) among different observations ($j \neq j'$) in any cluster ($i = 1, \ldots, k$). If the actual σ^2 should be infinite, the asymptotic covariance of β^* would become zero, and we would have accurate inference for β. This conclusion at the first instance appears to provide conflicting inference when compared to the traditional analysis where σ^2 is thought to be a dispersion or overdispersion parameter only. In fact, for the present model, σ^2 plays an intra-cluster association role which is similar to the role played by the intra-cluster correlation in the linear mixed model. Consequently, as the intra-cluster association increases, it is reasonable to expect that β^* will have smaller and smaller asymptotic covariance.

On the other hand, the asymptotic covariance of β^* can be reduced by increasing the number of distinct clusters and cluster size, as well as by increasing the difference of fixed effect covariates among different observations in any cluster. For highly clustered correlated

count data (σ^2 is large), the asymptotic covariance of β^* will significantly increase by using the same or similar values for the fixed effect covariates among different observations in any cluster. In another words, the asymptotic covariance of β^* will get larger, especially if x_{ij} and $x_{ij'}$ $(j \neq j', j, j' = 1, \ldots, n_i)$ are very close to each other for any i, as apparently shown in 3.41. On the contrary, for nearly independent count data (σ^2 is near zero), the asymptotic covariance of β^* will not be affected much by using the same or similar values of the fixed effect covariates among different observations in any cluster. Therefore, unless the actual σ^2 is small, fixed effect covariates should be designed to have values as different as possible among different observations in any cluster. The simulation study in Chapter 5 supports the above findings.

3.4.2 When σ^2 is Unknown

When σ^2 is unknown , by exploiting the score equations 3.15 and 3.16, we have the following result.

Theorem 2 *If the rank of* $E\left(\partial \begin{pmatrix} U_1(\beta) \\ U_2(\sigma^2) \end{pmatrix} / \partial \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}\right)$ *is* $p+1$, $Var \begin{pmatrix} U_1(\beta) \\ U_2(\sigma^2) \end{pmatrix}$ *is positive definite, and* σ^2 *is unknown, then the approximate likelihood estimates* β^* *and* σ^{2*} *are both consistent, and*

$$\sqrt{k}\left(\begin{pmatrix} \beta^* \\ \sigma^{2*} \end{pmatrix} - \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}\right)$$

is asymptotically $(k \to \infty)$ *distributed as multivariate normal with mean zero for* $\beta^* - \beta$ *and mean below zero for* $\sigma^{2*} - \sigma^2$, *and* $(p+1 \times p+1)$ *covariance matrix*

$$\begin{pmatrix} -E\frac{\partial U_1(\beta)}{\partial \beta}/k & -E\frac{\partial U_1(\beta)}{\partial \sigma^2}/k \\ -E[\frac{\partial U_2(\sigma^2)}{\partial \beta^\top}]/k & -E\frac{\partial U_2(\sigma^2)}{\partial \sigma^2}/k \end{pmatrix}^{-1} \times \begin{pmatrix} \frac{Var(U_1(\beta))}{k} & \frac{Cov(U_1(\beta),U_2(\sigma^2))}{k} \\ \frac{[Cov(U_2(\sigma^2),U_1^\top(\beta))]}{k} & \frac{Var(U_2(\sigma^2))}{k} \end{pmatrix}$$
$$\times \begin{pmatrix} -E\frac{\partial U_1(\beta)}{\partial \beta}/k & -E[\frac{\partial U_2(\sigma^2)}{\partial \beta}]/k \\ -E\frac{\partial U_1^\top(\beta)}{\partial \sigma^2}/k & -E\frac{\partial U_2(\sigma^2)}{\partial \sigma^2}/k \end{pmatrix}^{-1} \tag{3.43}$$

whose elements are specified in the following proof of this theorem.

Proof: It is apparent that the first-order derivative of $U_1(\beta)$ and $U_2(\sigma^2)$ is continuous. Therefore, using the Taylor expansion and ignoring the high order terms, we have

$$\sqrt{k}\left(\begin{pmatrix} \beta^* \\ \sigma^{2*} \end{pmatrix} - \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}\right) \approx \begin{pmatrix} -\frac{\partial U_1(\beta)}{\partial \beta}/k & -\frac{\partial U_1(\beta)}{\partial \sigma^2}/k \\ -(\frac{\partial U_2(\sigma^2)}{\partial \beta^\top})/k & -\frac{\partial U_2(\sigma^2)}{\partial \sigma^2}/k \end{pmatrix}^{-1} \begin{pmatrix} U_1(\beta)/k^{1/2} \\ U_2(\sigma^2)/k^{1/2} \end{pmatrix}. \tag{3.44}$$

By using 1.8, 1.9, 1.10, 3.5 and 3.6 as in the proof of Theorem 1, we have

$$EU_1(\beta) = 0 \tag{3.45}$$

and

$$Var(U_1(\beta)) = \exp(\frac{\sigma^2}{2}) \sum_{i=1}^{k} (L_i - \frac{l_i l_i^\top}{\mu_i^*}). \tag{3.46}$$

Because

$$\begin{aligned} \frac{\alpha}{\lambda} - \frac{Ey_i^*}{\mu_i^*} &= \frac{\alpha}{\lambda} - \frac{\alpha + \sum_{j=1}^{n_i} Ey_i}{\mu_i^*} \\ &= \frac{\alpha}{\lambda} - \frac{\alpha + \sum_{j=1}^{n_i} \exp(x_{ij}^\top \beta)\frac{\alpha}{\lambda}}{\mu_i^*} \\ &= \frac{\alpha}{\lambda} - \frac{\alpha}{\lambda} = 0, \end{aligned}$$

we have

$$
\begin{aligned}
EU_2(\sigma^2) &= \alpha'(\sigma^2)\sum_{i=1}^{k}[E\psi(y_i^*) - \psi(\alpha) + \log(\frac{\lambda}{\mu_i^*})] \\
&\quad + \lambda'(\sigma^2)\sum_{i=1}^{k}[\alpha/\lambda - Ey_i^*/\mu_i^*] \\
&= \alpha'(\sigma^2)\sum_{i=1}^{k}[E\psi(y_i^*) - \log\mu_i^* - (\psi(\alpha) - \log\lambda)] \\
&\quad \begin{cases} = 0 & \text{if } \sigma^2 \text{ is zero} \\ \approx 0 & \text{if } \sigma^2 \text{ is near zero} \\ < 0 & \text{if } \sigma^2 \text{ is away from zero.} \end{cases}
\end{aligned} \tag{3.47}
$$

Thus according to the central limit theorem,

$$
\begin{pmatrix} U_1(\beta)/k^{1/2} \\ U_2(\sigma^2)/k^{1/2} \end{pmatrix}
$$

has an asymptotically normal distribution with mean

$$
\begin{pmatrix} 0 \\ E(U_2(\sigma^2)) \end{pmatrix}
$$

and covariance

$$
\begin{pmatrix} \frac{Var(U_1(\beta))}{k} & \frac{Cov(U_1(\beta),U_2(\sigma^2))}{k} \\ \frac{[Cov(U_2(\sigma^2),U_1^T(\beta))]}{k} & \frac{Var(U_2(\sigma^2))}{k} \end{pmatrix},
$$

where $Var(U_2(\sigma^2))$ and $Cov(U_1(\beta), U_2(\sigma^2))$ depend on the expectation of the functions of $\psi(\alpha + \sum_{j=1}^{n_i} y_{ij})$ which usually does not have simple expression. On the other hand, according to the large number theory, we have,

$$
\begin{pmatrix} -\frac{\partial U_1(\beta)}{\partial\beta}/k & -\frac{\partial U_1(\beta)}{\partial\sigma^2}/k \\ -[\frac{\partial U_2(\sigma^2)}{\partial\beta^T}]/k & -\frac{\partial U_2(\sigma^2)}{\partial\sigma^2}/k \end{pmatrix} \xrightarrow{p} \begin{pmatrix} -E\frac{\partial U_1(\beta)}{\partial\beta}/k & -E\frac{\partial U_1(\beta)}{\partial\sigma^2}/k \\ -E[\frac{\partial U_2(\sigma^2)}{\partial\beta^T}]/k & -E\frac{\partial U_2(\sigma^2)}{\partial\sigma^2}/k \end{pmatrix} \tag{3.48}
$$

where

$$E(-\frac{\partial U_1(\beta)}{\partial \beta}) = Var(U_1(\beta)). \tag{3.49}$$

$$\begin{aligned} E(-\frac{\partial U_1(\beta)}{\partial \sigma^2}) &= E(-\frac{\partial U_2(\sigma^2)}{\partial \beta}) \\ &= \sum_{i=1}^{k} \frac{\mu_i^* \alpha'(\sigma^2) - E y_i^* \lambda'(\sigma^2)}{\mu_i^{*2}} l_i \\ &= \sum_{i=1}^{k} \frac{\mu_i^* \alpha'(\sigma^2) - [\alpha + \sum_{j=1}^{n_i} \exp(x_{ij}^T \beta) \frac{\alpha}{\lambda}] \lambda'(\sigma^2)}{\mu_i^{*2}} l_i \\ &= \sum_{i=1}^{k} \frac{\alpha'(\sigma^2) - \lambda'(\sigma^2) \frac{\alpha}{\lambda}}{\mu_i^*} l_i \\ &= 0.5\alpha \sum_{i=1}^{k} \frac{l_i}{\mu_i^*}. \end{aligned} \tag{3.50}$$

and $E(-\frac{\partial U_2(\sigma^2)}{\partial \sigma^2})$ depends on the expectation of $\psi'(\alpha + \sum_{j=1}^{n_i} y_{ij})$ which usually does not have simple expression. Combining 3.44-3.48 yields the proof of this theorem. □

Theorem 2 indicates that the approximate likelihood estimates of β are almost asymptotically efficient (in the sense that they tend to be efficient as the σ^2 goes to zero) for small σ^2, and are always asymptotically unbiased no matter how large the actual σ^2 is. As opposed to case for β^*, the behaviour of σ^{*2}, the approximate likelihood estimator of σ^2, is quite different. For small σ^2, σ^{*2} is almost asymptotically unbiased and almost efficient for σ^2 (in the sense that it tends to be asymptotically unbiased and efficient as the σ^2 goes to zero). As the actual σ^2 gets larger, the 'working' distribution of the random effects γ_i deviates further from the normal distribution, leading to larger kurtosis than that of the normal distribution with the same mean and variance. That is, the 'working' distribution of γ_i is more peaked around its center than the density of a normal curve with the same

mean and variance when its variance is far away from zero. Consequently, for large σ^2, σ^{*2} is asymptotically negatively biased for σ^2.

Chapter 4

Two Recent Approximate Methods of Estimation

We now spell out in detail the formulae for the estimators of β, σ^2 and γ_i $(i = 1, \ldots, k)$ in the Poisson mixed model based on the estimating function approach suggested by Waclawiw and Liang (1993) and on the penalized quasi-likelihood approach suggested by Breslow and Clayton (1993), and also point out their main drawbacks.

4.1 Penalized Quasi-Likelihood Method

For the Poisson mixed model defined in Chapter 1, the random effects γ_i $(i = 1, \ldots, k)$ are assumed to be normally distributed as in 1.6. Therefore, the quasi-likelihood function is the same as the likelihood function of y_{ij} $((j = 1, \ldots, n_i), i = 1, \ldots, k)$ for (β, σ^2) which is

defined by

$$\begin{aligned} L(\beta,\sigma^2,y) &= \prod_{i=1}^{k}\int\prod_{j=1}^{n_i} f(y_{ij}\mid\gamma_i)f(\gamma_i)d\gamma_i \\ &= c\prod_{i}^{k}\int\frac{1}{\sigma}\exp\{\sum_{j=1}^{n_i} y_{ij}(x_{ij}^{\top}\beta+\gamma_i) \\ &\quad -\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta+\gamma_i)-\frac{\gamma_i^2}{2\sigma^2}\}d\gamma_i \end{aligned} \tag{4.1}$$

where $c = 1/[(2\pi)^{k/2}\prod_{i=1}^{k}\prod_{j=1}^{n_i} y_{ij}]$. Breslow and Clayton (1993) applied Laplace's method for the above integral approximation. Denote

$$h(\gamma_i) = -\sum_{j=1}^{n_i} y_{ij}(x_{ij}^{\top}\beta+\gamma_i)+\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta+\gamma_i)+\frac{\gamma_i^2}{2\sigma^2}. \tag{4.2}$$

Let $h'(\gamma_i)$ and $h''(\gamma_i)$ represent the first- and second- order derivatives of $h(\gamma_i)$ with respect to γ_i, then

$$\begin{aligned} h(\gamma_i) &= h(\hat{\gamma}_i)+\frac{1}{2}(\gamma_i-\hat{\gamma}_i)^2h''(\hat{\gamma}_i)+o(|\gamma_i-\hat{\gamma}_i|) \\ &\approx h(\hat{\gamma}_i)+\frac{1}{2}(\gamma_i-\hat{\gamma}_i)^2h''(\hat{\gamma}_i) \end{aligned} \tag{4.3}$$

where $\hat{\gamma}_i$ denotes the posterior mode of γ_i computed from

$$h'(\gamma_i) = -\sum_{j=1}^{n_i} y_{ij}+\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta+\gamma_i)+\frac{\gamma}{\sigma^2}=0, \tag{4.4}$$

and

$$h''(\hat{\gamma}_i) = \sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta+\hat{\gamma}_i)+\frac{1}{\sigma^2}. \tag{4.5}$$

Ignoring the multiplicative constant c, putting 4.3 into 4.1 yields the approximate profile log likelihood

$$l(\beta,\sigma^2) \approx -\frac{k}{2}\log(\sigma^2)-\sum_{i=1}^{k}\{h(\hat{\gamma}_i)+\frac{1}{2}\log(h''(\hat{\gamma}_i))\}$$

$$
\begin{aligned}
&= \sum_{i=1}^{k}\{-\frac{1}{2}\log[1+\sigma^2\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)] \\
&\quad +\sum_{j=1}^{n_i} y_{ij}(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)-\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)-\frac{\hat{\gamma}_i^2}{2\sigma^2}\} \\
&= -\frac{1}{2}\sum_{i=1}^{k}\log[1+\sigma^2\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)]-\sum_{i=1}^{k}h(\hat{\gamma}_i). \qquad (4.6)
\end{aligned}
$$

Differentiating 4.6 with respect to σ^2 generates the score equation of σ^2

$$
\begin{aligned}
g_3(\sigma^2) &= \sum_{i=1}^{k}\hat{\gamma}_i^2-\sigma^4\sum_{i=1}^{k}\frac{\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)}{1+\sigma^2\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)} \\
&= \sum_{i=1}^{k}\hat{\gamma}_i^2-\sigma^2\sum_{i=1}^{k}[1-\frac{1}{1+\sigma^2\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)}] \\
&= \sum_{i=1}^{k}\hat{\gamma}_i^2-k\sigma^2+\sigma^2\sum_{i=1}^{k}[\frac{1}{1+\sigma^2\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\hat{\gamma}_i)}]=0. \qquad (4.7)
\end{aligned}
$$

After further approximation, Breslow and Clayton (1993) suggested to estimate β and γ_i jointly by maximizing Green's (1987) penalized log likelihood

$$
\sum_{i=1}^{k}\log[f(y_i\mid\gamma_i)f(\gamma_i)]\propto\sum_{i=1}^{k}\{\sum_{j=1}^{n_i}y_{ij}(x_{ij}^{\mathsf{T}}\beta+\gamma_i)-\sum_{j=1}^{n_i}\exp(x_{ij}^{\mathsf{T}}\beta+\gamma_i)-\frac{\gamma_i^2}{2\sigma^2}\}. \qquad (4.8)
$$

Differentiating 4.8 with respect to β and γ_i leads to the score equations

$$
g_1(\beta)=\sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_{ij}-\exp(x_{ij}^{\mathsf{T}}\beta+\gamma_i)]x_{ij}=0, \qquad (4.9)
$$

and

$$
g_2(\gamma_i)=\sum_{j=1}^{n_i}[y_{ij}-\exp(x_{ij}^{\mathsf{T}}\beta+\gamma_i)]-\frac{\gamma_i}{\sigma^2}=0 \qquad (4.10)
$$

for β and γ_i $(i=1,\ldots,k)$ respectively.

Breslow and Clayton (1993) proposed solving the score functions 4.9 and 4.10 for β and γ_i jointly through the modified Fisher scoring algorithm. This requires one to compute the

inverse of the $(k+p)\times(k+p)$ Fisher information matrix, and is usually very difficult to compute due to large k. Moreover, the Fisher scoring algorithm usually converges slower than the Hessian scoring algorithm. Breslow and Clayton (1993) further suggested making degrees-of-freedom adjustments through the modified restricted maximum likelihood. But for large k, this restricted maximum likelihood has little difference from the usually maximum likelihood.

Therefore, our algorithm is slightly different from Breslow and Clayton (1993), although we follow their main idea. We use the Newton Raphson iteration with the Hessian rather than Fisher scoring for the above score functions 4.7, 4.9 and 4.10 in the iterative way. Let $\hat{\beta}^{(m)}$, $\hat{\gamma}_i^{(m)}$ and $\hat{\sigma}^{2(m)}$ be the estimates of β, γ_i and σ^2 at the mth iteration. Then the improved estimates of these parameters are obtained at the $(m+1)$st iteration by using

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} - \left(\frac{\partial g_1}{\partial \beta^T}\right)^{-1} g_1\Big|_{\beta=\hat{\beta}^{(m)},\gamma_i=\hat{\gamma}_i^{(m)}}$$

$$\hat{\gamma}_i^{(m+1)} = \hat{\gamma}_i^{(m)} - \left(\frac{\partial g_2}{\partial \gamma_i}\right)^{-1} g_2\Big|_{\beta=\hat{\beta}^{(m)},\gamma_i=\hat{\gamma}_i^{(m)},\sigma^2=\hat{\sigma}^{2(m)}}$$

and

$$\hat{\sigma}^{2(m+1)} = \hat{\sigma}^{2(m)} - \left(\frac{\partial g_3}{\partial \sigma^2}\right)^{-1} g_3\Big|_{\beta=\hat{\beta}^{(m)},\hat{\gamma}_i=\hat{\gamma}_i^{(m)},\sigma^2=\hat{\sigma}^{2(m)}}$$

where the Hessian matrices are given by

$$\frac{\partial g_1}{\partial \beta^T} = -\sum_{i=1}^{k}\sum_{j=1}^{n_i}\{\exp(x_{ij}^T\hat{\beta}^{(m)}+\hat{\gamma}_i^{(m)})x_{ij}x_{ij}^T\}$$

$$\frac{\partial g_2}{\partial \gamma_i} = -\sum_{j=1}^{n_i}\exp(x_{ij}^T\hat{\beta}^{(m)}+\hat{\gamma}_i^{(m)}) \quad -\frac{1}{\sigma^{2(m)}}$$

and

$$\frac{\partial g_3}{\partial \sigma^2} = -k + \sum_{i=1}^{k} \left[1 + \hat{\sigma}^{2(m)} \sum_{j=1}^{n_i} \exp(x_{ij}^T \hat{\beta}^{(m)} + \hat{\gamma}_i^{(m)}) \right]^{-2}$$

respectively. The cycles of iteration continue until convergence is achieved.

Note that by using the actual β and γ_i in the approximate profile likelihood based score equation 4.7, one may obtain the simple form

$$\sigma^2 = \frac{1}{k} \sum_{i=1}^{k} \gamma_i^2 + \frac{1}{k} \sum_{i=1}^{k} \frac{\sigma^2}{1 + \sigma^2 \sum_{j=1}^{n_i} \exp(x_{ij}^T \beta + \gamma_i)} \tag{4.11}$$

for the estimation of σ^2. Now, for large k, the first term in the right side of the equation 4.11 converges to σ^2. But, the second term converges to a certain positive value rather than zero. Consequently, the penalized quasi-likelihood approach yields an inconsistent estimator of σ^2. Furthermore, the inconsistent estimator may cause instability in the estimation of other parameters, especially in the prediction of the random effects γ_i $(i = 1, \ldots, k)$.

4.2 Generalized Estimating Function Method

Assuming an initial value for σ^2, Waclawiw and Liang (1993) used the marginal estimating equation approach of Liang and Zeger (1986) [also see Zeger, Liang and Albert (1988)] to achieve estimates for the fixed effects β. More specifically, in this approach, β^I is the solution of the following marginal score equation

$$U(\beta, \sigma^2) = \sum_{i=1}^{k} \frac{\partial u_i^T(\beta, \sigma^2)}{\partial \beta} V_i^{-1}(\beta, \sigma^2)[y_i - u_i(\beta, \sigma^2)] = 0 \tag{4.12}$$

for β, where $a_i(\beta,\sigma^2)$ is the marginal mean vector of y_i, and $V_i(\beta,\sigma^2)$ is the marginal covariance matrix of y_i. It then easily follows from 1.8, 1.9 and 1.10 that

$$\begin{aligned} a_i(\beta,\sigma^2) &= E(y_i) \\ &= \exp(\frac{\sigma^2}{2}) \begin{pmatrix} \exp(x_{i1}^\top\beta) \\ \vdots \\ \exp(x_{in_i}^\top\beta) \end{pmatrix}, \end{aligned} \tag{4.13}$$

and

$$\begin{aligned} V_i(\beta,\sigma^2) &= Var(y_i) \\ &= Cov(E(y_i \mid \gamma_i)) + E(Cov(y_i \mid \gamma_i)) \\ &= [\exp(2\sigma^2) - \exp(\sigma^2)] \begin{pmatrix} \exp(x_{i1}^\top\beta) \\ \vdots \\ \exp(x_{in_i}^\top\beta) \end{pmatrix} \Big(\exp(x_{i1}^\top\beta), \ldots, \exp(x_{in_i}^\top\beta)\Big) \\ &\quad + \exp(\frac{\sigma^2}{2}) \begin{pmatrix} \exp(x_{i1}^\top\beta) & 0 & 0 & \ldots & 0 \\ 0 & \exp(x_{i2}^\top\beta) & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \exp(x_{in_i}^\top\beta) \end{pmatrix}. \end{aligned} \tag{4.14}$$

The following Lemma is useful in order to compute the inverse of the above marginal covariance matrix.

Lemma 2 *Let A be a nonsingular $p \times p$ matrix and let u and v be two vectors with p components. Then*

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u} \tag{4.15}$$

Using Lemma 2 leads to the following inverse of the marginal covariance matrix

$$V_i^{-1}(\beta,\sigma^2) = \exp(-\frac{\sigma^2}{2})\begin{pmatrix} \exp(-x_{i1}^{\top}\beta) & 0 & 0 & \dots & 0 \\ 0 & \exp(-x_{i2}^{\top}\beta) & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \exp(-x_{in_i}^{\top}\beta) \end{pmatrix}$$
$$-\frac{\exp(\sigma^2)-1}{1+\exp(\frac{\sigma^2}{2})[\exp(\sigma^2)-1]\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta)} 1_{n_i\times 1}1_{n_i\times 1}^{\top}, \qquad (4.16)$$

where

$$1_{n_i\times 1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Substituting 4.13 and 4.16 in the marginal score function 4.12 yields the estimating functions

$$U(\beta,\sigma^2) = \sum_{i=1}^{k}\{\sum_{j=1}^{n_i}(y_{ij}x_{ij}) - \frac{\alpha+\sum_{j=1}^{n_i}y_{ij}}{\lambda+\sum_{j=1}^{n_i}\exp(x_{ij}^{\top}\beta)}\sum_{j=1}^{n_i}(\exp(x_{ij}^{\top}\beta)x_{ij})\} = 0 \qquad (4.17)$$

for β, where

$$\alpha = \frac{1}{\exp(\sigma^2)-1} \qquad (4.18)$$

and

$$\lambda = \frac{1}{\exp(\frac{\sigma^2}{2})(\exp(\sigma^2)-1)} \qquad (4.19)$$

Combining the concepts of estimating functions and linear Bayesian theorems, Liang and Waclawiw (1990) then extended the Stein estimator outside the classical framework to the more general situation such as where unbiased or even finite moment estimates of parameters may not exist. Waclawiw and Liang (1993) used this approach for random

effects γ_i. Considering the conditional score function of the Poisson mixed model 1.4

$$g_i(y_i, \gamma_i, \beta) = \frac{\partial \log f(y_i \mid \gamma_i)}{\partial \gamma_i} = \sum_{j=1}^{n_i} [y_{ij} - \exp(x_{ij}^{\top}\beta + \gamma_i)] = 0, \tag{4.20}$$

they introduced a class of linear estimating functions of the responses for γ_i, that is,

$$\dot{g}_i(y_i, \gamma_i, \beta) = \sum_{j=1}^{n_i} [a_{ij}y_{ij} + b_i - \exp(x_{ij}^{\top}\beta + \gamma_i)]. \tag{4.21}$$

Following the optimal criterion of Godambe (1960) and Ferreira (1982), they determined the optimal a_{ij} and b_i by minimizing

$$\begin{aligned} R(\dot{g}_i) &= E(\dot{g}_i / E(\frac{\partial \dot{g}_i}{\partial \gamma_i}))^2 \\ &\propto E\{\sum_{j=1}^{n_i} [a_{ij}y_{ij} + b_i - \exp(x_{ij}^{\top}\beta + \gamma_i)]\}^2. \end{aligned} \tag{4.22}$$

Differentiation of the above equation 4.22 with respect to b_i and a_{ij} yields

$$\begin{aligned} \frac{\partial R}{\partial b_i} &= 2E\{\sum_{j=1}^{n_i} [a_{ij}y_{ij} + b_i - \exp(x_{ij}^{\top}\beta + \gamma_i)]\} \\ &= 2\sum_{j=1}^{n_i} \{(a_{ij} - 1)\exp(x_{ij}^{\top}\beta + \frac{\sigma^2}{2}) + b_i\} = 0 \end{aligned} \tag{4.23}$$

and

$$\begin{aligned} \frac{\partial R}{\partial a_{ik}} &= 2E\{y_{ik}\sum_{j=1}^{n_i} [a_{ij}y_{ij} + b_i - \exp(x_{ij}^{\top}\beta + \gamma_i)]\} \\ &= 2\exp(x_{ik}^{\top}\beta + \frac{\sigma^2}{2})\{\sum_{j=1}^{n_i} [(a_{ij} - 1)\exp(x_{ij}^{\top}\beta + \frac{3\sigma^2}{2})] + a_{ik} + n_i b_i\} = 0. \end{aligned} \tag{4.24}$$

Combination of equations 4.23 and 4.24 leads to

$$a_{ik}^{*} = \frac{\sum_{j=1}^{n_i} \exp(x_{ij}^{\top}\beta)}{\lambda + \sum_{j=1}^{n_i} \exp(x_{ij}^{\top}\beta)} \tag{4.25}$$

and

$$b_i^* = \frac{a_{ik}^* \alpha}{n_i}, \qquad k = 1, \ldots, n_i, \tag{4.26}$$

where α and λ are given in formulas 4.18 and 4.19.

The Stein-type estimator for γ_i can then be computed from

$$\hat{g}_i^* = \sum_{j=1}^{n_i} [a_{ij}^* y_{ij} + b_i^* - \exp(x_{ij}^T \beta + \gamma_i)] = 0, \tag{4.27}$$

that is

$$\begin{aligned} \gamma_i^\dagger &= \log \frac{\sum_{j=1}^{n_i} (a_{ij}^* y_{ij} + b_i^*)}{\sum_{j=1}^{n_i} \exp(x_{ij}^T \beta)} \\ &= \log \frac{\sum_{j=1}^{n_i} y_{ij} + \alpha}{\sum_{j=1}^{n_i} \exp(x_{ij}^T \beta) + \lambda}. \end{aligned} \tag{4.28}$$

Further, Waclawiw and Liang (1993) suggested the following approximation for the estimation of σ^2

$$\begin{aligned} \sigma^{\dagger 2} &\approx E(\gamma_i^{\dagger 2} + E(\gamma_i^\dagger - \gamma_i)^2 \\ &\approx \frac{\sum_{i=1}^k \gamma_i^{\dagger 2}}{k} + \frac{1}{k} E[\hat{g}_i^* / E(\frac{\partial \hat{g}_i^*}{\partial \gamma_i})]^2 \\ &= \frac{\sum_{i=1}^k \gamma_i^{\dagger 2}}{k} + \frac{1}{k} \sum_{i=1}^k \{ \frac{\exp(\sigma^2) - 1}{1 + [\exp(\sigma^2) - 1] \sum_{j=1}^{n_i} \exp(x_{ij}^T \beta + \sigma^2/2)} \\ &\quad + \frac{[\exp(\sigma^2) - 1]^2 [\exp(\sigma^2) + 1]}{[1 + (\exp(\sigma^2) - 1) \sum_{j=1}^{n_i} \exp(x_{ij}^T \beta + \sigma^2/2)]^2} \}. \end{aligned} \tag{4.29}$$

The approximation involved in the process of deducing the variance formula 4.29 may not be suitable for the Poisson mixed model. One main drawback of this estimating formula for σ^2 is that as $\gamma_i^\dagger$ approaches to true γ_i, the right side of 4.29 does not converge to σ^2. This is because $\sum_{i=1}^k \gamma_i^2 / k$ converges to σ^2 for large k, but each component of the second term

converges to certain positive value. This shows that the right side of 4.29 converges to a quantity which is more than σ^2. Thus, similar to the penalized quasi-likelihood method, the estimating function approach also yields inconsistent estimates of σ^2.

Given an initial value of σ^2, the estimating equations 4.17 for β are solved by using the Newton Raphson iteration procedure to obtain $\beta^{\dagger}$. One then directly computes the Stein-type estimators $\gamma_i^{\dagger}$ $(i = 1, \ldots, k)$ by using $\beta^{\dagger}$ and the initial value of σ^2 in 4.28. Next, in order to avoid a large load of computation, the computation for $\sigma^{\dagger 2}$ is completed from the following approximate formula of the variance 4.29 for σ^2

$$\sigma^{\dagger 2} \approx \frac{\sum_{i=1}^{k} \gamma_i^{\dagger 2}}{k} + \frac{1}{k} \sum_{i=1}^{k} \{ \frac{\exp(\sigma^2) - 1}{1 + [\exp(\sigma^2) - 1] \sum_{j=1}^{n_i} \exp(x_{ij}^{\top} \beta + \sigma^2/2)} \} \qquad (4.30)$$

again by using the Newton Raphson iteration procedure. In this approximate formula of 4.29, the ignored term is always positive. Thus the estimate of σ^2 would be a little less from the approximation 4.30 than from 4.29 which would yield more inconsistent estimate for σ^2.

The value of $\sigma^{\dagger 2}$ is further applied to obtain improved estimates of β. Next the improved estimates of β as well as the value of $\sigma^{\dagger 2}$ are used to obtain the improved estimates of γ_i $(i = 1, \ldots, k)$. Further application of the improved estimates of β and γ_i in 4.30 yields the improved estimate of σ^2. These cycles of iterations continue until convergence is achieved.

In summary, the application of the Poisson mixed model has been hampered by the difficulty of computation in evaluating the marginal likelihood of the parameters involved. Many approximate approaches have recently been proposed, for example, the penalized quasi-likelihood approach of Breslow and Clayton (1993), and the generalized estimating

function approach of Waclawiw and Liang (1993). But these methods, as we have just shown, produce inconsistent inference for the variance component. Furthermore, both the penalized likelihood and estimating function methods need the iteration among the three types of parameters, and thus usually involve a large load of computation.

Chapter 5

Simulation Study

5.1 Simulation Design

In the simulation study, we used the Poisson mixed model 1.4 along with $\log\{E(y_{ij}|\gamma_i)\} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + \gamma_i$ where $p = 4$, $\beta_1 = 2.5$, $\beta_2 = -1.0$, $\beta_3 = 1.0$ and $\beta_4 = 0.5$, and for all $i = 1, \ldots, k$.

$$x_{ij1} = 1, \text{ for } j = 1, \ldots, n_i; \quad x_{ij2} = \begin{cases} 1, \text{ for } j = 1, \ldots, \frac{n_i}{2} \\ 0, \text{ for } j = \frac{n_i}{2} + 1, \ldots, n_i; \end{cases}$$

$$x_{ij3} = j - \frac{n_i + 1}{2}; \quad j = 1, \ldots, n_i; \text{ and } x_{ij4} = x_{ij2} x_{ij3}.$$

Further, taking $k = 50$ and 100, the γ_i's were independently generated from a normal distribution with the mean 0 and variance σ^2. Five σ^2 values: $\sigma^2 = 0.1$, 0.25, 0.50, 0.75, 1.0 and two cluster sizes $n_i = 4$ and 6 were considered. The responses $(y_{i1}, \ldots, y_{in_i})$ for each cluster i were generated from realizations of the Poisson mixed model 1.4 with the

mean and variance equal to $\exp\{\beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + \gamma_i\}$. The simulated data y_{ij}, $j = 1, \ldots, n_i$; $i = 1, \ldots, k$, and the covariates x_{iju}, $u = 1, \ldots, p$; $j = 1, \ldots, n_i$; $i = 1, \ldots, k$ were used to compute the estimates of the fixed effect parameters β, the variance component σ^2 of the random effects, and the random effects γ_i $(i = 1, \ldots, k)$, based on all three approaches discussed in Chapter 3 and Chapter 4. The simulation was repeated 5,000 times in order to obtain the mean values and standard errors of the parameter estimates.

5.2 Estimates of β and σ^2

Tables 5.1-5.4 report the simulated mean values and standard errors of the estimates of β_1, β_2, β_3, β_4 and σ^2 computed by: (1) the proposed two-step approximate likelihood approach (AL), (2) the penalized quasi-likelihood approach (PQL) of Breslow and Clayton (1993), and (3) the generalized estimating function approach (GEF) of Waclawiw and Liang (1993).

It is clear from the table that the proposed method performs extremely well in estimating σ^2 as compared to the PQL and GEF approaches. The standard error as well as the absolute value of the bias of the AL estimate are much smaller than those of both the PQL and the GEF estimates for σ^2. These results confirm the fact that the AL yields consistent estimates for σ^2, whereas both the PQL and GEF do not. When σ^2 is near zero, the standard error as well as the absolute value of the bias of the AL estimate are very small for σ^2 , supporting that the AL estimate is almost efficient for small σ^2. Moreover, for all three approaches, the standard error as well as the absolute value of the bias of the estimation of σ^2 get larger as σ^2 increases, reflecting the fact that for all three approaches, the estimate of σ^2 will become

more and more asymptotically biased as the actual σ^2 becomes larger and larger.

On the other hand, all three approaches perform very well in estimating β_2, β_3 and β_4, revealing that the estimates of the fixed effect parameters are usually consistent and asymptotically unbiased. When σ^2 is near zero, the AL performs very well in estimating β_1 with very minor bias, both the PQL and GEF do not, indicating that for small σ^2, the AL yields almost efficient estimation for the fixed effect parameters, whereas both the PQL and GEF may not. Furthermore, for all three approaches, as the actual σ^2 increases, the standard error of the estimate will decrease for β_2, β_3 and β_4, and increase for β_1 in general. This is because the x_{ij1} $(j = 1, \ldots, n_i; i = 1, \ldots, k)$ uniformly have the same value 1, whereas x_{ij2}, x_{ij3} and x_{ij4} are not same for all $j = 1, \ldots, n_i$. These simulation results demonstrate that the asymptotic covariance of the estimates for the fixed effect parameters will become smaller and smaller in general as the variance component σ^2, an index of the intra-cluster association, increases, and can be noticeably reduced by assigning the values of the fixed effect covariates as different as possible among different observations in any cluster. However, if the fixed effect covariate has the same or nearly equal values among different observations in any cluster, the asymptotic variance of the estimate for the corresponding fixed effect parameter may increase as σ^2 gets larger.

Finally, for all three approaches, the simulation results display that the estimates of the fixed effect parameters and the variance component become better, in the sense that their standard errors as well as the absolute values of their bias decrease, in general as the number of clusters k and the cluster size n_i $(i = 1, \ldots, k)$ increase.

5.3 Prediction of the Random Effects

Table 5.5 lists the simulated total mean square error (MSE) of the random effect predictors based on the AL, PQL and GEF approaches. Let γ_{is}^{*} be the AL estimator of the random effect γ_i in the sth ($s = 1, \ldots, 5000$) simulation. Then the total MSE of the AL predictors is defined by $\sum_{i=1}^{k}\left\{\sum_{s=1}^{5000}(\gamma_{is}^{*}-\gamma_i)^2/5000\right\}$, where $k = 50$ or 100 is the number of independent clusters . Similarly, the total MSE of the PQL and GEF predictors are defined by $\sum_{i=1}^{k}\left\{\sum_{s=1}^{5000}(\hat{\gamma}_{is}-\gamma_i)^2/5000\right\}$ and $\sum_{i=1}^{k}\left\{\sum_{s=1}^{5000}(\gamma_{is}^{\dagger}-\gamma_i)^2/5000\right\}$ respectively. It is clear from the table that the GEF approach is inferior to the PQL and AL approaches in the prediction of the random effects. Between the PQL and AL approaches, the total MSE of the AL predictors are generally much smaller than those of the PQL predictors, for small $k = 50$. Thus, the proposed AL approach performs much better than its competitors in the prediction of the random effects. For large $k = 100$, the AL approach performs better than the PQL for small σ^2, but the PQL approach appears to perform better for large σ^2. Thus the proposed AL approach is always best for small σ^2.

As well, Table 5.5 exhibits that the AL yields extremely good estimates for the random effects when the actual σ^2 is small, but both the PQL and GEF do not, reflecting that when the actual σ^2 is near zero, the AL produces almost optimal estimation for the random effects, whereas both the PQL and GEF may not. The simulation also displays that, for all three approaches, the estimates of the random effects will have smaller total MSE in general as the cluster size n_i $(i = 1, \ldots, k)$ increases and the number of clusters k decreases. That is, the estimates of the random effects will become more accurate as the relative clustered size (n_i/k) increases. All these simulation results are coincident to the analytic ones which are pointed out in Chapter 4.

Table 5.1: Comparison of Simulated Mean Values and Standard Errors (SE) of the Regression Estimates and Variance Components of Random Effects for Selected Values of σ^2: $k = 50$; $n_i = 1$ $(i = 1, \ldots, k)$: True Values of the Regression Parameters: $\beta_1 = 2.5$, $\beta_2 = -1.0$, $\beta_3 = 1.0$ and $\beta_4 = 0.5$: 5000 Simulations.

$k = 50$	$n_i = 1$		Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.10	AL	Mean	0.130	2.507	-0.986	1.001	0.534
		SE	0.010	0.047	0.175	0.035	0.224
	PQL	Mean	0.178	2.275	-0.994	1.001	0.517
		SE	0.015	0.038	0.175	0.035	0.222
	GEF	Mean	0.379	2.422	-0.995	1.001	0.516
		SE	0.144	0.074	0.175	0.035	0.222
0.25	AL	Mean	0.302	2.488	-0.996	1.000	0.514
		SE	0.015	0.044	0.169	0.033	0.217
	PQL	Mean	0.422	2.387	-0.998	1.000	0.511
		SE	0.026	0.035	0.169	0.033	0.214
	GEF	Mean	0.398	2.352	-1.007	1.000	0.493
		SE	0.035	0.055	0.169	0.033	0.215

$k = 50$	$n_i = 4$		Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.50	AL	Mean	0.515	2.519	-0.993	1.000	0.520
		SE	0.014	0.042	0.150	0.030	0.189
	PQL	Mean	0.871	2.573	-0.997	1.000	0.512
		SE	0.046	0.032	0.150	0.030	0.188
	GEF	Mean	1.358	2.258	-0.994	1.000	0.516
		SE	0.281	0.175	0.150	0.030	0.189
0.75	AL	Mean	0.681	2.636	-0.997	1.000	0.512
		SE	0.015	0.085	0.137	0.027	0.174
	PQL	Mean	1.378	2.754	-1.001	1.000	0.506
		SE	0.070	0.029	0.137	0.027	0.174
	GEF	Mean	1.572	2.242	-0.999	1.000	0.510
		SE	0.390	0.154	0.137	0.027	0.173
1.00	AL	Mean	0.812	2.817	-1.000	1.000	0.506
		SE	0.016	0.139	0.127	0.025	0.159
	PQL	Mean	1.932	2.926	-1.000	1.000	0.504
		SE	0.094	0.027	0.127	0.025	0.158
	GEF	Mean	1.821	2.151	-0.999	1.000	0.507
		SE	0.376	0.143	0.127	0.025	0.158

Table 5.2: Comparison of Simulated Mean Values and Standard Errors (SE) of the Regression Estimates and Variance Components of Random Effects for Selected Values of σ^2: $k = 50$; $n_i = 6$ $(i = 1, \ldots, k)$; True Values of the Regression Parameters: $\beta_1 = 2.5$, $\beta_2 = -1.0$, $\beta_3 = 1.0$ and $\beta_4 = 0.5$; 5000 Simulations.

$k = 50$	$n_i = 6$		Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.10	AL	Mean	0.129	2.507	-0.986	1.000	0.532
		SE	0.006	0.030	0.153	0.014	0.167
	PQL	Mean	0.157	2.194	-0.996	1.000	0.513
		SE	0.008	0.025	0.153	0.014	0.166
	GEF	Mean	0.601	2.304	-0.997	1.000	0.513
		SE	0.198	0.110	0.153	0.014	0.166
0.25	AL	Mean	0.300	2.488	-0.994	1.000	0.517
		SE	0.009	0.028	0.140	0.013	0.154
	PQL	Mean	0.395	2.314	-0.997	1.000	0.510
		SE	0.015	0.022	0.134	0.013	0.152
	GEF	Mean	0.880	2.241	-1.006	1.000	0.495
		SE	0.411	0.169	0.140	0.013	0.153

$k = 50$	$n_i = 6$		Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.50	AL	Mean	0.514	2.521	-0.994	1.000	0.517
		SE	0.009	0.028	0.124	0.012	0.136
	PQL	Mean	0.837	2.501	-0.999	1.000	0.508
		SE	0.026	0.021	0.124	0.012	0.135
	GEE	Mean	1.745	2.023	-1.001	1.000	0.503
		SE	0.143	0.068	0.125	0.012	0.136
0.75	AL	Mean	0.677	2.600	-0.995	1.000	0.514
		SE	0.010	0.053	0.116	0.011	0.125
	PQL	Mean	1.330	2.681	-1.000	1.000	0.505
		SE	0.038	0.019	0.116	0.011	0.125
	GEE	Mean	1.699	2.175	-1.000	1.000	0.505
		SE	0.498	0.179	0.117	0.011	0.127
1.00	AL	Mean	0.809	2.782	-0.998	1.000	0.507
		SE	0.011	0.135	0.107	0.010	0.116
	PQL	Mean	1.872	2.854	-0.999	1.000	0.505
		SE	0.051	0.018	0.106	0.010	0.116
	GEE	Mean	1.787	2.198	-0.997	1.000	0.508
		SE	0.402	0.119	0.106	0.010	0.116

Table 5.3: Comparison of Simulated Mean Values and Standard Errors (SE) of the Regression Estimates and Variance Components of Random Effects for Selected Values of σ^2; $k = 100$; $n_i = 4$ $(i = 1, \ldots, k)$; True Values of the Regression Parameters: $\beta_1 = 2.5$, $\beta_2 = -1.0$, $\beta_3 = 1.0$ and $\beta_4 = 0.5$; 5000 Simulations.

$k = 100$	$n_i = 4$		Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.10	AL	Mean	0.102	2.468	-0.987	1.000	0.527
		SE	0.007	0.034	0.129	0.025	0.162
	PQL	Mean	0.140	2.222	-0.995	1.000	0.510
		SE	0.009	0.028	0.128	0.025	0.161
	GEF	Mean	0.311	2.396	-0.995	1.000	0.510
		SE	0.116	0.055	0.128	0.025	0.161
0.25	AL	Mean	0.241	2.422	-0.993	1.000	0.519
		SE	0.009	0.033	0.126	0.024	0.156
	PQL	Mean	0.320	2.283	-0.997	1.000	0.510
		SE	0.016	0.026	0.126	0.024	0.156
	GEF	Mean	0.372	2.300	-1.004	1.000	0.498
		SE	0.117	0.054	0.126	0.024	0.156

$k = 100$ $n_i = 4$			Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.50	AL	Mean	0.436	2.410	-0.992	1.000	0.517
		SE	0.010	0.031	0.117	0.023	0.147
	PQL	Mean	0.649	2.409	-0.997	1.000	0.508
		SE	0.027	0.025	0.117	0.023	0.146
	GEF	Mean	0.871	2.343	-1.001	1.000	0.499
		SE	0.199	0.153	0.117	0.023	0.147
0.75	AL	Mean	0.591	2.435	-0.995	1.000	0.512
		SE	0.011	0.029	0.109	0.021	0.136
	PQL	Mean	1.020	2.542	-0.999	1.000	0.504
		SE	0.040	0.023	0.109	0.021	0.136
	GEF	Mean	1.466	2.135	-0.990	1.000	0.523
		SE	0.076	0.071	0.110	0.021	0.138
1.00	AL	Mean	0.722	2.602	-0.995	1.000	0.510
		SE	0.011	0.114	0.100	0.020	0.058
	PQL	Mean	1.429	2.675	-0.997	1.000	0.508
		SE	0.057	0.021	0.101	0.020	0.125
	GEF	Mean	1.654	2.169	-0.992	1.000	0.517
		SE	0.201	0.159	0.101	0.020	0.126

Table 5.4: Comparison of Simulated Mean Values and Standard Errors (SE) of the Regression Estimates and Variance Components of Random Effects for Selected Values of σ^2: $k = 100$; $n_i = 6$ $(i = 1, \ldots, k)$; True Values of the Regression Parameters: $\beta_1 = 2.5$, $\beta_2 = -1.0$, $\beta_3 = 1.0$ and $\beta_4 = 0.5$; 5000 Simulations.

$k = 100$ $n_i = 6$			Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.10	AL	Mean	0.101	2.469	-0.990	1.000	0.525
		SE	0.004	0.022	0.109	0.010	0.118
	PQL	Mean	0.119	2.138	-1.000	1.000	0.506
		SE	0.005	0.017	0.108	0.010	0.117
	GEF	Mean	0.508	2.291	-1.000	1.000	0.506
		SE	0.166	0.097	0.108	0.010	0.117
0.25	AL	Mean	0.242	2.425	-0.994	1.000	0.518
		SE	0.005	0.022	0.103	0.010	0.112
	PQL	Mean	0.294	2.207	-0.999	1.000	0.509
		SE	0.009	0.017	0.103	0.010	0.112
	GEF	Mean	0.812	2.223	-1.002	1.000	0.502
		SE	0.272	0.144	0.104	0.010	0.112

$k = 100$	$n_i = 6$		Estimates of Parameters				
Actual σ^2	Method		σ^2	β_1	β_2	β_3	β_4
0.50	AL	Mean	0.434	2.412	-0.994	1.000	0.513
		SE	0.006	0.021	0.098	0.009	0.105
	PQL	Mean	0.615	2.337	-0.999	1.000	0.503
		SE	0.015	0.016	0.098	0.009	0.105
	GEF	Mean	1.264	2.082	-0.997	1.000	0.507
		SE	0.256	0.068	0.099	0.009	0.105
0.75	AL	Mean	0.588	2.437	-0.994	1.000	0.515
		SE	0.007	0.019	0.091	0.008	0.098
	PQL	Mean	0.975	2.470	-0.999	1.000	0.505
		SE	0.021	0.015	0.091	0.008	0.098
	GEF	Mean	1.629	2.068	-0.989	1.000	0.524
		SE	0.135	0.103	0.091	0.008	0.100
1.00	AL	Mean	0.718	2.568	-0.994	1.000	0.511
		SE	0.007	0.112	0.086	0.008	0.093
	PQL	Mean	1.374	2.604	-0.997	1.000	0.505
		SE	0.028	0.014	0.086	0.008	0.093
	GEF	Mean	1.824	2.078	-0.990	1.000	0.519
		SE	0.086	0.071	0.086	0.008	0.093

Table 5.5: Comparison of Total Mean Square Errors of the Random Effect Predictions for Selected Values of σ^2: $k = 50$ and 100; $n_i = 4, 6$ $(i = 1, \ldots, k)$; True Values of the Regression Parameters: $\beta_1 = 2.5$, $\beta_2 = -1.0$, $\beta_3 = 1.0$ and $\beta_4 = 0.5$; 5000 Simulations.

Number of	Size of		Actual Variance Component σ^2				
Clusters	Clusters	Method	0.10	0.25	0.50	0.75	1.00
50	4	AL	0.600	0.688	0.842	2.308	7.342
		PQL	3.151	1.349	1.022	3.986	9.841
		GEF	9.628	1.426	24.709	18.651	17.553
	6	AL	0.221	0.245	0.311	0.978	5.355
		PQL	4.898	2.016	0.277	1.941	6.576
		GEF	20.309	22.504	43.612	26.479	16.995
100	4	AL	1.266	1.899	2.305	2.193	4.646
		PQL	8.974	6.145	2.435	1.912	4.921
		GEF	19.725	11.156	33.773	65.370	57.546
	6	AL	0.531	1.039	1.290	1.003	2.442
		PQL	13.537	9.116	3.240	0.723	1.768
		GEF	40.450	55.662	75.170	86.041	79.548

Chapter 6

Conclusions and Some Suggestions

The proposed two-step approximate likelihood approach is demonstrated to produce consistent estimates for both the fixed effect parameters and the variance component in the Poisson mixed model. When the actual variance component is near zero, our estimates are almost efficient for both the fixed effect parameters and the variance component, and are almost optimal for the random effects. When the actual variance component is away from zero, our estimates are always asymptotically unbiased for the fixed effect parameters, whereas our estimate is asymptotically negatively biased for the variance component.

Results of our limited simulation study also support that the two-step approximate likelihood approach usually performs better than the penalized quasi-likelihood (PQL) approach of Breslow and Clayton (1993) and the estimating function (GEF) approach of Waclawiw and Liang (1993) in estimating the three types of parameters of the Poisson mixed model. Specifically, the proposed approach performs slightly better than the PQL and GEF ap-

proaches in estimating the fixed effect parameters and at the same time leads to substantial improvement over the PQL and GEF approaches in estimating the random effects and their variance, especially for small σ^2. Between the PQL and GEF approach, the PQL approach was found to be superior to the GEF approach in estimating all the parameters.

Note that the arbitrary use of the Poisson mixed model for clustered count data, however, may yield invalid inference. It is very important to think through whether the assumptions of the Poisson mixed model, such as the conditional independence, are reasonable for a specific problem before using this model. Sometimes, it may be necessary to include previous responses or time variables in the fixed effect covariates and/or make special transformations for some fixed effect covariates in order to use the Poisson mixed model well. Not all clustered count data meet the assumptions of the Poisson mixed model.

Several models for discrete clustered data have recently been proposed, for example, the mixed model (Stiratelli, Laird and Ware 1984), the marginal model (Liang and Zeger 1986) and the "mixed parameter" model (Fitzmaurice and Laird 1993). But all of these models have some drawbacks. The advantages and disadvantages of these methods are discussed in detail in Liang and Qaqish (1992) as well as Fitzmaurice, Laird and Rotnitzky (1993).

Most models for clustered data studies may require one to do a large amount of computation, and/or may even be impossible to produce good inference because of the difficulty of computation. For example, the generalized linear mixed models have been hampered by the need for numerical integration to evaluate the marginal likelihood. This thesis proposes a two-step approximate likelihood approach for the Poisson Mixed Model with univariate

random effects. This approach produces the consistent estimation for both the fixed effect parameters and the variance of random effects, and has been shown by theory and simulation to be superior to two competitors in estimating all three types of parameters. However, this approach may be difficult to develop for the Poisson mixed model with multivariate random effects, and for generalized linear mixed models.

Owen's empirical likelihood approach in nonparametric models was generalized into semi-parametric models by Qin and Lawless (1994). This approach may be further developed for clustered data studies when the actual distribution of data is difficult to specify well.

Bibliography

Bartlett, M.S., and Kendall, D.G. (1946). The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation. *Journal of the Royal Statistical society*, Series B, 8, 128-138.

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

Breslow, N.E., and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88, 9-25.

Brillinger, D.R. (1986). The Natural Variability of Vital Rates and Associated Statistics (with discussion). *Biometrics*, 42, 693-734.

Clayton, D., and Kaldor, J. (1987). Empirical Bayesian Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. *Biometrics*, 43, 671-681.

Crouch, E.A.C., and Spiegelman, D. (1990). The Evaluation of Integrals of the Form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$: Application to Logistic-Normal Models. *Journal of the American Statistical Association*, 85, 464-469.

Dean, C.B., (1992). Testing for Overdispersion in Poisson and Binomial Regression Models. *Journal of the American Statistical Association*, 87, 451-457.

Dean, C.B., and Lawless, J.F. (1989). Tests for Detecting Overdispersion in Poisson Regression Models. *Journal of the American Statistical Association*, 84, 467-472.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood From Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B. 39, 1-38.

Ferreira, P.E. (1982). Estimating Equations in the Presence of Prior Knowledge. *Biometrika*, 69, 667-669.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* 80, 141-151.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science* 8 284- 309.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, 398-409.

Godambe, V.P. (1960). An Optimum Property of Regular Maximum Likelihood Equation. *Annals of Mathematical Statistics*, 31, 1208-1211.

Goodwin, E.T. (1949). The Evaluation of Integrals of the Form $\int_{-\infty}^{\infty} f(x) \exp^{-x^2} dx$. Proceedings of the Cambridge Philosophical Society, 45, 241-245.

Green, P.J. (1987). Penalized Likelihood for General Semi-parametric Regression Models. *International Statistical Review*, 55, 245-259.

Greenwood, M. and Yule, G.U. (1920). An enquiry into the nature of frequency distributions to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of*

the Royal Statistical Society, A, 83, 255-279.

Harvey, A.C., and Fernandes, C. (1989). Time Series Models for Count or Qualitative Observations. *Journal of Business and Statistics*, 7, 407-422.

Johnson, N.L., and Kotz, S. (1970). Continuous Univariate Distributions-1. Boston: Houghton Mifflin.

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

Lawless, J.F. (1987a). Negative Binomial Regression Models. *Canadian Journal of Statistics*, 15, 209-226.

Lawless, J.F. (1987b). Regression Methods for Poisson Process Data. *Journal of the American Statistical Association*, 82, 808-815.

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

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73, 13-22.

Liang, K.Y., and Waclawiw, M.A. (1990). Extension of the Stein Estimating Procedure Through the Use of Estimating Functions. *Journal of the American Statistical Association*, 85, 435-440.

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

Prasad, N.G.N., and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 85, 163-171.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist*, 22, 300-325.

Ripley, B.D., and Kirkland, M.D. (1990). Iterative Simulation Methods. *Journal of Computational and Applied Mathematics*, 31, 165-172.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). Variance Components. New York:John Wiley & Sons.

Seber, G.A.F., and Wild, C.J. (1989). Nonlinear Regression. New York:John Wiley & Sons.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random-Effects Models for Serial Observations with Binary Response. *Biometrics*, 40, 961-971.

Van der Laan, C.G., and Temme, N.M. (1984). Calculation of Special Functions: the Gamma Function, the Exponential Integrals and Error-like Functions. Amsterdam: Centre for Mathematics and Computer Science.

Waclawiw, M.A., and Liang, K.Y. (1993). Prediction of Random Effects in the Generalized Linear Model. *Journal of the American Statistical Association*, 88, 171-178.

West, M., Harrison, P.J., and Migon, H.S. (1985). Dynamic Generalized Linear Models and Bayesian Forecasting (with discussion). *Journal of the American Statistical Association*, 80, 73-97.

Wu, C.F. (1983). On the Convergence Properties of the EM Algorithm. *Annals of Statistics*, 11, 95-103.

Zeger, S.L., and Karim, M.R. (1991). Generalized Linear Models with Random Effects: A Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86, 79-86.

Zeger, S.L, Liang, K.Y., and Albert, P.S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, 44, 1049-1060.