INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Generalized Linear Mixed Models: Development and Comparison of Different Estimation Methods

Kerrie P Nelson

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2002

Program Authorized to Offer Degree: Statistics

UMI Number: 3062995

UMI Microform 3062995

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Kerne Nelson

Date 5th August 2002

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Kerrie P Nelson

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Chair of Supervisory Committee:

Brian Leroux

Reading Committee:

Norman Breslow

Brian Leroux

Michael Perlman

Date: August 5, 2002

University of Washington

Abstract

Generalized Linear Mixed Models: Development and Comparison of Different Estimation Methods

by Kerrie P Nelson

Chair of Supervisory Committee:

Professor Brian Leroux

Biostatistics

The use of generalized linear mixed models is growing in popularity in the modelling of correlated data. To date, methods available are either computationally intensive or asymptotically biased. The following work examines the performance of three methods through the use of simulation studies: maximum likelihood, approximate maximum likelihood and iterative bias correction. The effects of sample size, the true values of parameters and the distribution of the random effects on the standard errors, bias and mean-squared errors of the resulting estimates are investigated. An improvement to the iterative bias correction method has been proposed to increase the method's computational efficiency.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Brian Leroux for all of his advice and support (and sense of humour!) throughout the last three years.

I would also like to thank my other committee members, in particular Norm Breslow, Michael Perlman, Dean Bilheimer and Jon Wakefield for their very helpful input over the last few years.

And many thanks go to my parents, Peter and Emily Nelson, my family and friends for their support.

In addition, many thanks go to Kristin Sprague for her support, and the many people who kindly allowed me to use their computers including Marina Meila, Mike Heroux, Center for Statistics and the Social Sciences, the Biostatistics Department and the MSCC computing center.

Chapter 1

INTRODUCTION

Over the last decade, the class of generalized linear mixed models, commonly known as "GLMM's" or random effects models, has become an increasingly popular modelling approach in a regression setting for correlated, clustered and overdispersed data (McCulloch 1997).

Developed from a background of generalized linear models and linear mixed models, the generalized linear mixed model can be used to model a wide spectrum of non-normally distributed dependent data in a variety of research areas, including medical clinical trials involving longitudinal data (Zeger 1988), spatial data (Breslow and Clayton 1993, Leroux 2000), breeding studies (Tempelman 1998, Engel 1998), epidemiologic studies, and many others (Breslow and Clayton 1993, Crowder 1978). They are especially applicable in situations where the focus of the research is to make inferences at an individual subject level rather than a population level (Zeger 1988).

However, the practical use of these models in real-life situations has thus far proved to be challenging due to the complexity of the likelihood functions involved. Methods developed have tended to be either computationally efficient, with the resulting parameter estimates prone to bias, or computationally intensive but with the estimates being more exact (Booth and Hobert 1999). Current possibilities for fitting generalized linear mixed models include a range of maximum likelihood algorithms, approximate maximum likelihood techniques involving Taylor's series approximations, Bayesian methods (Zeger and Karim 1991), non-parametric techniques (Aitken 1999), and more recently, a method based on the method

of moments (Jiang 1998), an iterative bias correction method (Kuk 1995), and a method based upon indirect inference (Mealli and Rampichini 1999).

In the following work, the next chapter describes the background of the generalized linear mixed model, including the development of techniques for model-fitting. For each of the chapters 3–5, a section within Chapter 2 is devoted to describing the background material, including the work to date carried out by other researchers in that area.

Chapter 3 examines the performance of exact maximum likelihood estimators in the generalized linear mixed model setting for a single time series of counts. Due to the computational intensity required in finding the exact maximum likelihood estimates, which cannot usually be carried out analytically, there has been only a limited amount of work done in this area. This chapter investigates the effect of sample size and other variants such as the size of the true regression coefficients and variance components on the resulting parameter estimates using simulation studies in a more thorough manner than previously seen.

In chapter 4, the performance of approximate maximum likelihood estimates using the penalized-quasi likelihood (PQL) are examined for the generalized linear mixed model structure described in chapter 2, using simulation studies in a similar manner to chapter 3.

Chapter 5 is devoted to the development and investigation of the iterative bias correction method initially proposed by Kuk (1995). The development of this technique involves using simple methods to get starting estimates of the parameters. This iterative bias correction method is then compared to other choices of methods currently being used. Conditions for the validity of the method and the calculation of standard errors are described here, in particular for the generalized linear mixed model structure introduced in chapter 2.

The final section in chapter 5 investigates the performance properties of the estimates of the regression coefficients and variance components found using the iterative bias correction method using simulation studies as in chapters 3 and 4.

A comparison of the performance properties of all three methods for finding estimates of the

regression coefficients and variance components is made in chapter 6, based on the results found earlier in chapters 3–5. Of particular interest is the comparison of the biases of the parameter estimates for each of the three methods and the mean-squared errors associated with the parameter estimates.

One assumption underlying the generalized linear mixed model is that the random effects come from a normal distribution. Chapter 7 examines the impact of non-normally distributed random effects on the three methods for a model based on a single time series set of count data.

Chapter 8 provides a set of conclusions and recommendations, and also potential areas of future research arising from the current work.

Chapter 2

BACKGROUND

2.1 Introduction

The classical linear regression scenario which has formed the basis for most analyses of continuously distributed data, has long since been built upon and generalized in many different ways to incorporate the modelling of a larger variety of data types (Nelder and Wedderburn 1972). The classical linear model takes the form:

$$y = X\beta + e,$$

where the errors contained in the error vector e are independent and identically distributed as Normal $(0, \sigma^2)$, X is an $n \times p$ matrix, the ith row representing the ith obsrvation, the jth column consisting of the values of a covariate variable. The regression coefficients $\beta_0, \beta_1, \ldots, \beta_{p-1}$ form the vector β, while the responses for the n observations are found in the $n \times 1$ vector $y^t = \{y_1, \ldots, y_n\}^t$. The relationship between the mean response and the explanatory variables is asssumed to be linear.

One extension to the linear regression scenario allows the response variable to be non-continuously distributed, for example, to have a discrete, such as a Poisson or binomial, distribution. This suggests that a non-linear relationship between the response variable and explanatory variables may be more appropriate for modelling purposes than the linear relationship assumed in the linear regression setting. Modelling of non-continuously distributed data can be carried out within the framework of generalized linear models.

2.2 *Generalized Linear Models*

These models are comprised of three main components: the random component, the systematic component and the link function (McCullagh & Nelder 1989). The constant variance assumption of the classical regression model is replaced with a mean-variance relationship assumption. For a set of independent random variables $Y_1, ..., Y_n$ forming the response vector, the model is as follows:

- random component: the response variables Y_i, $i = 1, ..., n$ are assumed to come from a one-parameter exponential family distribution, with density function $f(y_i \,|\theta_i; \phi)$.

- systematic component: A linear predictor η_i for the ith observation is formed by

$$\eta_i \;=\; x_i^t \beta.$$

- link function: the random and systematic components are associated by a link function g to form the overall model, i.e., the link function relates the linear predictor η_i to μ_i in the following manner:

$$g(\mu_i) = \eta_i = x_i^t \beta,$$

  where g is a monotone, differentiable function and $\mu_i = \mathrm{E}(y_i)$, the expected value of the ith observation. For certain types of data such as Poisson and binomial there is a special link function called a "canonical" link function which occurs when $\theta_i = \eta_i$. Possible canonical link functions include $g(\mu_i) = \log \mu_i$ which is commonly used for Poisson count data, and $g(\mu_i) = \mathrm{logit}(\mu_i)$ which is frequently used as a link function for binomial data. A linear regression model uses the identity link function $g(\mu_i) = \mu_i = E(y_i)$. The linear regression model can be considered a special case of the generalized linear model.

Generalized linear models (GLM's) can be fitted easily using maximum likelihood algorithms including iterative weighted least squares, Fisher scoring, and Newton-Raphson. Further

details can be found in the comprehensive text, "Generalized linear models", by McCullagh and Nelder, 1989.

2.3 Correlated Data Analysis

Generalized linear models have "unified regression methodology for a wide variety of discrete, continuous and censored response data that can be assumed to be independent" (Zeger and Karim 1991). However, in many studies carried out today, responses are clustered in some way, and this dependence between responses needs to be accounted for in order to correctly assess the relationship of the response Y with the explanatory variables of interest X_j, $j = 0, \ldots, p-1$.

For example, longitudinal studies are designed to investigate changes over time for some characteristic (for example, blood pressure, temperature) that is measured for each study participant, so the measurements collected for any participant cannot generally be considered independent. In genetic epidemiology, responses for members of one family will be correlated. In sample surveys, responses from members in the same family or town may be correlated, and in spatial data collection, measurements made in one region (such as a county) may be correlated to those in neighbouring regions. So dependence between observations can occur through measurements made over time, or perhaps through some common spatial, genetic, or environmental link. These are just a few of the many possible scenarios where the assumed independence of observations in many statistical modelling techniques would not be satisfied. If the data does contain some type of dependency, usage of these methods assuming independence may lead to erroneous results, such as biasing the variance estimated for the parameter estimates, which in turn could invalidate the test statistics and confidence intervals (Watier, Richardson and Hemon 1997, Zeger and Liang 1992).

Methods for dealing with correlated data can be broadly divided into two types, depending on whether the response data is normally or non-normally distributed. Generally, methods tend to be more complex than those for independent data due to the need to account for

the correlation structure, and a normally-distributed response variable is simpler to model than a non-normally distributed response variable.

There are many possible strategies for modelling correlated data with a normally-distributed response within the regression framework. Some methods include univariate and multivariate ANOVA for repeated measures data, in which all subjects receive all treatments in a randomized order (especially useful for designed experiments and balanced data), the use of derived variables which incorporate the use of univariate techniques, analyzing each time point separately, and fitting a linear mixed model to the data. Figure 2.1 displays each of these methods.

Using a modelling approach is advantageous in that it allows the inclusion of continuous and factor (qualitative) predictors in a regression modelling framework as well as flexible modelling of the covariance structures, and can be used for both continuous and discrete response data (Diggle, Liang and Zeger 1994), when compared to the other methods mentioned that tend to be more restrictive. For example, ANOVA and MANOVA allow only factor variables and no continuous predictors.

When the data is correlated and non-normally distributed, there are two main modelling approaches that can be used: a marginal approach (Liang and Zeger 1986, Zeger and Liang 1986), and a conditional approach (Breslow and Clayton 1993). The marginal approach is particularly appropriate when one is interested in conclusions based at the population level; for example, in a medical clinical trial, the main focus is on the average difference between the control and treatment groups. In a logistic setting, the marginal model parameters describe the ratio of the population odds. Marginal models have been considered by Diggle, Liang and Zeger (Diggle, Liang and Zeger 1994) to be natural analogues for correlated data of generalized linear models for independent data. Methods for fitting marginal models include generalized estimating equations, "GEE's" (an extension of quasi-likelihood to the analysis of dependent data), and estimating function approaches (Liang and Zeger 1986). The remaining work in this thesis focuses on the conditional approach.

When the objective of a study is to make inference about individuals rather than the population average, a conditional model, otherwise known as a subject-specific or random-effects model is more useful. In contrast to the marginal model setting, in the logistic model setting, the conditional model parameters describe the ratio of an individual's odds. For example, (Neuhaus and Segal 1996) one could estimate how much an individual's probability of experiencing respiratory symptoms in a medical study changes in response to changes in environmental conditions. In a seed study (Crowder 1978) where four different combinations of two treatments were applied to a total of twenty-one plates of seeds, a population approach is more appropriate when the effects of the factors on germination are of interest, while a conditional approach would be more helpful if selecting plates of seeds with particularly high germination rates was the focus.

The interpretation of the results from a marginal model and conditional model differ (Zeger, Liang and Albert 1988). If a study was carried out on subjects each measured at a number of different time points, and at each time point, subjects were asked whether they had a respiratory infection (yes/no response) and whether they were currently smoking (yes/no covariate), a possible population model interpretation would be "that there is a 5% difference in prevalence of infection between smokers and non-smokers." A conditional model could lead to an interpretation such as "if individual i starts smoking at time 2 after time 1, we estimate his probability of infection to change from 10% to 20%".

2.3.1 Linear mixed models

Multivariate normal linear models have been used by applied statisticians since the 1930's (Palmgren 2000), but it wasn't until much later, in 1982, that Laird and Ware (1982), based on some ideas introduced by Harville (1977), formally defined a family of models for serial measurements that included repeated-measures models as a special case, leading to the formation of the linear mixed model. These models, commonly known as linear mixed models ("LMM"s), can also be written as general linear models, and are appropriate to use when the outcome variable is continuously distributed. They take the general form

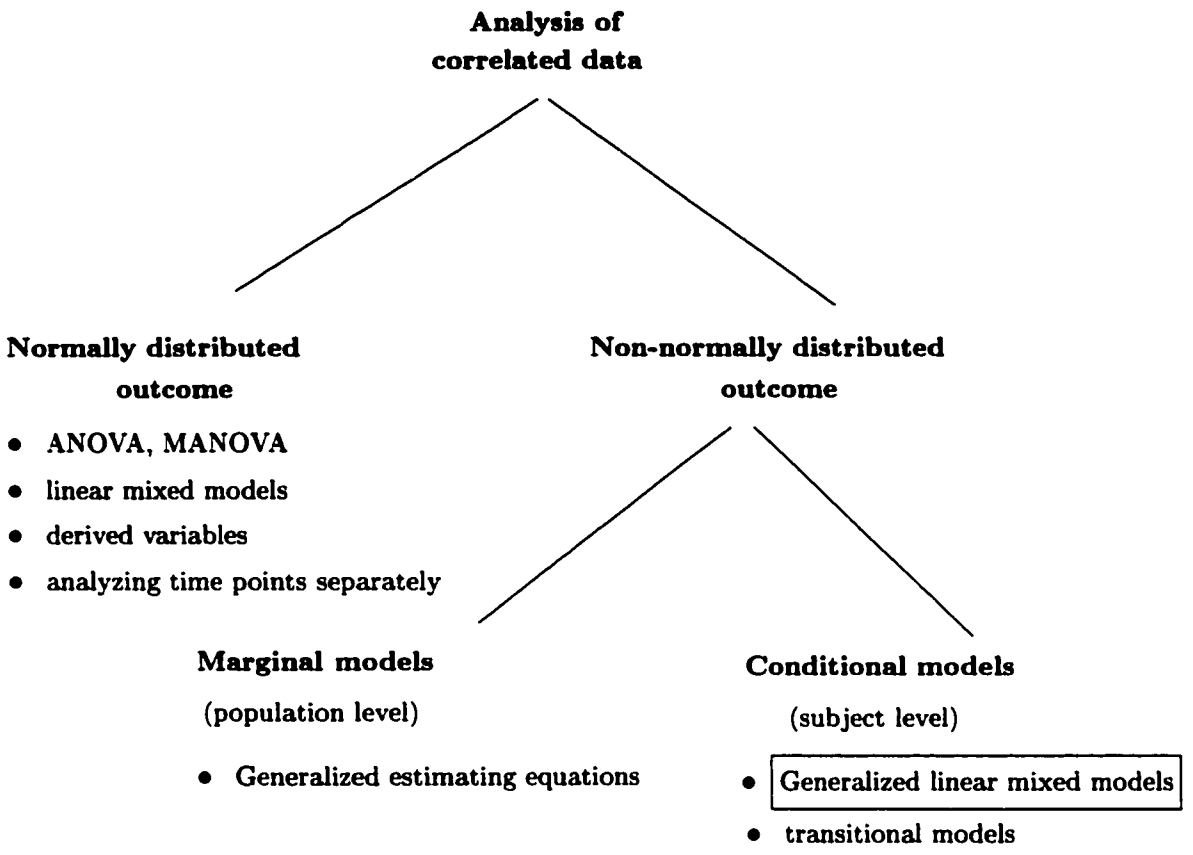

Figure 2.1: Methods for analysis of correlated data

$$y = X\beta + Zu + e \quad \text{where} \quad \begin{aligned} y_{n\times 1} &= \text{vector of responses} \\ X_{n\times p} &= \text{fixed effects design matrix} \\ \beta_{p\times 1} &= \text{fixed effects vector of coefficients} \\ Z_{n\times q} &= \text{design matrix for the random effects} \\ u_{q\times 1} &= \text{random effects vector} \sim (0, D) \\ D_{q\times q} &= \text{random effects covariance matrix} \\ e_{n\times 1} &= \text{error vector} \sim (0, R). \end{aligned}$$

The random effects structure Zu can be considered as a technique to incorporate correlation between observations at a subject or cluster level, with the errors e accounting for any residual error at an individual observation level. For one-level clustered data models, where

m independent clusters are observed, a linear mixed model structure can be applied to each cluster, with the fixed effect coefficients β constant across clusters. A covariance matrix for clustered data accounts for between-cluster and within-cluster variation. If the random effects structure is set up correctly, the responses within the ith cluster with n_i observations, $y_{i1}, y_{i2}, \cdots, y_{in_i}$ can be considered conditionally independent, i.e. all the correlation in the data can be explained by the random effects u such that $cov(y_i | u_i) = cov(e_i)$.

ANOVA and MANOVA can be considered special cases of LMM's. In addition, LMM's can be used to model non-clustered data such as time series data, spatial data, data which contains crossed random effects, and also for data containing no independent clusters.

One common estimation technique for regression coefficients of an LMM is weighted least squares (equivalent to maximum likelihood). Estimation of the covariance matrix for the random effects can be carried out using maximum likelihood, restricted maximum likelihood (Laird and Ware 1982) and method of moments. For the covariance of $\hat{\beta}$, model-based or empirical methods can be used. Many computer packages such as Splus, SAS, R, and others contain procedures and functions to analyze continuously-distributed correlated data in a regression setting.

2.3.2 The form of generalized linear mixed models

While LMM's can accommodate repeated-measures and longitudinal outcomes for continuously distributed data, and generalized linear models can model non-continuous outcome data but assume independence of the observations, the natural extension of these methods is the class of generalized linear mixed models (GLMM's), as shown in Figure 2.2. These can model data that is both non-continuously distributed and correlated, and have proved to be a very useful and flexible modelling technique in a wide variety of situations.

In the early 1980's (Diggle, Liang and Zeger 1994 and McCulloch 1994), researchers were developing generalized linear mixed models as a way to account for overdispersion in data (Williams 1982, Breslow 1984). Overdispersion is often seen in Poisson or binomial data. It

occurs when an assumed mean-variance relationship, such as mean = variance = μ as for the Poisson case is not satisfied, an important covariate is omitted, or the covariate(s) are measured with error (Follmann and Lambert 1989). For example, data collected by Ochi and Prentice (Ochi and Prentice 1984) on radiation was considered to be overdispersed because the variance was affected by individual differences in susceptibility to radiation damage, or perhaps substantial random errors in the estimated radiation exposure levels. Not accounting for overdispersion can lead to underestimates of standard errors associated with regression parameters (Ochi and Prentice 1984). Stiratelli, Laird and Ware (1984) and Zeger, Liang and Albert (1988) developed GLMM's for modelling the dependence seen in binary and other outcome variables for longitudinal, clustered, and repeated-measures studies. Further applications of GLMM's include shrinkage estimates of parameters in spatial studies (Leroux 2000, Clayton and Kaldor 1987) and meta-analysis (Aitken 1999, Berkey 1998, Platt et al 1999).

Generalized linear mixed models fall under the framework of conditional models, which also include transitional models. The general form of a GLMM is very similar to a LMM except for a non-linear link function:

$$g(\mu^u) = \eta^u = X\beta + Zu \quad \text{where} \quad \begin{aligned} y_{N\times 1} &= \text{vector of responses} \\ \mu^u &= E(y|u) \\ var(y|u) &= V(\mu^u). \end{aligned}$$

The assumptions behind the model are (Diggle, Liang and Zeger 1994):

1) The conditional distribution of Y given the random effects u follows a distribution from the exponential family with density $f(y|\, u, \beta, D)$;

2) Given the random effects u_i, the repeated measurements or correlated observations (perhaps within a cluster) $Y_{i1}, ..., Y_{in_i}$ are independent;

3) The u_i are *iid* with density function $f(0, D)$, which is usually multivariate normal for generalized linear mixed models.

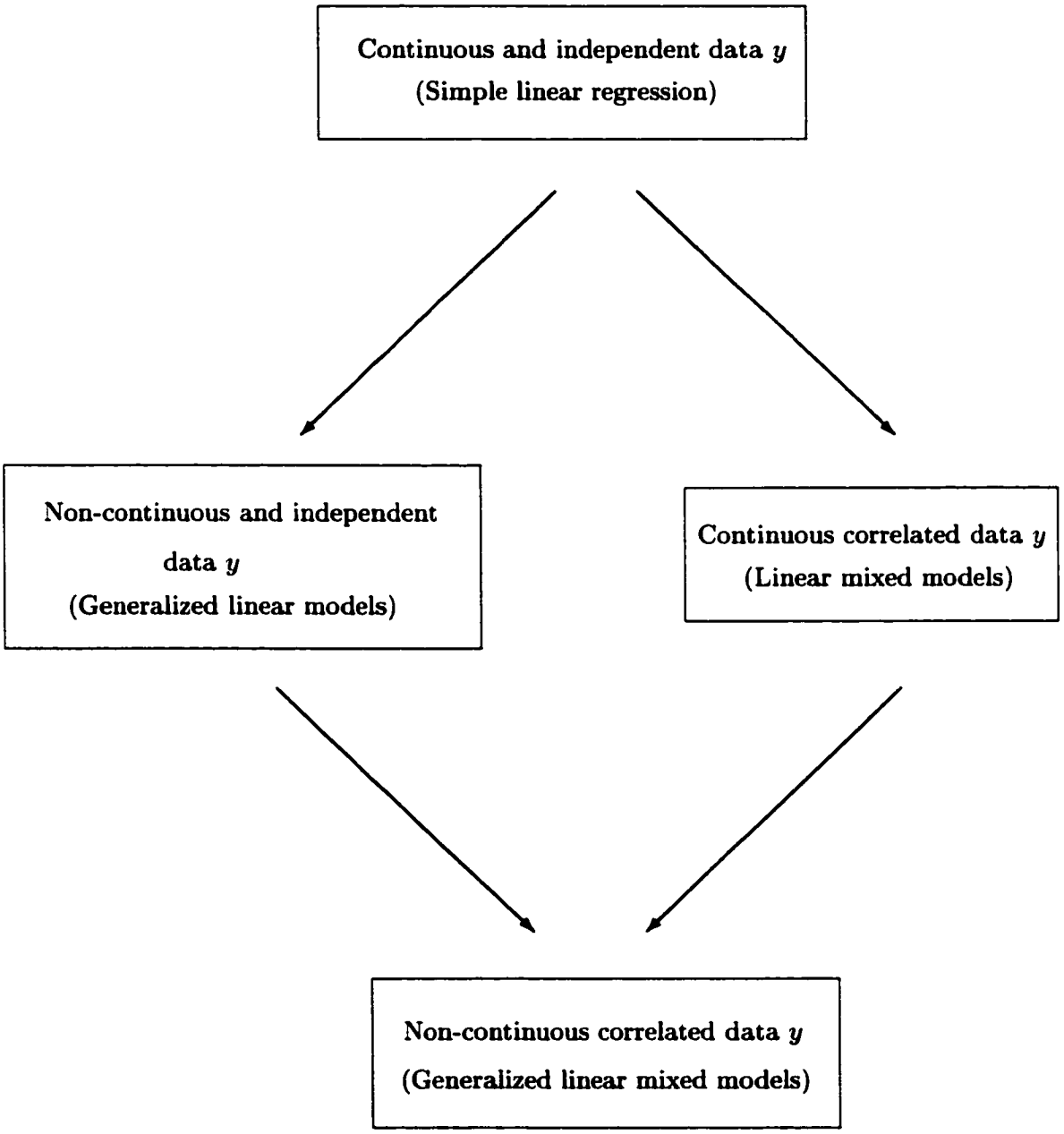

Figure 2.2: The development of generalized linear mixed models

The covariance matrix D is used to incorporate the correlation structure present in the data, for example, the data may come from a time-series dataset and an auto-regressive AR(1) or higher-order autoregressive correlation structure may be appropriate.

An example of a generalized linear mixed model for Bernoulli data, with a random intercept is a model motivated by data collected in an Indonesian children's health study on the effects of vitamin A intake on the probability of getting respiratory disease (Diggle, Liang and Zeger 1994). The model used here assumes that every child has their own probability for getting respiratory disease (thus a random effect for each child is appropriate) but that the effect of vitamin A on the probability for getting respiratory disease is the same for every child . The model is

$$\text{logit } \Pr(Y_{ij} = 1 \mid u_i) \quad = \quad \beta_0 + \beta_1 x_{ij} + u_i,$$

where the random intercepts, u_i are *iid* Normal$(0, D)$, $i = 1, ..., n$, $\beta = (\beta_0, \beta_1)$ is the vector of fixed effects, x_{ij} is a covariate measuring whether child i is vitamin A deficient, and Y_{ij} is an indicator response variable indicating whether a child had a respiratory infection at time t_{ij}. An additional assumption for the model used here is that given u_i, the repeated observations for the ith child are independent of one another. The intercept coefficient β_0 would be interpreted as the log odds of respiratory infection for a typical child with random effect $u_i = 0$ and $x_{ij} = 0$. The parameter β_1 is the log odds ratio for respiratory infection when a child is vitamin A deficient relative to when that same child is not (Diggle, Liang and Zeger 1994).

2.3.3 The fitting of generalized linear mixed models

The traditional form of estimation, maximum likelihood, that is commonly used for linear mixed models and generalized linear models, has so far proved to be limited in its usefulness for estimation in generalized linear mixed models due to the complexity of the likelihood functions for these models. In particular, when the model contains a large number of random effects, crossed random effects (Breslow and Clayton 1993), or random effects that are not

independent (Zeger 1988), the marginal likelihood which is necessary for calculation of the log likelihood can involve very high-dimensional integrals (Kuk and Cheng 1999) that are often intractable analytically.

Consequently, much effort has been put in over the last decade and is continuing at present, to develop strategies to fit generalized linear mixed models. Current methods available can be categorized broadly as follows:

- conditional likelihood methods
- approximate maximum likelihood methods
- Bayesian methods
- non-parametric methods
- exact maximum likelihood methods
- iterative bias-correction method
- indirect inference method
- approximate method based on method of moments.

2.3.4 Conditional likelihood

In this method (McCulloch and Nelder 1989), the random effects u_i are treated as a set of nuisance parameters and the regression coefficients β are estimated using the conditional likelihood of the data given the sufficient statistics for the random effects. This likelihood is fairly simple to maximize for simple forms of generalized linear mixed models, such as those with a random intercept, where observations are grouped together, with each group having its own unique random effect. One limitation of the conditional likelihood approach is that it can only be used to model within-cluster effects (effects for observations with the same random effect), and not between-cluster effects, since the random effects have been

conditioned out.

2.3.5 Approximate maximum likelihood methods

Several inference procedures have been developed using analytical approximations to the likelihood to get around the problem of intractable high-dimensional integrals. One main approach, penalized quasi-likelihood (PQL) was developed by Breslow and Clayton (1993). This approach involves integrating first- and second-order Taylor series expansions of the likelihood integral. Various other methods have been derived in different ways with alternative motivations to arrive at essentially the same estimating equations, and yield the same approximate estimates. These methods include IRREML for GLMM's (Schall 1991, Engel and Keen 1994), Laird (1978), MAP (maximum a posteriori) (Gianola and Foulley 1983, Harville and Mee 1984, Stiratelli, Laird and Ware 1984), pseudo-likelihood (Wolfinger and O'Connell 1993), REML or BLUP for GLMM's (McGilchrist 1994), and MAPHL (Maximum adjusted profile h-likelihood) (Lee and Nelder 1996).

As mentioned previously, the calculation of the marginal likelihood $f(y;\theta)$ for use in maximum likelihood is usually very challenging due to the intractability of the integrals involved in integrating out the random effects. The main idea behind PQL (Breslow and Clayton 1993) is to make a first-order Laplace approximation to an integrated likelihood of the data y, leading to a set of estimating equations that can be solved easily by iteratively fitting a linear mixed model to a modified dependent variable. The resulting equations can also be derived using a method of joint maximization.

PQL has been implemented as a computationally efficient method to fit generalized linear mixed models in a number of computer packages, including the GLIMMIX macro and NLMIXED procedure in SAS, MlWin (London Institute of Education) which implements penalized quasi-likelihood, marginal quasi-likelihood and Bayesian Gibbs sampling methods, HLM (Scientific Software International) which can fit two- and three-level hierarchical models, and GENSTAT which implements an IRREML macro for hierarchical generalized

linear models. The development of these computer packages has increased the popularity of generalized linear mixed models as a modelling tool for non-statisticians in a large variety of situations, including animal breeding (Gilmour, Anderson and Rae 1985), education (Goldstein 1986), and environmental studies (Millar and Willis 1999). The PQL method is described in more detail in §3.1.

2.3.6 Bayesian methods

An alternative approach for parameter estimation for generalized linear mixed models is the use of Bayesian methods. Zeger and Karim were the first to cast the generalized linear random effects model in a Bayesian framework in their 1991 paper and use a Monte Carlo method, the Gibbs sampler, to avoid the often intractable numerical integration issues, although their methods are not exact. Since then, Bayesian methods have been commonly used for data modelling in this context (Dellaportas and Smith 1993, Tan et al 1999, Wang et al 2000), especially with the development of computer software such as BUGS "Bayesian Inference using Gibbs Sampling" (http://www.mrc-bsu.cam.ac.uk/bugs/).

The use of Bayesian methods in the generalized linear mixed model setting is especially common in breeding studies (Moreno, Sorensen et al 1997, Tempelman 1998, Lee 2000), where it has been mentioned that parameter estimates from the Bayesian approach were comparable with those from non-Bayesian approaches. Millar and Willis (1999) comment that a Bayesian approach could also be a worthwhile approach in their study which is modelling data using generalized linear mixed models, especially with the use of non-informative priors to help allay the concerns of many critics.

Malec, Sedransk et al. (1997) also make use of hierarchical Bayesian modelling techniques in the analysis of nationwide survey data involving a very large dataset. Karim and Zeger (1992) apply the Bayesian methodology to a well-known dataset involving crossed random effects.

To show the Bayesian formulation for a generalized linear mixed model, Zeger and Karim use

a clustered data model structure, where the data is composed of a response y_{ij} from an exponential family distribution and a vector of p predictors x_{ij} for observations $j = 1, 2, ..., n_i$ within clusters $i = 1, ..., I$. The form of the model is:

$$\eta_{ij} = x_{ij}^t \beta + z_{ij}^t u_i,$$

where u_i is assumed to be multivariate Gaussian with mean 0 and variance D.

The performance of Bayesian methods has been noted by many researchers. Breslow and Clayton (1993) comment on the flexibility of the Bayesian approach for full assessment of the uncertainty in the estimated random effects and functions of model parameters, and the potential drawbacks including the computational intensity required in using these methods (Karim and Zeger 1992) and questions about when the sampling process has achieved equilibrium.

McCulloch (1997) refers to other researchers who have suggested using a Bayesian paradigm with flat or diffuse priors to approximate maximum likelihood estimates as a general approach to difficult maximum likelihood problems. However, McCulloch comments that though the numerator in such computations is the same as for the maximum likelihood calculations, this will often be inappropriate for models with random effects since the posterior distribution may not exist for diffuse priors, and that this may not be detected when using computational techniques such as the Gibbs sampler, and wrong estimates can result. Since then, some researchers have further investigated the use of alternative priors to overcome these difficulties (Daniels 1999, Natarajan and Kass 2000, Browne and Draper 2000), and further developed Gibbs sampling techniques (Hoijtink 2000).

More details on Bayesian hierarchical modelling can be found in "Bayesian Data Analysis" by Gelman, Carlin, Stern and Rubin (1995).

2.3.7 Non-parametric techniques

The term 'non-parametric' can describe a variety of situations. For generalized linear mixed models, it could be referring to methods that use some type of non-parametric smoothing technique for the estimation of the parameters and random effects; removing the common normal distributional assumption for the random effects; or thirdly, not making an assumption regarding the distribution of the response y, which is often assumed to be a member of the exponential family distribution. In a generalized linear mixed model setting, the majority of non-parametric analysis carried out focuses on removing or weakening distributional assumptions for the random effects. This is motivated by the possible sensitivity of parameter estimation to the assumption of a specific parametric model for the random effects, such as the usual Gaussian distribution. Some refer to models that remove or weaken the assumptions for the random effects as "semi-parametric" models (Neuhaus and Lesperance). A number of authors (Neuhaus and Segal 1996, Butler and Louis 1992) demonstrate in some generalized linear mixed model examples that misspecifying the random-effect distribution has little effect on the fixed-effects estimates. However, a paper by Heckman and Singer (1984) suggested that substantial changes in parameter estimates can occur with quite small changes in the mixing distribution specification. This is a current ongoing area of research.

Kiefer and Wolfowitz (1965) were possibly the first researchers to consider nonparametric maximum likelihood estimation of a mixing distribution. This paper set the scene for later research into the area of nonparametric distributions for random effects models, including a paper by Simar in 1976.

A number of researchers have investigated non-parametric methods for the linear mixed model (Madger and Zeger 1996), and also for the generalized linear mixed model, in particular, finding nonparametric techniques for estimating the distribution of the random effects. Follmann and Lambert (1989) use nonparametric maximum likelihood to estimate a discrete distribution of a random slope in a logistic generalized linear mixed model. Bartlett and Sutradhar (1999) provide a semi-parametric solution to finding parameter estimates in a generalized linear mixed model, which involves a two-step joint estimating equations approach. In the first step, an estimating function based approach is used to obtain the

estimates of the random effects, and in the second step, the first two moment based joint estimating equations for the regression parameters and the variance component of the random effects are constructed in the case of a generalized linear mixed model with a single random effect.

A semi-parametric Bayesian approach to generalized linear mixed models is proposed by Kleinman and Ibrahim (1998) where the usual normal distribution prior on the random effects is replaced with a non-parametric prior followed by a Dirichlet process prior on that general distribution.

Aitken (1999) reported on an EM algorithm developed for nonparametric maximum likelihood regression in generalized linear models with variance component structure. The algorithm is initially derived as a form of Gaussian quadrature assuming a normal mixing distribution, but with only slight variation, it can be used for a completely unknown mixing distribution giving a straightforward method for the fully nonparametric maximum likelihood estimation of this distribution.

While Aitken's method provides a nonparametric distribution for the random effects which is discrete, and therefore perhaps not totally realistic, Tao, Palta et al. (1999) propose a semiparametric mixed effects regression model where the common assumption of Gaussian random effects is relaxed by using a predictive recursion method to provide a nonparametric smooth density estimate. This approach does not yield direct estimates of individual random effects but is feasible for fitting semiparametric mixed models on quite large datasets.

Walker and Mallick (1997) consider a Bayesian nonparametric approach to analysing hierarchical generalized linear models.

The performance of these different methods is well-documented within these papers for specific examples; however, not a lot has been done to date to compare the performance of these methods with the more commonly used parametric approaches.

2.3.8 Exact maximum likelihood methods

Because bias may be present in the approximate maximum likelihood methods mentioned above, there has been ample motivation for investigating methods for finding the exact maximum likelihood estimators, despite the intractability of the integrals involved (Booth and Hobert 1999). For generalized linear mixed models that have a simple random effect structure, such as a random intercept, numerical integration techniques that provide exact solutions can be used for full maximum likelihood estimation, such as some form of Gaussian quadrature (Anderson and Aitken 1985), which is widely regarded as computationally intensive (Aitken 1999). Gaussian quadrature methods for many different generalized linear mixed model structures can now be implemented in SAS using the NLMIXED procedure.

Recently, Lesaffre and Spiessens (2001) tested out the Gauss-Hermite method (based on Gaussian quadrature points) for a very simple example, and found that it gave valid results only when a high number of quadrature points was used. They suggest that the adaptive Gaussian quadrature procedure as implemented in the SAS procedure NLMIXED usually works better, but that in their experience with even relatively simple models, convergence to a global maximum can be difficult to obtain.

Through the use of the EM (Dempster, Laird and Rubin 1977) and Newton-Raphson (NR) algorithms (Tanner 1991), and various simulation techniques (Metropolis-Hastings algorithm, importance, rejection and Gibbs sampling), a number of methods were developed in the 1990's to obtain exact maximum likelihood estimates for both simple and more complex generalized linear mixed models. Generally these methods are computationally very demanding, and the availability of more powerful and cheaper computers has led to these methods becoming popular in the fitting of generalized linear mixed models, although the complexity of the programming involved is still daunting for most.

In 1994 McCulloch described a Monte-Carlo Expectation-Maximization algorithm (MCEM) implementing Gibbs sampling that can handle simple and complex fixed and random effects structure, but is restricted to datasets that have a binary response with a probit link

function. McCulloch extended this method in 1997 to a more general MCEM algorithm that incorporates the Metropolis-Hastings algorithm, allowing for a broader range of link functions and response distributions. In addition, McCulloch also proposed a Monte-Carlo Newton-Raphson algorithm (MCNR), and adapted simulated maximum likelihood (SML) methods (as developed by Geyer and Thompson (1992) and Gelfand and Carlin (1993)), and also a hybrid algorithm combining MCNR and SML for use in at least some generalized linear mixed models.

Booth and Hobert in their 1999 paper proposed two new implementations of the EM algorithm for maximum likelihood fitting of generalized linear mixed models, which use rejection and importance sampling. They suggest that their methods can be considerably more efficient than those based on Monte-Carlo Markov-chain algorithms, such as McCulloch's MCEM algorithm. However, they point out their methods may break down when the intractable integrals in the likelihood are of a high dimension.

Geyer and Thompson (1992) and Gelfand and Carlin (1993) developed the use of simulation to directly approximate the likelihood, described by Kuk and Cheng as the "functional approach" (Kuk and Cheng 1999). These methods were applied in the context of generalized linear mixed models by McCulloch (1997), who also compared them to his MCEM algorithm, and found that this functional approximation approach often performed poorly. Kuk and Cheng (1999) similarly compare these algorithms and describe the functional approach as more ambitious, but giving an approximation which is local in nature.

Quintana, Liu et al. (1999) proposed a new Monte Carlo EM algorithm to compute maximum likelihood estimates in the context of random effects models. The algorithm involves the construction of efficient sampling distributions for the Monte-Carlo implementation of the E-step, together with a reweighting procedure that allows repeated use of a sample of random effects. They tested their method on one simulated binomial dataset similar in structure to one of McCulloch's (1997) examples:

$$\text{logit}(p_{ij}) \;=\; \beta x_{ij} + u_i, \text{ where } u_i \sim \text{ iid } N(0, \sigma^2).$$

They suggested that the introduction of the reweighting step might represent considerable savings in computational effort, compared with McCulloch's algorithm.

Since then, other EM-type algorithms have also been developed. Van Dyk (2000) proposed an approach involving two or more nestings of the EM algorithm that he found to be computationally more efficient than the other Monte-Carlo approaches for at least one generalized linear mixed model structure.

Molenberghs and Goetghebeur (1997) and Galecki, Ten Have et al. (2001) proposed a faster alternative to available EM approaches under a multivariate generalized logistic model with a composite link function, which may be useful in the generalized linear mixed model setting.

Currently these methods are not available in computer packages, except for adaptive quadrature, which is available in the procedure NLMIXED in SAS for a range of generalized linear mixed models. Computer code generated by researchers tends to be fairly specialized.

2.3.9 Iterative bias correction method

Kuk (1995) remarks that it is rather difficult to obtain asymptotically unbiased estimators for general random effects models, and has developed a general method to adjust inconsistent estimators to result in estimators that are asymptotically unbiased. The method is motivated by an iterative bias correction and does not require approximate linearization of the model. He carried out a small simulation study to show that this method, which adjusts initial estimates found using a method such as BLUP, can lead to estimates that are nearly unbiased even for the variance components, and consistent, at least for a binary example shown. For this example at least, Kuk's estimates of the regression coefficients and variance components are unbiased, and perform better than both McGilchrist's approximate residual maximum likelihood method, PQL and bias-corrected PQL (CPQL) methods. Kuk mentions that the tradeoff for correcting the downward bias of σ^2 is in 10-20% larger standard errors, but in general, the standard errors appear comparable to other methods. Since the publication of Kuk's paper in 1995, a few researchers have applied this technique in certain

fields. More detail is provided on Kuk's method in Chapter 5.

2.3.10 Indirect inference method

A flexible approach, indirect inference, originally developed by Gourieroux, Monfort et al. (1993), was recently extended for use in the generalized linear mixed model setting by Mealli and Rampichini (1999). Estimation using the indirect inference method consists of two main steps. In the first step, an approximated (auxiliary) model, chosen as it is easier to handle than the original model, is used to derive estimates of some auxiliary parameters. In the second step, simulations are used to correct the discrepancy of the auxiliary parameters from the original ones. The simulations are run for the model of interest with fine tuning of the parameters until the estimates of the auxiliary parameters using the simulated data are close with regard to some defined criteria to those using the original dataset. Mealli and Rampichini comment that indirect inference appears to be a more general procedure that encompasses the iterative bias correction method. The performance of the indirect inference method is discussed below in §2.4.6.

2.3.11 Approximate method based on method of moments

Jiang (1998) developed a simple method based on simulated moments to estimate the regression coefficients and variance components in a generalized linear mixed model. In the method, a set of sufficient statistics is first found from the density function and a set of estimating equations is then obtained by equating sample moments of the sufficient statistics to their expectations. However, while the integrals involved with these expectations are usually of much lower dimension (equalling the number of sources of random effects) than those of integrals involved in the likelihood function, analytic evaluation of the integrals may still not be feasible; thus simulated moments are used as an approximation. Jiang notes that such a method has been studied in econometrics. The performance of this method is discusssed below in §2.4.6.

2.4 Sampling Performance of Different Estimation Techniques

2.4.1 Introduction

The use of generalized linear mixed models to model correlated data is growing in popularity. However, to date, there has been limited investigation into the properties and performance of the estimators for regression coefficients and variance components for the different methods used in estimation for generalized linear mixed models. These methods include exact maximum likelihood methods such as McCulloch's MCEM algorithm, approximate maximum likelihood methods, and the iterative bias correction method.

While it is generally assumed that exact maximum likelihood estimates are asymptotically unbiased, little is known about the performance of exact maximum likelihood in smaller to moderate-sized samples, in the generalized linear mixed model setting. The same problems occur in linear mixed models when the random effects are crossed. The lack of investigation generally can be attributed to the complexity and intractability of the mathematics involved in attempting to prove results theoretically, and the intensity of the computing effort required if simulation studies are used. As Millar and Willis (1999) aptly pointed out, "in situations of small or moderate sample sizes there would be no guarantee that the MLE's had good properties and it would remain desirable to perform a simulation study, but this would be computationally prohibitive".

Furthermore, while approximate likelihood methods are attractive to use due to their computational efficiency, the more thorough investigations that have been done on these methods suggests there can be serious bias present in the resulting estimates of the regression coefficients and variance components. Other methods such as the iterative bias correction method developed by Kuk have also been considered up to now to be computer intensive, and consequently, little has been done to test their performance.

The following section presents an overview of work done to date in investigating the performance properties of the commonly-used exact and approximate maximum likelihood

methods, iterative bias correction, and to a lesser degreee, other methods such as Bayesian methods, indirect inference, and Jiang's method based on simulated moments.

2.4.2 Sampling performance of approximate maximum likelihood estimators

Out of all the methods currently used in fitting generalized linear mixed models, much of the effort that has been spent in investigating the performance properties has focused on approximate methods, due to the computational efficiency and availability of these methods in SAS, Genstat and MLn. Researchers have especially looked at the modelling of correlated data with a binary response variable, as this has many applications in research areas such as animal breeding and medical studies.

In 1993, Breslow and Clayton applied PQL to the Crowder seed data (Crowder 1978). The seed data, as described earlier in §2.3 consists of the outcome variable y_i as the proportion of seeds that germinated on plate i ($i = 1, ..., 21$), and two covariates: seed variety (two types) and type of root extract (two types) arranged in a factorial design. The generalized linear mixed model used to account for the overdispersion present in the model was:

$$\text{logit } \Pr(Y_i = 1 \mid u_i) \quad = \quad \beta_0 + \beta_1 * \text{type}_i + \beta_2 * \text{extract}_i + u_i.$$

Because of the simplicity of the model, Gaussian quadrature was also applied to get exact maximum likelihood estimates. The parameter estimates and their standard errors from this analysis as presented in Table 2.1 show some differences between the exact maximum likelihood estimates and the PQL estimates, with the maximum likelihood estimate of the variance component being 20% smaller.

Breslow and Clayton then went on to examine the effect of the size of the binomial denominator on the estimated regression coefficients and variance components in two sets of simulations for a simple correlated binomial data example. The data consisted of 100 clusters of 7 observations each, with 200 and 100 datasets, respectively, for the two sets of simulations. The form of the model was:

Table 2.1: Summary of seed data analysis results (Breslow and Clayton 1993).

Parameter	GLM	PQL	Maximum Likelihood
β_0	-0.430 (0.114)	-0.375 (0.182)	-0.389 (0.166)
β_1	-0.270 (0.155)	-0.363 (0.228)	-0.347 (0.215)
β_2	1.065 (0.144)	1.012 (0.224)	1.029 (0.205)
σ	-	0.352 (0.118)	0.295 (0.112)

$$\text{logit } \Pr(Y_{ij} = 1 \mid u_i) \;=\; \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 x_i t_j + u_i^0 + u_i^1 t_j.$$

Different sets of simulations were carried out for the binomial denominators $1, 2, 4$ and 8, two covariates x_i and t_j, where $x_i = 1$ for half the sample and 0 for the other half, and $t_j = j - 4$ for $j = 1, 2, ...7$. The true regression coefficients used were $(-2.5, 1, -1, -0.5)$. For the first set of simulations, the random effects were independently and identically distributed with covariance matrix D_1 (as shown below), while the second set of simulations the random effects had dispersion matrix D_2. The matrices D_1 and D_2 are:

$$D_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } D_2 = \begin{pmatrix} 0.50 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

The results for the first set of simulations (with a single component of variance) showed that the often substantial downward bias present in the estimated variance components for small binomial denominators of 1 or 2, improved as the binomial denominator became larger, and the bias present in all of the estimated regression coefficients for small denominators showed some improvement as the binomial denominator increased to 8. A summary of these simulation results, as presented in Breslow and Clayton (1993) is presented in Table 2.2 below. The results given are the average parameter estimates based on 200 replications

for the first set of simulations, and 100 replications for the second set of simulations for each binomial denominator.

Similar effects were seen for the second set of simulations where two variance components were involved, with the serious downward bias in the variance components generally improving as the binomial denominator increased, and similarly for the estimated regression coefficients. These results also are summarised in Table 2.2 below.

Table 2.2: Results for Breslow and Clayton's binary simulations using PQL.

Binomial denominator	σ_1^2	σ_{12}	σ_2^2	β_0	β_1	β_2	β_3
First set of simulations based on D_1							
1	0.68	–	–	-2.31	0.93	-0.94	-0.42
2	0.79	–	–	-2.36	0.95	-0.90	-0.46
4	0.82	–	–	-2.38	0.96	-0.93	-0.46
8	0.90	–	–	-2.46	0.98	-0.94	-0.48
true value	1.00	–	–	-2.50	1.00	-1.00	-0.50
Second set of simulations based on D_2							
1	0.35	-0.05	0.15	-2.31	0.93	-0.94	-0.42
2	0.43	-0.04	0.17	-2.36	0.95	-0.90	-0.46
4	0.36	-0.00	0.20	-2.38	0.96	-0.93	-0.46
8	0.41	-0.01	0.22	-2.46	0.98	-0.94	-0.48
true value	0.50	0.00	0.25	-2.50	1.00	-1.00	-0.50

These approximate maximum likelihood methods are computationally efficient and in many situations provide close to exact maximum likelihood estimates, especially when the true distribution of the random effects is normal, and when the variance components in a binary response data situation are small. Unfortunately under other circumstances (such as when

the data is binary and consists of matched pairs, the random effects have a large variance, or the random effects do not follow a normal distribution), these methods can lead to estimates of the β coefficients and variance components which are biased, often underestimating the true values of the variance components (Engel and Buist 1998, Breslow and Lin 1995, Lin and Breslow 1996, Booth and Hobert 1999, Aitken 1999, Rodriguez and Goldman 1995).

Neuhaus and Segal (1996) suggest the attentuation of the variance components closely corresponds to what one may expect from fitting population-averaged models to non-continuous correlated data. Breslow and Lin (1995) and Lin and Breslow (1996) made bias corrections to the PQL method which greatly extend the range of parameter values for which the approximate estimation procedures have satisfactory asymptotic properties (Breslow and Lin 1995), while not completely reducing the bias in some situations (Neuhaus and Segal 1996). Other researchers including Goldstein and Rashbash (1996) have also made improvements to approximate maximum likelihood methods.

Some authors (Neuhaus and Segal 1996) consider that these approximate methods even with bias corrections, may require modification before their asymptotic bias can be competitive with more exact mixed effects model methods that provide consistent estimation. Engel mentioned that the scope for reduction of bias in the estimation of heritability for binary data in animal breeding is thought to be slim (Engel 1998, Engel and Buist 1998). Many authors (Booth and Hobert 1999, Engel 1998) have considered that the relative simplicity of the PQL (or IRREML) approaches keep these as attractive alternatives when compared to the more computer intensive exact maximum likelihood methods.

The estimating equations that the PQL method is based upon are the REML (Residual or Restricted Maximum Likelihood) equations under the normal theory linear model, and consequently, as Breslow and Clayton remark, the PQL approximations are likely to improve as the individual y_i become more normally distributed, such as when the denominator of a binomial proportion increases (as was seen in the above simulation studies), or the mean of Poisson observations increases. Overall, Breslow and Clayton comment that the simulation results were encouraging as regards the ability of PQL to render approximately correct

inferences on regression coefficients in hierarchical models. However, they found that bias in the parameter estimates was an issue, particularly when the binomial denominator was small.

In 1994 McGilchrist developed a method for estimation in generalized linear mixed models that extended the best linear unbiased predictor (BLUP) methods of Henderson (1963). The idea is to use BLUP to obtain approximate maximum likelihood estimates of the regression coefficients and variance components for data that are correlated and non-continuous. While McGilchrist's intention was not to do extensive simulations, he carried out two small simulation studies on a binomial-logit model of 30 observations to compare the performance of his two methods, the above mentioned approximate maximum likelihood method (McGilchrist-1), and a method based on the full likelihood of the data (McGilchrist-2), increasing the value of the variance component between the two sets of simulations (each consisting of 100 datasets) from 1.0 to 2.0. The form of this model was:

$$\text{logit } \Pr(Y = 1 \mid u) = X\beta + Zu,$$

where $u = (u_1, u_2, ..., u_{15})$ are iid $N(0, \sigma^2)$. A summary of the results from his simulation studies are presented in Table 2.3. As can be seen from this table, the approximate maximum likelihood estimates (McGilchrist-1) of the variance components were less biased than the estimates from McGilchrist's other method in both sets of simulations. For both McGilchrist's approximate maximum likelihood method and his other method based on the full likelihood (McGilchrist-2), the regression parameters β_0 and β_1 are estimated in the same way, but differ slightly in the results due to the different estimates of σ^2 produced. When the value of the variance is increased from 1.0 to 2.0, the bias appears to worsen, though it is difficult to draw any firm conclusions based on one hundred simulations for each set.

Table 2.3: Results for McGilchrist's binary simulations.

Simulation set	Method	Average estimated parameter values (se's) σ^2	β_0	β_1
Set 1:	McGilchrist-2	0.705 (0.439)	0.249 (0.309)	0.100 (0.037)
	McGilchrist-1	0.907 (0.540)	0.253 (0.314)	0.101 (0.038)
	True value	1.000	0.200	0.100
Set 2:	McGilchrist-2	1.336 (0.640)	0.109 (0.343)	0.097 (0.048)
	McGilchrist-1	1.681 (0.795)	0.111 (0.350)	0.099 (0.049)
	True value	2.000	0.200	0.100

2.4.3 Sampling performance of bias-corrected approximate maximum likelihood estimators

After the recognition of the often substantial bias present in the PQL method, Breslow and Lin developed the PQL method further. They published three papers on their corrected PQL (CPQL) methods, the first in 1995 for generalized linear mixed models with a single component of variance (Breslow and Lin 1995), another in 1996 for generalized linear mixed models with multiple sets of random effects (Lin and Breslow 1996a), and a third paper (Lin and Breslow 1996b) describing the PQL correction procedure for analysis of correlated binary data in logistic-normal models, and a comparison with a Bayesian Gibbs sampling approach. In their 1995 paper, Breslow and Lin derive general expressions for the asymptotic bias present in the PQL estimators of the regression coefficients and variance component for a generalized linear mixed model with a single component of variance, along with other approximate estimators (Solomon and Cox 1992, Liu and Pierce 1993) which involve first and second-order Laplace expansions of the integrated likelihood for use in inference on the odds ratio in a simple random effects model for a series of two by two tables.

The bias present in the estimated regression coefficients, as σ^2 goes to zero, is of the order σ^2

for PQL, of the order $(\sigma^2)^2$ for the first-order Laplace and Solomon-Cox methods, while of a smaller magnitude for the second-order Laplace and Solomon-Cox methods. The corrected PQL method (CPQL) involves subtracting the linear bias term from the estimated regression coefficients. Performance of the corrected PQL method can be summarized as follows: for σ^2 less than 0.5, there is little bias in the regression coefficients (also true for uncorrected PQL and first-order Laplace methods, with slightly greater bias for the first and second-order Solomon-Cox estimators). With regard to the variance components, CPQL (and the second-order Laplace method) worked quite well for σ less than 0.5. It should be observed that these results were noted for correlated binary data in a small range of regression coefficients and variance component values. Breslow and Lin comment that the correlated data problem for which the numerical work was carried out could be considered close to a "worst case scenario" for approximate procedures, due to success probabilities being close to 0 or 1, and that substantially better performance can be expected when larger binomial denominators are used.

The 1996 paper by Lin and Breslow focused on correcting the PQL method for generalized linear mixed models with multiple sets of random effects. Instead of a correction factor as for a single component of variance, a correction matrix was developed. Numerical results for the matched pairs correlated data problem suggested that CPQL often overcompensated for the bias in the estimated regression coefficients, especially for moderate values of these and the dispersion parameter. Aside from comparing the CPQL method to other previously reported analyses on a well-known salamander dataset involving crossed effects, Lin and Breslow carried out some simulation studies comparing the various PQL methods and REML for 1000 simulations based on the salamander data for two different sample sizes, $n = 360$ and $n = 720$, and two sets of variance components, $(\sigma_f^2, \sigma_m^2) = (0.5, 0.5)$ and $(\sigma_f^1, \sigma_m^2) = (1.69, 1.50)$. The logit model used for the simulations takes the form

$$\text{logit}\, P\left(y_{ij} = 1 \mid u_i^f, u_j^m\right) = x_{ij}^t \beta + u_i^f + u_j^m, \quad \text{with} \quad i, j = 1, \ldots, 60,$$
$$\beta^t = (1.06, -3.05, -0.72, 3.77).$$

Table 2.4 below presents the results from these simulations under the different approximate

methods used.

Table 2.4: Results for Lin and Breslow's salamander data simulations.

Sample size	Method	Average estimated parameter values					
		σ_1^2	σ_2^2	β_0	β_1	β_2	β_3
$n = 360$	PQL	0.33	0.32	0.94	-2.73	-0.64	3.38
	PQL with CPQL σ^2	0.46	0.46	0.96	-2.78	-0.66	3.44
	First-order CPQL	0.46	0.46	1.17	-3.31	-0.79	4.11
	Second-order CPQL	0.46	0.46	1.01	-2.95	-0.69	3.65
	REML	0.55	0.54	1.09	-3.14	-0.74	3.88
$n = 720$	PQL	0.30	0.29	0.92	-2.66	-0.62	3.29
	PQL with CPQL σ^2	0.43	0.42	0.94	-2.71	-0.63	3.35
	First-order CPQL	0.43	0.42	1.13	-3.21	-0.76	3.97
	Second-order CPQL	0.43	0.42	1.01	-2.94	-0.69	3.64
	REML	0.52	0.51	1.06	-3.06	-0.71	3.79
	True value	0.50	0.50	1.06	-3.05	-0.72	3.77

The above table suggests that any of the corrected PQL methods work fairly well when estimating the variance components, while PQL is seriously biased as expected when the sample size is $n = 360$. When the sample size is doubled to $n = 720$, the bias seen in the variance components for each of the PQL methods becomes slightly worse. The REML estimates of variance components are almost unbiased overall, but have much larger mean-squared errors attached to them than the PQL variance component estimates. The estimated regression coefficients are really only reasonably estimated by the second-order CPQL method, while the other PQL methods lead to fairly biased estimates. The REML method slightly overestimated the regression coefficients. When the variance components are increased to $(\sigma_1^2, \sigma_2^2) = (1.69, 1.50)$, both the first-order and second-order CPQL methods failed, and Lin

and Breslow recommend using only the CPQL version of the variance components with the original PQL method to estimate the regression coefficients.

Lin and Breslow (1996b) carried out extensive simulation studies of 200 datasets each on binary observations with one hundred clusters of seven observations each for different binomial denominators $m = 1$ and 8. The model used is:

$$\text{logit } \Pr(Y_{ij} = 1 \mid u_i) \quad = \quad \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 x_i t_j + u_i,$$

where $t_j = j - 4$ for $j = 1, .., 7$, and $x_i = 0$ for half of the sample and 1 for the other half. Several tables of results are presented in the paper. A sample of some of the results is as follows:

Table 2.5: Average parameter estimates for a binomial model (Lin and Breslow 1996b).

	Parameter	σ^2	β_0	β_1	β_2	β_3
m	true value	1.00	-2.50	1.00	-1.00	0.50
1	PQL	0.66	-2.28	0.93	-0.89	0.43
	PQL(σ^2_{CPQL})	0.88	-2.32	0.94	-0.90	0.44
	1st CPQL	0.88	-2.66	1.08	-1.03	0.51
	2nd CPQL	0.88	-2.40	0.97	-1.00	0.49
	Gibbs Sampler	1.21	-2.67	1.07	-0.96	0.49
8	PQL	0.92	-2.43	0.98	-0.97	0.48
	PQL(σ^2_{CPQL})	0.95	-2.43	0.98	-0.97	0.48
	1st CPQL	0.95	-2.81	1.13	-1.12	0.57
	2nd CPQL	0.95	-0.62	0.27	-0.90	0.37

Lin and Breslow concluded from this paper that the correction works best when the variance components are small and the sample size is reasonably large. Both first- and second-order corrections reduce the bias in the PQL regression coefficient estimators for small values of

the variance components, especially when the values of the regression coefficients are large and the data binary. The first-order correction often overcompensates slightly for the bias in the PQL regression coefficient estimators. Both corrections, especially the second-order one, break down for large variance components and large binomial denominators (see last line of Table 2.5).

2.4.4 Sampling performance of approximate maximum likelihood methods by other authors

Other authors have carried out simulation studies to compare a range of methods. Goldstein and Rasbash (1996) carried out a simulation study of 200 datasets on a binary data model with two separate random effects, and three nested levels for the model as follows

$$\text{logit } p_{ijk} = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + u_k, \qquad \text{where}$$

$$u_{jk} \sim N(0, \sigma^2_{u_2}) \quad \text{and}$$
$$u_k \sim N(0, \sigma^2_{u_3})$$

to compare the performance of MQL (marginal quasi-likelihood) with first- and second-order corrected PQL. MQL is a method described by Breslow and Clayton (1993) alongside PQL which is helpful to calculate estimates of covariate effects on population averages rather than specific subjects. The main difference between the two methods is that the MQL estimating equations do not contain the random effects terms in the linear predictor, while the PQL estimating equations do.

The results are given in the table below. Goldstein and Rasbash found that second-order corrected PQL outperformed the first-order corrected PQL and MQL methods in terms of bias but not standard errors, and the variance components still showed some negative bias.

An earlier study carried out by Rodriguez and Goldman (1995) examined binary data models using a large number of simulations with varying hierarchical data structures. Their results

Table 2.6: Summary of Goldstein and Rasbash (1996) results.

Parameter	True value	MQL	CPQL (1st-order)	CPQL (2nd-order)
β_0	0.665	0.512 (0.010)	0.548 (0.011)	0.660 (0.014)
β_1	1.0	0.738 (0.012)	0.795 (0.013)	0.965 (0.015)
β_2	1.0	0.745 (0.006)	0.805 (0.006)	0.968 (0.008)
β_3	1.0	0.767 (0.014)	0.837 (0.015)	1.002 (0.019)
σ_{u_2}	1.0	0.119 (0.010)	0.457 (0.006)	0.802 (0.011)
σ_{u_3}	1.0	0.748 (0.004)	0.800 (0.005)	0.968 (0.007)

revealed substantial biases in the estimates of the fixed effects and the variance components or both whenever the random effects were sufficiently large, or the number of observations within a given level of clustering was small.

Engel and Buist (1998) carried out simulation studies on a binary data model in the context of an animal breeding study to investigate the performance of the approximate maximum likelihood method IRREML, corrected and uncorrected for bias, and a third method using alternative weights within the IRREML algorithm. Corrected and uncorrected IRREML procedures worked very well when large numbers of sires and offspring per sire were included in the model, while more bias was present for a moderate number of fixed effects. Engel and Buist suggest that for many statistical problems with a relatively small number of fixed effects and a large number of random effects, the Breslow and Lin correction factor is a useful asset. However, in animal breeding studies where large numbers of fixed effects are common, the correction factor seems to be of limited use.

Neuhaus and Segal (1996) investigated the performance of approximate maximum likelihood for a matched pairs data example based on pulmonary function using a binomial model, where the goal of the analysis was to examine whether an individual's propensity to experience respiratory symptoms changed with exposure to ozone. The bias-corrected PQL

estimates are the closest to the true maximum likelihood estimates (found using EGRET), but are attenuated. In this matched-pairs example, PQL exhibits substantial bias for the variance component.

Table 2.7: Approximate maximum likelihood results for a study of ozone exposure on respiratory morbidity.

Parameter	Maximum Likelihood	PQL	CPQL
β_0	-2.69 (0.79)	-1.52 (0.34)	-1.92
β_1	1.61 (0.63)	0.93 (0.40)	1.15
σ^2	6.78	1.26	
ρ	0.44		

A study carried out by Sutradhar and Qu (1998) compared three approaches for the analysis of a count data set, a refreshing change from the more commonly-seen binomial data models. The three approaches are PQL, a proposed likelihood method based on a small σ^2-based approximate likelihood function, and methods of Waclawiw and Liang (1993) based on so-called Stein-type estimating functions. Fairly extensive simulations were carried out for two different cluster sizes, four and six, and a range of σ^2 values from 0.1 to 1.0. The results for their example demonstrate that the fixed effects are estimated similarly well by all three methods. The variance components are estimated with least bias, with the proposed likelihood approximation method exhibiting the least bias, and the Waclawiw approach being the most biased.

PQL and other approximate procedures are proving to be popular methods for estimating parameters in the generalized linear mixed model setting, especially as they are significantly more computer efficient, provide good estimates in many situations, and can be fairly simple to program.

2.4.5 Sampling performance of exact maximum likelihood methods

In his 1997 paper, McCulloch simulated one hundred datasets for a binary data model, and compared his regression coefficient and variance component estimates using the methods described above with those using PQL. The model used was:

$$\text{logit}(p_{ij}) \;=\; \beta_0 + \beta_1 x_{ij} + u_i, \text{ where } u_i \sim \text{ iid } N(0, \sigma^2).$$

These results are summarized in Table 2.5.

Table 2.8: Results for McCulloch's (1997) binary simulations.

Method	Average estimated parameter values (standard errors) σ^2	β_1
PQL	0.96 (0.13)	4.630 (0.05)
MCEM	1.41 (0.07)	4.990 (0.01)
MCNR	1.39 (0.09)	4.990 (0.02)
SML	1.14 (0.15)	4.420 (0.05)
MCNR + SML (hybrid)	1.41 (0.09)	4.446 (0.02)
MCNR + 2SML (hybrid)	1.42 (0.10)	4.443 (0.02)
True value	1.50	5.000

These simulation results suggest that while PQL is biased in an attenuated manner for both the beta coefficients and variance component, the MCEM and MCNR algorithms, which are close to the exact maximum likelihood estimates (allowing for Monte-Carlo error), are less biased and also have smaller standard errors. The SML and hybrid methods do not appear to contribute any improvement over the MCEM and MCNR estimation procedures.

2.4.6 *Sampling performance of the iterative bias correction method and other methods*

Kuk's original paper (1995) described the iterative bias correction method with BLUP used for estimation of starting values. A binomial example given in McGilchrist (1994) was used in Kuk's paper. This example is based on a simulation study involving 200 sets of simulations. The results are presented in Table 2.12 along with McGilchrist's and Jiang's results. The IBC method yielded less biased results than both of McGilchrist's methods for the variance component σ^2, and close to unbiased estimates for the regression coefficients.

Goldstein also carried out some small simulation studies consisting of one hundred datasets in his paper (Goldstein 1995) for the binomial model

$$\text{logit}(p_{ij}) \;=\; \beta_0 + \beta_1 x_{ij} + u_i, \text{ where } u_i \sim \text{ iid } N(0, \sigma^2)$$

while comparing the performance of the iterative bias correction method, available in the advocated computer package MLn, with other currently available methods including first- and second-order PQL and MQL. The true parameter values are all equal to 1. The results are presented in Table 2.9. In this example, the IBC estimates are the least biased with similar standard errors to the other methods. The parameter estimates calculated using second-order PQL were only slightly more biased than those for IBC.

Table 2.9: Average parameter estimates for a binomial data model using MQL, PQL and IBC (Goldstein 1996). Standard errors are given in parentheses.

Parameter	True value	1st-order MQL	1st-order PQL	2nd-order PQL	Kuk's IBC
β_0	1	0.89 (0.03)	0.88 (0.03)	1.07 (0.04)	1.05 (0.04)
β_1	1	0.91 (0.03)	0.88 (0.03)	1.10 (0.04)	1.07 (0.04)
σ^2	1	0.49 (0.03)	0.49 (0.04)	0.93 (0.07)	0.98 (0.06)

Moreno, Sorensen et al. (1997) compared the use of Kuk's iterative bootstrap bias correction

method, in conjunction with using the mode of the joint posterior distribution for the initial estimates, with two other methods in a simulation study. The study was conducted to study frequentist properties of three estimators of the variance component in a binary data model based on animal breeding. They found that the iterative bootstrap bias correction method, in contrast to a method based on Gibbs sampling, leads to unbiased estimates of the variance components in all cases studied as shown in Table 2.10 below, but also found that the iterative bias correction method can lead to estimates that fall outside the parameter space.

Table 2.10: Results for Moreno and Sorensen's (1997) data simulations. (Mean-squared errors are given in parentheses).

No. of fixed levels	No. of random levels	True variance	Average estimated parameter values: Kuk's method (m.s.e)	Average estimated parameter values: Gibbs sampler method (m.s.e)
100	500	0.2	0.204 (0.0051)	0.211 (0.0041)
100	500	0.5	0.512 (0.0074)	0.503 (0.0850)
900	250	0.2	0.177 (0.0065)	0.221 (0.0053)
900	250	0.5	0.479 (0.0128)	0.475 (0.0128)
500	250	0.5	0.496 (0.0757)	0.743 (0.1976)

Recently, Mealli and Rampichini (1999) carried out a simulation study comparing four methods: MQL, indirect inference, IBC and CPQL. Their model was a two-level logit model with a single variance component. Four scenarios were examined, for 1000 subjects, split into either $K = 200$ subjects with 5 observations each, or $K = 50$ subjects with 20 observations each, at two different variance component values, $\sigma^2 = 0.5$ and 1.0. The random effects u_k were distributed as $N(0, \sigma^2)$. One hundred simulations were carried out for each scenario. The model was:

$$\text{logit } \Pr(Y_{ik} = 1 \mid u_k) \quad = \quad \beta_0 + \beta_1 x_{ik} + u_k.$$

The results are presented in Table 2.11.

Table 2.11: Average parameter estimates and Monte-Carlo standard errors in parentheses for Mealli and Rampichini's (1999) simulations.

Cluster sizes	Para-meter	True value	Average estimated parameter values (se's)			
			MQL	Indirect inference	Iterative bootstrap	CPQL
	β_0	1	0.836 (0.14)	0.987 (0.19)	1.012 (0.16)	0.990 (0.16)
50 subjects	β_1	1	0.856 (0.08)	1.002 (0.11)	1.024 (0.11)	1.009 (0.10)
$n = 20$	σ^2	1	0.695 (0.17)	1.049 (0.36)	1.119 (0.32)	0.993 (0.31)
per subject	β_0	1	0.933 (0.13)	1.026 (0.16)	1.027 (0.14)	1.025 (0.14)
	β_1	1	0.932 (0.08)	1.020 (0.10)	1.019 (0.10)	1.002 (0.11)
	σ^2	0.5	0.406 (0.12)	0.520 (0.22)	0.535 (0.20)	0.499 (0.19)
	β_0	1	0.860 (0.09)	0.985 (0.11)	1.020 (0.11)	1.001 (0.11)
200 subjects	β_1	1	0.856 (0.09)	1.007 (0.11)	1.005 (0.11)	0.985 (0.11)
$n = 5$	σ^2	1	0.571 (0.14)	1.024 (0.15)	1.007 (0.32)	0.875 (0.29)
per subject	β_0	1	0.870 (0.07)	0.960 (0.09)	0.962 (0.11)	0.950 (0.10)
	β_1	1	0.888 (0.07)	0.960 (0.09)	0.976 (0.09)	0.957 (0.09)
	σ^2	0.5	0.321 (0.14)	0.520 (0.28)	0.515 (0.28)	0.424 (0.23)

These results suggest that the three methods, indirect inference, iterative bootstrap, and PQL all perform similarly well for the first two scenarios, with the variance components being more biased in estimation by PQL in the second two sets of simulations when there are only 5 observations per each of 200 subjects.

Jiang (1998), in testing his simple method based on simulated moments, carried out some simulation studies using the same data structures as McGilchrist (1994) and Lin and Breslow (1996). These results are presented in Tables 2.12 and 2.13, respectively.

Table 2.12: Results for McGilchrist (1994), Kuk (1995) and Jiang's (1998) binary simulations.

Simulation set	Method	Average estimated parameter values		
		σ^2	β_0	β_1
Set 1:	McGilchrist-2	0.705 (0.439)	0.249 (0.309)	0.100 (0.037)
	McGilchrist-1	0.907 (0.540)	0.253 (0.314)	0.101 (0.038)
	Kuk (1995)	0.992 (0.603)	0.194 (0.263)	0.100 (0.035)
	Jiang (1998)	0.953	0.197	0.101
	True value	1.000	0.200	0.100
Set 2:	McGilchrist-2	1.336 (0.640)	0.109 (0.343)	0.097 (0.048)
	McGilchrist-1	1.681 (0.795)	0.111 (0.350)	0.099 (0.049)
	Kuk (1995)	1.904 (0.955(	0.204 (0.364)	0.100 (0.045)
	Jiang (1998)	1.979	0.180	0.104
	True value	2.000	0.200	0.100

Inspection of Jiang's results for the McGilchrist model suggest that Jiang's method performs better with respect to bias than McGilchrist's methods, but not quite as well as Kuk's new method. For the salamander dataset simulations, Jiang's variance component estimates are seriously biased. However, the regression coefficient estimates are practically unbiased, and performed better than the CPQL methods, but had larger standard errors attached to them. Jiang points out that a motivation for his method is its computational efficiency, and in addition, the improvement seen in the performance of the estimators, considering both bias and standard errors as the sample size increases.

Neuhaus and Lesperance (1996) looked at the estimation problem from a different perspec-

Table 2.13: Results for Lin and Breslow's (1996b) salamander data simulations.

Sample size	Method	Average estimated parameter values					
		σ_1^2	σ_2^2	β_0	β_1	β_2	β_3
$n = 360$	PQL	0.33	0.32	0.94	-2.73	-0.64	3.38
	PQL with σ^2_{CPQL}	0.46	0.46	0.96	-2.78	-0.66	3.44
	First-order CPQL	0.46	0.46	1.17	-3.31	-0.79	4.11
	Second-order CPQL	0.46	0.46	1.01	-2.95	-0.69	3.65
	Jiang (1998)	0.58	0.59	1.07	-3.13	-0.73	3.87
	REML	0.55	0.54	1.09	-3.14	-0.74	3.88
$n = 720$	PQL	0.30	0.29	0.92	-2.66	-0.62	3.29
	PQL with σ^2_{CPQL}	0.43	0.42	0.94	-2.71	-0.63	3.35
	First-order CPQL	0.43	0.42	1.13	-3.21	-0.76	3.97
	Second-order CPQL	0.43	0.42	1.01	-2.94	-0.69	3.64
	Jiang (1998)	0.49	0.49	1.05	-3.05	-0.73	3.79
	REML	0.52	0.51	1.06	-3.06	-0.71	3.79
	True value	0.50	0.50	1.06	-3.05	-0.72	3.77

tive, and investigated the efficiency of likelihood methods using simulation for estimating the regression parameters of mixed effects logistic regression models for three approaches. These are: a conditional likelihood approach, a parametric approach with the random effects distribution assuming a parametric form, and a non-parametric approach where the random effects distribution is left unspecified. Neuhaus and Lesperance simulated 200 datasets of 200 pairs of binary data according to a mixed effects model with a range of values for within-pair correlation of the covariate $\rho_x = \text{corr}(X_{i1}, X_{i2})$ for the ith cluster. The model took the following form:

$$\text{logit } P(Y_{ij} = 1 \mid X_{ij}, u, \beta) = \beta_0 + \beta_1 X_{ij} + u_i, \quad \text{with} \quad \beta = 1.0,$$
$$u_i \sim \text{Normal}(0, 4).$$

They found that the conditional likelihood approach provides more variable estimators of β than the parametric and non-parametric approaches for all values of the correlation coefficient, and the observed efficiency decreased as the correlation increased.

The above analyses by a range of different researchers has shown some investigation into the properties of approximate maximum likelihood estimates with regard to changing the sample size and variance components, though only in a correlated binary data setting. Little has been done on examining the properties of exact maximum likelihood estimates at all. It is clear that the new methods by Kuk and Jiang may hold some potential for providing unbiased estimation. These methods would need to be examined for a wider range of generalized linear mixed model structures.

2.5 Motivating Example

2.5.1 Polio incidence data

In 1988 Zeger reported an analysis of the monthly number of cases of poliomyelitis for the years 1970 - 1983 (as collected by the U.S. Centers for Disease Control). One central idea in modelling this sequence was detection of a possible decreasing trend over time. Zeger used an estimating equation approach analogous to quasi-likelihood to estimate the regression coefficients in a regression model for this time series of count data. His initial analysis has since sparked a great deal of interest in this particular dataset, and since the publication of the 1988 paper, a number of researchers including Wang and Puterman (2001), Davis, Dunsmuir and Ying Wang (2000), Kuk and Cheng (1999), Chan and Ledolter (1995) have also analyzed the data using a variety of modelling techniques.

The data as reported by Zeger (1988) are listed in Appendix A, and are displayed in Figure 2.3. The random effects were modelled assuming a normal autoregressive structure with a

lag of 1, i.e. an AR(1) process $\{u_t\}$ satisfying

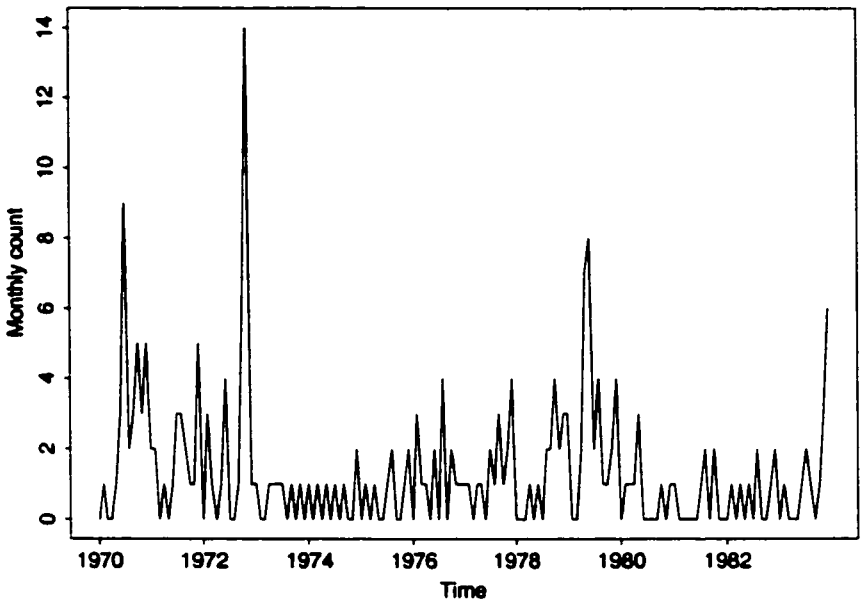

Figure 2.3: Monthly counts of polio in the USA 1970-1983.

$$u_t = \rho u_{t-1} + \epsilon_t \qquad \text{where} \quad \epsilon_t \sim iid \text{ Normal } (0, \sigma_\epsilon^2),$$
$$t = 1, 2, ..., 168,$$

with the correlation coefficient ρ assumed to remain constant over the time period, and the time series to come from a stationary process. In addition, the observations y_t are assumed to be conditionally independent, i.e.,

$$y_t \,|\, u_t \;\sim\; \text{Poisson}(\mu_t),$$

where

$$\log \mu_t \;=\; x_t^t \beta + u_t,$$

with the time trend and seasonality in the data modelled by linear and trigonometric components as follows:

$$x_t = \left(1, \frac{t}{1000}, \cos\left(\frac{2\pi t}{12}\right), \sin\left(\frac{2\pi t}{12}\right), \cos\left(\frac{2\pi t}{6}\right), \sin\left(\frac{2\pi t}{6}\right)\right).$$

It is also assumed that the distribution of the random effects u_t does not depend on x_t. The likelihood function for this data does not have a simple closed form, due to the unobserved random effects and intractable integrals which leads to difficulty in finding the form of the marginal likelihood. The density function for a single response y is

$$\begin{aligned} f(y;\theta) &= \int f(y,\, u;\, \theta)\, du \\ &= \int f(y|\, u;\, \theta)\, h(u;\, \theta)\, du \\ &= \int e^{\{y\, log\mu - \mu - logy!\}} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-u^2}{2\sigma^2}}\, du. \end{aligned}$$

Zeger's modelling of the data involved an estimating equation approach, a time-series analogue of quasi-likelihood as discussed by McCullagh (1983). The estimating equations are shown in Zeger (1988). Both Kuk and Cheng (1999), and Chan and Ledolter (1995) used Monte-Carlo Markov chain methods to fit the above latent process model to the data. Gibbs sampling was used by both groups to simulate random effect vectors $u^{(1)}, ..., u^{(N)}$ (where N is the number of Monte-Carlo simulations used), and a one-step MCEM algorithm was used to get the maximum likelihood parameter estimates. Both groups used an ordinary Poisson generalized linear model fit for the starting value for β and initial values of ρ and σ^2 were 0 and 1.0, respectively.

Wang and Puterman (2001) used a Markov Poisson regression method in which the coefficient vector β depends on the state of an unobserved stationary Markov chain with a finite number of states. Davis, Dunsmuir and Wang (2000) used the polio data to illustrate results from a newly-developed approach to diagnose the existence of a latent stochastic process in the mean of a Poisson regression model. The results from the different approaches for the original polio dataset are displayed in Chapter 6.

At present, procedures available in SAS such as NLMIXED and GLIMMIX are unable to model the specialized correlation structure present in a single time series dataset.

Chapter 3

EXACT MAXIMUM LIKELIHOOD ESTIMATION IN GENERALIZED LINEAR MIXED MODELS

3.1 *Fitting Generalized Linear Mixed Models using Exact Maximum Likelihood Methods*

A number of algorithms have been developed to find exact maximum likelihood estimates for generalized linear mixed models. These include an MCEM algorithm developed by McCulloch (1997), as described in §2.3.5, and also an MCEM algorithm developed by Kuk and Cheng (1997, 1999). Before describing the algorithms in more detail, it will be helpful to establish some notation that will be used throughout the remaining work.

3.1.1 Some notation

Let $f(y_i \mid u; \theta)$ be the conditional density of the observed data y_i given the unobserved random effects, where y_i comes from an exponential family distribution,

$$f(y_i \mid u_i; \theta) \;=\; exp\left\{\frac{y_i\,\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right\},$$

for canonical parameter η_i, given by $\eta_i = x_i^t\beta + u_i$, cumulant function $b(\eta_i)$, dispersion parameter $a(\phi)$ and $\theta = (\beta,\ D,\ \phi)$. Some examples of exponential family distributions commonly used in generalized linear mixed models include the normal, binomial and Poisson distributions with their canonical link functions, as outlined in the following table (McCullagh and Nelder 1989).

Table 3.1: Exponential family distributions commonly used in GLMM's.

Response data y	η	$b(\eta)$	$a(\phi)$	$\mu(\eta)$	$c(y,\phi)$
normal	identity	η^2	σ^2	η	$-\frac{1}{2}(\frac{y^2}{\phi}+\log(2\pi\phi))$
Binomial	$\log(\frac{\mu}{1-\mu})$	$\log(1+e^{\eta})$	$\frac{1}{m}$	$\frac{e^{\eta}}{1+e^{\eta}}$	$\log(\frac{m}{my})$
Poisson	$\log\mu$	$\exp(\eta)$	1	$\exp(\eta)$	$-\log y!$

The marginal likelihood function based on the observed data y can be obtained by integrating out the unobserved random effects u from the joint density function $f(y,u;\theta)$:

$$L(\theta,\, y) \;=\; \int f(y,\, u;\, \theta)\, du = \int f(y|\, u;\, \theta)\, h(u;\, \theta)\, du, \tag{3.1}$$

where $h(u;\theta)$ is the density function of the random effects, and $f(y|\,u;\,\theta) = \prod_{i=1}^{n} f(y_i|\,u;\theta)$. The forms of the log likelihood function and its derivatives, as based on Equation (3.1) above are:

$$l(\theta,y) \;=\; log\, L(\theta;\, y)$$

$$\begin{aligned} l'(\theta;y) &= \text{vector of first derivatives of } l(\theta;y) \text{ with respect to the components of } \theta \\ &= E\{l'(\theta;y,u)|y,\theta\} \quad \text{(Kuk and Cheng 1999)} \end{aligned}$$

$$\begin{aligned} l''(\theta;y) &= \text{matrix of second derivatives of } l(\theta;y) \text{ with respect to the components of } \theta \\ &= E\{l''(\theta;y,u)|y,\theta\} + E\{l'(\theta;y,u)\, l'^{t}(\theta;y,u)|y;\theta\} - l'(\theta;y)\, l'^{t}(\theta;y). \end{aligned}$$

3.1.2 The MCEM algorithm

The general MCEM algorithm (McCulloch 1994, 1997) is a Monte-Carlo implementation of the EM algorithm developed by Dempster, Laird and Rubin in 1977, which is used for finding exact maximum likelihood estimates. It has been successfully used in a wide variety of situations.

The EM algorithm is particularly well-suited to research problems where there is incomplete data (Dempster, Laird and Rubin 1977, Louis 1982) and it is difficult to maximize the observed log likelihood function directly, whereas the log likelihood function based on the complete data can be maximized easily (Kuk and Cheng 1997). The unobserved random effects, u, in generalized linear mixed models can be formulated as the missing data, with the complete data as $W = (y, u)$, so the EM algorithm is especially appropriate for fitting these models.

A variant of the MCEM algorithm was proposed by Lange (1995) which avoids the usual EM iterations within iterations by carrying out a one-step EM procedure. He proved that his algorithm is locally equivalent to the conventional EM algorithm (Kuk and Cheng 1999).

The vector of parameters is defined as $\theta = (\beta,\ D,\ \phi)$, consisting of the regression coefficients, variance components and dispersion parameter, respectively. The general structure of the Monte-Carlo EM algorithm for finding the exact maximum likelihood estimates for generalized linear mixed models is as follows:

Step (1). Choose starting values $\theta^{(0)} = (\beta^{(0)}, D^{(0)},\ \phi^{(0)})$. Set the iteration number $m = 0$.

Step (2). Simulate the random effects vectors $u^{(1)}, ..., u^{(N)}$ from their conditional density $f(u|\ y, \theta^{(m)})$ (described in more detail below).

Step (3). E-step:

Given $\tilde{\theta}^{(m)}$, the current estimate of the parameter vector θ, the conditional expected value of the complete data log-likelihood is obtained as follows:

$$Q(\theta,\ \tilde{\theta}^{(m)}) = E\left[\, l(\theta\ ;\ W)\,|\,y; \tilde{\theta}^{(m)}\right] = \int l(\theta;\ y, u)\ f(u\,|\,y; \tilde{\theta}^{(m)})\ du$$

where $l(\theta, y, u) = \log f(u, y\ ;\ \theta)$.

Step (4). M-step:

In this step, $Q(\theta,\ \tilde{\theta}^{(m)}) = E[l(\theta; W)\,|\,y;\ \theta^{(m)}]$ is maximized as a function of θ to obtain the updated parameter estimate vector $\tilde{\theta}^{(m+1)}$. With the exclusion of the simplest forms of GLMM's, the expectation cannot be computed in closed form due to the high dimensionality of the integrals involved in the above expression. When the expectation cannot be carried out analytically, the following Monte-Carlo approximation can be used:

$$\tilde{Q}(\theta,\ \tilde{\theta}^{(m)}) \ = \ \frac{1}{N}\sum_{j=1}^{N} l(\theta;\ y, u^{(j)}), \tag{3.2}$$

where N is the number of Monte-Carlo simulations carried out and the random effects vectors $u^{(j)}$'s are generated in Step (2). This approximation can be maximized by a single Newton iteration, following the spirit of the EM-gradient algorithm:

EM Gradient (Lange's):

$$\begin{aligned}\tilde{\theta}^{(m+1)} &= \tilde{\theta}^{(m)} - \left(\tilde{Q}''(\theta,\tilde{\theta}^{(m)})|_{\theta=\tilde{\theta}^{(m)}}\right)^{-1} \tilde{Q}'(\theta,\tilde{\theta}^{(m)})|_{\theta=\tilde{\theta}^{(m)}} \\ &= \tilde{\theta}^{(m)} + (\tilde{I}_1)^{-1}\,\tilde{l}'(\tilde{\theta}^{(m)};\ y),\end{aligned}$$

where the following Monte-Carlo approximations are used:

$$\tilde{Q}'(\theta,\tilde{\theta}^{(m)}) \equiv \tilde{l}'(\tilde{\theta}^{(m)};\ y) = \frac{1}{N}\sum_{j=1}^{N} l'(\tilde{\theta}^{(m)};\ y, u^{(j)}) \ \text{ and}$$

$$\tilde{Q}''(\tilde{\theta}^{(m)},\theta) = \ -\tilde{I}_1 = \ \frac{1}{N}\sum_{j=1}^{N} \tilde{l}''(\tilde{\theta}^{(m)}; y, u^{(j)}).$$

It is also useful to define $\tilde{I}_2$ here for use in the next section:

$$\tilde{I}_2 \ = \ \frac{1}{N}\sum_{j=1}^{N} l'(\tilde{\theta}^{(m)}; y, u^{(j)})\ l'^{t}(\tilde{\theta}^{(m)}; y, u^{(j)}) - \tilde{l}'(\tilde{\theta}^{(m)}; y)\,\tilde{l}'^{t}(\tilde{\theta}^{(m)}; y)$$

Set $m = m + 1$.

Step (5). Compare the (old) $\tilde{\theta}^{(m)}$ with the (new) $\tilde{\theta}^{(m+1)}$. If convergence is achieved, $\tilde{\theta}^{(m+1)}$ is declared the maximum likelihood estimator; otherwise return to step 2.

McCulloch (1997) shows how the log-likelihood function of the complete data $f(u, y; \theta)$ can be split into two parts: the first part involving only the regression coefficients, $f(y|u, \beta, \phi)$, and the second involving only the variance component(s), $f(u|D)$. This can make the maximization problem simpler in Step (4) (the M-step), especially for more complex random effect structures. The EM Gradient equation for $\tilde{\theta}^{(m+1)}$ as displayed in Step (4) can be used to maximize the log likelihood with respect to β. Closed form maximum likelihood equations can be used for estimating the variance components, or a Newton-Raphson algorithm may be used when closed form solutions do not exist. For the seed data, for example, where $u_i \sim \text{Normal}(0, D)$,

$$\tilde{D}^{(m+1)} = \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{i=1}^{n}\frac{u_i^{2(j)}}{n}\right).$$

3.1.3 The MCNR algorithm

The general Monte-Carlo Newton-Raphson method (MCNR) (Burden and Faires 1985, Tanner 1991, Penttinen 1984), which is a Monte-Carlo implementation of the Newton-Raphson procedure, is considered to be a viable alternative for fitting generalized linear mixed models because it has a faster rate of convergence than the MCEM algorithm (Kuk and Cheng 1999). However, the MCNR algorithm can also behave erratically and is less stable generally due to the use of the Monte Carlo gradient vector and Hessian matrix $-\tilde{l}''(\theta, y) = \tilde{I}_1 - \tilde{I}_2$ which is sometimes not positive definite. Kuk and Cheng (1997) suggested a half-stepping procedure to help provide stability in fitting generalized linear mixed models, which they found to perform quite well.

The basic structure of the Monte-Carlo NR algorithm for finding the exact maximum likelihood estimates for generalized linear mixed models is as follows. It is very similar to the EM procedure above:

Step (1). Choose starting values $\theta^{(0)} = (\beta^{(0)}, D^{(0)}, \phi^{(0)})$. Set the iteration number $m = 0$.

Step (2). Simulate the random effects $u^{(1)}, ..., u^{(N)}$ from their conditional density $f(u|y, \tilde{\theta}^{(m)})$.

Step (3). The analytical form of a Newton-Raphson iteration is given by:

$$\textbf{Newton–Raphson:} \quad \tilde{\theta}^{(m+1)} = \tilde{\theta}^{(m)} - \{l''(\tilde{\theta}^{(m)}; y)\}^{-1} l'(\tilde{\theta}^{(m)}; y)$$

The Monte-Carlo Newton-Raphson procedure is an approximation of the analytical form of the Newton–Raphson procedure that is used when we cannot evaluate analytically the conditional expectations $l''(\theta^{(m)}; y)$ and $l'(\theta^{(m)}; y)$ in the above expression. The following Monte-Carlo approximation can be used:

$$\{\tilde{l}''(\theta^{(m)}; y)\}^{-1} = \tilde{I}_1 - \tilde{I}_2$$

leading to the following Monte-Carlo form of the Newton-Raphson iteration:

$$\textbf{Newton-Raphson:} \quad \tilde{\theta}^{(m+1)} = \tilde{\theta}^{(m)} + (\tilde{I}_1 - \tilde{I}_2)^{-1} \tilde{l}'(\tilde{\theta}^{(m)}; y)$$

Set $m = m + 1$.

Step (4). Compare the (old) $\tilde{\theta}^{(m)}$ with the (new) $\tilde{\theta}^{(m+1)}$. If convergence is achieved, $\tilde{\theta}^{(m+1)}$ is declared the maximum likelihood estimator; otherwise return to step 2.

As can be seen, the MCEM and MCNR algorithms actually lead to very similar procedures for finding the maximum likelihood estimates in a generalized linear mixed model. A comprehensive outline of the relationships between the various Monte-Carlo techniques was given by Kuk and Cheng (1999), and is presented in Figure 3.1.

The iterative Monte-Carlo likelihood approach was proposed by Geyer and Thompson (1992). The aim behind this approach is to approximate the whole likelihood function, using a ratio of the likelihood $L(\theta; y)$ to $L(\theta_o; y)$, relative to a prechosen point θ_o. Then simulating random effects vectors $u^{(1)}, ..., u^{(N)}$, the likelihood ratio can be estimated using

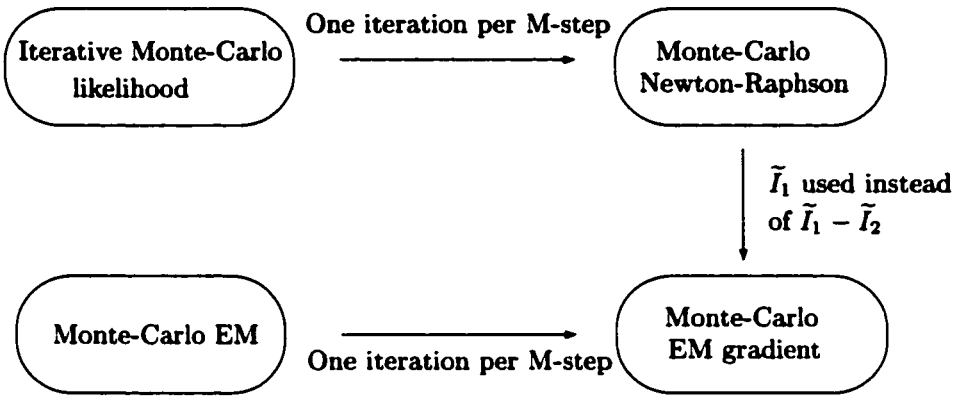

Figure 3.1: Relationships between the Monte-Carlo methods.

Monte-Carlo methods.

3.1.4 Simulation of the random effects

Because it is not possible to directly sample the random effects from their conditional distribution $f(u|y)$ in step 2 at any particular iteration in the above algorithm, the random effects $u^{(1)}, ..., u^{(N)}$ (where N is the number of Monte-Carlo simulations to be performed), are obtained through simulation using Markov chain methods. McCulloch (1994) was the first author to use Markov Chain Monte-Carlo methods in the form of Gibbs sampling to compute maximum likelihood estimates. McCulloch then generalized his 1997 algorithm and used a Metropolis-Hastings algorithm to sample the random effects (McCulloch 1997).

Since then, other authors (Kuk and Cheng 1997, 1999, Booth and Hobert 1999) have suggested alternative Monte-Carlo methods. Thus, the current choices are the Metropolis-Hastings algorithm, importance sampling techniques, rejection sampling, and Gibbs sampling. Only the Metropolis-Hastings algorithm has been implemented for use in the current study, as it takes a simple form as described by McCulloch (1997). This algorithm is described in more detail below.

3.1.5 The Metropolis-Hastings algorithm

The simple algorithm called the Metropolis-Hastings algorithm is a Markov Chain Monte-Carlo method, first proposed by Metropolis et al (1953), and then generalized by Hastings (1970). It can be used in the context of fitting generalized linear mixed models to obtain dependent samples of the random effects u from their (target) conditional distribution $f(u|y)$ which is difficult or impossible to sample from directly.

The dependence of the samples comes from the structure of Markov chains in that when a sequence of random variables $X_0, X_1, X_2, \ldots$ is generated, the next state X_{t+1} which is sampled from $Pr(X_{t+1}|\, X_t)$, has a dependence on the previous observation X_t. An assumption is that the probabilities of moving from observation to observation in the chain remains the same over time, i.e., $Pr(X_{t+1}|\, X_t)$ does not depend on t.

In a general setting, to obtain a single (dependent) sample, $X_0, X_1, \ldots,$ the Metropolis-Hastings algorithm has a simple structure as follows:

Step 1. Initialize starting values X_0 and $t = 0$.

Step 2. Iterate:

Sample a candidate value X^* from the proposal distribution $q(X^* \mid X)$, which may or may not depend on the current value $X_t = X$.

Calculate the acceptance probability α to determine whether to accept X^* as the next value in the Markov chain, where π is the target distribution:

$$\alpha \;=\; \min\left(1,\; \frac{\pi(X^*)\, q(X \mid X^*)}{\pi(X)\, q(X^* \mid X)}\right)$$

The next state X_{t+1} becomes

$$X_{t+1} = \begin{cases} X^* \text{ (candidate value)} & \text{with probability } \alpha \\ X \text{ (current value)} & \text{with probability } 1-\alpha. \end{cases}$$

Repeat loop for desired sample size.

The beauty of this algorithm is that the proposal distribution $q(X^*|\ X)$ is arbitrary. However, some forms will lead to more efficient sampling strategies than others (Hastings 1970). Some recommended guidelines (Gelman, Carlin et al. 1995) include selecting a distribution that is easy to sample from, choosing candidate values that are not rejected too frequently, and ensuring that the distance moved is a reasonable distance in the parameter space (otherwise the Markov chain moves around the entire distribution too slowly).

Applying this algorithm to generating simulated samples of random effects vectors, u, is a little more involved, depending on the complexity of the random effects structure. For the generalized linear mixed model structure in the example in §2.3 where only a random intercept u_i is modelled, a random effects vector is generated at each step of the Markov-chain. The random effects vector u is $(u_1, u_2, ..., u_n)$, where n is the number of subjects or clusters. For example, if we wish to generate a single random effect for each of n subjects, there will be N random effects vectors (one for each Monte-Carlo simulation), where $u_i^{(j)}$ is the random effect for the ith subject at the jth Monte-Carlo simulation:

Monte-Carlo Simulation:

$$\begin{array}{cccccc} 1 & 2 & 3 & \ldots & N \\ \begin{pmatrix} u_1^{(1)} \\ u_2^{(1)} \\ u_3^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ u_n^{(1)} \end{pmatrix} & \begin{pmatrix} u_1^{(2)} \\ u_2^{(2)} \\ u_3^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ u_n^{(2)} \end{pmatrix} & \begin{pmatrix} u_1^{(3)} \\ u_2^{(3)} \\ u_3^{(3)} \\ \cdot \\ \cdot \\ \cdot \\ u_n^{(3)} \end{pmatrix} & \begin{matrix} \cdot\ \cdot \\ \cdot\ \cdot \\ \cdot\ \cdot \\ \cdot\ \cdot \\ \cdot\ \cdot \\ \cdot\ \cdot \end{matrix} & \begin{pmatrix} u_1^{(N)} \\ u_2^{(N)} \\ u_3^{(N)} \\ \cdot \\ \cdot \\ \cdot \\ u_n^{(N)} \end{pmatrix} \end{array}$$

Hastings (1970) suggests three different methods for using this algorithm when the target distribution is multi-dimensional, as is the situation when we have n random effects to sample at the jth Monte-Carlo iteration:

1. Choose n candidate values u_i^* and compare the new values to the original u_i values

$i = 1, ..., n$ all at once, i.e. $(u_1^*, u_2^*, ..., u_n^*)$ versus $(u_1, u_2, ..., u_n)$.

2. Choose one of the n elements at random, say the ith element, and choose a candidate value only for this element, keeping the remaining elements the same, so that we are checking only one new u_i^* at a time, i.e. $(u_1, u_2, .., u_i^*, ..., u_n)$ versus $(u_1, u_2, .., u_i, ..., u_n)$.

3. Sequentially choose the $1, ..., n$ values of the vector one at a time to test a new candidate value where the preceding values may or may not be changed to their new value depending on whether they were accepted or not.

$$\begin{pmatrix} u_1^* \\ u_2 \\ \cdot \\ u_i \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{pmatrix} \text{ vs } \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ u_i \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{pmatrix} \text{ and then } \begin{pmatrix} u_1 \\ u_2^* \\ \cdot \\ u_i \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{pmatrix} \text{ vs } \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ u_i \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{pmatrix} \text{ etc.}$$

The third method has been used in the current work as it appears to work more efficiently than the first method and is comparable to the second method. The acceptance probability was neatly formulated by McCulloch (1997) for the generalized linear mixed model scenario as follows:

Aim: To sample a random effects vector $u = (u_1, ..., u_n) \sim f(u|\, y)$ for one Monte-Carlo simulation, where

- target distribution π is $f(u|\, y)$
- proposal distribution is $q(u^*|\, u)$.

Algorithm:

(a) Initialize $u = (u_1, u_2, ..., u_n) = (0, 0, ..., 0)$

(b) Randomly generate $u^* \sim q(u^*|u)$ so that

$$u^* = (u_1, u_2, .., u_{i-1}, u_j^*, u_{i+1}, ..., u_n) \quad \text{and}$$

$$u = (u_1, u_2, .., u_{i-1}, u_i, u_{i+1}, ..., u_n)$$

(c) Accept u^* with probability α, where

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{f(u^*\,|\,y)\,q(u\,|\,u^*)}{f(u\,|\,y)\;\;q(u^*\,|\,u)}\right\} \\
&= \min\left\{1, \frac{f(y|\,u^*)\,f(u^*)\,q(u\,|\,u^*)}{f(y\,|\,u)\;\;f(u)\;\;\;q(u^*\,|\,u)}\right\} \\
&= \min\left\{1, \frac{f(y\,|\,u^*)}{f(y\,|\,u)}\right\} \\
&= \min\left\{1, \exp\left(\sum_{i=1}^{n} y_i((\eta_i^* - \eta_i) - (b(\eta_i^*) - b(\eta_i))\right)\right\}
\end{aligned}$$

where the cancellation of the candidate distributions on the numerator and denominator occur since $q(u\,|\,u^*) = f(u)$ and $q(u^*|\,u) = f(u^*)$.

The value $\tilde{u}_i$ becomes

$$\tilde{u}_i = \begin{cases} u_j^* \text{ (candidate value)} & \text{with probability } \alpha \\ u_i \text{ (current value)} & \text{with probability } 1-\alpha. \end{cases}$$

Repeat (b) and (c) sequentially for each element of the vector u.

3.1.6 Estimation of the parameter standard errors

The observed information matrix $\mathcal{I}$ provides estimates of the standard errors for both the regression coefficients and the variance components at the final MCEM iteration, after the

parameters have converged. If there are regression coefficients $\beta_0, ..., \beta_p$ and two variance components ρ and σ^2, as for the polio incidence data, then the observed information matrix is $\{\tilde{l}''(\hat{\theta};\, y)\}^{-1} = \tilde{I}_1 - \tilde{I}_2$, where

$$\tilde{I}_1 = \begin{pmatrix} \frac{\partial^2 l(\theta;y,u)}{\partial\beta_0\partial\beta_0} & \frac{\partial^2 l(\theta;y,u)}{\partial\beta_0\partial\beta_1} & \cdots & & 0 & 0 \\ \frac{\partial^2 l(\theta;y,u)}{\partial\beta_0\partial\beta_1} & \frac{\partial^2 l(\theta;y,u)}{\partial\beta_1\partial\beta_1} & & \cdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & & \\ & & \cdots & \frac{\partial^2 l(\theta;y,u)}{\partial\beta_p\partial\beta_p} & 0 & 0 \\ 0 & \cdots & & 0 & \frac{\partial^2 l(\theta;y,u)}{\partial\rho\partial\rho} & \frac{\partial^2 l(\theta;y,u)}{\partial\rho\partial\sigma^2} \\ 0 & \cdots & & 0 & \frac{\partial^2 l(\theta;y,u)}{\partial\sigma^2\partial\rho} & \frac{\partial^2 l(\theta;y,u)}{\partial\sigma^2\partial\sigma^2} \end{pmatrix}.$$

The complete-data log likelihood function can be separated into two components, one for the regression coefficients and the other for the variance components. The first and second derivatives of the complete-data log likelihood function with respect to the regression coefficients can easily be derived as follows.

The first derivatives are

$$\frac{\partial \log \mathrm{L}(\theta;\, y,\, u)}{\partial\beta_j} = x_j^t(y-\mu)$$

and the second derivatives are

$$\frac{\partial^2 \log \mathrm{L}(\theta;\, y,\, u)}{\partial\beta_i\partial\beta_j} = \left[-X^t X e^{X\beta+u}\right]_{ij}.$$

The second derivative(s) for the variance component(s) will depend on their distributional form and assumed correlation structure, but can be calculated from their likelihood $h(u \mid \theta)$.

3.2 The Performance Properties of Maximum Likelihood

3.2.1 Measurement of performance criteria

Simulation studies were carried out to investigate the performance of exact maximum likelihood estimators. Of particular interest here are the following issues:

- While it is presumed that maximum likelihood estimates are asymptotically unbiased, the amount of bias present in the parameter estimates in small to medium-sized samples has not been well-established. Simulation studies will be used to examine the effect of sample size on the levels of bias present in maximum likelihood estimates for sample sizes of $n = 50, 100$ and 250.

- The effects of the values of the correlation coefficient ρ and variance σ_u^2 of the random effects on the resulting bias and mean-squared error in the maximum likelihood estimates will be investigated. For example, for a fixed σ_u^2, does the value of ρ impact the level of bias? The values of ρ investigated were 0, 0.25, 0.5, 0.75, and for σ_ϵ^2, 0, 0.5, 1.0, and 2.5.

3.2.2 Simulation studies setup

The polio incidence data example forms the basis for the generalized linear mixed model under investigation, i.e.,

$$\log \mu_t = x_t^t \beta + u_t.$$

The random effects are modelled assuming a normal autoregressive structure with a lag of one, i.e., an AR(1) process $\{u_t\}$ satisfying

$$u_t = \rho u_{t-1} + \epsilon_t \qquad \text{where} \quad \rho \text{ is the correlation coefficient,}$$
$$\epsilon_t \sim iid \text{ Normal } (0, \sigma_\epsilon^2),$$
$$t = 1, 2, ..., n,$$

and the y_t's are conditionally independent, i.e., $y_t \mid u_t \sim \text{Poisson}(\mu_t^u)$.

The trend and seasonality in the data are modelled with linear and trigonometric components:

$$x_t^t = \left(1, \frac{t}{1000}, \cos\left(\frac{2\pi t}{12}\right), \sin\left(\frac{2\pi t}{12}\right), \cos\left(\frac{2\pi t}{6}\right), \sin\left(\frac{2\pi t}{6}\right)\right).$$

For each scenario (each combination of sample size n, correlation coefficient ρ and variance σ_u^2 as presented in Table 3.2), two hundred simulated datasets were fitted using McCulloch's MCEM algorithm as shown in Figure 3.2.

The datasets were generated from the population model based on the true parameter values by first generating a set of random effects u_t from MVN(0,D) where D has an AR(1) correlation structure. The covariates used were the same as Zeger (1988) used in the true polio dataset as described in §5.5.1. A log link function was used, thus $\mu_t^u = \exp(\eta_t^u) = \exp(x_t^t\beta + u_t)$, and a dataset based on each μ^u was then randomly drawn as $y_{sim} \sim \text{Poisson}(\mu_t^u)$.

Table 3.2: Parameter values used in simulation studies.

Parameter	Range of values examined
n	50, 100, 250, 500
ρ	0, 0.25, 0.5, 0.75
σ_ϵ^2	0, 0.5, 1.0, 2.5
β	(1, 0, 0, 0, 0, 0)

Within the Metropolis-Hastings algorithm, the proposal distribution $q(u|\, u^*)$ is chosen to

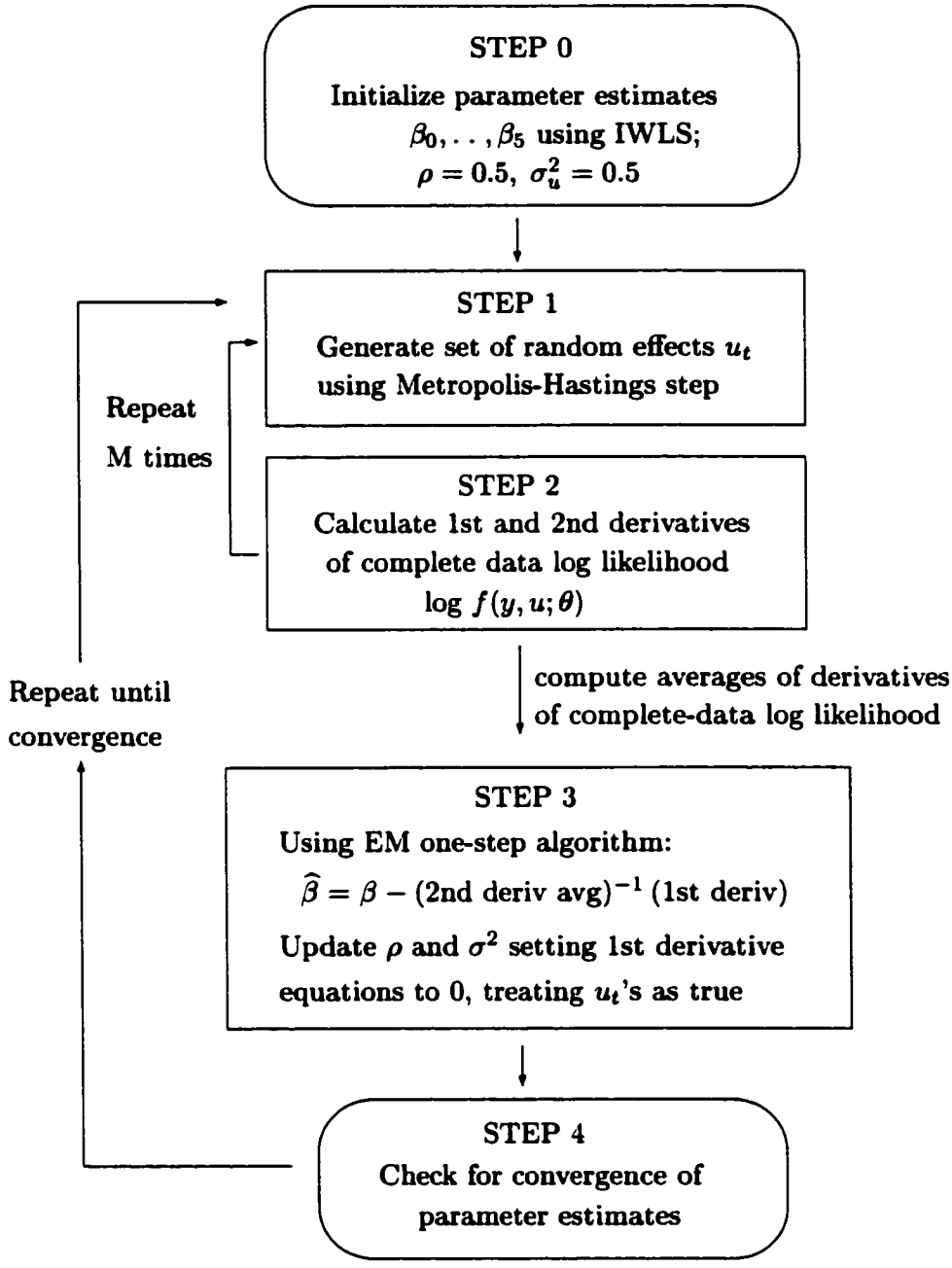

Figure 3.2: The MCEM algorithm for the polio incidence data.

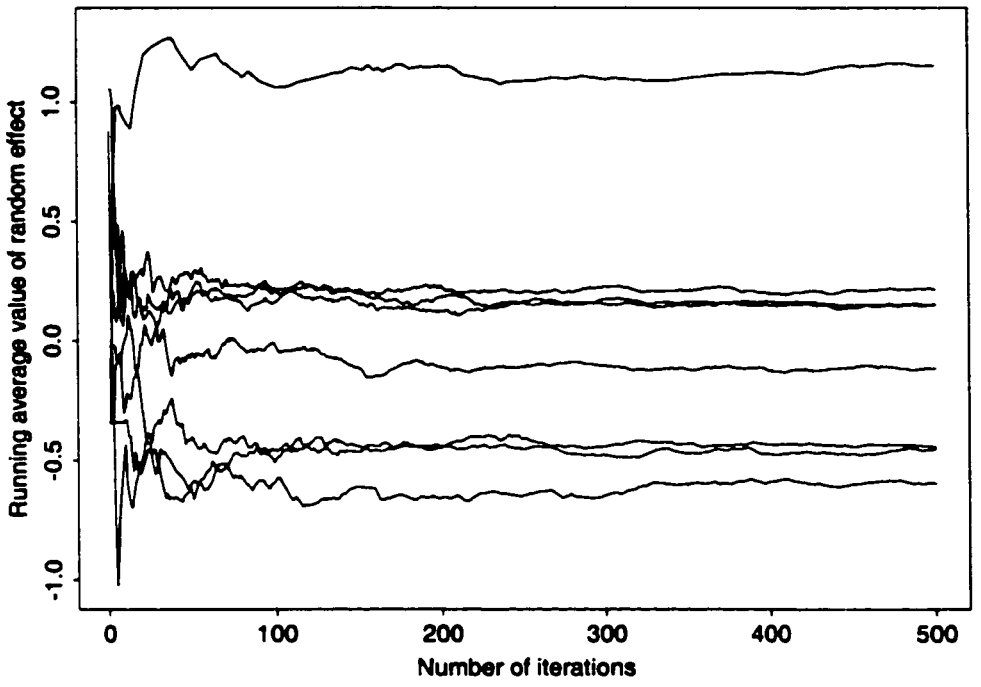

Figure 3.3: Simulation of random effects using Metropolis-Hastings algorithm.

be Normal$(0, D)$, thus the form of the acceptance probability α used is (as shown in §3.5.1):

$$\alpha = \min\left\{1,\ exp\left(\sum_{i=1}^{n} y_i((\eta_i^* - \eta_i) - (b(\eta_i^*) - b(\eta_i))\right)\right\}$$

Shown in Figure 3.3 is a sample of the simulated random effects using the Metropolis-Hastings algorithm.

The number of Monte-Carlo iterations carried out at each step varied from iteration to iteration. As pointed out by Booth and Hobert, a smaller number of Monte-Carlo iterations can be carried out in the early iterations to increase computational efficiency, since a high level of accuracy is not required until nearer to convergence. For the simulations carried out here, the numbers of Monte-Carlo iterations presented in Table 3.3 were found to work efficiently.

Table 3.3: Number of Monte-Carlo iterations used in simulation studies.

MCEM iteration number	Number of Monte-Carlo iterations
$1 \rightarrow 5$	1500
$6 \rightarrow 10$	5000
$11 \rightarrow 20$	10000
$21 \rightarrow$ onwards	70000

If the calculated information matrix was found not to be positive definite, then an additional 10000 Monte-Carlo iterations were added at each MCEM iteration until the information matrix became positive definite (necessary for calculation of the corresponding standard errors of the estimates). This was not required in the early iterations, but became necessary as the parameter estimates got closer to convergence.

Starting values used for $\beta = (\beta_0, \beta_1, ..., \beta_5)$ were found using IWLS, an algorithm commonly used for estimating the regression coefficients in a generalized linear model. These values are commonly used by other researchers and provide starting values that are in a range close to the final estimated maximum likelihood values. However, it should be noted that a range of starting values were tried out, and it was found that the values used made little difference in the efficiency of the algorithm, as it would tend to take larger steps to start with if the starting values used were a lot further away from the maximum likelihood estimates. Starting values for ρ and σ^2 were set at 0.5.

Strict convergence criteria were used within the MCEM algorithm to achieve a high level of accuracy of parameter estimates θ. For the first thirty iterations, the convergence criterion was:

$$\Delta = \frac{|\hat{\theta}_{new} - \hat{\theta}_{old}|}{|\hat{\theta}_{old}|} < 0.0001.$$

Since the standard error of an estimate describes to a large degree how accurate an estimate

one will be able to get from the method used to fit the data, the convergence criterion was changed to:

$$\Delta = \frac{|\hat{\theta}_{new} - \hat{\theta}_{old}|}{\text{se}(\hat{\theta}_{old})} < 0.0001$$

at the thirtieth iteration. If $\Delta < 0.0001$ within the first thirty iterations, or less than 0.01 for iterations after that, then $\hat{\theta}$ was considered to have converged.

3.2.3 Estimation of the variance components

As mentioned above in §3.6.1 , the second derivative equations for the variance components for the polio incidence data are found using the joint density function $f(u;\, \rho, \sigma^2)$ as below:

$$f(u\,;\, \rho, \sigma_\epsilon^2) \;=\; \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{n}{2}}}\,(1-\rho^2)^{\frac{1}{2}}\, e^{-\frac{1}{2\sigma_\epsilon^2}[(1-\rho^2)u_1^2 \,+\, \sum_{t=2}^{n}(u_t - \rho u_{t-1})^2]}.$$

The log density function is then:

$$\log f(u\,;\, \rho, \sigma_\epsilon^2) \;=\; \frac{-n}{2}\log(2\pi\sigma_\epsilon^2) \,+\, \frac{1}{2}\log(1-\rho^2) \,-\, \frac{1}{2\sigma_\epsilon^2}[(1-\rho^2)u_1^2 + \sum_{t=2}^{n}(u_t - \rho u_{t-1})^2].$$

The first derivatives of $\log f(u; \rho, \sigma_\epsilon^2)$ with respect to each of the variance components is:

$$\frac{\partial \log f(u;\, \rho, \sigma_\epsilon^2)}{\partial \rho} \;=\; -\frac{\rho}{1-\rho^2} + \frac{\rho u_1^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_\epsilon^2}\sum_{t=2}^{n} u_t u_{t-1} - \frac{1}{\sigma_\epsilon^2}\sum_{t=2}^{n}\rho u_{t-1}, \qquad (3.3)$$

$$\frac{\partial \log f(u;\, \rho, \sigma_\epsilon^2)}{\partial \sigma_\epsilon^2} \;=\; \frac{-n}{2\sigma_\epsilon^2} + \frac{(1-\rho^2)u_1^2}{2(\sigma_\epsilon^2)^2} + \frac{1}{2(\sigma_\epsilon^2)^2}\sum_{t=2}^{n}(u_t - \rho u_{t-1})^2 \qquad (3.4)$$

The second derivatives for the variance components take the following form:

$$\frac{\partial^2 \log f(\rho, \sigma_\epsilon^2 ; y)}{\partial \sigma_\epsilon^2 \sigma_\epsilon^2} = \frac{n}{2(\sigma_\epsilon^2)^2} - \frac{u_1^2(\rho^2)}{(\sigma_\epsilon^2)^3} - \frac{1}{2(\sigma_\epsilon^2)^2} \sum_{t=2}^{n} (u_t - \rho u_{t-1})^2,$$

$$\frac{\partial^2 \log f(\rho, \sigma_\epsilon^2 ; y)}{\partial \rho \partial \sigma_\epsilon^2} = \frac{-u_1^2 \rho}{(\sigma_\epsilon^2)^2} + \frac{\rho}{(\sigma_\epsilon^2)^2} \sum_{t=2}^{n} u_{t-1}^2 - \frac{1}{(\sigma_\epsilon^2)^2} \sum_{t=2}^{n} (u_t - \rho u_{t-1}),$$

$$\frac{\partial^2 \log f(\rho, \sigma_\epsilon^2 ; y)}{\partial \rho \partial \rho} = \frac{-1}{(1-\rho^2)} - \frac{2\rho^2}{(1-\rho^2)^2} + \frac{u_1^2}{\sigma_\epsilon^2} - \frac{1}{(\sigma_\epsilon^2)^2} \sum_{t=2}^{n} u_{t-1}^2.$$

3.2.4 Least squares estimates of the variance components

To solve the equations for ρ and σ_ϵ^2 in Step 3 of the EM algorithm, closed form approximate solutions were found using least-squares (Box and Jenkins 1994). As Box and Jenkins (1994) point out, the term $\frac{1}{2}\log(1-\rho^2)$ is relatively small for moderate to large samples, so that in the calculation of the least squares estimates, it can be omitted.

Using Equation (3.2) in Step 3 of the MCEM algorithm (the M-step) for the variance components involved averaging the log-likelihood and then solving for ρ and σ_ϵ^2 by setting equations (3.4) and (3.5) to zero, and omitting the term $\frac{1}{2}\log(1-\rho^2)$. This led to the following estimates of ρ and σ_ϵ^2:

$$\tilde{\rho} = \frac{\sum_{j=1}^{N} \sum_{i=1}^{n-1} u_i^{(j)} u_{i+1}^{(j)}}{\sum_{j=1}^{N} \sum_{i=2}^{n-1} u_i^{2(j)}},$$

$$\tilde{\sigma}_\epsilon^2 = \frac{1}{Nn} \sum_{j=1}^{N} \left[\sum_{i=1}^{n-1} (u_{i+1}^{(j)} - \hat{\rho} u_i^{2(j)} + (1-\hat{\rho}^2) u_1^{2(j)} \right],$$

where N is the number of Monte-Carlo iterations carried out at each MCEM step, and n is the sample size of the dataset.

3.2.5 Results

Estimation of regression coefficients using exact maximum likelihood methods proved to be in accordance with the theoretical and empirical results found by many researchers before in non-GLMM modelling situations. Overall, there was little or no bias present in the estimates of the regression coefficients, while the estimation of the variance components, in particular, ρ produced interesting results as discussed in more detail below. Tables 3.4 – 3.5 provide details on the average estimates for ρ, σ^2, β_0, β_1 and β_4. The other regression coefficients β_2, β_3 and β_5 were omitted as their results were very comparable to those for β_4. The associated standard errors were calculated in two ways: the standard deviation of the parameter estimates from the two hundred simulations for any particular variance component ("theoretical"), and the average of the standard errors for the two hundred simulations ("observed"). Mean-squared errors (m.s.e's) for the same parameter estimates in each regression scenario are presented in Tables 3.8 – 3.9.

Convergence issues

An attractive feature of maximum likelihood is its presumed consistent estimation of regression coefficients and variance components. However, a major disadvantage in its use here, where there are no closed form solutions and some type of iterative algorithm such as EM has to be used in fitting the data, is the extreme computational intensity required in fitting even one dataset. The time required to fit a dataset using the MCEM algorithm to a satisfactory convergence is lengthy, requiring many hours, perhaps days, even using an efficient computing programming language such as C and a fast computer. The computational time required to obtain a satisfactory convergence is approximately exponential, mainly due to the extra computing time required in the Metropolis-Hastings step in the simulation of the random effects. A small dataset, for example of sample size $n = 50$, usually converges satisfactorily within an hour or so. The time to convergence also increased as the values of the true variance component ρ increased, as these datasets displayed less information in the data due to a higher correlation between consecutive observations. Regarding the number

of MCEM iterations, many of the datasets for $n = 50$ satisfactorily converged in $30 - 40$ iterations, while many datasets of a sample size of $n = 250$ required over 100 iterations, even perhaps as many as 300 iterations to satisfactorily converge. The number of iterations required for convergence was dependent to a large degree on the values of the variance componenents: as ρ got larger, more iterations were required to reach satisfactory convergence, and the number of Monte-Carlo iterations within each MCEM step also was required to be larger.

None of the estimates of σ_u^2 were less than 0.001 when the true value of σ_u^2 was zero for any of the sample sizes tested, although if these simulations had been allowed to run many more iterations, they may have reached this lower level. At a sample size of $n = 250$, 99% of the datasets estimated σ_u^2 at less than 0.1.

Due to the slowness of convergence in many of the sets of simulations, especially where $\rho = 0.75$ and $\sigma_\epsilon^2 = 1$ or 2.5, the number of MCEM steps for fitting any particular dataset was restricted to a maximum of sixty. Consequently, satisfactory convergence was reached in only 70-80% of the datasets when $n = 50$. Due to a high rate of non-convergence in the largest sample size investigated, where $\rho = 0.75$ and $\sigma_u^2 = 5.714$, these results are not reported in the tables.

Other researchers have suggested faster algorithms such as the Newton-Raphson algorithm to increase the rate of convergence, or the use of importance/rejection sampling techniques for generation of the random effects. Their methods have certain disadvantages however, mainly a lack of stability (Kuk and Cheng 1999). A Newton-Raphson algorithm was attempted in the current study and found to be unstable. Because this study required fitting a large number of datasets, the stability of the Expectation-Maximization algorithm was advantageous, if somewhat slow.

The regression coefficient β_0

The intercept, β_0, was estimated on average over the 200 simulations with very little bias (bias = average estimated value - true value), for all combinations of the variance components ρ and σ_u^2. Even for smaller sample sizes such as $n = 50$ the bias was minimal. The variability of the two hundred β_0 estimates for each set of simulations was low, especially when the variance components were small. This variability increased with increasing ρ and σ^2. Figure 3.4 displays an example of this increasing variability in the estimated value of β_0 where ρ is kept fixed at 0.25 and σ_ϵ^2 increases from 0 to 2.5. A similar pattern was also seen for the other values of ρ.

The regression coefficient β_1

This regression coefficient, β_1 estimated the time trend present in the data, and was estimated with little bias, but with large variability present in the two hundred estimates, with some larger values often leading to a slightly biased mean value, although boxplots displayed little bias. The large variability was especially noticeable in the smallest sample size tested, $n = 50$, and decreased substantially as the sample size increased to $n = 250$. Larger variability was also observed for increasing values of σ_u^2, as displayed in Figure 3.5.

The other regression coefficients

In a similar manner to β_0 the estimates of the remaining regression coefficients, β_2 to β_5 also exhibited very little bias (all less than 5%), even for the smallest sample size examined, $n = 50$, and for all combinations of the variance components. Any small biases present were frequently negative. As the variance components became larger, the variability seen in the two hundred estimates also increased in a similar manner to β_0.

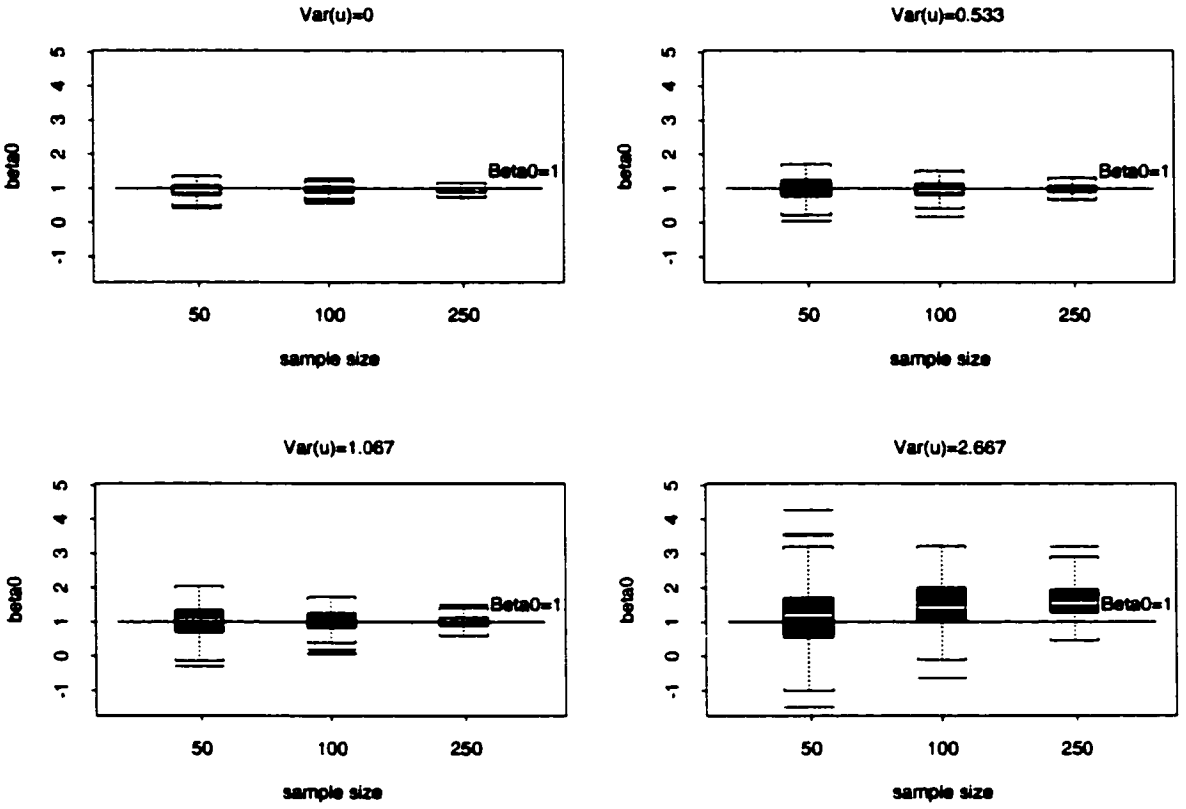

Figure 3.4: Estimation of β_0 for varying σ_u^2 and fixed $\rho = 0.25$.

The variance components

The variance component σ^2 was generally estimated with no or little bias. However, ρ displayed significant negative bias in certain circumstances as described below.

When the true value of the variance was zero, σ_u^2 was slightly overestimated on average, (bias less than 0.07) due to the restrictive lower bound of 0 imposed on its estimation. For other combinations of the variance components, σ^2 tended to be underestimated for the smallest sample size of 50 (at worst, the bias was close to -30%), but showed significant

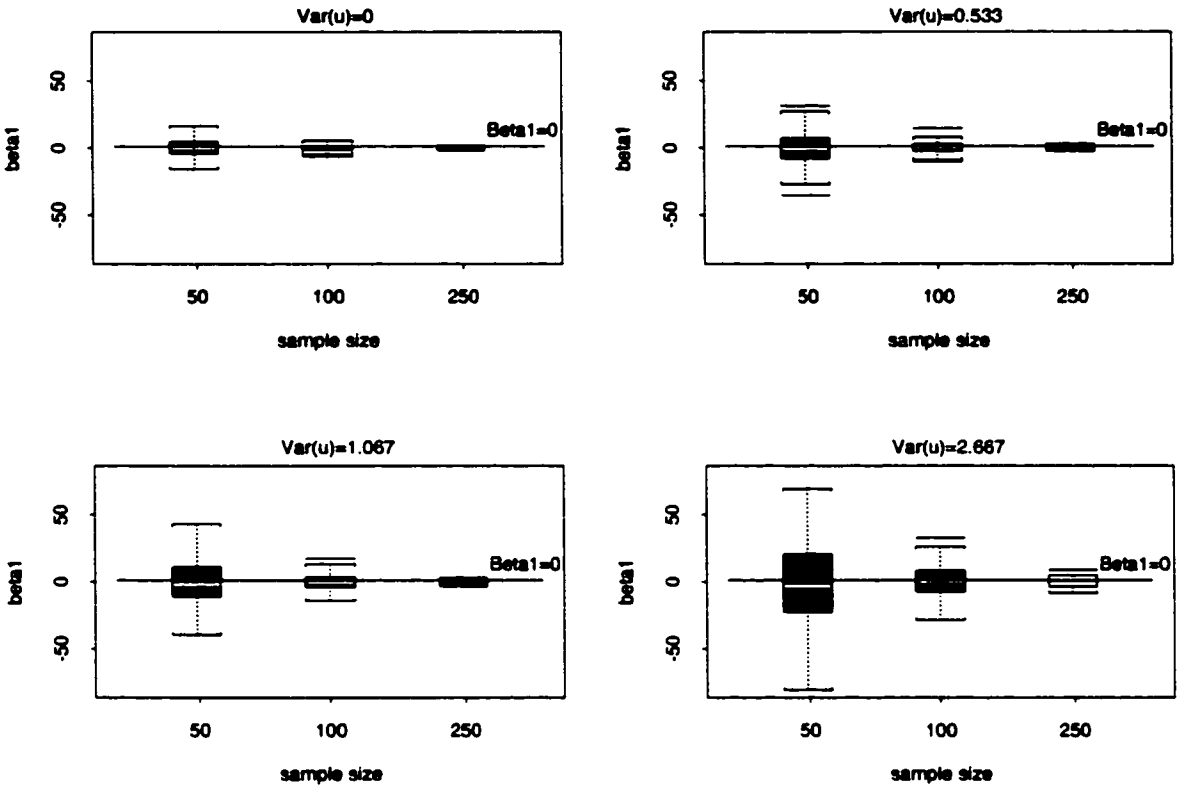

Figure 3.5: Estimation of β_1 for varying σ_u^2 and fixed $\rho = 0.25$.

improvement to yield minimal bias (less than 5% as the sample size increased, regardless of the true combination of σ_u^2 and ρ involved. This suggests that the true value of ρ has little impact on the estimation of σ^2.

The estimation of ρ yielded some interesting results. Larger values of ρ led to more severe underestimation of ρ in terms of absolute bias but less severe in terms of bias as a percentage relative to the true value of ρ. The smallest sample size, $n = 50$, especially exhibited large negative biases as large as 90%. As the sample size increased, the average estimate of ρ improved, with the level of bias dropping significantly, although corresponding 95%

confidence intervals for the mean value still did not cover the true value of ρ. Here, the impact of the size of the variance, σ_u^2, was perhaps surprising: as the value of σ_u^2 became larger, the bias associated with the estimation of ρ became correspondingly smaller, as seen in Figure 3.6. The reason is probably that a larger variability in the (unobserved) random effects provides more information and a wider range of values to assist in the estimation of ρ than a smaller variance. Further discussion regarding the bias observed in the estimation of the variance components can be found in the next section.

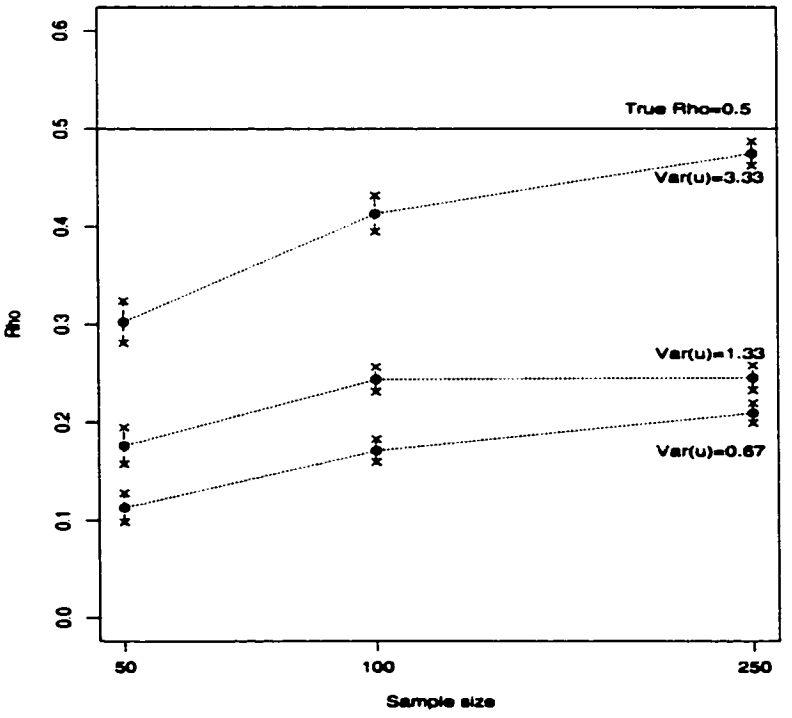

Figure 3.6: Average estimate and corresponding 95% confidence interval for $E(\hat{\rho})$ for varying σ_u^2 and fixed $\rho = 0.5$.

There have been few results reported by other researchers on the performance of exact maximum likelihood methods in a generalized linear mixed model setting with a Poisson link function. However, as discussed in §2.4.5, small simulation studies carried out for binary

data by McCulloch (1997) reported underestimation of a single variance component σ^2.

The standard errors, both theoretical and observed, were moderate in value for the regression coefficients $\beta_2, ..., \beta_5$ and ρ, ranging from 0.03 to 0.17 for the smallest sample size of 50 for ρ. The standard errors decreased with larger sample sizes. The standard errors associated with the estimation of σ_u^2 tended to be large, especially for larger σ_u^2 and ρ values, reaching as high as 4.9 when $\sigma_u^2 = 5.71$. Larger standard errors were also observed for the regression coefficients β_0 and β_1, especially for a sample size of 50.

Mean-squared errors reflected the increased standard errors for σ_u^2, β_0 and β_1 observed for larger variance component values. The increasing bias observed in the estimation of ρ for small values of σ_u^2 led to larger mean-squared errors for ρ, because bias is included as a squared term in the mean-squared error.

Discussion of the bias of the variance components estimates

The question is raised as to the cause of the noticeable negative bias present in the estimation of ρ, and also σ_u^2 using maximum likelihood. This bias is severe in small-sized samples and with little improvement, even at a sample size of $n = 250$. This issue is of interest, particularly as it is presumed that maximum likelihood estimation will lead to asymptotically normal and consistent estimates of the regression coefficients and variance components in the generalized linear mixed model setting. There are a number of possible explanations for the severe bias present in the estimation of the variance components in the polio incidence model:

One explanation is that one of the regularity conditions underlying the standard proof of the asymptotic normality and consistency of maximum likelihood estimates is that the observations are independent and identically distributed (Ferguson 1996). However, the polio model consists of observations that are all correlated with one another under the AR(1) correlation structure of the random effects, which leads to exponential decay in the correlation. While the correlation present in the data is a departure from the independence

assumption, the fact that the decay is exponential suggests that the departure from this assumption is not too dramatic, and that it is plausible that the polio model estimates will still be asymptotically normal and consistent. While this has not yet been formally proved for correlated data in the framework of generalized linear mixed models, an outline sketch of such a theorem and proof is as follows (partly based on a proof by Cramér presented in Ferguson (1996) for *iid* observations):

Theorem: Let Y_1, Y_2, ..., Y_n be distributed according to the model described in §2.5.1, with parameter $\theta = (\beta_0, \beta_1, ..., \beta_5, \rho, \sigma_u^2)$. Let θ_0 denote the true value of the parameter.

It may be shown that there exists a strongly consistent sequence $\widehat{\theta}_n$ of roots of the likelihood equation $\frac{\partial}{\partial\theta}\log L_n(\theta)$ such that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma),$$

where Σ is the inverse of the observed information matrix, $\mathcal{I}(\theta_0)^{-1}$.

Outline of proof: The proof will be based on Cramér's approach, and can be split into two parts: the first showing the existence of consistent roots, and the second part, asymptotic normality.

(i) Existence of consistent roots: The existence of a strongly consistent sequence $\hat{\theta}_n$ of roots of $\frac{\partial}{\partial\theta}\log L_n(\theta) = 0$ follows from an (as of yet unproven) extension of Theorem 17 in Ferguson (1996) for correlated observations, which shows that $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$..

(ii) Asymptotic normality: For any fixed y, expand $\frac{\partial}{\partial\theta}\log L_n(\hat{\theta}_n)$ about θ_0 using a Taylor's series expansion:

$$l'(\hat{\theta}_n) = l'(\theta_0) + (\hat{\theta}_n - \theta_0)l''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l'''(\theta_n^*)$$

where θ_n^* lies between θ_0 and $\hat{\theta}_n$. Setting the left-hand-side to zero, as assumed, we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{(\frac{1}{\sqrt{n}})l'(\theta_0)}{-\frac{1}{n}l''(\theta_0) - (\frac{1}{2}n)(\hat{\theta}_n - \theta_0)l'''(\theta_n^*)}$$

It is necessary to show that:

- $(\frac{1}{\sqrt{n}})l'(\theta_0) \xrightarrow{\text{d}} \text{Normal}(0, \mathcal{I}(\theta_0))$,
- $-\frac{1}{n}l''(\theta_0) \xrightarrow{\text{p}} \mathcal{I}(\theta_0)$, and that
- $\frac{1}{n}l'''(\theta_n^*)$ is bounded in probability.

In order to show the first, since the data is correlated, a strong mixing central limit theorem, such as Theorem 3.3.1 as presented in Guyon (1995) might possibly be used in place of the standard maximum likelihood central limit theorem as used by Cramér to show that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{a.s.}} \text{Normal}(0, \mathcal{I}(\theta_0))$.

This is merely an outline of a possible proof showing the asymptotic normality and consistency of maximum likelihood estimates for correlated data using generalized linear mixed models.

A second possible explanation is that if the estimates of the regression coefficients and variance components can be shown as above to be asymptotically normal and consistent, then the severe bias may be attributed to either small-sample behavior of maximum likelihood, or to the failure of the Metropolis-Hastings algorithm to converge to the true distribution of the random effects. The first of these, the small-sample behavior, is the most likely explanation of the bias, which is particualarly severe in the simulations of sample size $n = 50$. Small improvements are observed in the estimation of both ρ and σ^2 as the sample size increases: however these improvements are small, and it would appear that significantly larger sample sizes would be required before the unbiased property of maximum likelihood estimators is observed. This could be shown through the use of simulation studies, which are outside the scope of the current work.

Table 3.4: Maximum likelihood average estimated parameter values: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Average estimated parameter values over 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
true value	N/A	0	1	0	0
50	-0.023	0.058	0.945	0.069	-0.0048
100	-0.009	0.057	0.957	0.088	-0.001
250	-0.006	0.066	0.947	0.072	0.007
true value	0	0.5	1	0	0
50	-0.059	0.383	1.007	-0.320	-0.014
100	-0.032	0.444	1.004	-0.356	-0.010
250	-0.016	0.478	0.995	-0.255	-0.008
true value	0.25	0.533	1	0	0
50	0.017	0.372	1.012	-0.034	-0.028
100	0.017	0.468	0.981	0.497	-0.000
250	0.083	0.508	0.976	0.244	-0.009
true value	0.5	0.667	1	0	0
50	0.113	0.441	1.025	-0.775	-0.018
100	0.171	0.570	0.970	0.656	0.008
250	0.209	0.634	0.996	-0.074	-0.029
true value	0.75	1.143	1	0	0
50	0.272	0.672	1.033	-0.505	-0.028
100	0.362	0.919	0.974	1.312	0.004
250	0.406	1.052	1.096	-5.731	0.008
true value	0	1	1	0	0
50	-0.075	0.800	1.012	0.096	-0.040
100	-0.041	0.916	1.008	-0.052	-0.018
250	-0.018	0.949	1.017	-0.094	-0.011
true value	0.25	1.067	1	0	0
50	0.038	0.807	1.029	-0.439	-0.037
100	0.089	0.940	1.020	-0.057	-0.024
250	0.116	1.003	1.024	-0.172	-0.011

Table 3.5: Maximum likelihood average estimated parameter values: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Average estimated parameter values over 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
true value	0.5	1.333	1	0	0
50	0.176	0.957	1.029	-0.239	-0.038
100	0.244	1.143	0.993	0.586	-0.001
250	0.249	1.158	1.044	-0.064	-0.023
true value	0.75	2.286	1	0	0
50	0.371	1.502	1.156	-1.046	-0.027
100	0.474	1.985	1.235	0.165	-0.032
250	0.517	2.183	1.403	-1.324	-0.023
true value	0	2.5	1	0	0
50	-0.007	2.298	1.336	-1.363	-0.073
100	0.027	2.536	1.246	0.882	-0.020
250	0.084	2.575	1.540	-0.474	-0.008
true value	0.25	2.667	1	0	0
50	0.105	2.206	1.214	-1.864	-0.076
100	0.210	2.732	1.459	0.108	-0.062
250	0.254	2.828	1.720	-1.457	-0.052
true value	0.5	3.33	1	0	0
50	0.268	2.645	1.319	-2.494	-0.112
100	0.413	3.950	1.702	0.221	-0.088
250	0.475	4.245	2.028	-1.131	-0.036
true value	0.75	5.714	1	0	0
50	0.483	5.134	1.567	-1.923	-0.086
100	–	–	–	–	–
250	–	–	–	–	–

Table 3.6: Maximum likelihood standard errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample	Theoretical (observed) standard errors over 200 simulations				
size n	ρ	σ_u^2	β_0	β_1	β_4
Param.	N/A	0	1	0	0
50	0.032 (0.828)	0.025 (0.132)	0.181 (0.205)	6.714 (4.533)	0.125 (0.133)
100	0.019 (0.584)	0.018 (0.078)	0.113 (0.134)	1.910 (1.811)	0.082 (0.085)
250	0.016 (0.338)	0.013 (0.058)	0.083 (0.088)	0.569 (0.557)	0.059 (0.059)
Param.	0	0.5	1	0	0
50	0.082 (0.327)	0.163 (0.161)	0.280 (0.256)	9.216 (8.360)	0.195 (0.174)
100	0.056 (0.189)	0.123 (0.122)	0.200 (0.186)	3.389 (3.114)	0.125 (0.127)
250	0.038 (0.114)	0.085 (0.084)	0.200 (0.119)	0.807 (0.804)	0.081 (0.083)
Param.	0.25	0.533	1	0	0
50	0.084 (0.322)	0.169 (0.158)	0.329 (0.258)	11.240 (8.516)	0.197 (0.174)
100	0.065 (0.189)	0.123 (0.180)	0.226 (0.193)	3.872 (3.221)	0.145 (0.131)
250	0.045 (0.113)	0.078 ()	0.136 (0.124)	1.035 (0.840)	0.088 (0.085)
Param.	0.5	0.667	1	0	0
50	0.107 (0.354)	0.204 (0.190)	0.446 (0.295)	15.320 (9.132)	0.210 (0.183)
100	0.084 (0.228)	0.148 (0.237)	0.307 (0.229)	5.242 (3.587)	0.151 (0.144)
250	0.046 (0.114)	0.092 (0.128)	0.189 (0.141)	1.396 (0.945)	0.113 (0.092)
Param.	0.75	1.143	1	0	0
50	0.141 (0.452)	0.322 (0.257)	0.803 (0.382)	27.010 (8.401)	0.229 (0.218)
100	0.114 (0.335)	0.323 (0.421)	0.585 (0.326)	9.991 (2.868)	0.150 (0.170)
250	0.078 (0.258)	0.195 ()	0.397 (0.310)	2.443 (0.686)	0.101 (0.130)
Param.	0	1	1	0	0
50	0.102 (0.229)	0.258 (0.267)	0.366 (0.355)	12.337 (11.161)	0.236 (0.245)
100	0.076 (0.148)	0.210 (0.209)	0.250 (0.250)	4.443 (4.401)	0.172 (0.190)
250	0.044 (0.091)	0.132 (0.150)	0.155 (0.155)	1.088 (0.993)	0.105 (0.106)
Param.	0.25	1.067	1	0	0
50	0.106 (0.231)	0.290 (0.262)	0.436 (0.341)	15.063 (11.262)	0.266 (0.235)
100	0.077 (0.147)	0.231 (0.211)	0.315 (0.255)	5.532 (3.886)	0.193 (0.178)
250	0.049 (0.090)	0.135 (0.170)	0.181 (0.177)	1.253 (1.040)	0.116 (0.113)

Table 3.7: Maximum likelihood standard errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample	Theoretical (observed) standard errors over 200 simulations				
size n	ρ	σ_u^2	β_0	β_1	β_4
Param.	0.5	1.333	1	0	0
50	0.135 (0.276)	0.389 (0.519)	0.644 (0.403)	21.559 (12.067)	0.291 (0.263)
100	0.091 (0.147)	0.272 (0.421)	0.456 (0.313)	7.989 (4.751)	0.219 (0.217)
250	0.089 (0.150)	0.318 (0.202)	0.436 (0.309)	3.002 (1.612)	0.207 (0.187)
Param.	0.75	2.286	1	0	0
50	0.154 (0.456)	0.903 (0.483)	1.254 (0.732)	42.546 (10.780)	0.373 (0.406)
100	0.119 (0.194)	1.053 (0.421)	0.974 (0.493)	16.982 (2.577)	0.295 (0.229)
250	0.088 (0.123)	0.599 ()	0.626 (0.372)	4.078 (0.136)	0.193 (0.264)
Param.	0	2.5	1	0	0
50	0.136 (0.212)	0.937 (0.519)	0.774 (0.503)	26.038 (7.300)	0.534 (0.341)
100	0.101 (0.145)	0.665 ()	0.522 (0.389)	9.078 (6.210)	0.373 (0.268)
250	0.083 (0.094)	0.622 ()	0.426 (0.275)	2.815 (0.079)	0.268 (0.168)
Param.	0.25	2.667	1	0	0
50	0.135 (0.202)	0.954 (0.579)	0.892 (0.256)	29.993 ()	0.530 (0.381)
100	0.120 (0.143)	0.996 (0.462)	0.712 (0.432)	11.775 ()	0.447 (0.254)
250	0.084 (0.088)	0.686 ()	0.505 (0.223)	3.073 ()	0.286 (0.154)
Param.	0.5	3.33	1	0	0
50	0.147 (0.215)	0.823 (0.721)	1.229 (0.655)	41.431 ()	0.561 (0.448)
100	0.134 (0.156)	1.961 (0.533)	1.007 (0.492)	16.304 ()	0.527 (0.298)
250	0.089 (0.096)	1.447 ()	0.744 (0.681)	4.790 ()	0.364 (0.161)
Param.	0.75	5.714	1	0	0
50	0.166 (0.288)	4.236 (0.859)	2.235 (0.607)	74.519 ()	0.775 (0.408)
100	–	–	–	–	–
250	–	–	–	–	–

Table 3.8: Maximum likelihood mean-squared errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Maximum likelihood mean-squared errors based on 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
Param.	N/A	0	1	0	0
50	0.002	0.004	0.036	38.123	0.016
100	0.000	0.004	0.015	3.656	0.007
250	0.000	0.005	0.010	0.329	0.004
Param.	0	0.5	1	0	0
50	0.010	0.173	0.078	85.032	0.038
100	0.004	0.212	0.040	11.514	0.016
250	0.002	0.236	0.014	0.654	0.007
Param.	0.25	0.533	1	0	0
50	0.061	0.054	0.108	126.339	0.039
100	0.040	0.019	0.051	15.055	0.021
250	0.030	0.007	0.0191	1.073	0.008
Param.	0.5	0.667	1	0	0
50	0.161	0.093	0.200	235.373	0.044
100	0.115	0.031	0.095	27.584	0.023
250	0.087	0.010	0.036	1.950	0.014
Param.	0.75	1.143	1	0	0
50	0.248	0.325	0.647	729.973	0.053
100	0.164	0.155	0.343	100.259	0.022
250	0.124	0.046	0.167	7.282	0.010
Param.	0	1	1	0	0
50	0.016	0.106	0.133	152.218	0.057
100	0.008	0.051	0.062	19.746	0.030
250	0.002	0.020	0.024	1.192	0.011
Param.	0.25	1.067	1	0	0
50	0.056	0.152	0.191	227.086	0.072
100	0.032	0.070	0.100	30.602	0.038
250	0.020	0.022	0.033	1.600	0.014

Table 3.9: Maximum likelihood mean-squared errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Maximum likelihood mean-squared errors based on 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
Param.	0.5	1.33	1	0	0
50	0.123	0.293	0.415	464.864	0.086
100	0.074	0.110	0.208	64.170	0.048
250	0.071	0.132	0.192	9.019	0.043
Param.	0.75	2.286	1	0	0
50	0.168	1.429	1.596	1811.256	0.140
100	0.090	1.199	1.004	288.416	0.088
250	0.062	0.369	0.554	18.376	0.038
Param.	0	2.5	1	0	0
50	0.019	0.919	0.711	679.846	0.290
100	0.011	0.443	0.333	83.182	0.140
250	0.014	0.392	0.473	8.150	0.072
Param.	0.25	2.667	1	0	0
50	0.039	0.914	0.841	903.086	0.287
100	0.016	0.997	0.717	138.662	0.204
250	0.007	0.497	0.773	11.567	0.084
Param.	0.5	3.333	1	0	0
50	0.075	2.344	1.611	1722.744	0.327
100	0.025	4.225	1.508	265.879	0.286
250	0.009	2.925	1.610	24.224	0.133
Param.	0.75	5.714	1	0	0
50	0.099	23.464	5.318	5556.789	0.608
100	–	–	–	–	–
250	–	–	–	–	–

Chapter 4

APPROXIMATE MAXIMUM LIKELIHOOD ESTIMATION IN GENERALIZED LINEAR MIXED MODELS

4.1 The Penalized Quasi-likelihood Method

In Chapter 1 §1.3.2, PQL was introduced briefly. Below is a description of the method in more detail. Breslow and Clayton (1993) derived the PQL method by making some modifications to the Laplace approximation of the integrated likelihood. The log-penalized quasi-likelihood as defined by Breslow and Clayton (1993) is:

$$\kappa(u) = -\sum_{i=1}^{n} l_i(\beta; u) \; - \; u^t D^{-1} u/2$$

where

$$l_i(\beta; u) \propto \int_{y_i}^{\mu_i^u} \frac{a_i(y_i - w)}{\phi \nu(w)} \, dw$$

defines the conditional log-quasi-likelihood of β given u, where β is the vector of regression coefficients, u is the vector of random effects, θ the vector of variance components, and $D = D(\theta)$ is the covariance matrix of the random effects u. By maximizing the log penalized quasilikelihood, (β, u) can be estimated for fixed θ. The integrated quasi-likelihood used to estimate (β, u) (Lin and Breslow 1996) is:

$$\begin{aligned} L(\beta, \theta) = e^{l(\beta,\theta)} \quad &\propto \quad |D|^{-1/2} \int \exp\left\{ \sum_{i=1}^{n} l_i(\beta, u) \; - \; \frac{1}{2} u^t D^{-1} u \right\} du \\ &= \quad c\,|D|^{-1/2} \int e^{-\kappa(u)} \, du. \end{aligned}$$

PQL may be motivated by making a first-order Laplace approximation to the above integrated quasi-likelihood. By making a quadratic expansion of $-\kappa(u)$ about its maximum point $\tilde{u} = \tilde{u}(\beta, \theta)$, the first-order Laplace approximation (Lin and Breslow 1996b, equation 10) is:

$$l_{L_1}(\beta, \theta) = -\frac{1}{2} \log |I + Z^t \widetilde{W} Z D| + \sum_{i=1}^{n} l_i(\beta, \tilde{u}) - \frac{1}{2} \tilde{u}^t D^{-1} \tilde{u}$$

where $\tilde{u} = \tilde{u}(\beta, \theta)$ is the solution to $-\kappa'(u) = Z^t r_u - D^{-1} u = 0$, r_u is an $n \times 1$ vector of residuals $a_i(y_i - \mu_i^u)/\phi$, and $\widetilde{W}$ is W^u (an $n \times n$ diagonal matrix with variances $a_i \nu(\mu_i^u)/\phi$ on the diagonal) evaluated at $u = \tilde{u}$. Z is the design matrix for the random effects (with rows z_i) and D is the covariance matrix of the random effects.

Assuming that the first term of l_{L_1} varies slowly with β for fixed θ, Breslow and Clayton (1993) set

$$l_P(\beta, \theta) = \sum_{i=1}^{n} l_i(\beta; \tilde{u}) - \frac{1}{2} \tilde{u}^t D^{-1} \tilde{u}$$

and the value that maximizes $l_P(\beta, \theta)$ for fixed θ is the PQL estimate of the regression coefficients, $\hat{\beta}_P(\theta)$. The score equations are:

$$\frac{\delta l_P(\beta, \theta)}{\delta \beta} = \sum_{i=1}^{n} \frac{(y_i - \mu_i^u) x_i}{\phi a_i \nu(\mu_i^u) g'(\mu_i^u)} = 0$$

$$\frac{\delta l_P(\beta, \theta)}{\delta u} = \sum_{i=1}^{n} \frac{(y_i - \mu_i^u) z_i}{\phi a_i \nu(\mu_i^u) g'(\mu_i^u)} = D^{-1} u$$

where x_i is the ith row of the design matrix X. These equations were derived by Stiratelli, Laird and Ware (1984) for logistic regression of binary data. A Fisher scoring algorithm can be used (Green 1987) to solve these equations as an iterated weighted least squares (IWLS) problem involving a working dependent variable Y (where $Y_i = \eta_i^u + (y_i - \mu_i^u) g'(\mu_i^u)$) and a weight matrix $V = W^{-1} + ZDZ^t$ updated at each iteration. The solution to these score equations can be expressed as the iterative solution to the system:

$$\begin{bmatrix} X^tWX & X^tW \\ Z^tWX & I + Z^tWZD \end{bmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} = \begin{bmatrix} X^tWY \\ Z^tWY \end{bmatrix}$$

where $u = D\nu$. This was derived by Harville (1977) for best linear unbiased estimation (BLUE) of β and u in a linear mixed model setting, $Y = X\beta + Zu + \epsilon$, where $\epsilon \sim$ Normal$(0, W^{-1})$ and $u \sim$ Normal$(0, D)$, with ϵ and u independent. This is equivalent to solving for β using:

$$(X^tV^{-1}X)\beta = X^tV^{-1}Y$$

and then finding the random effects u using

$$\hat{u} = D\hat{\nu} = DZ^tV^{-1}(Y - X\hat{\beta}).$$

To find estimates of the variance components, the log-quasi-likelihood approximated by Breslow and Clayton (1993),

$$l(\theta) \approx -\frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)^tV^{-1}(y - X\beta)|_{\beta=\hat{\beta}_P(\theta)}$$

can be differentiated with respect to each of the variance components, although in practice, the REML version of this likelihood function is used, as shown in Breslow and Clayton (1993), equation (13). If the interest is focused on the asymptotic bias of the variance component estimates, the simpler standard maximum likelihood equations (Harville 1977) can be used. These standard maximum likelihood estimating equations for θ are defined as follows:

$$\tilde{U}(\theta_j) = \frac{1}{2}\left[(Y - X\beta)^tV^{-1}\frac{\partial V}{\partial \theta_j}V^{-1}(Y - X\beta) - tr\left(V^{-1}\frac{\partial V}{\partial \theta_j}\right)\right]\Big|_{\beta=\hat{\beta}_P(\theta)} = 0 \qquad (4.1)$$

and the Fisher information matrix $\mathcal{I}$ has as the jkth component:

$$\mathcal{I}_{jk} = \frac{1}{2}tr\left(P\frac{\partial V}{\partial \theta_j}P\frac{\partial V}{\partial \theta_k}\right), \qquad (4.2)$$

where

$$P = V^{-1} - V^{-1}X(X^tV^{-1}X)^{-1}X^tV^{-1}.$$

A very common technique that is used to find updated estimates of the variance components is the Fisher scoring algorithm. This involves taking the maximum likelihood score equations for the variance components and the expected Fisher information matrix as described above in equations (4.1) and (4.2) and iteratively updating the value of the variance components until convergence. Letting θ_i represent the ith variance component, and θ the vector of variance components, the Fisher scoring algorithm is as follows:

Step 1. Initialize $\theta = \widehat{\theta}^{(k)}$ where $k = 0$.

Step 2. Repeat until convergence: $\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} + \left[I(\theta^{(k)})\right]^{-1} U(\theta^{(k)})$, where

$$U(\theta^{(k)}) = \tfrac{1}{2}\left[(Y - X\beta)^t V^{-1} \tfrac{\delta V}{\delta \theta_i} V^{-1} (Y - X\beta) - tr\left(P \tfrac{\delta V}{\delta \theta_i}\right) \right]$$

$$I(\theta^{(k)})_{ij} = \tfrac{1}{2}\, tr\left(P \tfrac{\delta V}{\delta \theta_i} P \tfrac{\delta V}{\delta \theta_j}\right)$$

and $P = V^{-1} - V^{-1} X (X^t V^{-1} X)^{-1} X^t V^{-1}$ (following Harville 1977).

A more detailed description of the PQL method can be found in Breslow and Clayton (1993) and in Lin and Breslow (1996). The algorithm presented in Figure 4.1 can be applied in practice in fitting a dataset.

4.1.1 Practical issues in implementing PQL

The Newton-Raphson procedure, when used to estimate the variance components, can be unstable. The instability can arise in three ways:

- The updated information matrix is not positive definite,
- An updated variance component falls outside of its parameter boundaries.,

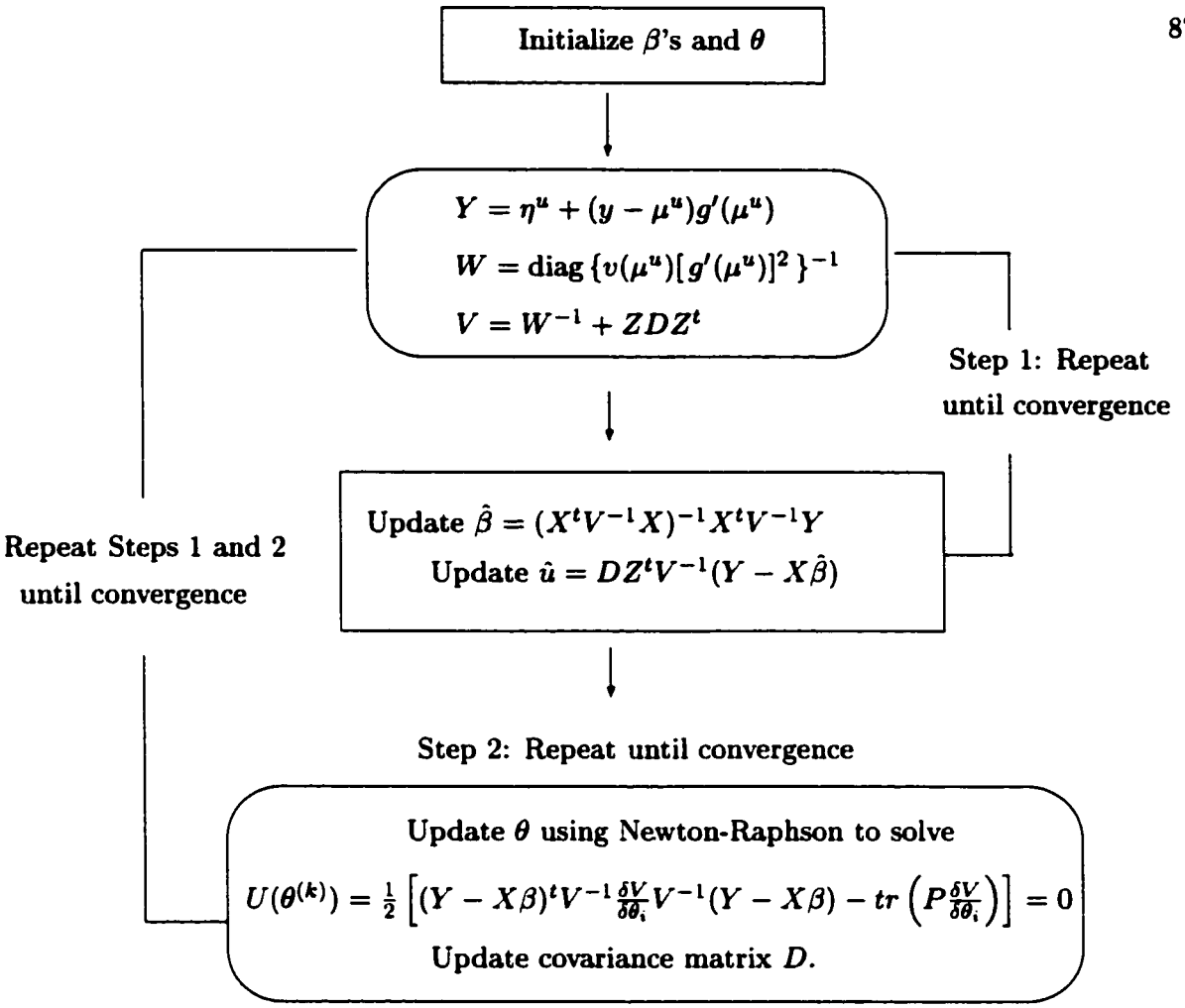

Figure 4.1: Algorithm for finding approximate maximum likelihood estimates using PQL.

- The true value of one or more of the variance components is extreme, i.e. close to the edge of the parameter space.

Possibilities for improving the stability of the Newton-Raphson algorithm within the PQL setting were explored. When the expected information matrix is not positive definite at any particular step, a half-stepping procedure can be iteratively applied to the updated variance components, and the information matrix recalculated until the matrix becomes positive definite. The half-stepping procedure used involves iteratively halving the value of $\left[I(\theta^{(k)})\right]^{-1} U(\theta^{(k)})$ at any particular step and after updating the previous variance components, recalculating the expected information matrix. This is repeated, getting successively

smaller values of $\left[I(\theta^{(k)})\right]^{-1} U(\theta^{(k)})$ until the expected information matrix becomes positive definite. Another type of half-stepping iteratively doubles the diagonal elements of the expected information matrix until the matrix became positive definite.

An alternative approach for finding the updated estimates of the variance components is to use an Expectation-Maximization algorithm instead of a Newton-Raphson algorithm. The EM algorithm (Dempster, Laird and Rubin 1977), while having a linear rate of convergence in contrast to the quadratic rate of the Newton-Raphson procedure, is known to be much more stable, and may perform better in the above mentioned situations where the Newton-Raphson procedure is particularly unstable. Some work was done investigating the use of EM as an alternative to the Newton-Raphson algorithm for estimating the variance components, but it ended up not being required, as the half-stepping procedures implemented within the Newton-Raphson algorithm proved to provide sufficient stability.

4.2 Penalized Quasi-likelihood for the Polio Incidence Model

The model used for the polio data is $\log \mu_t = x_t^t \beta + u_t$, where $u_t \sim MVN(0, D)$,

and the covariance matrix D is $\sigma_u^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots \\ \rho & 1 & \rho & \rho^2 & \dots \\ \rho^2 & \rho & 1 & \rho & \dots \\ & & & \dots & \\ & & & & \dots \\ & & \dots & \rho^2 & \rho & 1 \end{pmatrix}$,

or alternatively, u_1 is Normal $(0,\ \sigma_u^2)$ where $\sigma_u^2 = \frac{\sigma_\epsilon^2}{1-\rho^2}$, and $u_t = \rho\, u_{t-1} +$ Normal $(0,\ \sigma_\epsilon^2)$.

The basic outline of the PQL algorithm for the Polio data is as follows:

Step 0: Initialize ρ, σ_u^2, β, η^u, μ^u, where $\mu_t^u = e^{x_t^t\beta+u_t}$ and $\eta_t^u = x_t^t\beta + u_t$.

Step 1: Updating β and u :

(i) Adjusted dependent variable $Y = \eta^u + (y - \mu^u)\, g'(\mu^u) = \eta^u + \dfrac{(y-\mu^u)}{\mu^u}$

(ii) $W = \text{diag}\,\{v(\mu^u)\,[g'(\mu^b)\,]^2\}^{-1}$

$$= \begin{pmatrix} \mu_1^u & 0 & 0 & 0 & \ldots \\ 0 & \mu_2^u & 0 & 0 & \ldots \\ 0 & 0 & \mu_3^u & 0 & \ldots \\ & & & \ldots & \\ 0 & 0 & 0 & 0 & \mu_n^u \end{pmatrix}$$

(iii) $V = W^{-1} + ZDZ^t$, (where Z is the $n \times n$ identity matrix, and n is the number of observations in the dataset).

$$= \begin{pmatrix} \frac{1}{\mu_1^u} & 0 & 0 & 0 & \ldots \\ 0 & \frac{1}{\mu_2^u} & 0 & 0 & \ldots \\ 0 & 0 & \frac{1}{\mu_3^u} & 0 & \ldots \\ & & & \ldots & \\ 0 & 0 & 0 & 0 & \frac{1}{\mu_n^u} \end{pmatrix} + \sigma_u^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \ldots \\ \rho & 1 & \rho & \rho^2 & \ldots \\ \rho^2 & \rho & 1 & \rho & \ldots \\ & & & \ldots & \\ & & & & \ldots \\ & & & \rho^2 & \rho \quad 1 \end{pmatrix}$$

$$= \sigma_u^2 \begin{pmatrix} \frac{1}{\mu_1^u}+1 & \rho & \rho^2 & \rho^3 & \ldots \\ \rho & \frac{1}{\mu_2^u}+1 & \rho & \rho^2 & \ldots \\ \rho^2 & \rho & \frac{1}{\mu_3^u}+1 & \rho & \ldots \\ & & & \ldots & \\ \rho^4 & \rho^3 & \rho^2 & \rho & \frac{1}{\mu_n^u}+1 \end{pmatrix}$$

(iv) $\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y$

(v) $\hat{u} = D\, Z^t V^{-1} (Y - X\,\hat{\beta})$

Step 2: Update the variance components ρ, σ_u^2, which are the solutions to the following score equations:

$$U_\rho(\rho, \sigma_u^2) = -\frac{1}{2}\left[(Y - X\beta)^t V^{-1} \frac{\partial V}{\partial \rho} V^{-1} (Y - X\beta) - tr\left(P\frac{\partial V}{\partial \rho}\right) \right] = 0 \quad (1)$$

$$U_{\sigma_u^2}(\rho, \sigma_u^2) = -\frac{1}{2}\left[(Y - X\beta)^t V^{-1} \frac{\partial V}{\partial \sigma_u^2} V^{-1} (Y - X\beta) - tr\left(P\frac{\partial V}{\partial \sigma_u^2}\right) \right] = 0 \quad (2)$$

where $P = V^{-1} - V^{-1} X (X^t V^{-1} X)^{-1} X^t V^{-1}$ and

$$\frac{\partial V}{\partial \rho} = \begin{pmatrix} 0 & 1 & 2\rho & 3\rho^2 & \dots \\ 1 & 0 & 1 & 2\rho & \dots \\ 2\rho & 1 & 0 & 1 & \dots \\ & & & \dots & \\ \dots & & 2\rho & 1 & 0 \end{pmatrix} \quad \text{and} \quad \frac{\partial V}{\partial \sigma_u^2} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots \\ \rho & 1 & \rho & \rho^2 & \dots \\ \rho^2 & \rho & 1 & \rho & \dots \\ & & & \dots & \\ & & & & \dots \\ & & & \rho^2 & \rho & 1 \end{pmatrix}.$$

Since there is no closed form solution to these equations, a quasi-Newton algorithm was used iteratively to find solutions:

$$(\rho, \sigma_u^2)_{\text{updated}} = (\rho, \sigma_u^2)_{\text{current}} + \mathcal{I}^{-1}(\rho, \sigma_u^2)\ U(\rho, \sigma_u^2)$$

Here, $\mathcal{I}$ is the Fisher (expected) information matrix with:

$$(1,1)\text{st component} = \frac{1}{2} tr\left(P\frac{\partial V}{\partial \rho} P \frac{\partial V}{\partial \rho}\right)$$

$$(1,2)\text{nd component} = \frac{1}{2} tr\left(P\frac{\partial V}{\partial \rho} P \frac{\partial V}{\partial \sigma_u^2}\right)$$

$$(2,1)\text{nd component} = \frac{1}{2} tr\left(P\frac{\partial V}{\partial \sigma_u^2} P \frac{\partial V}{\partial \rho}\right)$$

$$(2,2)\text{nd component} = \frac{1}{2} tr\left(P\frac{\partial V}{\partial \sigma_u^2} P \frac{\partial V}{\partial \sigma_u^2}\right)$$

These steps are repeated until convergence of both the β and u, and the variance components.

4.2.1 Practical issues

The half-stepping procedure described in §4.1.1 was applied in the Newton-Raphson algorithm and was required quite often especially when the values of ρ or σ_u^2 or both were extreme ($\rho > 1$, $\sigma_u^2 < 0$).

When the true value of $\sigma_u^2 = 0$, the random effects u_t are zero in value. Consequently, the value of ρ is undefined. In this scenario, the Newton-Raphson algorithm (without any adjustment) is fairly unstable. The Newton-Raphson procedure usually converged when $\sigma_u^2 = 0$, frequently making use of half-stepping. The final updated value of ρ was fairly unstable (which is why the half-stepping was frequently necessary) and not representative of the maximum likelihood value of ρ at all.

When the true value of $\sigma_u^2 = 0$, the variance component σ_u^2 was estimated as being < 0.001 in 20% of the datasets fitted using PQL for a sample size of $n = 50$, increasing to 35% at a sample size of $n = 250$, with σ_u^2 less than 0.1 for every dataset at $n = 250$.

When the true value of ρ is particularly large, this too can lead to instability of the Newton-Raphson algorithm. This is due to the parameter being close to the boundary of the parameter space. Half-stepping of the variance components was applied frequently when this occurred.

4.2.2 Performance criteria and simulation studies

Simulation studies were carried out to evaluate the performance of approximate maximum likelihood estimators for the polio incidence data in an identical manner to the exact maximum likelihood estimators as in §3.2.2. Full details can be found in that section, but a summary of the model will be provided below.

The form of the model used to simulate the data is $\log \mu_t = x_t^t\beta + u_t$. The random effects follow an AR(1) process $\{u_t\}$ satisfying:

$$u_t = \rho u_{t-1} + \epsilon_t, \qquad \text{where} \quad \rho \text{ is the correlation coefficient,}$$
$$\epsilon_t \sim iid \text{ Normal } (0, \sigma_\epsilon^2),$$
$$y = 1, 2, ..., n,$$

and the y_t's are conditionally independent, i.e. $y_t \mid u_t \sim \text{Poisson}(\mu_t^y)$.

The trend and seasonality in the data were modelled with linear and trigonometric components, with the covariate vector

$$x_t^t = \left(1, \frac{t}{1000}, \cos\left(\frac{2\pi t}{12}\right), \sin\left(\frac{2\pi t}{12}\right), \cos\left(\frac{2\pi t}{6}\right), \sin\left(\frac{2\pi t}{6}\right)\right).$$

For each scenario (each combination of sample size n, correlation coefficient ρ and variance σ^2) a set of 200 simulations were carried out using the PQL algorithm as described in §3.2. The datasets were generated from the population model based on the true parameter values by first generating a set of random effects u_t from MVN(0,D) where D has an AR(1) correlation structure.

Table 4.1: Parameter values used in simulation studies.

Parameter	Range of values examined
n	50, 100, 250, 500
ρ	0, 0.25, 0.5, 0.75
σ_ϵ^2	0, 0.5, 1.0, 2.5
β	(1, 0, 0, 0, 0, 0)

Strict convergence criteria were used at three steps within the PQL program. The first involved the algorithm used to estimate the β coefficients within each PQL iteration. Here the convergence criterion was:

$$\Delta = |\hat{\beta}_{new} - \hat{\beta}_{old}| < 0.00000001.$$

For the estimation of the variance components using a Newton-Raphson algorithm within each PQL iteration, the convergence criterion was the same as for the beta coefficients, except less strict ($\Delta < 0.001$). Overall, when checking for overall convergence of $\theta = (\beta_0, ..., \beta_5, \rho, \sigma^2)$ between PQL iterations, the convergence criteria was

$$\Delta = |\hat{\theta}_{new} - \hat{\theta}_{old}| < 0.00001.$$

4.2.3 Results

The simulation studies yielded interesting results regarding the performance of PQL for a log-link generalized linear mixed model. In particular, the true values of the variance components and their combination played an important role in the estimation of both the regression coefficients and the variance components themselves. A full numerical analysis of the simulation results is presented in Tables 4.2 – 4.7 below. These include the mean estimate and corresponding average standard error (calculated as the average of the two hundred standard errors calculated from the information matrix $\mathcal{I}$ for each simulation) and the standard deviation of the two hundred parameter estimates. Results are presented for ρ, σ^2, β_0, β_1 and β_4. Results for the remaining regression coefficients β_2, β_3 and β_5 were omitted due to their close similarity to results for β_4. Mean-squared errors (m.s.e's) for the parameter estimates in each regression scenario are presented in Tables 4.5 – 4.6.

The regression coefficient β_0

The regression coefficient β_0 (the intercept) was estimated with a varying level of success; the level of bias in the average estimates depended on the true value of the variance components ρ and σ_u^2. When the true value of σ_u^2 was zero, β_0 was underestimated slightly with only

small bias (less than 0.05) with an improvement in the bias as the sample size increased. For all other values of σ_u^2, the intercept coefficient was positively biased (from 7–23%), with the bias getting gradually worse as σ_u^2 or ρ increased, or both increased. The bias was larger as the sample size increased, and corresponding 95% confidence intervals for the expected value of the estimator $E(\hat{\beta}_0)$, did not usually cover the true value of β_0. Figure 4.2 displays these effects clearly for a fixed value of $\rho = 0.5$ and increasing variance σ_u^2, where the true value is represented by the horizontal line at $\beta_0 = 1$. Examination of the same set of curves for other fixed values of ρ and varying σ_u^2 yielded the same pattern, as confirmed by the numerical summaries presented in Tables 4.2 - 4.4. These results are consistent with the findings in Leroux (2000) where a small simulation study conducted on a log-linear model for spatial data using PQL also overestimated the intercept term on average. Other researchers (Breslow and Clayton 1993, Raudenbush, Yang et al 2000, Mealli and Rampichini 1999), carrying out simulation studies for binomial data using PQL, had varying results in the estimation of β_0, though commonly found their PQL regression coefficients to be less in absolute value than the true value, or close to no bias. Neuhaus and Segal (1996) noted that the estimation of β_0 using PQL may bear similarity to the estimation of β_0 in marginal (population-averaged) models, which could lead to the observed positive bias present in the current analysis.

Improvements in the observed standard error (as measured by the standard deviation of the estimates from the two hundred simulations) and the average estimated standard error (as calculated from the information matrix for each simulated dataset) were seen as sample size increased. The estimated and observed standard errors were comparable for all sample sizes and combinations of the variance components. Breslow and Clayton (1993) also found that standard errors estimated for the regression coefficients in a simulation study carried out on binary data using PQL agreed reasonably well with the simulated standard errors.

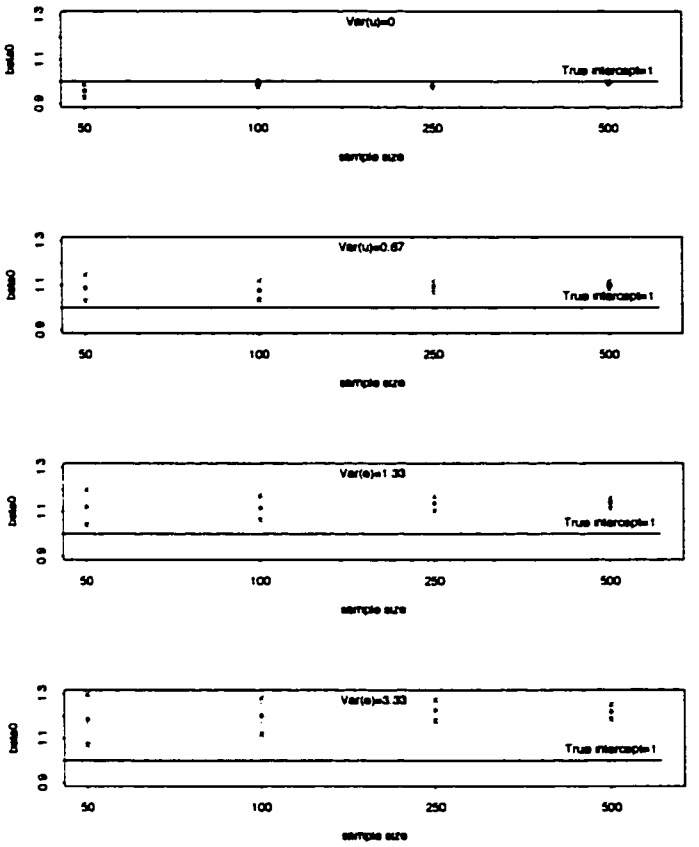

Figure 4.2: Average estimates and corresponding 95% confidence intervals of $E(\hat{\beta}_0)$ for varying σ_u^2 and fixed $\rho = 0.5$.

The regression coefficient β_1

The estimation of the regression coefficient used to model the time trend, β_1, was highly variable, especially with increasing σ_u^2, and to a lesser degree, ρ. The bias ranged from -0.97 for the combination of the most extreme values of the variance components at a sample size of 50, to 0.01 for $\rho = 0.25$ and $\sigma_u^2 = 0.533$ at a sample size of 100. However, the corresponding 95% confidence interval for $E(\hat{\beta}_1)$ almost always covered the true value of β_1. As for the intercept, both measures of standard error (theoretical and observed), which

were similar in size to each other, showed improvement with an increasing sample size, more drastically so than the standard errors associated with β_0.

The other regression coefficients

The remaining regression coefficients were estimated on average with minimal bias (less than 0.06) for all sample sizes from $n = 50$ to $n = 500$, and with correspondingly small mean-squared errors. The small level of bias is displayed clearly in Figure 4.3 for β_2 where $\rho = 0.5$ and $\sigma_u^2 = 3.33$.

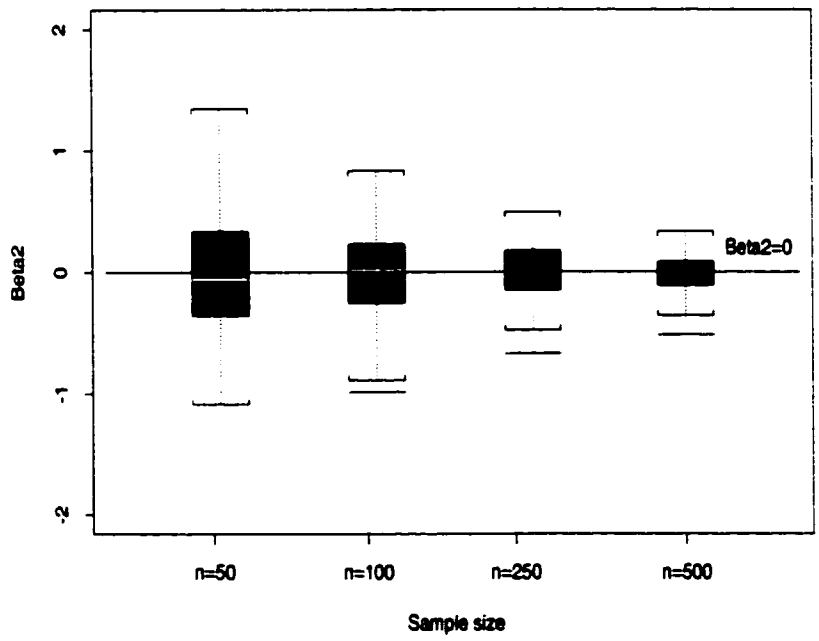

Figure 4.3: Boxplot showing variability reduction for β_2 estimation as sample size increases for $\sigma_u^2 = 3.33$, $\rho = 0.5$.

The variance components

The variance component σ_u^2 was generally estimated with negative bias which became more noticeable as the sample size increased. Most of the biases were in the vicinity of 10-12%, while the smallest was 3% for a sample size of 50 for the most extreme values of the variance components tested, and the largest bias was 11% when the variance of the random effects was 1.143 with $\rho = 0.75$. The exception was when the true value of σ_u^2 was zero, when σ_u^2 was always overestimated (since it has a lower bound of zero). In 20% of the datasets for $n = 50$ and 35% of the datasets for $n = 250$, σ_u^2 was estimated to be very close to zero (less than 0.001). At a sample size of $n = 250$, nearly 90% of the datasets estimated σ_u^2 to be less than 0.05.

An absolute bias of 0.6 was observed in the average estimate of σ_u^2 for the smallest sample size of 50 down to 0.01 at a sample size of 500. It should be noted that the corresponding 95% confidence intervals for $E(\hat{\sigma}_u^2)$ cover the true value in most cases. Figure 4.5 presents boxplots of the estimates for σ_u^2 for the fixed value of $\rho = 0.5$ and increasing sample size. This clearly demonstrates that as the level of variability of the random effects increases, the level of absolute bias of the average estimates also increases. However, the biases relative to the value of σ_u^2 as measured in percentages rose only slightly with increasing variance. This pattern was seen in similar plots for other values of ρ, as confirmed by the numerical summaries presented in Tables 4.2 - 4.3. The bias is also clearly seen in Figure 4.4, where the individual bias (parameter estimate - true value) of σ_u^2 for each of two hundred datasets for varying σ_u^2 and a fixed $\rho = 0.25$ at a sample size of $n = 250$ is presented.

The standard errors of σ_u^2 also increased with larger values of ρ and σ_u^2, as displayed in Tables 4.4 - 4.5.

As for the regression coefficients, the observed and estimated standard errors for σ_u^2 were comparable. Breslow and Clayton (1993) observed rather large discrepancies between the theoretical and observed standard errors of the variance components for small denominators in a logit GLMM, a situation which improved with larger denominators.

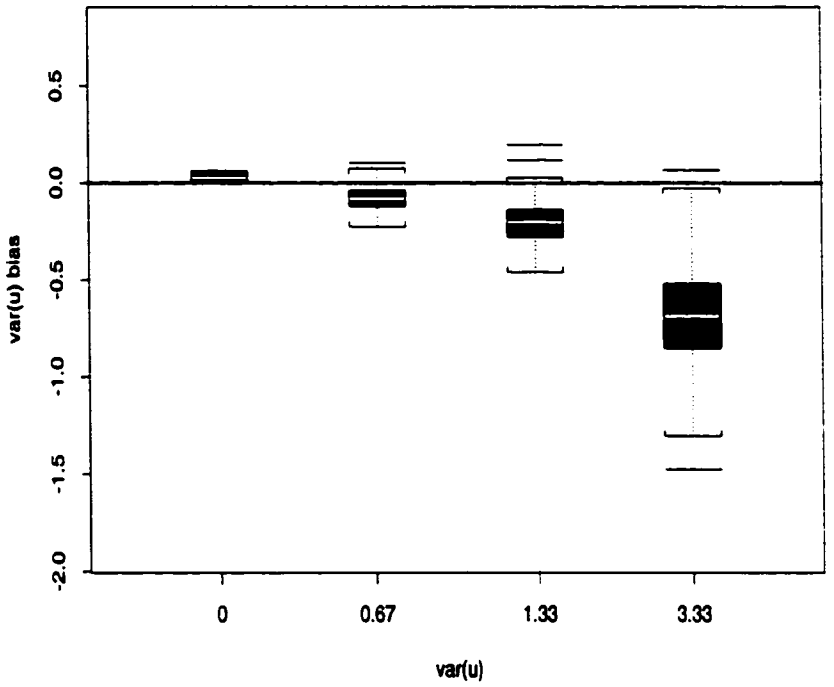

Figure 4.4: Boxplots of biases of individual dataset estimates of σ_u^2 for varying σ_u^2 using PQL.

The other variance component ρ was always underestimated on average, but with very little bias (always less than 0.05) regardless of the true values of ρ and σ_u^2. Improvements in already small biases were observed as sample size increased, as seen in Figure 4.6. The standard errors displayed for ρ in Tables 4.4 - 4.5 remain stable at values around 0.2 for a sample size of 50 and values around 0.08 for a sample size of $n = 250$.

There was a very clear effect of sample size on the variability of the PQL estimates of all regression coefficients and variance components for every combination of ρ and σ^2. The largest variability was predictably seen in the smallest sample size investigated, $n = 50$, decreasing quite substantially as the sample size increased to $n = 500$. A typical set of

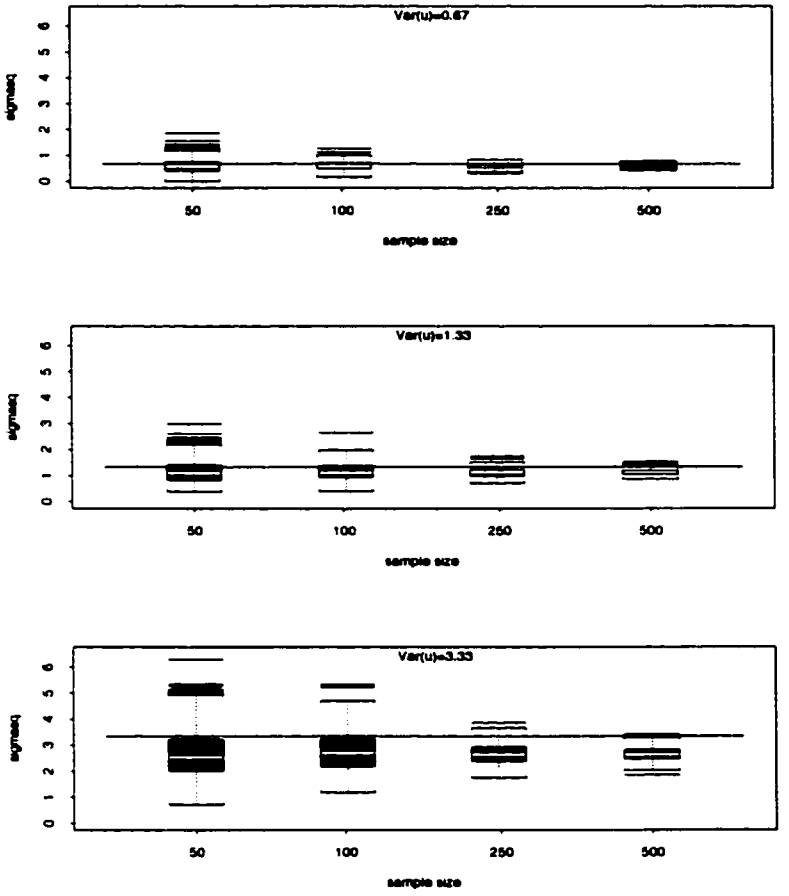

Figure 4.5: Boxplots of σ_u^2 for fixed $\rho = 0.5$.

boxplots showing this reduction in variability is shown in Figure 4.3.

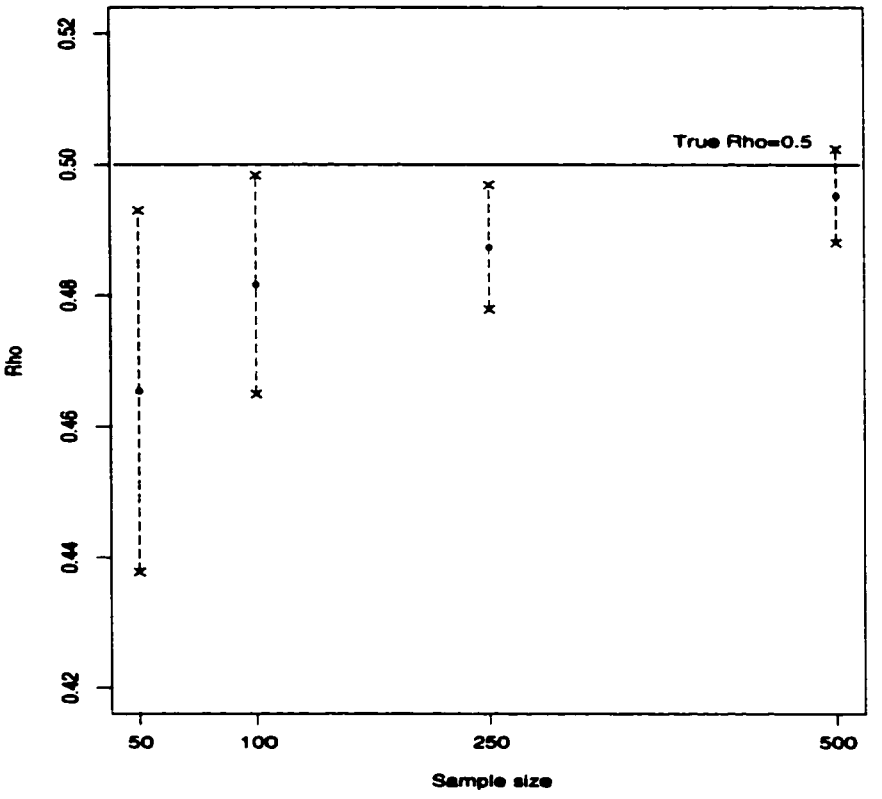

Figure 4.6: Plot showing the small bias in the estimation of ρ for $\sigma_u^2 = 1.33$ and $\rho = 0.5$.

Table 4.2: PQL average estimated parameter values: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Average estimated parameter values over 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
true value	N/A	0	1	0	0
50	-0.063	0.110	0.960	0.240	-0.002
100	0.252	0.039	0.991	-0.102	0.001
250	0.194	0.016	0.980	0.066	0.007
500	0.333	0.010	0.003	0.017	-0.003
true value	0	0.5	1	0	0
50	-0.003	0.441	1.073	-0.220	-0.0013
100	-0.013	0.438	1.094	-0.176	-0.010
250	-0.015	0.437	1.097	-0.049	-0.008
500	-0.007	0.436	1.103	-0.003	-0.004
true value	0.25	0.533	1	0	0
50	0.228	0.472	1.079	-0.089	-0.027
100	0.220	0.473	1.085	0.011	-0.015
250	0.236	0.469	1.091	-0.036	-0.010
500	0.246	0.465	1.103	-0.012	-0.006
true value	0.5	0.667	1	0	0
50	0.455	0.600	1.089	-0.964	-0.017
100	0.477	0.597	1.077	0.223	-0.020
250	0.498	0.583	1.094	-0.066	-0.013
500	0.495	0.589	1.100	-0.023	-0.003
true value	0.75	1.143	1	0	0
50	0.719	1.420	1.091	-1.027	-0.028
100	0.728	1.072	1.083	0.019	-0.029
250	0.741	1.015	1.103	-0.218	-0.010
500	0.748	1.020	1.098	-0.054	-0.007
true value	0	1	1	0	0
50	-0.012	0.843	1.104	0.1732	-0.040
100	-0.015	0.852	1.128	-0.046	-0.018
250	-0.009	0.831	1.141	-0.075	-0.011
500	-0.011	0.830	1.143	-0.011	-0.004
true value	0.25	1.067	1	0	0
50	0.222	0.907	1.116	-0.387	-0.040
100	0.236	0.909	1.126	-0.027	-0.025
250	0.245	0.890	1.137	-0.093	-0.013
500	0.242	0.891	1.140	-0.019	-0.004

Table 4.3: PQL average estimated parameter values: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Average estimated parameter values over 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
true value	0.5	1.33	1	0	0
50	0.465	1.169	1.121	-1.082	-0.031
100	0.482	1.165	1.119	0.072	-0.026
250	0.487	1.122	1.137	-0.121	-0.016
500	0.495	1.129	1.138	-0.039	-0.008
true value	0.75	2.286	1	0	0
50	0.713	2.546	1.123	-1.510	-0.022
100	0.733	2.074	1.123	-0.013	-0.022
250	0.739	1.963	1.153	-0.384	-0.016
500	0.743	1.955	1.150	-0.112	-0.006
true value	0	2.5	1	0	0
50	-0.008	2.001	1.181	-0.425	-0.035
100	-0.012	1.999	1.200	-0.015	-0.022
250	-0.004	1.960	1.220	-0.149	-0.013
500	-0.005	1.966	1.220	-0.043	-0.003
true value	0.25	2.667	1	0	0
50	0.236	2.182	1.182	-0.796	-0.044
100	0.238	2.144	1.192	0.166	-0.036
250	0.241	2.102	1.220	-0.172	-0.016
500	0.239	2.092	1.220	-0.048	-0.006
true value	0.5	3.33	1	0	0
50	0.466	2.716	1.185	-0.472	-0.057
100	0.476	2.758	1.199	-0.157	-0.046
250	0.484	2.660	1.227	-0.275	-0.021
500	0.487	2.643	1.220	-0.072	-0.013
true value	0.75	5.714	1	0	0
50	0.707	5.901	1.193	-0.972	-0.046
100	0.723	4.775	1.222	-0.028	-0.041
250	0.731	4.490	1.258	-0.460	-0.024
500					

Table 4.4: PQL standard errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample	Theoretical (observed) standard errors over 200 simulations				
size n	ρ	σ_u^2	β_0	β_1	β_4
Param.	N/A	0	1	0	0
50	0.786 (19.663)	0.182 (0.387)	0.184 (0.263)	6.408 (8.515)	0.128 (0.130)
100	0.800 (12.115)	0.050 (0.074)	0.124 (0.167)	2.080 (2.790)	0.083 (0.088)
250	0.800 (5.420)	0.020 (0.022)	0.083 (0.086)	0.573 (0.592)	0.059 (0.055)
500	0.073 (3.938)	0.013 (0.013)	0.059 (0.058)	0.197 (0.201)	0.038 (0.039)
Param.	0	0.5	1	0	0
50	0.275 (0.288)	0.163 (0.175)	0.270 (0.267)	8.810 (9.118)	0.189 (0.176)
100	0.178 (0.189)	0.109 (0.114)	0.190 (0.180)	3.232 (3.092)	0.119 (0.124)
250	0.115 (0.115)	0.072 (0.070)	0.112 (0.112)	0.767 (0.772)	0.077 (0.078)
500	0.088 (0.080)	0.045 (0.049)	0.076 (0.079)	0.259 (0.273)	0.060 (0.055)
Param.	0.25	0.533	1	0	0
50	0.275 (0.278)	0.244 (0.240)	0.317 (0.320)	10.980 (10.842)	0.190 (0.182)
100	0.173 (0.179)	0.123 (0.124)	0.227 (0.215)	3.963 (3.688)	0.138 (0.132)
250	0.107 (0.108)	0.078 (0.075)	0.136 (0.134)	0.938 (0.928)	0.082 (0.084)
500	0.069 (0.075)	0.046 (0.052)	0.092 (0.095)	0.318 (0.328)	0.057 (0.059)
Param.	0.5	0.667	1	0	0
50	0.269 (0.397)	0.273 (0.297)	0.432 (0.425)	14.763 (14.324)	0.200 (0.189)
100	0.144 (0.148)	0.170 (0.166)	0.311 (0.294)	5.305 (5.029)	0.140 (0.136)
250	0.073 (0.087)	0.092 (0.096)	0.182 (0.185)	1.293 (1.273)	0.086 (0.086)
500	0.063 (0.060)	0.064 (0.067)	0.125 (0.130)	0.425 (0.451)	0.065 (0.061)
Param.	0.75	1.143	1	0	0
50	0.170 (0.185)	1.624 (2.065)	0.730 (0.826)	24.199 (26.455)	0.207 (0.195)
100	0.100 (0.103)	0.425 (0.411)	0.546 (0.538)	9.372 (9.101)	0.139 (0.137)
250	0.056 (0.057)	0.196 (0.205)	0.328 (0.340)	2.331 (2.336)	0.085 (0.086)
500	0.037 (0.039)	0.132 (0.141)	0.233 (0.245)	0.795 (0.845)	0.061 (0.061)
Param.	0	1	1	0	0
50	0.228 (0.226)	0.254 (0.268)	0.339 (0.326)	11.328 (11.127)	0.214 (0.218)
100	0.160 (0.150)	0.182 (0.178)	0.230 (0.223)	4.096 (3.833)	0.158 (0.154)
250	0.087 (0.093)	0.111 (0.107)	0.140 (0.138)	0.978 (0.955)	0.095 (0.097)
500	0.072 (0.065)	0.074 (0.075)	0.094 (0.097)	0.314 (0.336)	0.066 (0.069)
Param.	0.25	1.067	1	0	0
50	0.219 (0.221)	0.318 (0.310)	0.400 (0.400)	13.652 (13.609)	0.239 (0.231)
100	0.140 (0.144)	0.217 (0.198)	0.286 (0.277)	5.008 (4.758)	0.174 (0.166)
250	0.082 (0.088)	0.122 (0.117)	0.164 (0.173)	1.174 (1.192)	0.107 (0.105)
500	0.060 (0.061)	0.083 (0.082)	0.114 (0.121)	0.383 (0.420)	0.072 (0.074)

Table 4.5: PQL standard errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Theoretical (observed) standard errors over 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
Param.	0.5	1.333	1	0	0
50	0.201 (0.196)	0.609 (0.592)	0.571 (0.570)	18.994 (19.165)	0.266 (0.243)
100	0.121 (0.124)	0.326 (0.286)	0.407 (0.393)	7.053 (6.720)	0.182 (0.174)
250	0.069 (0.074)	0.168 (0.161)	0.235 (0.243)	1.720 (1.676)	0.112 (0.110)
500	0.052 (0.051)	0.110 (0.113)	0.171 (0.174)	0.581 (0.600)	0.075 (0.078)
Param.	0.75	2.286	1	0	0
50	0.143 (0.162)	3.842 (3.128)	1.108 (1.067)	33.557 (34.561)	0.257 (0.250)
100	0.085 (0.092)	0.787 (0.759)	0.733 (0.740)	12.775 (12.502)	0.188 (0.176)
250	0.054 (0.053)	0.379 (0.380)	0.445 (0.465)	3.197 (3.196)	0.106 (0.110)
500	0.034 (0.036)	0.253 (0.257)	0.318 (0.331)	1.091 (1.142)	0.074 (0.078)
Param.	0	2.5	1	0	0
50	0.198 (0.193)	0.537 (0.548)	0.456 (0.464)	15.580 (15.855)	0.321 (0.311)
100	0.137 (0.129)	0.393 (0.361)	0.328 (0.316)	5.756 (5.435)	0.220 (0.219)
250	0.079 (0.079)	0.245 (0.217)	0.196 (0.197)	1.395 (1.361)	0.137 (0.138)
500	0.057 (0.055)	0.165 (0.152)	0.138 (0.139)	0.473 (0.480)	0.096 (0.098)
Param.	0.25	2.667	1	0	0
50	0.189 (0.191)	0.647 (0.657)	0.593 (0.591)	20.293 (20.114)	0.344 (0.338)
100	0.116 (0.125)	0.445 (0.410)	0.403 (0.404)	7.139 (6.933)	0.248 (0.240)
250	0.076 (0.076)	0.282 (0.243)	0.244 (0.251)	1.800 (1.735)	0.152 (0.151)
500	0.056 (0.053)	0.178 (0.168)	0.173 (0.177)	0.591 (0.611)	0.104 (0.107)
Param.	0.5	3.333	1	0	0
50	0.167 (0.174)	0.970 (1.006)	0.819 (0.818)	27.793 (27.628)	0.371 (0.352)
100	0.107 (0.111)	0.710 (0.620)	0.585 (0.582)	10.248 (9.963)	0.259 (0.255)
250	0.063 (0.067)	0.381 (0.352)	0.348 (0.363)	2.545 (2.505)	0.154 (0.160)
500	0.045 (0.046)	0.252 (0.243)	0.251 (0.257)	0.868 (0.887)	0.110 (0.113)
Param.	0.75	5.714	1	0	0
50	0.132 (0.150)	9.699 (7.115)	1.559 (1.580)	19.467 (51.284)	0.370 (0.367)
100	0.080 (0.088)	1.750 (1.641)	1.108 (1.089)	4.843 (18.426)	0.260 (0.259)
250	0.048 (0.050)	0.826 (0.831)	0.666 (0.685)	4.887 (4.708)	0.159 (0.161)
500	–	–	–	–	–

Table 4.6: PQL mean-squared errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Mean-squared errors for PQL based on 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
Param.	N/A	0	1	0	0
50	0.716	0.045	0.035	41.12	0.016
100	0.639	0.004	0.016	4.339	0.007
250	0.643	0.001	0.007	0.333	0.004
500	0.570	0.000	0.003	0.039	0.001
Param.	0	0.5	1	0	0
50	0.078	0.221	0.078	77.660	0.036
100	0.032	0.203	0.045	10.476	0.014
250	0.013	0.196	0.022	0.591	0.006
500	0.008	0.192	0.016	0.067	0.004
Param.	0.25	0.533	1	0	0
50	0.076	0.063	0.107	120.563	0.037
100	0.031	0.019	0.059	15.705	0.019
250	0.012	0.010	0.027	0.880	0.007
500	0.005	0.007	0.019	0.101	0.003
Param.	0.5	0.667	1	0	0
50	0.075	0.079	0.195	218.868	0.040
100	0.021	0.034	0.102	28.195	0.020
250	0.005	0.015	0.042	1.675	0.008
500	0.004	0.010	0.026	0.181	0.004
Param.	0.75	1.143	1	0	0
50	0.030	2.716	0.542	586.665	0.043
100	0.010	0.186	0.305	87.834	0.020
250	0.003	0.055	0.118	5.481	0.007
500	0.001	0.033	0.064	0.635	0.004
Param.	0	1	1	0	0
50	0.052	0.089	0.126	128.350	0.048
100	0.026	0.055	0.069	16.781	0.025
250	0.008	0.041	0.039	0.961	0.009
500	0.005	0.034	0.029	0.098	0.004
Param.	0.25	1.067	1	0	0
50	0.049	0.126	0.173	186.541	0.059
100	0.020	0.072	0.098	25.085	0.031
250	0.007	0.046	0.046	1.386	0.012
500	0.004	0.038	0.033	0.147	0.005

Table 4.7: PQL mean-squared errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Mean-squared errors for PQL based on 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
Param.	0.5	1.333	1	0	0
50	0.041	0.221	0.341	358.132	0.072
100	0.015	0.135	0.179777	49.754	0.034
250	0.005	0.073	0.074	2.973	0.013
500	0.003	0.054	0.048	0.339	0.006
Param.	0.75	2.286	1	0	0
50	0.022	14.829	1.052	1128.375	0.067
100	0.008	0.664	0.553	163.213	0.036
250	0.003	0.248	0.222	10.371	0.012
500	0.001	0.173	0.123	1.203	0.006
Param.	0	2.5	1	0	0
50	0.039	0.537	0.241	242.921	0.104
100	0.019	0.405	0.148	33.126	0.049
250	0.006	0.351	0.087	1.968	0.019
500	0.003	0.313	0.068	0.225	0.009
Param.	0.25	2.667	1	0	0
50	0.036	0.653	0.385	412.438	0.120
100	0.014	0.472	0.200	50.993	0.063
250	0.006	0.398	0.108	3.269	0.023
500	0.003	0.362	0.078	0.352	0.011
Param.	0.5	3.333	1	0	0
50	0.029	1.322	0.705	772.666	0.141
100	0.012	0.835	0.382	105.041	0.069
250	0.004	0.599	0.173	6.551	0.024
500	0.002	0.540	0.111	0.758	0.012
Param.	0.75	5.714	1	0	0
50	0.019	94.104	2.469	2587.751	0.139
100	0.007	3.944	1.277	378.958	0.069
250	0.003	2.181	0.510	24.091	0.026
500	–	–	–	–	–

Chapter 5

THE ITERATIVE BIAS CORRECTION METHOD

5.1 Background

The methods described in previous chapters for fitting generalized linear mixed models involve finding either approximate or exact maximum likelihood estimates. In 1995 Kuk published a paper proposing "a general method of adjusting any conveniently defined initial estimates to result in estimates which are asymptotically unbiased and consistent". This method is based upon a technique known as iterative bias correction which can supposedly be applied to any parametric model. Another potential advantage of this method is that it is assumed only that $f(y|\, u;\theta)$ and $h(u;\theta)$ are parametric, whereas other methods such as PQL currently used make the more restrictive assumption that $h(u;\theta)$ is normal.

5.1.1 Description of the iterative bias correction method

The iterative bias correction method initially finds estimates of regression coefficients and variance components using a method such as PQL/MQL or BLUP (Best Linear Unbiased Prediction). The method chosen to find the initial estimates may yield inconsistent and biased estimates in certain situations, such as when the true distribution of the random effects has a large variance. These initial estimates are denoted as $\widetilde{\theta}$, and are assumed to have an asymptotic limit of θ^* (which commonly may not be the true parameter values θ for one or more of the parameters). The asymptotic bias is then defined as the difference between θ and θ^*, i.e.

$$\begin{aligned} b(\theta) &= \theta^* - \theta \\ &\equiv h(\theta) - \theta, \end{aligned}$$

where θ^*, the asymptotic limit of $\widetilde{\theta}$, is considered to be a function of the true parameter θ. This relationship can be defined as $\theta^* = h(\theta)$, where $h(.)$ is $1-1$ and differentiable. The true values of the parameters are found in the $1 \times (p+r)$ vector θ. The initial estimation vector θ^* is of a similar dimension, where p is the number of regression components and r is the number of variance components to be estimated.

It is assumed that the initial estimates $\widetilde{\theta}$ are solutions to an estimating function written in vector form as:

$$\psi(\theta; y) = 0.$$

The estimating function $\psi(\theta; y)$ of dimension $(p+r) \times 1$ can also be written more explicitly as $\sum_{i=1}^{n} \psi(\theta; y_i)$, the sum of the estimating functions for the individual observations. For standard estimating equations under certain regularity conditions, Fahrmeir and Kaufman (1985) (and earlier, Haberman 1977) show that the initial estimates $\widetilde{\theta}$ are asymptotically normally distributed and consistent, i.e.

$$\sqrt{n}(\widetilde{\theta} - \theta^*) \xrightarrow{d} N(0, \Sigma),$$

where the asymptotic limit of the vector of initial estimates, θ^* is defined implicitly by $E_\theta[\psi(\theta^*, Y)] = 0$.

Hwever, for the example of the polio incidence data (following in §5.3), all the observations are correlated to some degree. An AR(1) correlation structure for the random effects means the strongest correlation is present between consecutive observations, with exponential decay for correlation between observations further apart. The presence of the correlation means that the independence of the observations assumption in the standard estimating equation

theory is not satisfied. However, the exponential decay of the correlation implies that the correlation present between most of the observations is not a dramatic departure from the independence assumption. While it has not yet been proven that estimates calculated from estimating equations based on correlated data with an autoregressive correlation structure are asymptotically normal and consistent, it is plausible that they will be.

Asymptotically, the covariance matrix Σ can be written as:

$$\begin{aligned}
\Sigma &= \lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} E_\theta(\psi'(\theta^*; Y_i)) \right\}^{-1} \frac{1}{n} E_\theta \left\{ \sum_{i=1}^{n} \psi(\theta^*; Y_i) \sum_{i=1}^{n} \psi^t(\theta^*; Y_i) \right\} \times \\
&\qquad \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} E_\theta(\psi'(\theta^*; Y_i)) \right\}^{-1} \right]^t \\
&= \lim_{n\to\infty} n \left\{ \sum_{i=1}^{n} E_\theta(\psi'(\theta^*; Y_i)) \right\}^{-1} E_\theta \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \psi(\theta^*; Y_i) \psi^t(\theta^*; Y_j) \right\} \times \\
&\qquad \left[\left\{ \sum_{i=1}^{n} E_\theta(\psi'(\theta^*; Y_i)) \right\}^{-1} \right]^t \\
&= \lim_{n\to\infty} n \{E_\theta(\psi'(\theta^*; Y))\}^{-1} E_\theta\{\psi(\theta^*; Y)\psi^t(\theta^*; Y)\} \left[\{E_\theta(\psi'(\theta^*; Y))\}^{-1} \right]^t,
\end{aligned}$$

where ψ' is a $(p+r) \times (p+r)$ matrix of derivatives of the components of ψ with respect to the components of θ^*, sometimes written as:

$$\psi'(\theta^*;y) = \begin{pmatrix} \frac{d\psi_0(\theta^*;y)}{d\theta_0^*} & \frac{d\psi_1(\theta^*;y)}{d\theta_0^*} & \cdots & \frac{d\psi_{(p+r)}(\theta^*;y)}{d\theta_0^*} \\ \frac{d\psi_0(\theta^*;y)}{d\theta_1^*} & \frac{d\psi_1(\theta^*;y)}{d\theta_1^*} & \cdots & \frac{d\psi_{(p+r)}(\theta^*;y)}{d\theta_1^*} \\ & \cdots & & \\ \frac{d\psi_0(\theta^*;y)}{d\theta_{(p+r)}^*} & \frac{d\psi_1(\theta^*;y)}{d\theta_{(p+r)}^*} & \cdots & \frac{d\psi_{(p+r)}(\theta^*;y)}{d\theta_{(p+r)}^*} \end{pmatrix}.$$

The expected values of the cross product terms in the middle expectation of Σ are zero, assuming the conditional independence of $Y_i|u$. It is assumed that θ^* exists, and as mentioned earlier, is a function of the true θ, i.e. $\theta^* = h(\theta)$, where the function $h(.)$ is assumed to be $1-1$ and differentiable. The $1 \times (p+r)$ vector $b(\theta)$ is defined as the bias of $\widetilde{\theta}$, and $\widehat{\theta}$ is defined as the $1 \times (p+r)$ vector of the final updated bias-corrected estimates of the parameters. The method works by repeatedly updating estimates of the bias vector $b(\theta)$ and $\widehat{\theta}$ iteratively until convergence of the parameter estimates, as follows:

$$\hat{b}^{(0)} = 0 \qquad \widehat{\theta}^{(0)} = \widetilde{\theta} - \hat{b}^{(0)}$$

$$\hat{b}^{(1)} = h(\widehat{\theta}^{(0)}) - \widehat{\theta}^{(0)} \qquad \widehat{\theta}^{(1)} = \widetilde{\theta} - \hat{b}^{(1)}$$

$$\hat{b}^{(2)} = h(\widehat{\theta}^{(1)}) - \widehat{\theta}^{(1)} \qquad \widehat{\theta}^{(2)} = \widetilde{\theta} - \hat{b}^{(2)}$$

$$\vdots \qquad\qquad\qquad \vdots$$

until at convergence,

$$\hat{b} = h(\widehat{\theta}) - \widehat{\theta} \qquad \widehat{\theta} = \widetilde{\theta} - \hat{b}$$

so that $\widetilde{\theta} = h(\widehat{\theta})$, or, writing the final estimates as a function of the initial estimates, $\widehat{\theta} = h^{-1}(\widetilde{\theta})$. Then, assuming that h is 1–1 and differentiable and by applying the multi-

variate delta theorem to $\sqrt{n}(\widetilde{\theta} - \theta^*) \xrightarrow{d} N(0, \Sigma)$, we get:

$$\sqrt{n}\,(\,h^{-1}(\widetilde{\theta}) - h^{-1}(\theta^*)\,) \xrightarrow{d} N(0,\, H\Sigma H^t), \quad \text{i.e.}$$

$$\sqrt{n}\,(\,\widehat{\theta} \;-\; \theta\,) \xrightarrow{d} N(0,\, H\Sigma H^t),$$

so that $\widehat{\theta}$ is an asymptotically unbiased estimator of θ. H is the matrix of the derivatives of the components of $\theta = h^{-1}(\theta^*)$ with respect to the components of θ^*, such that $H = \frac{\partial h^{-1}(\theta^*)}{\partial \theta^*} = \frac{\partial \theta}{\partial \theta^*}$, i.e.

$$H = \begin{pmatrix} \frac{\partial \theta_0}{\partial \theta_0^*} & \frac{\partial \theta_1}{\partial \theta_0^*} & \cdots & \frac{\partial \theta_{p+r}}{\partial \theta_0^*} \\ \frac{\partial \theta_0}{\partial \theta_1^*} & \frac{\partial \theta_1}{\partial \theta_1^*} & \cdots & \frac{\partial \theta_{p+r}}{\partial \theta_1^*} \\ & \cdots & & \cdots \\ \frac{\partial \theta_0}{\partial \theta_{p+r}^*} & \frac{\partial \theta_1}{\partial \theta_{p+r}^*} & \cdots & \frac{\partial \theta_{p+r}}{\partial \theta_{p+r}^*} \end{pmatrix}.$$

As $H = \frac{\partial h^{-1}(\theta^*)}{\partial \theta^*} = \frac{\partial \theta}{\partial \theta^*}$, the inverse of this matrix is:

$$H^{-1} = \left[\frac{\partial \theta}{\partial \theta^*}\right]^{-1} = \frac{\partial \theta^*}{\partial \theta}.$$

The function $h(\theta)$, defined implicitly above can be estimated through the use of simulation. In particular, since $h(\theta) = \theta^*$ is the asymptotic mean of $\widetilde{\theta}$, $h(\theta)$ can be estimated by generating a number of datasets, M, based on θ, which are each then fitted using the chosen starting method, yielding $\widetilde{\theta}_m$, $(m = 1, ..., M)$. The $\widetilde{\theta}_m$ are then averaged to yield an approximation of $h(\theta)$. Under special circumstances, Kuk shows that the final updated estimates calculated using the iterative bias correction method are also the maximum likelihood estimators. An algorithm to implement this method is presented in Figure 5.1.

5.1.2 Standard errors for iterative bias correction

While the method itself is fairly straightforward to implement, the calculation of the standard errors associated with the final bias-corrected estimates of θ is a little more tedious.

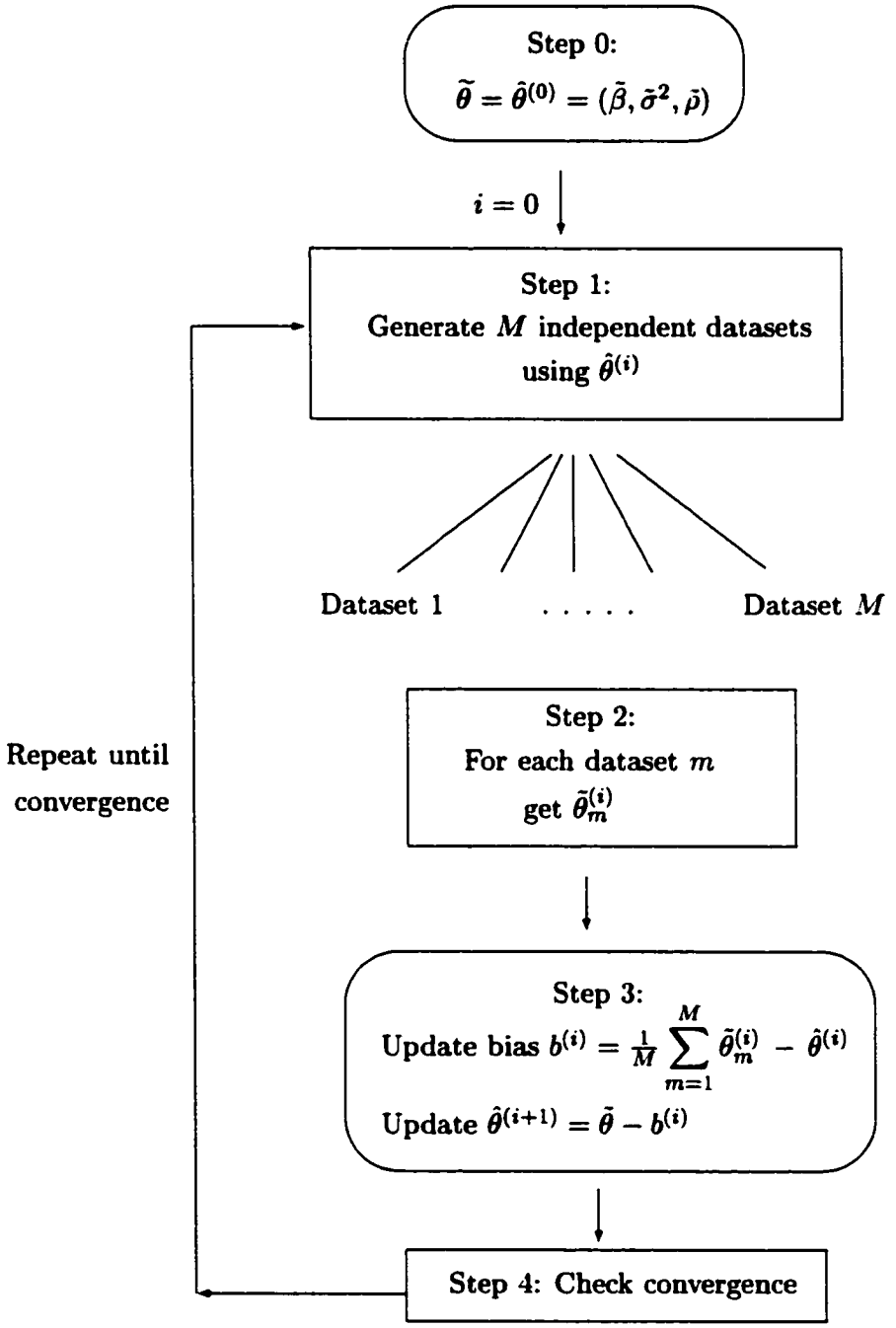

Figure 5.1: Algorithm for the iterative bias correction method.

The asymptotic covariance matrix of $\widehat{\theta}$, $\frac{H\Sigma H^t}{n}$, can be used in the calculation of the standard errors. The form of this matrix is:

$$\frac{H\Sigma H^t}{n} = \left[E_\theta\left\{\psi(\theta^*;Y)S^t(\theta;Y,U)\right\}\right]^{-1} E_\theta\left[\psi(\theta^*;Y)\psi^t(\theta^*;Y)\right] \times \left[\left(E_\theta\left\{\psi(\theta^*;Y)S^t(\theta;Y,U)\right\}\right)^{-1}\right]^t,$$

where $S(\theta;\, Y,U)$ is the vector of the derivatives of the complete-data likelihood function $f(y,u;\,\theta)$ with respect to the components of θ.

5.1.2.1 Derivation of the covariance matrix

The H matrix required as part of the covariance matrix can be derived by expanding the form of the derivative of $E_\theta[\psi(\theta^*;Y)] = 0$. The relationship $\theta^* = h(\theta)$ is defined implicitly by $E_\theta[\psi(\theta^*;Y)] = 0$. Setting the first derivative of $E_\theta[\psi(\theta^*;Y)] = 0$ yields

$$\begin{aligned}
& \frac{d}{d\theta}E_\theta[\psi(\theta^*;Y)] = 0 \\
\Rightarrow \quad & \frac{d}{d\theta}E_\theta[\psi(h(\theta);Y)] = 0 \\
\Rightarrow \quad & \frac{d}{d\theta}\int \psi(h(\theta);y)\, f_\theta(y;\,\theta)\, dy = 0 \\
\Rightarrow \quad & \int \frac{d}{d\theta}[\psi(h(\theta);y)\; f(y;\,\theta)]\, dy = 0 \\
\Rightarrow \quad & \int \left[\frac{d}{d\theta}\psi(h(\theta);y)\right] f_\theta(y;\,\theta)\, dy \;+\; \int \psi(h(\theta);y)\left[\int_u \frac{df(y,u;\,\theta)}{d\theta}\, du\right] dy = 0
\end{aligned}$$

$$\Rightarrow \quad \int \left[\frac{d}{d\theta}\psi(h(\theta);y)\right] f(y;\theta)\,dy \;+\; \int \psi(\theta^*;y)\left[\int_u \frac{df(y,u;\theta)}{d\theta}\,du\right] dy \;=0$$

(since $\theta^* = h(\theta)$)

$$\Rightarrow \quad \int \frac{d\psi(h(\theta);y)}{dh(\theta)}\,\frac{dh(\theta)}{d\theta}\,f(y;\theta)\,dy \;+\int \psi(\theta^*;y)\left[\int_u \frac{df(y,u;\theta)}{d\theta}\,du\right] dy = 0 \;\text{(chain rule)}$$

and since

$$\frac{\partial}{\partial\theta}\ln f_Y(y,u;\theta) \;=\; S^t(\theta;Y,U) \;=\; \frac{\frac{\partial}{\partial\theta}f_{Y,u}(y,u;\theta)}{f(y,u;\theta)}$$

$$\Rightarrow \frac{\partial f(y,u;\theta)}{\partial\theta} \;=\; S^t(\theta;Y,U)\,f(y,u;\theta),$$

we get

$$\int \left[\frac{d}{d\theta^*}\psi(\theta^*;y)\right] H^{-1}\,f_{y;\theta}(y)\,dy \;+\; \int_y\int_u \psi(\theta^*;y)\,S^t(\theta;Y,U)\,f(y,u;\theta)\,du\,dy = 0$$

$$\left(\text{as } H = \frac{dh^{-1}(\theta)}{d\theta} = \frac{1}{\frac{dh(\theta)}{d\theta}}\right)$$

$$\Rightarrow \int \left[\frac{d}{d\theta^*}\psi(\theta^*;y)\right] H^{-1}\,f_\theta(y;\theta)\,dy \;+\int_y\int_u \psi(\theta^*;y)S^t(\theta;Y,U)\,f(y|u;\theta)\,f(u;\theta)\,du\,dy = 0.$$

So $\int \left[\frac{\delta}{\delta\theta}\psi(\theta^*;y)\right] f_\theta(y;\theta)\,dyH^{-1} \;+E_\theta\left[\psi(\theta^*;Y)S^t(\theta;Y,U)\right]=0$

$$\Rightarrow E_\theta[\psi'(\theta^*;Y)]\,H^{-1} \;+\; E_\theta\left[\psi(\theta^*;Y)S^t(\theta;Y,U)\right]=0$$

and the matrix H is

$$H \;=\; -\left[E_\theta\left\{\psi(\theta^*;Y)\,S^t(\theta;Y,U)\right\}\right]^{-1} E_\theta\left[\psi'(\theta^*;Y)\right].$$

The matrix $H\Sigma H^t$ can then be written as

$$
\begin{aligned}
H\Sigma H^t &= \lim_{n\to\infty} \left[-E_\theta \left\{\psi(\theta^*;Y)\, S^t(\theta;Y,U)\right\}\right]^{-1} E_\theta \left[\psi'(\theta^*;Y)\right]\, n\, \left\{E_\theta(\psi'(\theta^*;Y))\right\}^{-1} \times \\
&\quad E_\theta \left[\psi(\theta^*;Y)\psi^t(\theta^*;Y)\right] \left[\left\{E_\theta(\psi'(\theta^*;Y))\right\}^{-1}\right]^t \left[E_\theta\{\psi'(\theta^*;Y)\}\right]^t \times \\
&\quad \left[-\left(E_\theta \left\{\psi(\theta^*;Y)\, S^t(\theta;Y,U)\right\}\right)^{-1}\right]^t \\
&= n \left[E_\theta \left\{\psi(\theta^*;Y)\, S^t(\theta;Y,U)\right\}\right]^{-1} E_\theta \left[\psi(\theta^*;Y)\psi^t(\theta^*;Y)\right] \times \\
&\quad \left[\left(E_\theta \left\{\psi(\theta^*;Y)\, S^t(\theta;Y,U)\right\}\right)^{-1}\right]^t .
\end{aligned}
$$

The covariance matrix of the parameter estimates $\widehat{\theta}$ is:

$$
\begin{aligned}
\frac{H\Sigma H^t}{n} &= \left[E_\theta \left\{\psi(\theta^*;Y) S^t(\theta;Y,U)\right\}\right]^{-1} E_\theta \left[\psi(\theta^*;Y)\psi^t(\theta^*;Y)\right] \times \\
&\qquad \left[\left(E_\theta \left\{\psi(\theta^*;Y) S^t(\theta;Y,U)\right\}\right)^{-1}\right]^t \qquad (5.1)
\end{aligned}
$$

The calculation of the final updated parameter estimates $\widehat{\theta}$ and the asymptotic covariance matrix is done through the use of simulation and is dependent to a large degree on the choice of the method used to get the initial parameter estimates $\tilde{\theta}$. An example of the calculation of the covariance matrix will be described more fully in the section below.

5.2 *Proposed Choice of Initial Estimators in the Iterative Bias Correction Method*

The choice of method used to get the starting values $\tilde{\theta}$ $(= \hat{\theta}^{(0)})$ has primarily been marginal quasi-likelihood (MQL) or PQL (for example, MLnWin provides a choice of MQL or PQL). Both of these methods (which may be inconsistent under certain conditions) are relatively fast (compared to full maximum likelihood methods involving Monte Carlo methods). However, when it is required to perform these methods on M simulated datasets at each iteration, where M may be anywhere between 10 and 1000, the method can be computationally

very intensive (Hoeschele and Tier 1995). Under these circumstances, the method possibly may not provide many advantages over the more well-established full maximum likelihood procedures.

I propose to reduce the computational intensity of the iterative bias correction method by selecting methods for the initial estimates that are computationally very efficient. This will avoid the computational intensity involved when an iterative approach such as PQL is used to find the initial estimators (§5.3.5). In particular, I propose to use iterated weighted least squares (IWLS) to obtain the initial estimates of the β coefficients, and closed-form method of moment equations for the initial estimators of the variance components. The ψ equations for the IWLS procedure for finding the maximum likelihood estimates of $\beta = (\beta_0, \beta_1, .., \beta_5)$ for the generalized linear model $g(\mu) = \eta = X\beta$ are given below. (Recall that the generalized linear model assumes that the data y is independent, and thus does not account for any correlation that may be present in the data.) The ψ equations are

$$\psi(\theta; y) = X^t(y - \mu) = 0,\ j = 1, ..., p.$$

An iterative algorithm that can be used to fit a dataset using IWLS is as follows:

Step (1). Initialize starting values $\beta_0, \beta_1, .., \beta_5$

Set $\eta = g(y)$.

Step (2). Repeat until convergence of the regression coefficients β:

$\mu = g^{-1}(\eta)$

$Z = \eta + (y - \mu)g'(\mu)$

$W = \text{diag}[\{(g'(\mu))^2\, a(\phi)\, V(\mu)\}^{-1}]$

$\beta = (X^tWX)^{-1}X^tWZ$

$\eta = X\beta.$

The method of moments (MoM) is perhaps the oldest method of finding point estimates, dating back to at least Karl Pearson in the late 1800's (Casella and Berger 1990). It has

the virtue of being quite simple to use and almost always yields some sort of estimate; however it is also well-known that these estimates can often be improved upon by some other estimation technique with regard to inference. The method of moment estimators are found by equating the first k sample moments to the corresponding k population moments.

However, for the purposes of finding estimates of the variance components in this bias-reduction method, the equations form an ideal starting estimation procedure, as their solutions, being closed-form, are quickly calculated, and the method is applicable for a wide range of models. Section §5.2.1 below shows comparisons of the time taken to fit a fixed number of datasets based on the polio incidence data, using both PQL and IWLS/MoM methods to find $\tilde{\theta}$.

5.2.1 Estimation of the standard errors for the parameter estimates

The covariance matrix can be estimated by replacing the true parameter values, θ and the asymptotic values of the initial estimators, θ^*, in equation (5.1) above, by $\hat{\theta}$ and $\tilde{\theta}$, respectively. The expectations can be approximated by the use of simulations. More explicitly, the first term and last term can be approximated using:

$$E_\theta\{\psi(\theta^*;Y)\,S^t(\theta;Y,U)\} \;\approx\; \frac{1}{M}\sum_{m=1}^{M}\{\psi^{(m)}(\tilde{\theta};Y)\,S^{t(m)}(\hat{\theta};Y,U)\}$$

and the middle expectation as:

$$E_\theta\left[\psi(\theta^*;Y)\psi^t(\theta^*;Y)\right] \;\approx\; \frac{1}{M}\sum_{m=1}^{M}\{\psi^{(m)}(\tilde{\theta};Y)\,\psi^{(m)\,t}(\tilde{\theta};Y)\},$$

where the mth simulated dataset ($m = 1, ..., M$) is generated using $\widehat{\theta}$, the final (converged) updated parameter estimates. Kuk mentions in his paper that there are some guidelines for the choice of M, the number of datasets used in the calculation of the standard errors, that are mentioned in Diggle and Gratton (1984) and Kuk and Chen (1992). The ψ equations for the variance component(s) will be based on the method of moment equations. (An example can be seen in the next section for the polio data).

Calculation of $S^t(\theta; Y, U)$ is based on the true parameter values, θ, and for the complete-data log-likelihood,

$$l(\theta;\, y, u) = \log f(y \mid u; \theta) + \log f(u; \theta).$$

The formula $S^t(\theta; Y, U)$ for β_j, the jth regression coefficient is:

$$S_j(\theta; Y, U) = X^t(y - \mu). \tag{5.2}$$

where $\mu = e^{X\tilde{\beta}+u}$. The formula $S^t(\theta; Y, U)$ for the variance component(s) can be found similarly by setting the first derivative of $\log f_u(u; \theta) = 0$.

5.3 The Iterative Bias Correction Method for the Polio Incidence Model

5.3.1 IWLS algorithm for estimation of the regression coefficients

The IWLS algorithm is the algorithm used for a generalized linear model with a Poisson link function, $g(\mu) = \log(\mu) = X\beta$, $V(\mu) = \mu$, $a(\phi) = 1$. The ψ equations take the form:

$$\psi_j(\tilde{\theta};\, y) = X^t(y - \tilde{\mu}) = 0 \quad \text{where } \tilde{\mu} = e^{\tilde{\eta}} = e^{X\beta}.$$

5.3.2 Method of moment equations for the variance components

The random effects are assumed to have an AR(1) correlation structure. Consequently, there are two variance components, ρ and σ^2. The method of moment equations for these are (Zeger 1988):

$$\tilde{\sigma}_Z^2 = \frac{\sum_{t=1}^{n} [(y_t - \tilde{\mu}_t)^2 - \tilde{\mu}_t]}{\sum_{t=1}^{n} \tilde{\mu}_t^2}$$

$$\tilde{\rho}_Z = \frac{1}{\tilde{\sigma}_Z^2} \frac{\sum_{t=2}^{n} (y_t - \tilde{\mu}_t)(y_{t-1} - \tilde{\mu}_{t-1})}{\sum_{t=2}^{n} \tilde{\mu}_t \tilde{\mu}_{t-1}}$$

where $\tilde{\mu} = e^{X_t \tilde{\beta}}$. The method of moment equations were derived initially by Zeger (1988) for a marginal model where the errors ϵ_t were assumed to come from a AR(1) distribution with $E(\epsilon_t) = 1$ and $\text{cov}(\epsilon_t, \epsilon_{t+1}) = \sigma^2 \rho_\epsilon(1)$, where ϵ_t is a latent process modelled by an AR(1) process. For Zeger's model, $y_t | \epsilon_t$ is assumed to follow a log linear model, where $E(y_t | \epsilon_t) = e^{X_t' \beta} \epsilon_t$. The marginal variance of the y_t's is $\text{var}(y_t) = \mu_t + \sigma^2 \mu_t^2$.

These equations were later adjusted by Chan & Ledolter (1995), who were interested in fitting the polio incidence data as a random effects model, using a multivariate normal distribution for the random effects with a zero expectation and AR(1) correlation matrix. These adjustments led to the following method of moment estimates for ρ and σ_u^2:

$$\tilde{\sigma}_u^2 = \log(\tilde{\sigma}_Z^2 + 1)$$

$$\tilde{\rho} = \frac{\log(\tilde{\rho}_Z (e^{\tilde{\sigma}_u^2} - 1) + 1)}{\tilde{\sigma}_u^2} = \frac{\log(\tilde{\rho}_Z \tilde{\sigma}_Z^2 + 1)}{\log(\tilde{\sigma}_Z^2 + 1)}$$

with an adjustment also being made to the intercept coefficient, β_0, subtracting $\tilde{\sigma}_u^2$ if it is to be compared to Zeger's estimated β_0. Zeger (1988) notes that these method of moment equations for σ^2 and ρ can lead to negative values of $\tilde{\sigma}^2$ and that $\tilde{\rho}$ is not constrained to the interval $(-1, 1)$.

The score equations for the variance components were calculated from the method of moment equations above. Defining $\tilde{\sigma}_u^2$ as the estimate of σ_u^2 using the initial estimation method,

method of moments, and $\hat{\sigma}_u^2$ as the final updated estimate of σ_u^2 after the iterative bias correction method has been carried out, the ψ equation for σ_u^2, is

$$\psi_{\sigma_u^2}(\tilde{\theta};\, y) \;=\; \tilde{\sigma}_u^2 - \log\,(\tilde{\sigma}_Z^2 + 1) = 0.$$

The ψ equation for ρ can be derived as:

$$\psi_{\rho}(\tilde{\theta};\, y) \;=\; \tilde{\rho} - \frac{\log(\,\tilde{\rho}_z\tilde{\sigma}_Z^2 + 1)}{\log(\,\tilde{\sigma}_Z^2 + 1)} = 0.$$

5.3.3 Estimation of the standard errors for the polio incidence data

Calculation of $S^t(\theta; Y, U)$ for the polio incidence model involved the use of equation (5.2) as displayed in §5.2.1 for β_j, $j = 1, ..., 5$, where $\log \mu = X\beta + u$. The components of $S^t(\theta; Y, U)$ for the variance component(s) can be found similarly from the first derivative of $\log f(u; \theta)$. These equations are based upon σ_ϵ^2 (from the error terms $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ as part of the AR(1) correlation structure of the random effects $u_t = \rho u_{t-1} + \epsilon_t$). Equations for σ_u^2 can be easily obtained using the formula $\sigma_\epsilon^2 = \sigma_u^2(1 - \rho^2)$. After the calculation of the final covariance matrix based on the final iterative bias correction updated estimates, the standard errors found for σ_ϵ^2 can be adjusted to those for σ_u^2 using the multivariate delta theorem. As also used in the MCEM algorithm used to find the exact maximum likelihood estimates, the formulae for the components of the score vectors are:

$$S^t_{\sigma_\epsilon^2}(\theta; Y, U) \;=\; -\frac{n}{2\sigma_\epsilon^2} + \frac{u_1^2(1-\rho^2)}{2\sigma_\epsilon^2} + \frac{1}{\sigma_\epsilon^2}\sum_{i=2}^{n}(u_i - \rho\, u_{i-1})^2,$$

$$S^t_{\rho}(\theta; Y, U) \;=\; -\frac{\rho}{1-\rho^2} + \frac{u_1^2\rho}{\sigma_\epsilon^2} + \sum_{i=2}^{n}\rho\,(u_{i-1} + u_i u_{i-1})^2.$$

As the number of variance components is $r = 2$, the form of the $(p+2) \times (p+2)$ matrix ψ' is:

$$\psi'(\theta^*;Y) = \begin{pmatrix} \frac{d\psi_0(\theta^*;y)}{d\beta_0^*} & \frac{d\psi_1(\theta^*;y)}{d\beta_0^*} & \cdots & \frac{d\psi_p(\theta^*;y)}{d\beta_0^*} & 0 & 0 \\ \frac{d\psi_0(\theta^*;y)}{d\beta_1^*} & \frac{d\psi_1(\theta^*;y)}{d\beta_1^*} & \cdots & \frac{d\psi_p(\theta^*;y)}{d\beta_1^*} & 0 & 0 \\ & \cdots & & \cdots & & \\ \frac{d\psi_0(\theta^*;y)}{d\beta_p^*} & \frac{d\psi_1(\theta^*;y)}{d\beta_p^*} & \cdots & \frac{d\psi_p(\theta^*;y)}{d\beta_p^*} & 0 & 0 \\ 0 & 0 & 0 & \cdots & \frac{d\psi_{(p+1)}(\theta^*;y)}{d\sigma_u^{2*}} & \frac{d\psi_{(p+2)}(\theta^*;y)}{d\sigma_u^{2*}} \\ 0 & 0 & 0 & \cdots & \frac{d\psi_{(p+1)}(\theta^*;y)}{d\rho^*} & \frac{d\psi_{(p+2)}(\theta^*;y)}{d\rho^*} \end{pmatrix}$$

5.3.4 Simulation study for the estimation of the standard errors

To check out how many datasets are required to get convergence of the standard error estimates, a small simulation study was carried out based on a randomly generated dataset of sample size $n = 250$, where the true values of the regression coefficients were $(\beta_0, ..., \beta_5) = (1, 0, 0, 0, 0)$ and $(\rho, \sigma_u^2) = (0.25, 0.5)$. For varying M in steps of 1000, the standard errors were found for a fixed $\hat{\theta}$ (based on the polio data model). This was repeated for five different random seed values. The plots for ρ, σ^2 and β_0 are presented below in Figure 5.2. These plots suggest that the standard errors for the regression coefficients require approximately $M = 15000$ datasets for stability, while the variance components require at least 10000 datasets. For the polio incidence model simulations carried out in this chapter, $M = 15000$ was used.

5.3.5 Gains in computational efficiency using the iterative bias correction method

It is of particular interest to investigate the potential gains in computational efficiency of using IWLS and MoM for the initial estimation of the regression coefficients and variance components respectively, as compared to PQL. This choice of using PQL or IWLS/MoM was tested for a range of sample sizes, with the variance components $(\rho, \sigma^2) = (0.25, 1.0)$. For each sample size, five datasets based on the polio incidence dataset were generated and

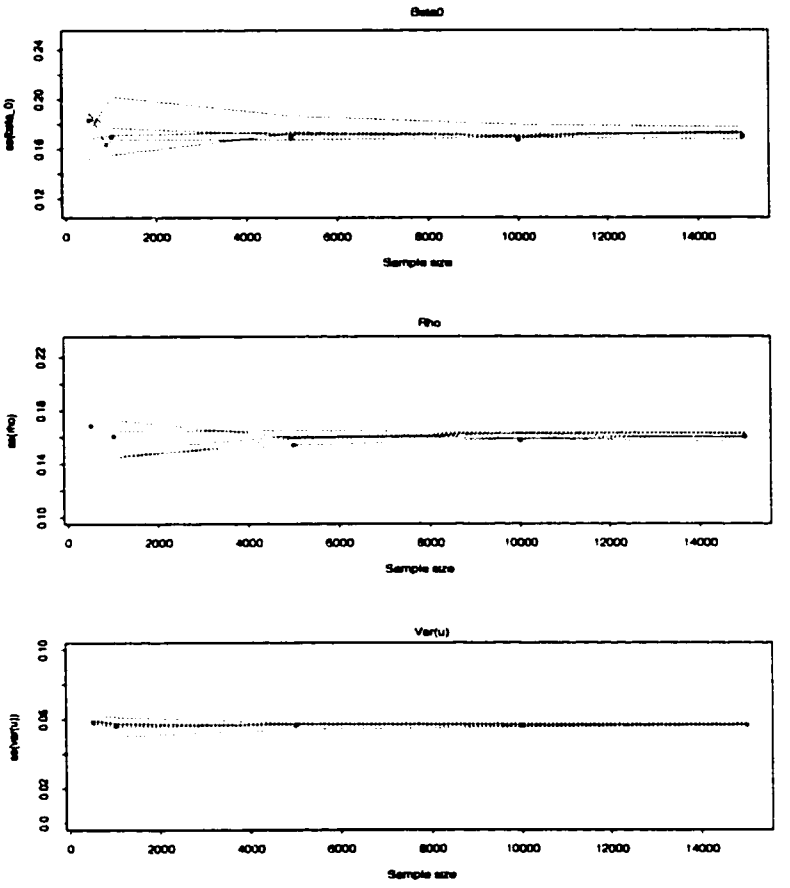

Figure 5.2: Plot showing the convergence of the standard errors for one dataset ($n = 250$) using the iterative bias correction method.

fitted to convergence using the two initial estimation procedures. Their average CPU time to convergence was calculated using a COMPAQ AlphaServer DS10 466 MHz computer with 1 Gigabyte memory. A significant time reduction was found in all situations by using the IWLS/MoM method.

Table 5.1: Average time taken (in minutes) to fit a dataset based on the polio incidence model using the iterative bias correction method with both PQL and IWLS/MoM as the initial estimation methods.

Sample size	IBC (PQL)	IBC (IWLS/MoM)
$n = 50$	309	8
$n = 100$	574	29
$n = 250$	6931	167

5.3.6 Simulation studies setup and performance criteria

In an identical manner to both exact maximum likelihood and approximate maximum likelihood estimators, simulation studies were carried out to evaluate the performance of the iterative bias correction estimators for the polio incidence model. Full details can be found in §3.2.2, but a summary of the model will be provided below.

The form of the model used to simulate the data is $\log \mu_t = x_t^t\beta + u_t$. The random effects follow an AR(1) process $\{u_t\}$ satisfying:

$$u_t = \rho u_{t-1} + \epsilon_t, \qquad \text{where} \quad \rho \text{ is the correlation coefficient,}$$
$$\epsilon_t \sim iid \text{ Normal } (0, \sigma_\epsilon^2),$$
$$y = 1, 2, ..., n,$$

and the y_t's are conditionally independent, i.e. $y_t \mid u_t \sim \text{Poisson}(\mu_t^u)$.

The trend and seasonality in the data were modelled with linear and trigonometric components, with the covariate vector

$$x_t^t = \left(1, \frac{t}{1000}, \cos\left(\frac{2\pi t}{12}\right), \sin\left(\frac{2\pi t}{12}\right), \cos\left(\frac{2\pi t}{6}\right), \sin\left(\frac{2\pi t}{6}\right)\right).$$

For each scenario (each combination of sample size n, correlation coefficient ρ and variance σ^2) a set of 200 simulations were carried out. The datasets were generated from the popu-

lation model based on the true parameter values by first generating a set of random effects u_i from MVN(0,D) where D has an AR(1) correlation structure.

Table 5.2: Parameter values used in simulation studies.

Parameter	Range of values examined
n	50, 100, 250
ρ	0, 0.25, 0.5, 0.75
σ_ϵ^2	0, 0.5, 1.0, 2.5
β	(1, 0, 0, 0, 0, 0)

Strict convergence criteria were used within the iterative bias correction computer program. The convergence criterion used in the estimation of the β coefficients within each IBC iteration was:

$$\Delta = |\hat{\beta}_{new} - \hat{\beta}_{old}| < 0.00001.$$

Similar convergence criteria were used in the estimation of the variance components.

5.3.7 Results

The iterative bias correction method used with IWLS/MoM led to parameter estimates that on average were unbiased and had mean-squared errors that were generally comparable to maximum likelihood and PQL, while being significantly less computationally intensive than methods currently available for maximum likelihood estimation. However, some convergence problems with the two variance components, ρ and σ_u^2, were observed in smaller sample sizes or in the presence of extreme values of the variance components, or both, as described in more detail below. Tables 5.3 – 5.4 present the average estimates of the parameters for a range of variance component combinations and Tables 5.5 – 5.6 show the theoretical and observed standard errors for the parameter estimates in each simulation scenario. Mean-squared errors (m.s.e's) for the parameter estimates are presented in Tables 5.7 – 5.8.

Convergence issues

The iterative bias correction method used with the IWLS/MoM was very stable and provided high proportions of converged datasets (greater than 98%) when the sample size was large at $n = 250$ and the variance component σ_u^2 was small (less or equal to 0.5) and for all values of ρ tested, i.e., 0 to 0.75. When the true value of the variance σ_ϵ^2 took the value 2.5, a lower proportion of converged datasets was observed (between 70% and 80%).

However, for the smallest sample size tested, at $n = 50$, the iterative bias correction method used with IWLS/MoM proved to be less stable. A lower proportion of converged datasets for most of the variance component combinations was observed, except for the smaller values of ρ and σ_u^2 (between 70% when $\rho = 0.75$ and $\sigma_u^2 = 5.714$ and 98% when $\rho = 0.25$ and $\sigma_u^2 = 0.5$). The lack of convergence was mainly due to the variance component σ_u^2 not converging. The other variance component, ρ, was much more stable and led to significantly fewer non-convergence issues. The method of moment estimators did not restrict the estimation of ρ to be within the bounds of -1 and 1. This led to occasional overestimation and underestimation of ρ, especially for a sample size of 50. The regression coefficients, β_0, β_1, ..., β_5, while influenced by the lack of convergence of σ_u^2 (in particular, β_0), converged with little difficulty in every dataset.

The convergence proportions showed a significant improvement with an increased sample size of 100 (greater than 95%) for all variance component combinations, except in the most extreme situations in which the convergence proportions were between 80% when $\rho = 0.75$ and $\sigma_u^2 = 5.714$, and 98% when $\rho = 0.25$ and $\sigma_u^2 = 0.5$.

With an increased sample size ($n = 250$), the convergence issues that plagued the smaller samples disappeared almost entirely, with all the convergence rates being above 90% except in the case of $\rho = 0.75$ and $\sigma_u^2 = 5.714$, where the convergence rates fell to around 75%. High ρ values, except where combined with a high σ_u^2 value, led to minor convergence issues for a sample size of 250.

For those simulation scenarios which suffered from high non-convergence (above 10%), the results have been omitted from the tables of results. In particular, many of the scenarios with $\rho \geq 0.5$ and $\sigma_u^2 > 1.7$ for a sample size of $n = 50$ were excluded, and for all sample sizes, the simulation scenario for $\rho = 0.75$ and $\sigma_u^2 = 5.714$ was excluded.

None of the estimates of σ_u^2 were less than 0.001 when the true value was zero for any of the sample sizes. However, over three-quarters of the estimates of σ_u^2 were less than 0.1 for a sample size of $n = 250$.

As the observed standard errors of the parameter estimates were also calculated through the use of simulation, these were also prone to non-convergence issues. The standard error calculations for the regression coefficients were very stable, though a little less so for β_1, while some non-converged standard errors were observed for the variance components. These were removed from the summaries of the observed standard errors. The proportion of non-converged standard errors of σ^2 was less than 10% except for the more extreme values of ρ and σ^2 where the non-convergence proportion rose to around 30%. This suggests that the method of simulation used in calculating the standard errors for the iterative bias correction method is reliable for small to moderate values of the variance components, but not so for more extreme values of the variance components.

The regression coefficient β_0

The intercept β_0 was estimated on average with very little bias (less than 3.5%) for all the simulation scenarios presented in the tables (reasons for omitted results are described in the previous section. These small levels of bias were observed for all the sample sizes tested, and no noticeable improvements on already small biases were seen with larger sample sizes. The corresponding standard errors for the estimates were moderate in value, and decreased to approximately half of their values with a change in sample size from 50 to 250, as expected from the asymptotic theory.

The regression coefficient β_1

The regression coefficient accounting for the time trend, β_1, was estimated on average with minimal bias in all simulation scenarios presented in the tables. Higher levels of variability than seen for the other regression coefficients were present for all sample sizes. The level of variability increased with larger values of ρ and σ_u^2.

The other regression coefficients

In a similar manner to both β_0 and β_1, the remaining regression coefficients $\beta_2, ..., \beta_5$ were also estimated with very little bias and low levels of variability for all sample sizes and simulation scenarios presented in the tables. Results for β_4 are displayed in the tables, and are representative of those for all of these other regression coefficients.

The variance components

The correlation coefficient ρ was estimated with very little bias on average for moderate values of σ_u^2 (its estimation is redundant for $\sigma_u^2 = 0$). The biases in the estimates of ρ ranged from -5% to 4.5%. The variability present in the two hundred estimates in any particular simulation scenario predictably decreased with an increase in sample size. Some outliers were present, especially for a sample size of 50. Figure 5.3 presents boxplots showing the estimates of ρ for increasing ρ and a fixed $\sigma_\epsilon^2 = 0.5$.

The other variance component, σ_u^2, was also estimated on average with little bias (less than 5%). Again, as with the estimation of ρ, there were outliers present, especially at a small sample size of 50. As the value of σ_u^2 increased, while keeping ρ fixed, the variability of the estimates increased. Figure 5.4 displays this effect for $\rho = 0.25$ and varying σ_u^2.

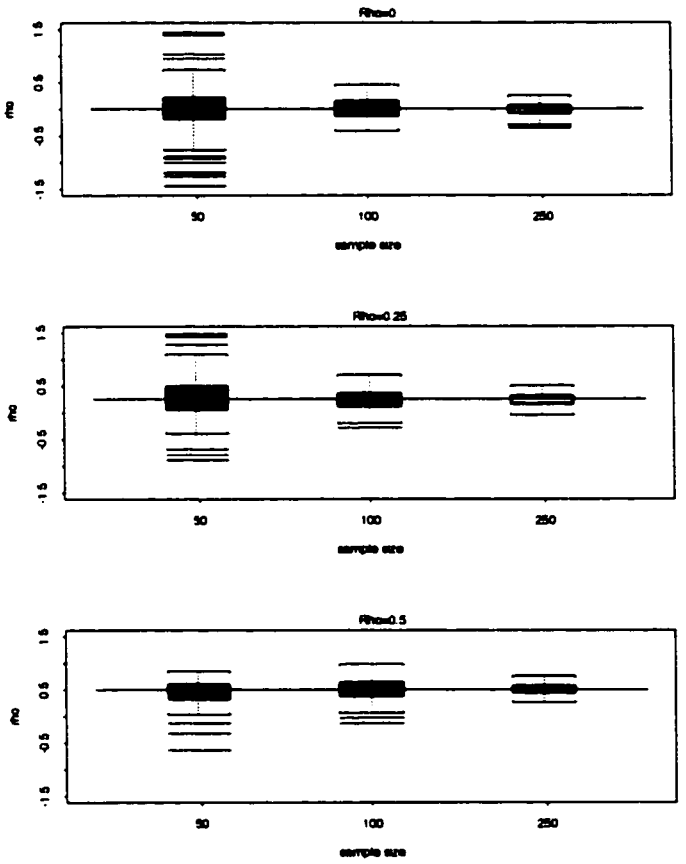

Figure 5.3: Estimation of ρ for increasing ρ and a fixed $\sigma_\epsilon^2 = 0.5$.

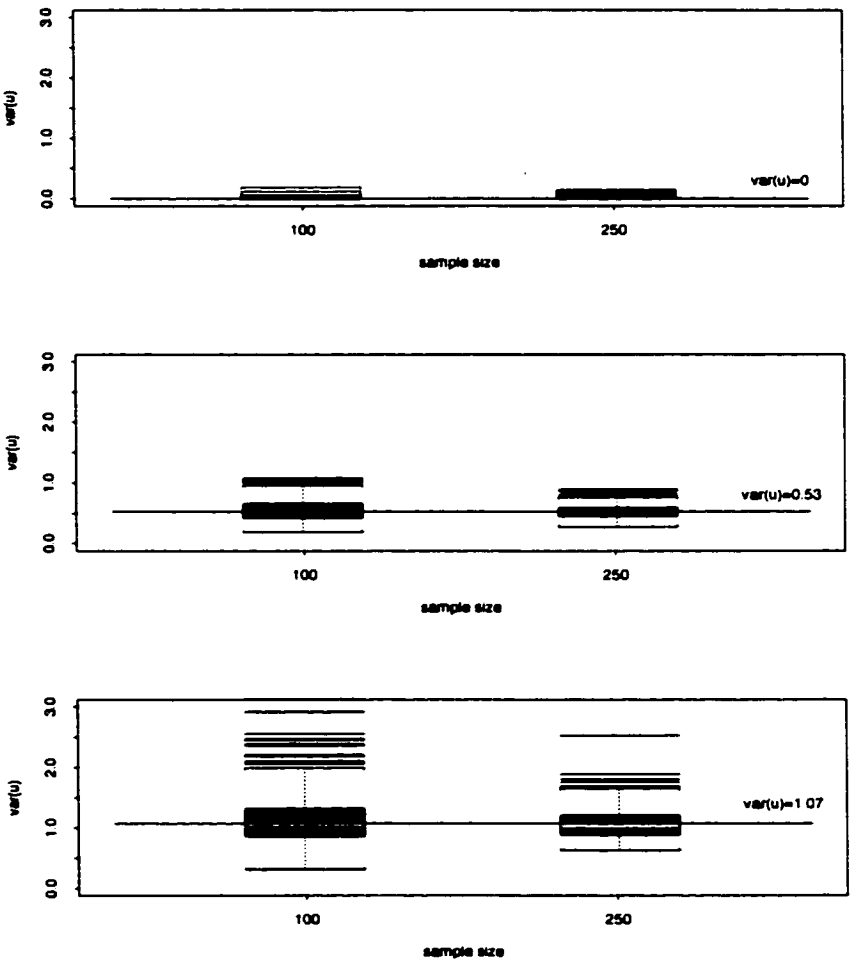

Figure 5.4: Estimation of σ_u^2 for a fixed $\rho = 0.25$.

Table 5.3: Iterative bias correction average estimated parameter values: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Average estimated parameter values over 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
true value	N/A	0	1	0	0
50	0.532	0.146	0.944	0.070	-0.005
100	0.320	0.093	0.966	-0.123	0.000
250	0.170	0.076	0.942	0.086	0.006
true value	0	0.5	1	0	0
50	0.002	0.531	0.997	-0.331	-0.025
100	0.007	0.517	0.998	-0.181	-0.011
250	-0.014	0.495	1.001	-0.086	-0.009
true value	0.25	0.533	1	0	0
50	0.278	0.573	1.015	-0.651	-0.037
100	0.239	0.549	0.991	0.002	-0.018
250	0.237	0.529	0.998	-0.094	-0.012
true value	0.5	0.667	1	0	0
50	0.463	0.678	1.067	-2.342	-0.029
100	0.493	0.733	0.973	0.185	-0.028
250	0.499	0.662	1.003	-0.146	-0.017
true value	0.75	1.143	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.750	1.222	1.000	-0.402	-0.014
true value	0	1	1	0	0
50	0.022	0.965	0.993	0.580	-0.048
100	0.006	1.058	0.972	0.155	-0.026
250	-0.003	0.995	1.012	-0.182	-0.010
true value	0.25	1.067	1	0	0
50	0.243	1.028	0.999	0.323	-0.039
100	0.261	1.166	0.969	-0.009	-0.044
250	0.246	1.053	1.016	-0.240	-0.017

Table 5.4: Iterative bias correction average estimated parameter values: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Average estimated parameter values over 200 simulations				
	ρ	σ^2	β_0	β_1	β_4
true value	0.5	1.333	1	0	0
50	–	–	–	–	–
100	0.485	1.322	1.021	-0.292	-0.047
250	0.499	1.378	1.004	-0.323	-0.016
true value	0.75	2.286	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.737	2.769	0.858	-0.491	0.000
true value	0	2.5	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.019	2.559	1.024	-0.491	-0.018
true value	0.25	2.667	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.243	2.754	1.129	-1.251	-0.039
true value	0.5	3.333	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.518	3.432	1.115	-1.149	-0.108
true value	0.75	5.714	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	–	–	–	–	–

Table 5.5: Standard errors for the iterative bias correction method: the first row in each section describes the true values of the parameters for that set of simulations.

Sample	Theoretical (observed) standard errors over 200 simulations				
size n	ρ	σ_u^2	β_0	β_1	β_4
Param.	N/A	0	1	0	0
50	1.088 (1.580)	0.058 (0.063)	0.182 (0.305)	6.712 (9.840)	0.124 (0.138)
100	1.035 (1.136)	0.035 (0.054)	0.121 (0.186)	1.988 (2.852)	0.082 (0.097)
250	1.026 (1.114)	0.033 (0.022)	0.084 (0.114)	0.567 (0.662)	0.060 (0.070)
Param.	0	0.5	1	0	0
50	0.407 (0.537)	0.251 (0.514)	0.294 (0.410)	9.698 (10.254)	0.200 (0.204)
100	0.189 (0.236)	0.169 (0.279)	0.215 (0.213)	3.570 ()	0.127 (0.141)
250	0.116 (0.126)	0.100 (0.121)	0.130 (0.129)	0.848 ()	0.082 (0.086)
Param.	0.25	0.533	1	0	0
50	0.344 (0.595)	0.296 (0.433)	0.341 (0.521)	11.820 (12.201)	0.203 (0.215)
100	0.189 (0.288)	0.180 (0.238)	0.245 (0.316)	4.281 (4.067)	0.150 (0.151)
250	0.115 (0.159)	0.102 (0.136)	0.153 (0.173)	1.009 (1.006)	0.090 (0.096)
Param.	0.5	0.667	1	0	0
50	0.221 (0.383)	0.274 (0.342)	0.451 (0.512)	16.095 ()	0.208 (0.252)
100	0.186 (0.307)	0.355 (0.170)	0.347 (0.389)	5.990 (5.824)	0.158 (0.167)
250	0.103 (0.176)	0.137 (0.099)	0.204 (0.217)	1.393 (1.470)	0.093 (0.100)
Param.	0.75	1.143	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.088 (0.225)	0.212 (0.151)	0.392 (0.431)	2.649 (2.806)	0.105 (0.130)
Param.	0	1	1	0	0
50	0.250 (0.276)	0.340 (0.379)	0.369 (0.425)	12.877 ()	0.234 (0.289)
100	0.188 (0.316)	0.381 (0.616)	0.289 (0.514)	4.894 (5.696)	0.190 (0.229)
250	0.109 (0.199)	0.226 (0.315)	0.196 (0.420)	1.211 (1.223)	0.113 (0.132)
Param.	0.25	1.067	1	0	0
50	0.286 (0.462)	0.560 (0.619)	0.447 (0.652)	15.803 (17.097)	0.272 (0.319)
100	0.181 (0.316)	0.497 (0.429)	0.361 (0.495)	6.055 (5.952)	0.207 (0.244)
250	0.112 (0.176)	0.261 (0.255)	0.230 (0.277)	1.435 (1.563)	0.125 (0.141)

Table 5.6: Standard errors for the iterative bias correction method: the first row in each section describes the true values of the parameters for that set of simulations.

Sample	Theoretical (observed) standard errors over 200 simulations				
size n	ρ	σ_u^2	β_0	β_1	β_4
Param.	0.5	1.333	1	0	0
50	–	–	–	–	–
100	0.163 (0.216)	0.461 (0.403)	0.440 (0.523)	8.195 ()	0.228 (0.266)
250	0.114 (0.233)	0.420 (0.359)	0.296 (0.447)	2.047 (2.643)	0.135 (0.196)
Param.	0.75	2.286	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.104 (0.378)	0.782 (0.738)	0.557 (0.642)	4.092 (4.812)	0.151 (0.219)
Param.	0	2.5	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.152 (0.304)	0.788 (0.824)	0.432 (0.678)	2.298 (2.374)	0.212 (0.242)
Param.	0.25	2.667	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.149 (0.243)	0.766 (0.508)	0.453 (0.582)	2.985 (3.504)	0.227 (0.414)
Param.	0.5	3.333	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.184 (0.314)	0.879 (0.419)	0.540 (0.626)	3.612 (3.995)	0.252 (0.317)
Param.	0.75	5.714	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	–	–	–	–	–

Table 5.7: Iterative bias correction mean-squared errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Iterative bias correction mean-squared errors based on 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
Param.	N/A	0	1	0	0
50	1.467	0.025	0.036	38.094	0.015
100	1.173	0.010	0.015	4.003	0.007
250	1.055	0.006	0.009	0.329	0.004
Param.	0	0.5	1	0	0
50	0.166	0.064	0.086	94.161	0.041
100	0.036	0.300	0.045	12.780	0.016
250	0.014	0.256	0.017	0.726	0.007
Param.	0.25	0.533	1	0	0
50	0.119	0.089	0.117	140.125	0.043
100	0.036	0.033	0.060	18.326	0.023
250	0.013	0.030	0.024	1.033	0.008
Param.	0.5	0.667	1	0	0
50	0.050	0.075	0.207	264.544	0.044
100	0.035	0.053	0.121	34.845	0.026
250	0.011	0.019	0.042	1.962	0.009
Param.	0.75	1.143	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.008	0.051	0.311	7.181	0.011
Param.	0	1	1	0	0
50	0.063	0.117	0.136	166.160	0.057
100	0.035	0.149	0.084	23.978	0.037
250	0.012	0.051	0.039	1.500	0.013
Param.	0.25	1.067	1	0	0
50	0.082	0.322	0.200	250.867	0.076
100	0.033	0.257	0.131	36.660	0.045
250	0.013	0.068	0.053	2.119	0.016

Table 5.8: Iterative bias correction mean-squared errors: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Iterative bias correction mean-squared errors based 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
Param.	0.5	1.333	1	0	0
50	-	-	-	-	-
100	0.027	0.212	0.194	67.248	0.054
250	0.014	0.178	0.089	4.292	0.019
Param.	0.75	2.286	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.011	0.612	0.311	16.751	0.023
Param.	0	2.5	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.024	0.625	0.187	5.520	0.045
Param.	0.25	2.667	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.022	0.594	0.222	10.473	0.053
Param.	0.5	3.333	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	0.034	0.783	0.304	14.363	0.075
Param.	0.75	5.714	1	0	0
50	–	–	–	–	–
100	–	–	–	–	–
250	–	–	–	–	–

Chapter 6

COMPARISON OF IBC, MAXIMUM LIKELIHOOD AND PQL FOR THE POLIO INCIDENCE MODEL

Chapters 3 to 5 presented the results for simulation studies, investigating the performance of three methods available (amongst others) for fitting generalized linear mixed models. The particular model of interest was a log linear model based on polio incidence data in the USA from 1970–1983. In this chapter, the results for the three methods are compared, with special regard to bias and mean-squared error. The bias is calculated as the average estimate from the two hundred converged datasets minus the true value for the parameter. The mean-squared error is calculated as the addition of squared bias and variance, where the variance is the sample variance of the two hundred estimates for the parameter of interest. Full details for the mean-squared errors calculated for each simulation scenario for each of the three methods are presented in Tables 6.1 - 6.5.

Computational efficiency

Of the three methods, maximum likelihood was significantly the most computationally intensive, and PQL was the most computationally efficient method. The iterative bias correction method falls somewhere in the middle. Figure 6.1 shows the average time taken (in minutes) to fit a dataset for the sample sizes $n = 50$, 100, and 250, where $\rho = 0.25$ and $\sigma_u^2 = 1.067$. The times are based on an average of the time taken to fit five different datasets, where time is measured as the CPU time taken on a COMPAQ AlphaServer DS10 466 MHz compute with 1 Gigabyte memory.

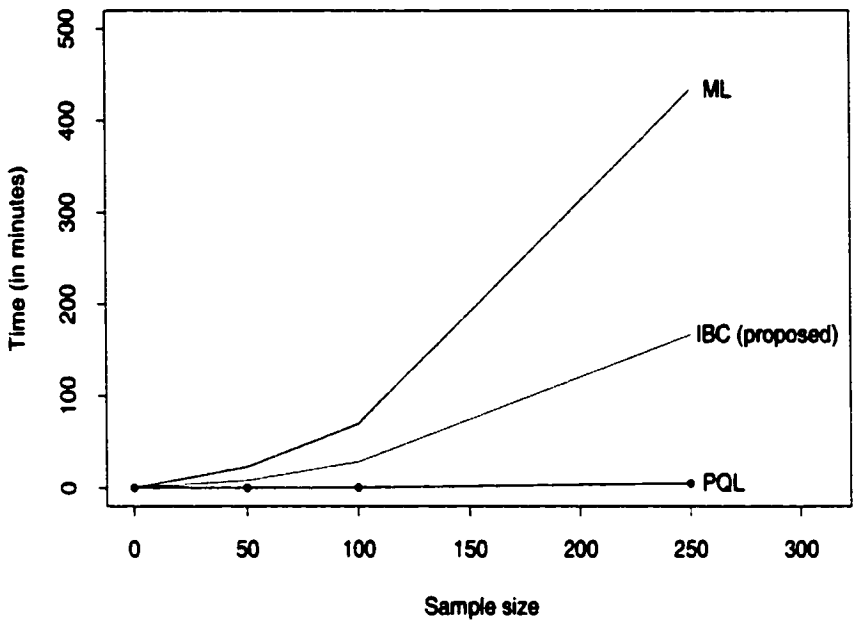

Figure 6.1: Time taken (in minutes) to fit a dataset using the three methods for $\rho = 0.25$ and $\sigma_u^2 = 1.07$.

The regression coefficient β_0

The iterative bias correction method estimated the intercept β_0 with little or no bias, while PQL and maximum likelihood overestimated β_0 for extreme values of ρ and σ^2. For a sample size of 250 (the largest sample size examined for all three methods), the biases for the average estimates of β_0 using maximum likelihood fluctuated from around -5% to 25%, while the estimates of β_0 using PQL and iterative bias correction had significantly less bias.

The mean-squared errors of β_0 decreased more rapidly for the estimates of β_0 calculated using maximum likelihood than for the corresponding mean-squared errors for PQL and IBC. Figure 6.2 clearly displays this trend for $\rho = 0.25$ and $\sigma_u^2 = 0.533$.

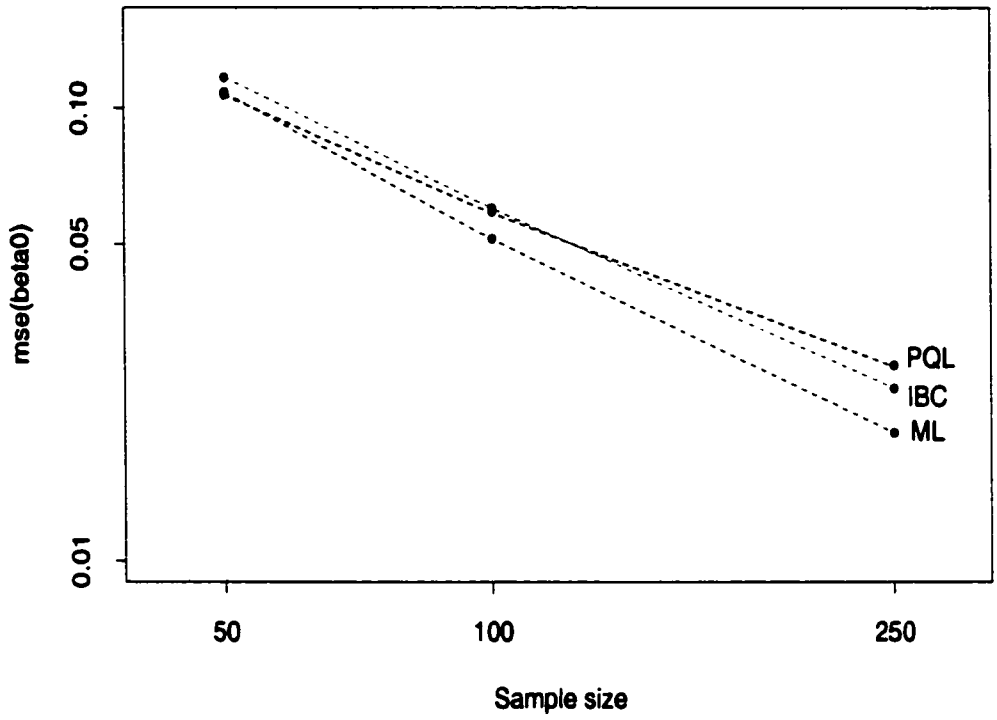

Figure 6.2: Mean-squared errors of β_0 for maximum likelihood, PQL and IBC where $\rho = 0.25$ and $\sigma_u^2 = 0.553$.

The mean-squared errors ranged from approximately 0.05–2.5 for the smallest sample size of 50, and 0.007–0.6 for the largest sample size of 250, showing the predictable decrease in mean-squared errors as the sample size increases.

Both PQL and maximum likelihood displayed increasing mean-squared errors for β_0 as either or both σ_u^2 and ρ increased. In contrast, the mean-squared errors for β_0 calculated using IBC were unaffected by changes in ρ and σ_u^2, although the mean-squared errors were slightly larger than those for both PQL and maximum likelihood. Figure 6.2 displays the mean-squared error of β_0 for varying σ_u^2 for a fixed $\rho = 0.5$. A similar trend was also

observed for the other values of ρ as displayed in the results presented in Tables 6.5 – 6.8.

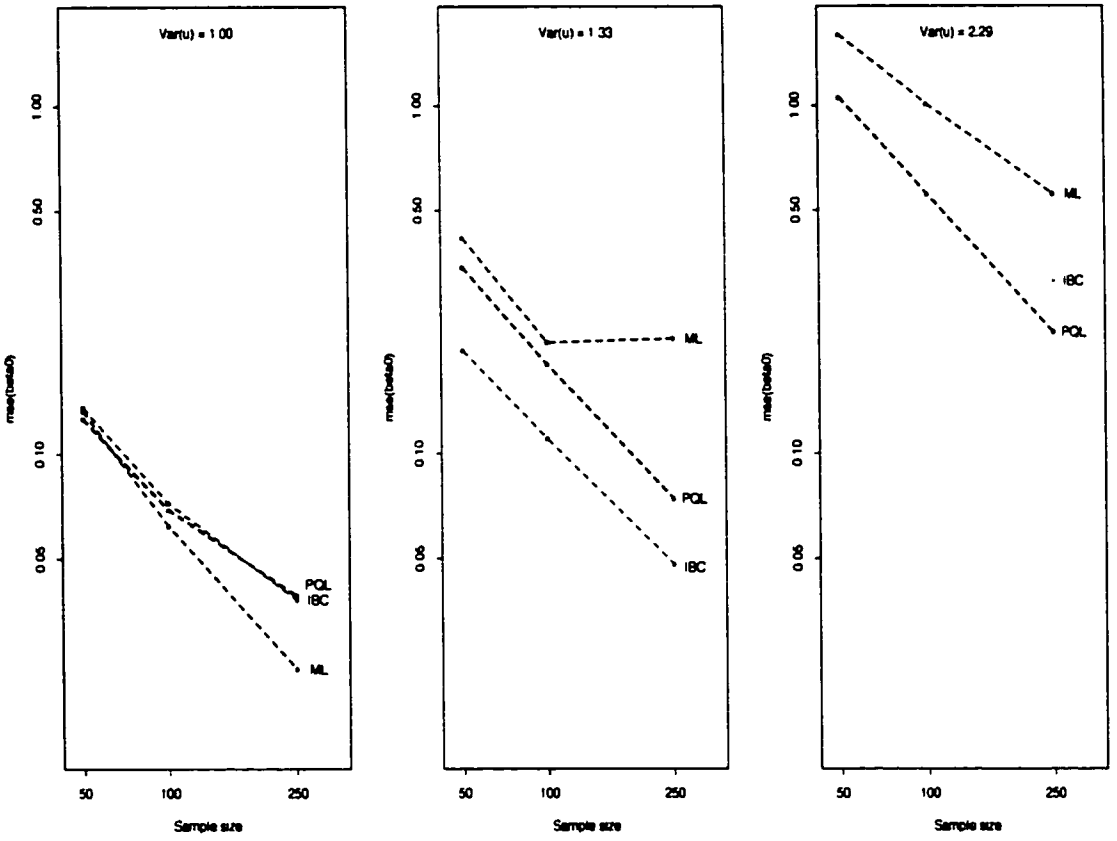

Figure 6.3: Mean-squared errors of β_0 for maximum likelihood, PQL and IBC for fixed $\rho = 0.25$ and varying σ_u^2.

Overall, maximum likelihood performed better, with lower mean-squared errors, when the variance σ_u^2 was small, while PQL performed better when the value of σ_u^2 was higher. IBC also performed in a manner comparable with PQL.

The regression coefficient β_1

Unlike the other regression coefficients, all three methods estimated β_1, the regression coefficient describing the time trend $\frac{t}{1000}$, with little bias and high variability. Boxplots in Figure 6.4. are presented of the two hundred estimates of β_1 for $\sigma_u^2 = 0.533$ and $\rho = 0.25$.

The high variability in the estimates was reflected in the significantly higher mean-squared errors of β_1 for each of the three methods in all the variance component combinations, although the mean-squared errors decreased substantially with increasing sample size.

In a similar manner to β_0, the mean-squared errors for all three methods' estimates of β_1 increased in value as the variance components, ρ and σ_u^2 increased.

The other regression coefficients

Each of the three methods estimated the remaining regression coefficients, β_2 to β_5, with very little or no bias on average, and the mean-squared errors for each of the three methods were almost identical for each combination of variance components, ranging in size from 0.007 – 0.29. A small decrease was observed in the mean-squared errors as the sample size increased for all three methods in a similar manner to the other regression coefficients. Mean-squared errors for β_4 are displayed in Tables 6.1 – 6.4 and reflect the mean-squared errors for the other regression coefficients, β_2, β_3, and β_5. All three methods would be recommended for the estimation of these regression coefficients for any of the sample sizes and variance component combinations tested.

The variance components

Both PQL and IBC provided the best estimation of the variance component ρ in terms of having the smallest bias for all sample sizes and combinations of the variance components. The estimates of ρ found using maximum likelihood exhibited large negative bias. This

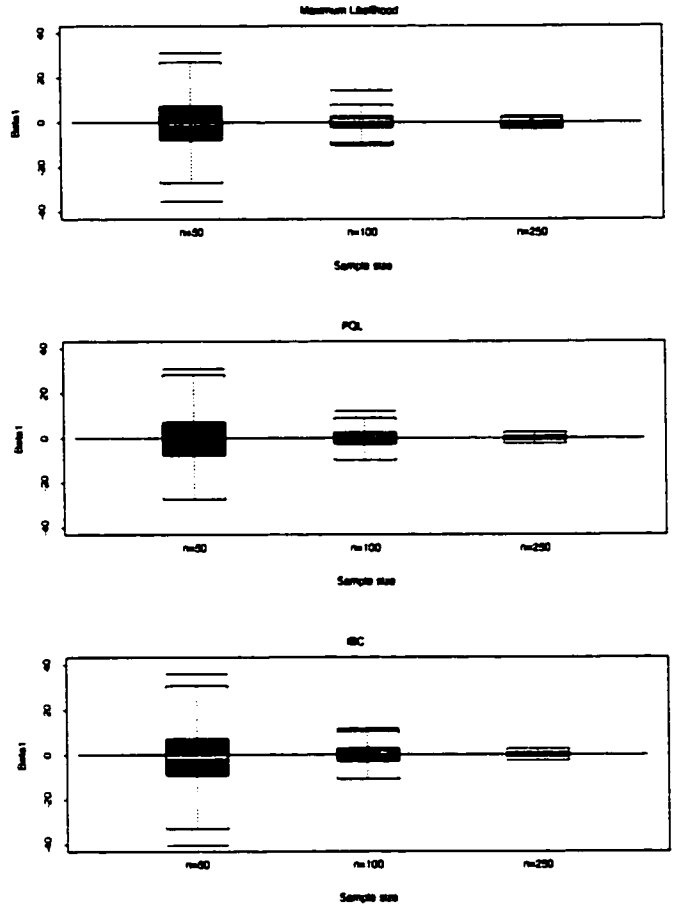

Figure 6.4: Boxplots of β_1 for maximum likelihood, PQL and IBC for $\sigma_u^2 = 0.533$ and fixed $\rho = 0.25$.

improved somewhat, but did not disappear entirely, when the sample size reached $n = 250$ (range: from -80% to -65% at sample size of 50; -30% to 2% at sample size of 250). The negative bias also improved as σ_u^2 got larger.

Iterative bias correction tended to have slightly larger mean-squared errors in the estimation of ρ, with PQL having the next largest mean-squared errors when ρ was smaller, and maximum likelihood having the next largest mean-squared errors when ρ was large. This effect is shown clearly in Figure 6.5 where the mean squared errors for ρ for both PQL and

ML are shown for increasing ρ. The sample size chosen was $n = 250$ and σ_ϵ^2 was fixed at 1 (i.e. the corresponding $\sigma_u^2 = 1, 1.07, 1.33, 2.29$).

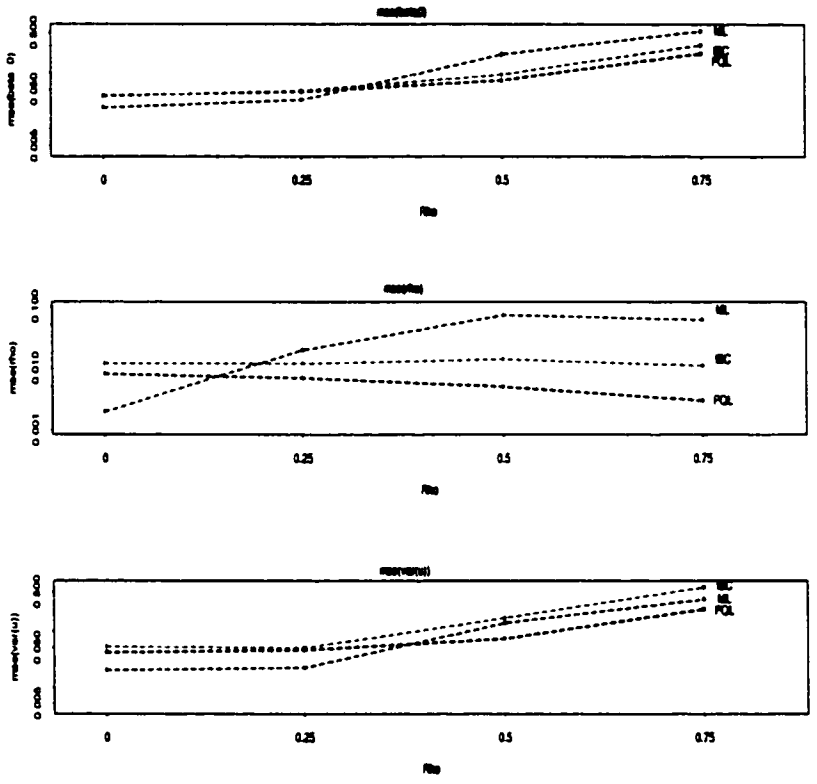

Figure 6.5: Mean-squared errors of $\hat{\rho}$ for IBC, PQL, and maximum likelihood for varying ρ where $\sigma_\epsilon^2 = 1$ and $n = 250$.

This can be explained easily by the mean-squared error formula, which is calculated as the addition of squared bias and variance of the two hundred estimates. The PQL estimates of ρ, while consistently displaying very little bias, exhibited high variability for $\rho = 0$ and 0.25 which decreased as ρ reached 0.5 and 0.75. Consquently the mean-squared errors for ρ became smaller when the true value of ρ was higher. On the other hand, the maximum likelihood estimates of ρ displayed only small bias at the lower levels of ρ which became larger as ρ increased. This contributed in a squared manner to the larger mean-squared errors observed for the estimates of ρ where the true value of ρ was large. Figure 6.6 shows

the breakdown of the mean-squared errors for both PQL and maximum likelihood into the variance and squared bias components. It is important to note that as ρ increased, σ_u^2 also increased and it was difficult to separate the effects of an increasing ρ with that of an increasing σ_u^2, as both occurred simultaneously. However to observe the effect of the variance, σ_u^2, alone, Figure 6.7 shows that for a fixed ρ-value of 0.5 and a sample size of $n = 250$ the mean-squared errors of ρ for both PQL and maximum likelihood decreased as σ_u^2 became larger. This attributes the increasing mean-squared errors of the maximum likelihood estimation of ρ observed in Figure 6.5 to the additional bias seen as ρ increases in value. A similar decreasing trend was also observed for other fixed values of ρ.

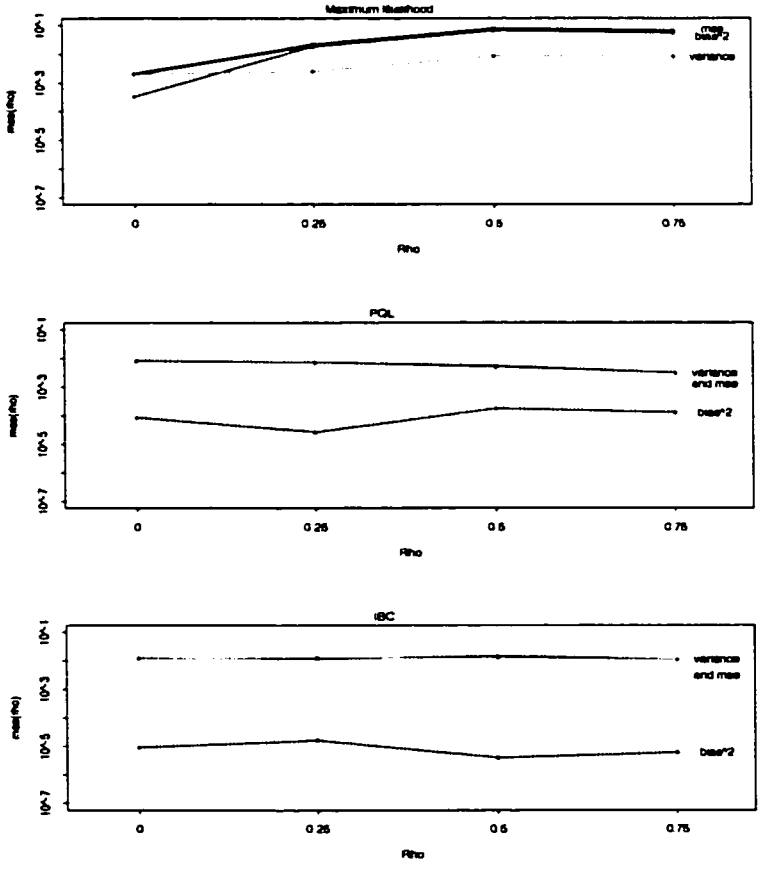

Figure 6.6: Mean-squared errors, variance and bias^2 of $\hat{\rho}$ for maximum likelihood, PQL and IBC for varying ρ and fixed $\sigma_\epsilon^2 = 1$ for $n = 250$.

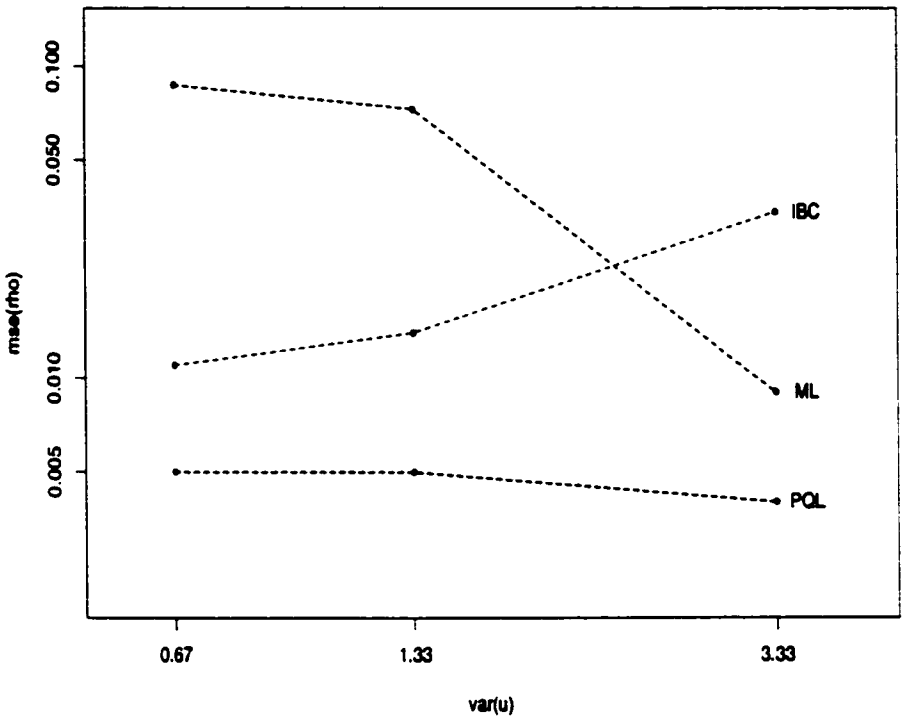

Figure 6.7: Mean-squared errors of ρ for PQL, maximum likelihood, and IBC for fixed $\rho = 0.5$ and varying σ_u^2.

The variance component σ_u^2 was estimated with little or no bias for maximum likelihood, with a tendency to slightly underestimate the true value, and with improvement as the sample size increased. In contrast, using PQL, σ_u^2 was underestimated, with bias that became worse with increasing sample size. However the mean-squared errors of σ_u^2 using PQL became smaller as the sample size increased despite the increasing severity of the bias, due to the compensating reduction in variability. For IBC, the mean-squared errors were a little larger for larger σ_u^2 than for the other two methods. Figure 6.7 displays the mean-squared errors of σ_u^2 for increasing sample size and a fixed ρ value of 0.25. The plot clearly shows that maximum likelihood performs better than both IBC and PQL in terms of mean-squared error for small σ_u^2, except where $\sigma_u^2 = 0$ and a sample size of 50, where

PQL performs much better. The maximum likelihood mean-squared errors improve more drastically with increased sample size compared with PQL and IBC. For $\sigma_u^2 = 2.67$, PQL outperforms maximum likelihood, and would be recommended for situations where σ_u^2 is large.

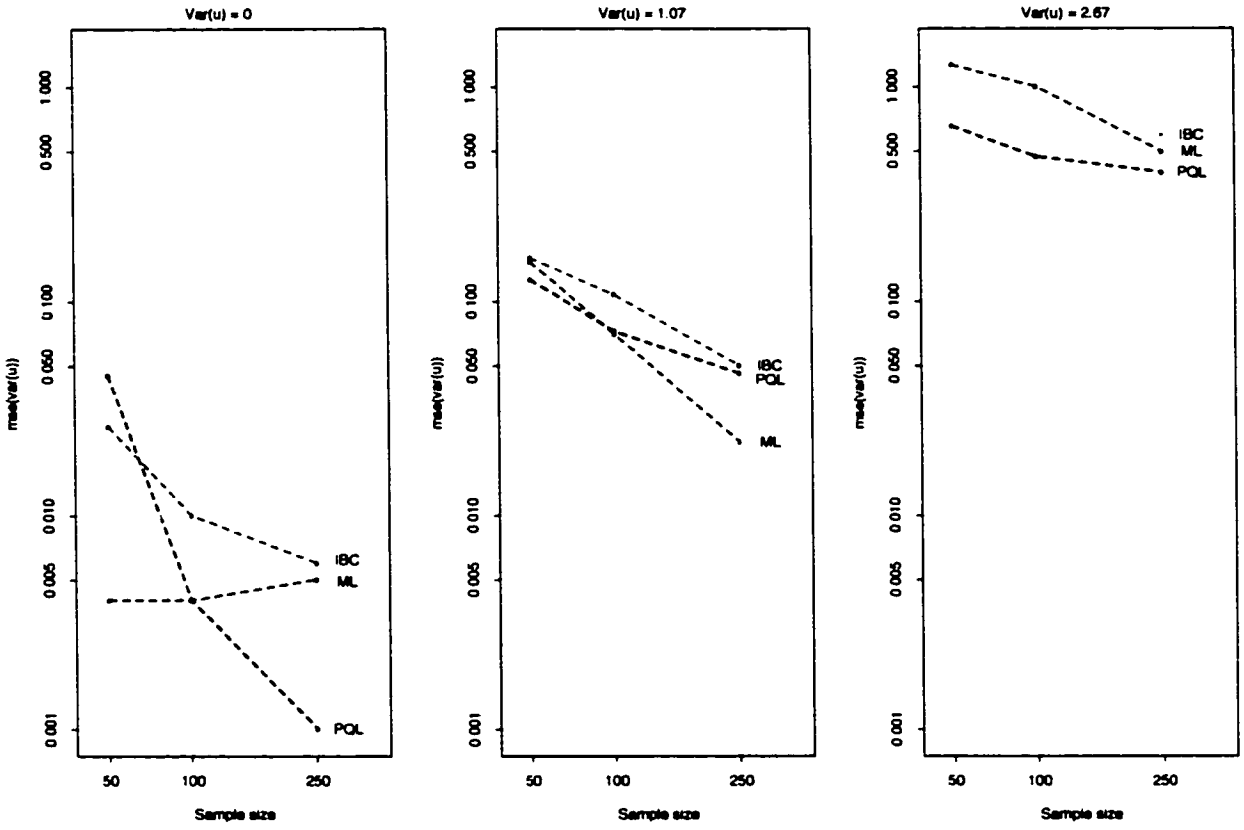

Figure 6.8: Mean-squared errors of σ_u^2 for maximum likelihood, PQL and IBC for fixed $\rho = 0.25$ and varying σ_u^2.

Each of the three methods is carried out using a different iterative process. The iterations to convergence of the variance components ρ and σ_u^2 for a dataset where $n = 100$, $\rho = 0.25$, and $\sigma_u^2 = 1.07$ are shown in Figure 6.9 for all three methods, maximum likelihood, PQL, and iterative bias correction. These plots suggest that the three methods converge

in very different ways to their final parameter estimates, with PQL providing the fastest convergence. The iterative bias correction method tends to jump around a lot more in its iterations, circling in towards the final converged values, while the MCEM algorithm tends to very quickly head onto a steady, slow course towards the final converged values.

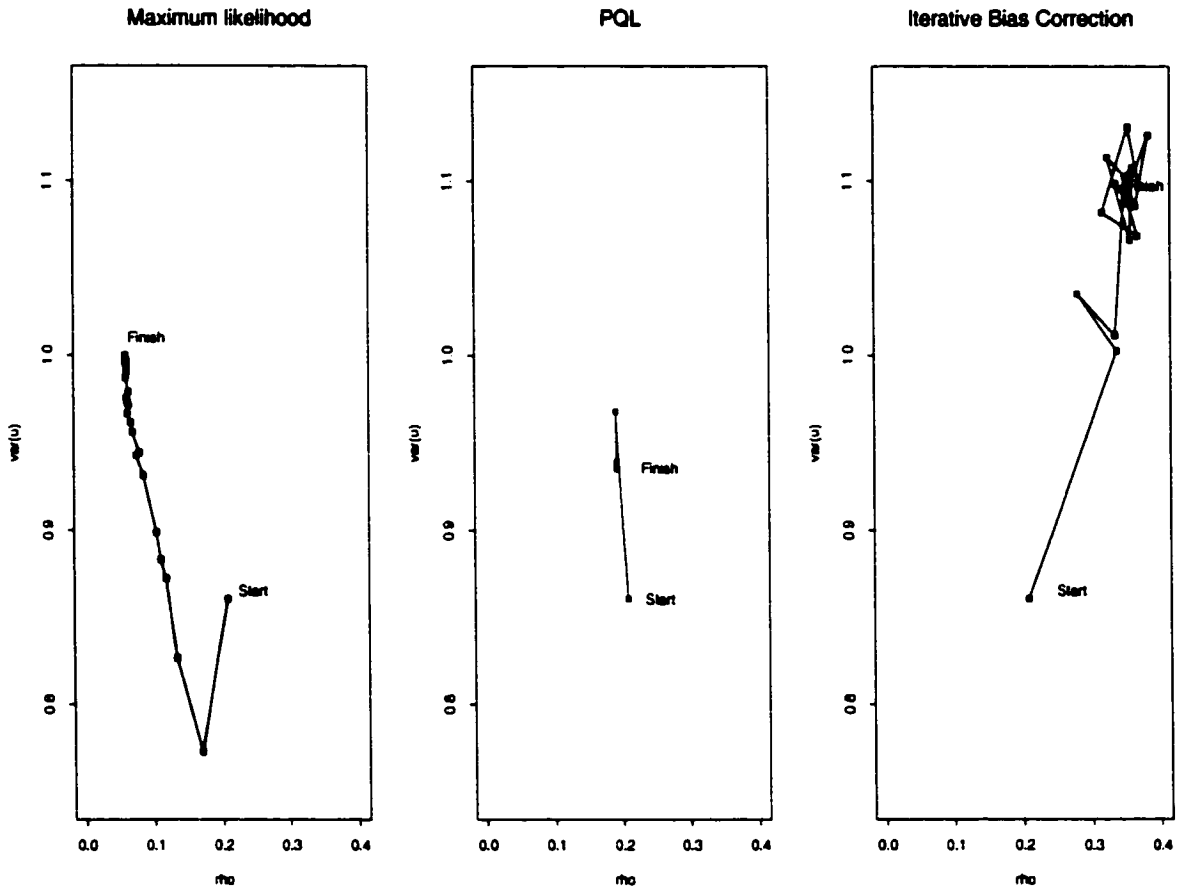

Figure 6.9: Iterations to convergence of ρ and σ_u^2 using the three methods for one dataset where $n = 100$, $\rho = 0.25$ and $\sigma_u^2 = 1.07$.

Analysis of the Original Polio dataset

As mentioned in chapter 2, a number of authors have analyzed the original polio incidence data. Below in Table 6.2 are the results from the analyses carried out. Various authors

estimated different forms of the variance component σ^2 including the variance σ_ϵ^2, however, the value of σ_u^2 is displayed in the table for ease of comparison between results.

Table 6.1: Results from modelling approaches of different researchers of the polio incidence data (Z: Zeger (1988); C&L: Chan and Ledolter (1995); K&C: Kuk and Cheng (1997); Mc: McCulloch (1997); B&C: Breslow and Clayton (1993); prop.: proposed).

	Estimated parameter values for each method					
Coeff.	GEE (Z)	MCEM (C&L)	MCEM (K&C)	MCEM (mine)	PQL (B&C)	IBC (prop.)
β_0	0.46 (0.13)	0.64 (0.13)	0.24 (0.29)	0.27 (0.19)	0.33 (0.29)	0.24 (0.06)
β_1	-4.35 (2.68)	-4.62 (1.38)	-3.79 (2.85)	-4.35 (1.96)	-3.46 (3.04)	-4.77 (0.63)
β_2	-0.11 (0.16)	0.15 (0.09)	0.16 (0.15)	0.15 (0.13)	0.16 (0.14)	0.13 (0.07)
β_3	-0.48 (0.17)	-0.50 (0.12)	-0.48 (0.17)	-0.51 (0.15)	-0.46 (0.16)	-0.52 (0.09)
β_4	0.20 (0.14)	0.44 (0.10)	0.42 (0.13)	0.42 (0.14)	0.40 (0.12)	0.46 (0.06)
β_5	-0.41 (0.14)	-0.04 (0.10)	-0.01 (0.12)	-0.04 (0.14)	-0.01 (0.12)	-0.08 (0.12)
ρ	0.82	0.90 (0.04)	0.67 (0.17)	0.10 (0.36)	0.70 (0.13)	0.55 (0.23)
σ_u^2	0.57	0.43	0.62 (0.29)	0.51 (0.21)	0.51 (0.17)	0.49 (0.19)

Table 6.2: Mean-squared errors for $n = 250$: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Mean-squared errors for the three methods based on 200 simulations				
	ρ	σ_u^2	β_0	β_1	β_4
true value	N/A	0	1	0	0
ML	N/A	0.005	0.010	0.329	0.004
PQL	N/A	0.001	0.007	0.333	0.004
IBC	N/A	0.006	0.009	0.341	0.004
true value	0	0.5	1	0	0
ML	0.002	0.236	0.014	0.654	0.007
PQL	0.013	0.196	0.022	0.591	0.006
IBC	0.014	0.256	0.017	0.726	0.007
true value	0.25	0.533	1	0	0
ML	0.030	0.007	0.019	1.073	0.008
PQL	0.012	0.010	0.027	0.880	0.007
IBC	0.013	0.030	0.024	1.033	0.008
true value	0.5	0.667	1	0	0
ML	0.087	0.010	0.036	1.950	0.014
PQL	0.005	0.015	0.042	1.675	0.008
IBC	0.011	0.019	0.042	1.969	0.009
true value	0.75	1.143	1	0	0
ML	0.124	0.046	0.167	7.282	0.010
PQL	0.003	0.055	0.118	5.481	0.007
IBC	0.008	0.051	0.311	7.181	0.011
true value	0	1	1	0	0
ML	0.002	0.020	0.024	1.192	0.011
PQL	0.008	0.041	0.039	0.961	0.009
IBC	0.012	0.052	0.038	1.486	0.013
true value	0.25	1.067	1	0	0
ML	0.020	0.022	0.033	1.600	0.014
PQL	0.007	0.046	0.046	1.386	0.012
IBC	0.012	0.050	0.048	2.105	0.016

Table 6.3: Mean-squared errors for $n = 250$: the first row in each section describes the true values of the parameters for that set of simulations.

Sample size n	Mean-squared errors for the three methods based on 200 simulations				
	ρ	σ^2	β_0	β_1	β_4
true value	0.5	1.333	1	0	0
ML	0.073	0.140	0.213	10.345	0.048
PQL	0.005	0.073	0.074	2.973	0.013
IBC	0.014	0.172	0.092	4.446	0.019
true value	0.75	2.286	1	0	0
ML	0.062	0.369	0.554	18.376	0.038
PQL	0.003	0.248	0.222	10.371	0.012
IBC	0.011	0.612	0.311	16.751	0.023
true value	0	2.5	1	0	0
ML	0.014	0.392	0.473	8.150	0.072
PQL	0.006	0.351	0.087	1.968	0.019
IBC	0.024	0.625	0.187	5.520	0.045
true value	0.25	2.667	1	0	0
ML	0.007	0.497	0.773	11.567	0.084
PQL	0.006	0.398	0.108	3.269	0.023
IBC	0.022	0.594	0.222	10.473	0.053
true value	0.5	3.333	1	0	0
ML	0.009	2.925	1.610	24.224	0.133
PQL	0.004	0.599	0.173	6.551	0.024
IBC	0.034	0.783	0.304	14.363	0.075
true value	0.75	5.714	1	0	0
ML	0.005	57.797	5.734	59.782	0.263
PQL	0.003	2.181	0.510	24.091	0.026
IBC	–	–	–	–	–

Chapter 7

PERFORMANCE OF GLMM FITTING METHODS WITH NON-NORMAL RANDOM EFFECTS

One assumption that is generally made when fitting generalized linear mixed models is that the random effects are normally distributed. If there is a single independent random effect, then the assumed distribution is Normal$(0, \sigma^2)$, while if there is some assumed correlation structure or more than one random effect, then a multivariate Normal$(0, D)$ distribution is assumed, where D is the variance-covariance matrix with a specified or unspecified correlation structure for the distribution.

When the random effects do not come from a normal distribution, the impact on the resulting parameter estimates has not yet been well investigated, although a number of researchers have examined this issue a little. The robustness of the parameter estimates to the assumed distribution of the random effects has been a controversial topic to date. In light of the possible impact a non-normal distribution may have on the resulting estimates, a number of researchers have proposed non-parametric solutions to the estimation of random effects. These were described earlier in §2.3.7.

Below, §7.1 discusses previous work done by researchers, and the various assumed distributions for the random effects that were investigated. Section §7.2 is a description of a non-normal random effects distribution that will be investigated here for the generalized linear mixed model based on the polio incidence model. The performance of PQL for the polio data where non-normal random effects are present is investigated in §7.3, the iterative bias correction method in §7.4, and exact maximum likelihood in §7.5.

7.1 Background

As mentioned above, a number of researchers have investigated the robustness of approximate maximum likelihood and other methods to the assumption made in generalized linear mixed models that the distribution of the random effects is normal. Neuhaus et al. (1992) investigated the effect of non-normally distributed random effects for a clustered binary data model.

They examined a range of random effects distributions with varying degrees of skewness and kurtosis on five hundred simulated datasets, each consisting of 100 clusters of size 5. The five distributions tested were: Gamma($\alpha = 0.5$, $\beta = 1$), Gamma($\alpha = 16$, $\beta = 1$), Student $t(\nu = 3)$, and the Normal distribution.

Their results showed little bias in the estimates of the regression coefficients even when the mixing distribution was highly skewed. There was also little bias in the estimation of the intercept for the symmetric mixing distributions, but larger bias for the highly skewed Gamma(($\alpha = 0.5$, $\beta = 1$) distribution. There was also large bias for the estimates of σ in all non-normal models. In addition, Neuhaus et al. examined the biases in the associated standard error estimates, and found large biases for the random efffects standard deviation and the intercept. Their conclusion was that the inferences made about the regression parameters in a logistic model and valid standard error estimates of the regression coefficients can be obtained.

These authors have found mixed results that appear promising for the robustness of the generalized linear mixed model to the assumption of normally-distributed random effects. However, Heagerty and Zeger (2000) mention that an often overlooked limitation of the conditional formulation for nonlinear models is that the interpretation of regression coefficients and their estimates can be highly sensitive to difficult-to-verify assumptions about the distribution of random effects, particularly the dependence of the latent variable distribution on covariates. They observed that regression parameters in conditionally specified models are more sensitive to random effects assumptions than their counterparts in the marginal

formulation.

McCulloch (1997) tested out an exponential distribution for the random effects for a binomial dataset with a single random effect using full maximum likelihood estimation assuming an exponential distribution or a normal distribution for the random effects, and PQL (which assumes a normal distribution for the random effects). These results are presented in Table 7.1.

Table 7.1: McCulloch (1997) results for non-normal random effects.

	Estimated bias (se's)		
Parameter (True value $\sigma^2 = 1$)	Maximum Likelihood (Exponential)	Maximum Likelihood (Normal)	PQL
Bias of σ^2	0.19 (0.05)	-0.58 (0.03)	-1.53 (0.01)
MSE of σ^2	0.93 (0.14)	0.56 (0.03)	2.53 (0.03)

The bias in the estimates of the random effects variance for maximum likelihood assuming normality of the random effects, when using exponentially distributed random effects, was quite large. When PQL was used, the bias was even larger for $\hat{\sigma}^2$.

Madger and Zeger (1996) investigated the effect of non-normal distributions of the random effects for linear mixed models. They tested three distributions, each with the same variance of four:

- Gaussian$(0, \sigma^2 = 4)$,
- A skewed distribution: 0.25 Gaussian(14,10) + 0.75 Chisq$(4, \frac{2}{\sqrt{109}})$,
- A discrete distribution with equal point masses placed on 2 and -2.

Four hundred datasets for each type of distribution were simulated. Their results suggested that their non-parametric method (as described in §2.3.7) outperformed a parametric method assuming that the random effects were normally distributed for both the skewed and discrete distributions, and performed only slightly less well than the parametric method when the random effects were normally distributed.

The assumptions underlying the specific method used in fitting the generalized linear mixed model will impact the robustness of the estimators to the assumption of normally distributed random effects. For example, exact maximum likelihood, approximate maximum likelihood and the iterative bias correction method may all be impacted in different ways in their estimation by non-normality of the random effects, depending on how the method uses the normality assumption. Below a comparison of the three methods will be made when the assumption of non-normal random effects is not kept. These comparisons will be made for the polio incidence model described below.

7.2 Non-Normal Random Effect Distributions for the Polio Incidence Model

A small set of simulation studies were carried out to compare the effect of a normal and non-normal random effects distribution on the estimation of the polio incidence model using the three different methods, exact and approximate maximum likelihood, and iterative bias correction.

The form of the model used to fit the polio incidence data is $\log \mu_t^u = x_t^t \beta + u_t$. The random effects follow an AR(1) process $\{u_t\}$ satisfying

$$u_t = \rho u_{t-1} + \epsilon_t, \qquad \text{where} \quad \rho \text{ is the correlation coefficient,}$$
$$\epsilon_t \sim iid \text{ Normal } (0, \sigma_\epsilon^2),$$
$$y = 1, 2, ..., n,$$

and the y_t's are conditionally independent, i.e. $y_t \mid u_t \sim \text{Poisson}(\mu_t^u)$.

The trend and seasonality in the data were modelled with linear and trigonometric compo-

nents, with the covariate vector

$$x_t^t = \left(1, \frac{t}{1000}, \cos\left(\frac{2\pi t}{12}\right), \sin\left(\frac{2\pi t}{12}\right), \cos\left(\frac{2\pi t}{6}\right), \sin\left(\frac{2\pi t}{6}\right)\right).$$

Setting

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & .. \\ \rho & 1 & \rho & \rho^2 & .. \\ \rho^2 & \rho & 1 & \rho & .. \\ \vdots & & \vdots & & .. \\ & & & \rho & 1 \end{pmatrix}$$

the distributional form of the random effects can also be written as:

$$u = (u_1, u_2, ..., u_{168}) \sim \text{ Multivariate Normal}(0, \sigma_u^2 \Sigma).$$

7.2.1 A Multivariate t-distribution for the polio incidence model

The non-normal distribution for the random effects that will be tested is a multivariate t-distribution with an AR(1) correlation structure (Gelman, Carlin et al. 1995) This distribution has heavier tails than the corresponding normal distribution, allowing for more extreme observations. A special relationship exists between a standardized multivariate normal random variable, $X \sim \text{MVN}(0, \Sigma)$ where Σ is the correlation matrix shown above (Johnson and Kotz 1972), and a random variable $v \sim \text{MV}t_\nu(0, \Sigma)$, where $\text{MV}t_\nu$ is a multivariate t-distribution with ν degrees of freedom. The random variable v_t, the tth random effect of v can be written as:

$$v_t = X_t \left(\sqrt{\frac{R}{\nu}}\right)^{-1},$$

with $R \sim \chi^2(\nu)$, independent of X_t. Then

$$v = (v_1, v_2, ..., v_{168}) \sim \text{ MV}t_\nu(0, \sigma_u^2\Sigma) \quad \text{with var}(v) = \frac{\nu}{\nu - 2}\Sigma.$$

To enable fair comparison between using the multivariate normal and multivariate t-distributions as random effect distributions, the variance of an individual random effect v_t from the multivariate t-distribution will be set to be equal to the variance of an individual random effect v_t from the multivariate normal distribution.

The variance for an individual random effect v_t from the multivariate t-distribution is calculated as:

$$\begin{aligned} var(v_t) &= var\left(X_t\left[\sqrt{\frac{R}{\nu}}\right]^{-1}\right) \\ &= var\left(\sqrt{\nu}\,\frac{X_t}{\sqrt{R}}\right) \\ &= \nu\, E\left(\frac{var(X_t)}{R}\,|\,R\right) + \nu\, var\left(\frac{E(X_t)}{R}\,|\,R\right) \\ &= \nu E\left(\frac{1}{R}\,|\,R\right) + \nu\, var(0\,|\,R) \\ &= \frac{\nu}{\nu-2} + 0, \quad \text{for } \nu > 2, \end{aligned}$$

as $\frac{1}{R}$ follows an inverted χ^2 distribution. In practice, it is easiest to simulate random effects $v_t$from a multivariate t-distribution using the corresponding random effects u_t already generated from a multivariate normal$(0, \sigma_u^2\,\Sigma)$ distribution. This allows for a more flexible range of possible multivariate-t_ν distributions, using the same values of the variance components ρ and σ_u^2 for both distributions as required. This is done by generating the v_t's using:

$$v_t = u_t\left(\sqrt{\frac{R}{\nu}}\right)^{-1} c$$

where c is a constant. The variance of v_t is:

$$\begin{aligned} \text{var}(v_t) &= \text{var}\left[u_t\left(\sqrt{\frac{R}{\nu}}\right)^{-1} c\right] \\ &= \sigma_u^2\,\frac{\nu}{\nu-2}\,c^2. \end{aligned}$$

In order to ensure var(v_t) =var(u_t) so that the two distributions have equal variances:

$$\sigma_u^2 = \sigma_u^2 \frac{\nu}{\nu - 2} c^2$$

$$\Rightarrow \frac{\nu - 2}{\nu} = c^2.$$

This allows for a wide range of possible values for the degrees of freedom ν for the multivariate t-distribution.

A scenario in particular that has been chosen for investigation is where $\rho = 0.25$, $\sigma_\epsilon^2 = 1 \Rightarrow$ $\sigma_u^2 = 1.067$ leading to a multivariate t-distribution with $\nu = 3$ d.f., and $c = \sqrt{\frac{1}{3}}$.

An example of the random effects generated for one dataset, where $n = 100$, $\rho = 0.25$, $\sigma_u^2 = 1.07$, and $\nu = 3$ is displayed in Figure 7.1. As anticipated, the random effects generated using the multivariate t-distribution exhibited more extreme values.

7.3 Results

Overall, under the multivariate t-scenario, the three methods estimated the regression coefficients with little bias: however, the variance components were generally severely underestimated on average, with σ^2 being estimated with significantly more variability even in a sample size of $n = 250$. The iterative bias correction method proved to be unstable under these circumstances with low rates (around 50-60%) of converged datasets (due to non-convergence of σ_u^2) in the smallest sample size, $n = 50$, with some improvement when the sample size increased to $n = 100$, and a high rate (around 90%) of converged datasets for $n = 250$. Thus, results for sample sizes of $n = 50$ and $n = 100$ for the iterative bias correction method are not presented in the following results. PQL had slightly lower rates of converged datasets for each sample size compared with their multivariate normal counterparts (98% compared with 100%). Individual datasets fitted using maximum likelihood each took significantly more iterations to converge. Tables 7.2 – 7.4 present details of the average estimates, theoretical and observed standard errors, and mean-squared errors, respectively, for all three methods for the parameters β_0, β_1, β_4 and ρ and σ_u^2. The remaining

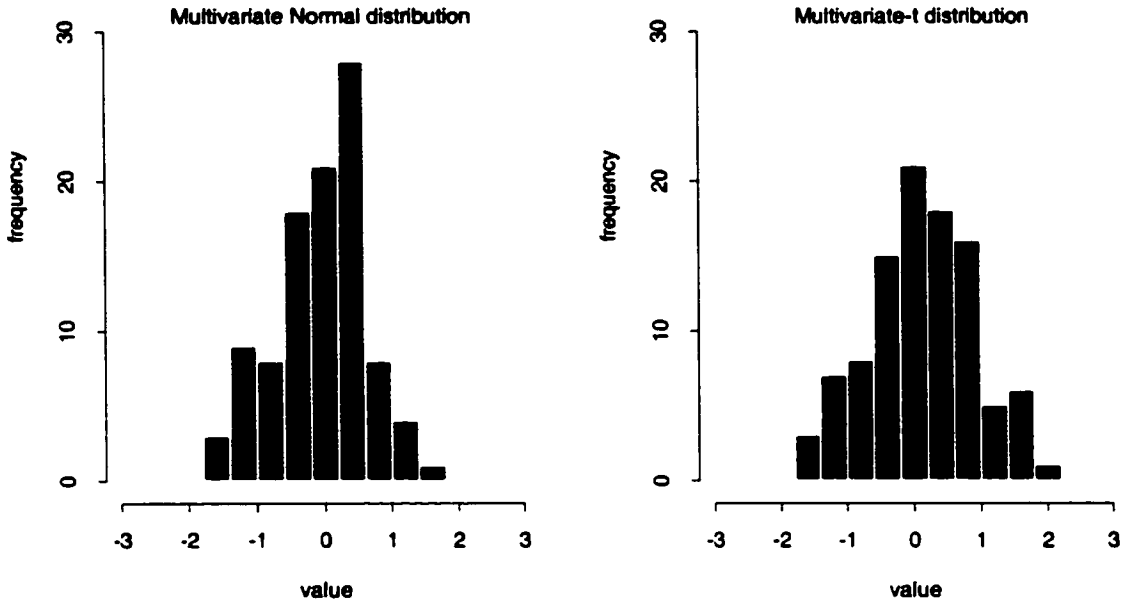

Figure 7.1: Plot of the multivariate normal and multivariate t random effects generated for one dataset with $n = 100$.

regression coefficients β_2, β_3 and β_5 were estimated in a similar manner to β_4, and have not been included in the tables of results.

The regression coefficients β_0 to β_5

When a multivariate t random effects distribution was used, the regression coefficients $\beta_0, ..., \beta_5$ were generally estimated by PQL with comparatively similar levels of bias to their corresponding estimates with a multivariate normal random effects distribution.

The intercept β_0 was estimated on average with more bias using maximum likelihood in

the smallest sample size $n = 50$, but similar bias to the corresponding multivariate normal scenario once a sample size of $n = 250$ was reached.

The time trend coefficient, β_1, was estimated with similar levels of bias to the corresponding multivariate normal scenarios using maximum likelihood. Higher variability was observed in the two hundred estimates of all the regression coefficients in the multivariate t simulation scenarios at the smallest sample size of $n = 50$ when compared to their corresponding multivariate normal estimates. This variability decreased to similar levels to the multivariate normal estimates for sample sizes of $n = 100$ and 250, except for β_0, which maintained higher variability, as reflected in theoretical standard errors up to fifteen times larger.

The estimation of β_0 using PQL proved to have similar levels of bias in the presence of multivariate t random effects, when compared with the multivariate normal setting for all the sample sizes examined. PQL also exhibited larger variability in the estimation of the regression coefficients in the multivariate t-setting at the smallest sample size of $n = 50$. Once a sample size of $n = 100$ was reached, the variability was comparable to the multivariate normal setting.

The iterative bias correction method at a sample size of $n = 250$ estimated β_0 with little bias in both multivariate t-and normal settings, with more variability present in the multivariate t-situation. The time trend coefficient, β_1 was estimated very similarly for both settings with little bias, as were the remaining regression coefficients.

The variance component ρ

Estimation of ρ proved to yield very similar results in terms of bias for both the multivariate t-and the multivariate normal distribution using all three methods. An exception was maximum likelihood at a sample size of $n = 250$, which underestimated ρ more with the multivariate t-distribution. Very little differences were observed in the average estimates of ρ at each sample size: however, the standard errors associated with the estimation of ρ in the multivariate t simulations were approximately 1.5 to 2 times as large as the corre-

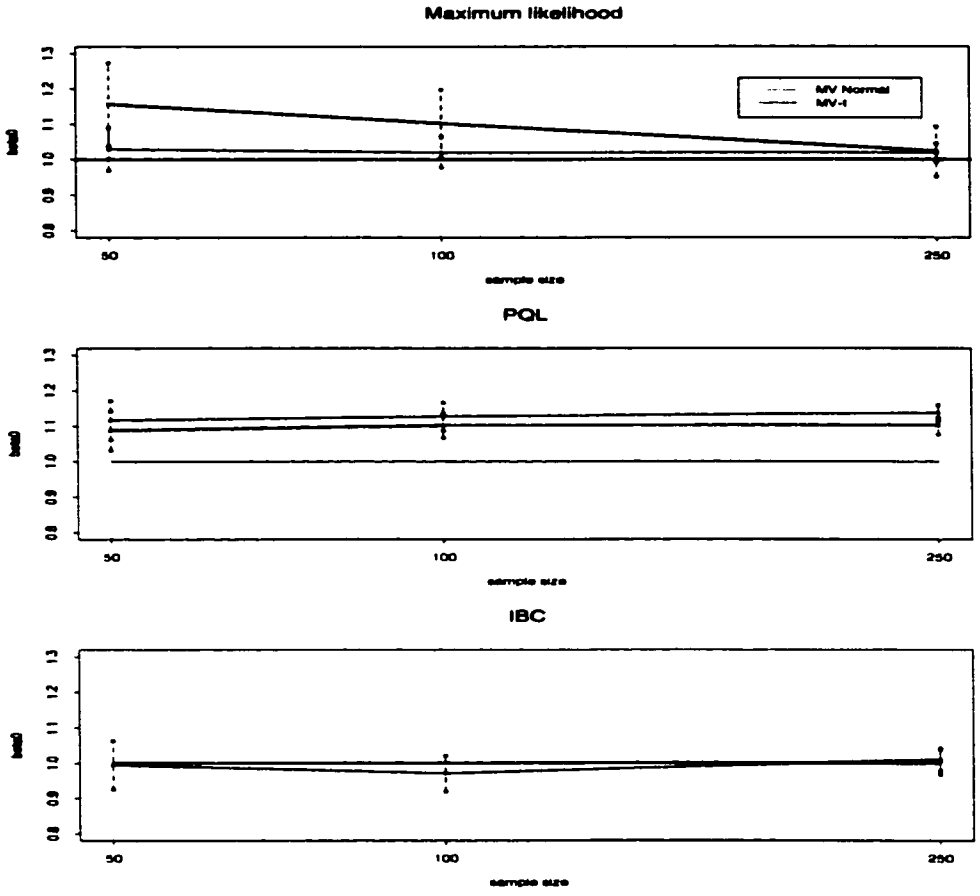

Figure 7.2: Average value and corresponding 95% confidence intervals for $E(\hat{\beta}_0)$ for $n = 250$ using both multivariate normal and multivariate t random effect distributions.

sponding standard errors in the multivariate normal simulations for each of the methods, particularly maximum likelihood, as displayed in Figure 7.3 below. These larger standard errors are reflected in mean-squared errors which were approximately 1.5 times as large as the corresponding mean-squared errors in the multivariate normal simulations.

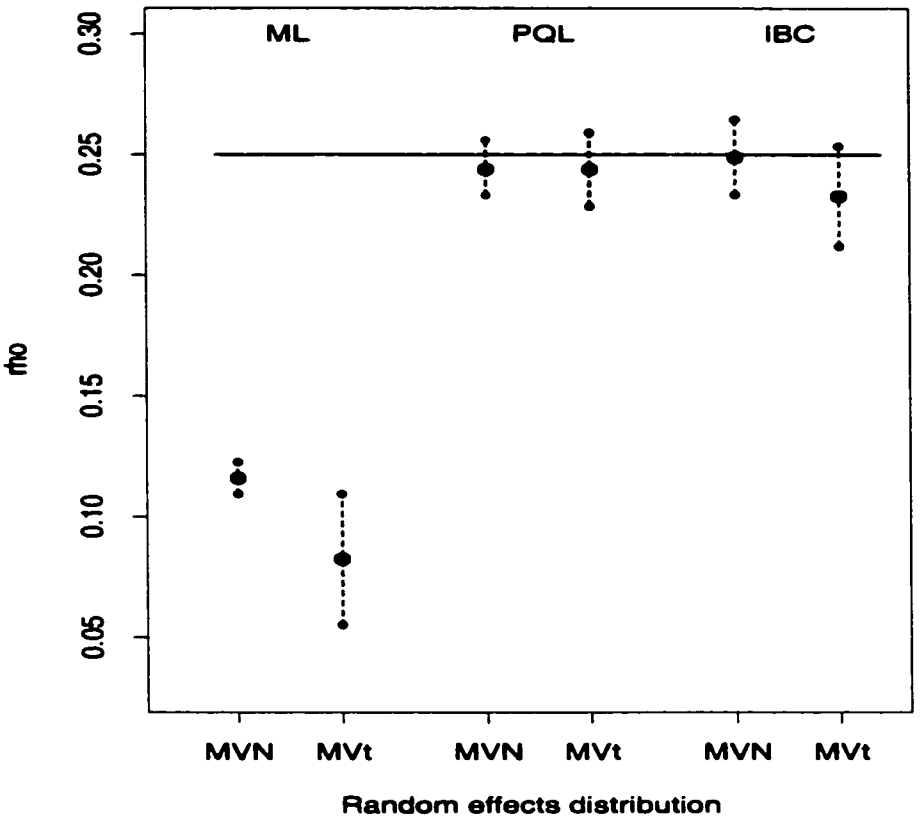

Figure 7.3: Average value and corresponding 95% confidence intervals for $E(\hat{\rho})$ for $n = 250$ using both multivariate normal and multivariate t random effect distributions.

The variance component σ_u^2

The other variance component, σ_u^2, was estimated with substantially higher levels of bias in the multivariate t simulations than for the multivariate normal simulations for each of the three methods, and at each sample size tested. In each multivariate t-scenario, more extreme observations were observed even at the largest sample size, $n = 250$, with the median values exhibiting larger negative bias (up to 35% larger) than observed for the corresponding multivariate normal simulations.

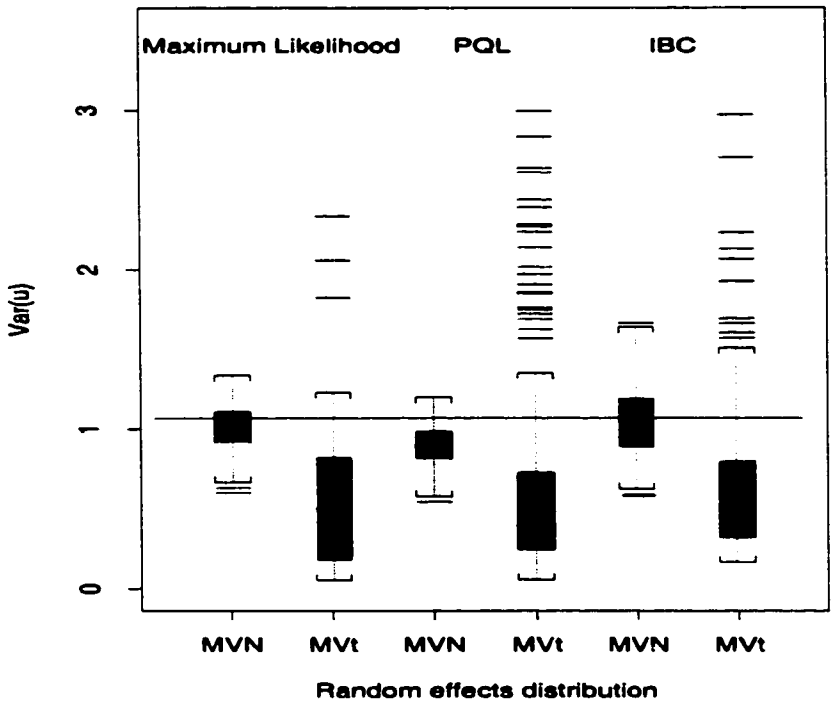

Figure 7.4: Multivariate normal and multivariate t random effect distribution results for σ_u^2 for $n = 250$.

Table 7.2: Average parameter estimates for multivariate normal and multivariate t random effects distributions.

Parameter	Sample size	Truth	ML		PQL		IBC	
			MVN	MV-t	MVN	MV-t	MVN	MV-t
β_0	50	1	1.029	1.155	1.116	1.087	0.999	–
	100		1.020	1.102	1.126	1.101	1.001	–
	250		1.024	1.022	1.137	1.101	1.016	1.003
β_1	50	0	-0.439	-1.576	-0.387	-0.353	0.323	–
	100		-0.057	0.290	-0.027	-0.111	-0.010	–
	250		-0.172	-0.196	-0.093	-0.019	-0.240	-0.008
β_4	50	0	-0.037	-0.139	-0.040	-0.034	-0.039	–
	100		-0.024	-0.003	-0.025	-0.017	-0.052	
	250		-0.011	-0.010	-0.013	-0.008	-0.017	-0.006
ρ	50	0.25	0.038	0.039	0.222	0.204	0.243	–
	100		0.089	0.073	0.236	0.224	0.254	–
	250		0.116	0.083	0.242	0.244	0.246	0.233
σ_u^2	50	1.07	0.807	0.894	0.907	0.859	1.028	–
	100		0.940	0.692	0.909	0.666	1.059	–
	250		1.003	0.642	0.890	0.708	1.053	0.659

Table 7.3: Theoretical (observed) standard errors for multivariate normal and multivariate t random effects distributions.

Parameter	n	ML		PQL		IBC	
		MVN	MV-t	MVN	MV-t	MVN	MV-t
$\beta_0 = 1$	50	0.44 (0.34)	0.84 (0.28)	0.40 (0.40)	0.40 (0.37)	0.44 (0.55)	-
	100	0.32 (0.26)	1.10 (0.60)	0.29 (0.28)	0.25 (0.23)	0.33 (0.44)	-
	250	0.18 (0.18)	2.84 (0.11)	0.16 (0.17)	0.17 (0.15)	0.22 (0.25)	0.17 (0.09)
$\beta_1 = 0$	50	15.06 (11.26)	18.59 (0.47)	13.65 (13.61)	14.54 (12.44)	15.27 ()	-
	100	5.53 (3.89)	5.21 (0.33)	5.01 (4.76)	4.14 (3.99)	6.04 ()	-
	250	1.253 (1.04)	1.81 (0.19)	1.17 (1.19)	1.07 (1.01)	1.43 ()	1.14 (0.60)
$\beta_4 = 0$	50	0.27 (0.24)	1.33 (0.19)	0.24 (0.23)	0.25 (0.21)	0.27 (0.29)	-
	100	0.19 (0.18)	0.19 (0.13)	0.17 (0.17)	0.15 (0.14)	0.20 (0.22)	-
	250	0.12 (0.11)	0.11 (0.08)	0.11 (0.11)	0.10 (0.09)	0.13 (0.14)	0.10 (0.06)
$\rho = 0.25$	50	0.11 (0.23)	0.143 (0.39)	0.22 (0.22)	0.30 (0.45)	0.27 (0.34)	-
	100	0.08 (0.15)	0.11 (0.26)	0.14 (0.14)	0.24 (0.52)	0.17 (0.21)	-
	250	0.05 (0.09)	0.15 (0.17)	0.08 (0.09)	0.11 (0.13)	0.11 (0.17)	0.07 (0.14)
$\sigma_u^2 = 1.07$	50	0.29 (0.26)	1.55 ()	0.32 (0.31)	1.10 (0.34)	0.40 (0.46)	-
	100	0.23 ()	0.69 (1.39)	0.22 (0.20)	0.82 (0.16)	0.33 (0.38)	-
	250	0.14 (0.14)	0.64 ()	0.16 (0.17)	0.87 (0.10)	0.22 (0.24)	0.641 ()

Table 7.4: Mean-squared errors for multivariate normal and multivariate t random effects distributions.

Parameter	Sample size	ML		PQL		IBC	
		MVN	MV-t	MVN	MV-t	MVN	MV-t
$\beta_0 = 1$	50	0.191	0.736	0.173	0.166	0.197	-
	100	0.100	0.364	0.098	0.071	0.110	-
	250	0.033	8.420	0.046	0.039	0.048	0.039
$\beta_1 = 0$	50	228.964	348.100	186.541	211.631	233.271	-
	100	31.717	27.182	25.085	17.165	26.500	-
	250	2.944	3.398	1.386	1.148	2.105	1.308
$\beta_4 = 0$	50	0.072	1.784	0.059	0.061	0.074	-
	100	0.038	0.036	0.031	0.022	0.042	-
	250	0.014	0.013	0.012	0.010	0.016	0.009
$\rho = 0.25$	50	0.056	0.065	0.049	0.092	0.075	-
	100	0.032	0.044	0.020	0.059	0.030	-
	250	0.020	0.043	0.007	0.012	0.012	0.014
$\sigma_u^2 = 1.07$	50	0.152	1.554	0.126	1.252	0.159	-
	100	0.070	0.686	0.072	0.832	0.107	-
	250	0.022	0.564	0.046	0.891	0.050	0.458

Chapter 8

CONCLUSIONS AND FUTURE WORK

The performance of three methods that can be used to fit generalized linear mixed models has been thoroughly examined using simulation results for a Poisson model with AR(1) correlation structure for the random effects using simulation studies. Performance was measured using averages, standard errors, and mean-squared errors.

Each method was shown to have strengths and weaknesses, and none of the three methods proved to be ideal for use in every situation for the Poisson data model used. Below are some conclusions and recommendations based on the work carried out.

- For a small sample size, around $n = 50$ to 150, either PQL or maximum likelihood are the methods of choice. The iterative bias correction method proved to be unstable in smaller sample sizes, with the variance components often not converging. Because only small sample sizes are involved, the time taken to fit a dataset using maximum likelihood is reasonable, and PQL is extremely fast.

- For moderate-sized samples, between $n = 150$ and 1000, either PQL or iterative bias correction can be recommended for use, as both methods provide estimates in a moderate time frame.

- PQL is the recommended method of fitting for large datasets ($n \geq 1000$), due to the computational intensity of the other two methods, iterative bias correction and maximum likelihood. However, if the level of bias in the parameter estimates, particularly the variance components, is an important consideration, the iterative bias correction method can be considered a viable alternative.

- Of the three methods, the iterative bias correction method provides the least biased estimates, particularly for the variance components. Both maximum likelihood and PQL displayed bias in a range of situations for either or both of the variance components, especially where the variance components were more extreme in value ($\sigma_u^2 \geq 1$, $\rho \geq 0.5$). However, the standard errors were often around 10% larger for the IBC method.

- All three methods proved to be less stable (with lack of convergence of ρ and σ_u^2) when the true values of the variance components were extreme ($\rho \geq 0.75$, $\sigma_u^2 \geq 2.5$).

- PQL and maximum likelihood generally had smaller mean-squared errors for the variance components than the iterative bias correction method. The IBC mean-squared errors were often around 10%–15% larger than the corresponding mean-squared errors for PQL and maximum likelihood.

- None of the three methods was particularly robust to non-normality of the random effects; the estimation of σ_u^2 was particularly affected, with substantially higher negative bias when the random effects came from a multivariate t-distribution with three degrees of freedom. In addition, datasets fitted using maximum likelihood took longer to converge, and the other two methods, PQL and iterative bias correction, were both less stable.

Future Work

The investigation of the performance of the three methods for fitting the polio incidence model raised a number of interesting questions that have much potential as topics for future research.

- In chapter two, the question was raised regarding the cause of the bias present in the maximum likelihood estimation of the variance component ρ, especially in small

samples. Of interest will be to examine the asymptotic behavior of the maximum likelihood estimators in a generalized linear mixed model, where there is correlation present between the observations. This issue could be examined both theoretically and through the use of simulation studies.

- The performance and comparison of the three methods in the work carried out has been for count data with an AR(1) random effects correlation structure. Similar investigation of the performance and comparison of the three methods for a binary data model would be of interest, especially as the use of binary outcomes is very common.

- The MCEM algorithm used here to obtain maximum likelihood estimates of the parameters uses a Metropolis-Hastings algorithm to generate random effects vectors from their conditional distribution $f(u|y)$. This is computationally very intensive, so consequently, Booth and Hobert (1999) investigated the use of faster sampling techniques, importance and rejection sampling, to generate random effects vectors. Of interest is to investigate the use of these alternatives, and other sampling techniques, in an effort to reduce the computational intensity of the algorithm.

- Bayesian methods are becoming increasingly popular for fitting hierarchical datasets, which can include many types of generalized linear mixed models. Much work has been done in this area, and it is of interest is to compare current Bayesian techniques with maximum likelihood, PQL, and iterative bias correction.

BIBLIOGRAPHY

[1] M. Aitken. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128, 1999.

[2] D.A. Anderson and M. Aitken. Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B*, 47:203–210, 1985.

[3] P. Armitage and G. Berry. *Statistical methods in medical research.* Blackwell Scientific Publications Ltd (Oxford), 1987.

[4] R.F. Bartlett and B.C. Sutradhar. On estimating equations for parameters in generalized linear mixed models with application to binary data. *Environmetrics*, 10:769–784, 1999.

[5] C.S. Berkey and D.C. Hoaglin et al. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, 17:2537–2550, 1998.

[6] J.G. Booth and J.P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, 61(1):265–285, 1999.

[7] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis.* Prentice-Hall, Inc., 1994.

[8] N.E. Breslow. Extra-Poisson variation in log-linear models. *Applied Statistics*, 33:38–44, 1984.

[9] N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

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

[11] W.J. Browne and D. Draper. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, 15(3):391–420, 2000.

[12] R.L. Burden and J.D. Faires. *Numerical Analysis.* Prindle, Weber & Schmidt Publishing Company, Boston, 1985.

[13] P.R. Burton and K.J. Tiller et al. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (glmms) and Gibbs sampling. *Genetic Epidemiology*, 17:118–140, 1999.

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

[15] G. Casella and R.L. Berger. *Statistical Inference.* Duxbury Press, Wadsworth, Inc., 1990.

[16] Centers for Disease Control. *Morbidity and Mortality Weekly Report Annual Summary: Reported Morbidity and Mortality in the U.S.* 1970-1983, Volume 18-31, No. 54.

[17] K.S. Chan and J. Ledolter. Monte-Carlo EM estimation for time-series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252, 1995.

[18] S. Chib and R. Winkelmann. Markov Chain Monte-Carlo analysis of correlated count data. *Journal of Business and Economic Statistics*, 19(4):428–435, 2001.

[19] D. Clayton and J. Rasbash. Estimation in large crossed random-effect models by data augmentation. *Journal of the Royal Statistical Society, Series A*, 162:425–436, 1999.

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

[21] M.J. Crowder. Beta-binomial ANOVA for proportions. *Applied Statistics*, 27:34–37, 1978.

[22] M.J. Daniels. A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27(3):567–578, 1999.

[23] M.J. Daniels and C. Gatsonis. Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine*, 16:2311–2325, 1997.

[24] R.A. Davis, W.T.M. Dunsmuir, and Y. Wang. On autocorrelation in a Poisson regression model. *Biometrika*, 87(3):491–505, 2000.

[25] A.P Dempster, N.M Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 39:34–37, 1977.

[26] P.J. Diggle and R.J. Gratton. Monte Carlo methods of inference for implicit statistical models (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, 46:193–227, 1984.

[27] P.J. Diggle, K.-Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data.* Oxford University Press, 1994.

[28] B. Engel. A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal*, 40:141–154, 1998.

[29] B. Engel and W. Buist. Bias reduction of approximate maximum likelihood estimates for heritability in threshold models. *Biometrics*, 54:1155–1164, 1998.

[30] B. Engel and A. Keen. A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, 48:1–22, 1994.

[31] L. Fahrmeir and H. Kaufman. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[32] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized linear Models*. Springer-Verlag, 1994.

[33] T.S. Ferguson. *A course in large sample theory*. Chapman & Hall, 1996.

[34] D.A. Follmann and D. Lambert. Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84(405):295–300, 1989.

[35] A.T. Galecki and T.R. Ten Have et al. A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models. *Computational Statistics and Data Analysis*, 35:265–281, 2001.

[36] A.E. Gelfand and B.P. Carlin. Maximum likelihood estimation for constrained- or missing-data problems. *Canadian Journal of Statistics*, 21:303–311, 1993.

[37] A. Gelman and J.B. Carlin et al. *Bayesian Data Analysis*. Chapman & Hall, 1995.

[38] J. Geweke. Bayesian inference in econometric models using Monte Carlo integration (STMA V31 3135). *Econometrica*, 57:1317–1339, 1989.

[39] C.J. Geyer and E.A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, 54:657–699, 1992.

[40] D. Gianola and J.L. Foulley. Sire evaluation for ordered categorical-data with a threshold-model. *Genetics Selection Evolution*, 15(2):201–223, 1983.

[41] A.R. Gilmour, R.D. Anderson, and A.L. Rae. The analysis of binomial data by a generalized linear mixed model. *Biometrika*, 72:593–599, 1985.

[42] V.P. Godambe. *Estimating functions*. Oxford University Press, 1991.

[43] V.P. Godambe and C.C. Heyde. Quasi-likelihood and optimal estimation. *International Statistical Review*, 55:231–244, 1987.

[44] H. Goldstein. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73:43–56, 1986.

[45] H. Goldstein. Consistent estimators for multilevel generalised linear models using an iterated bootstrap. *Multilevel Modelling Newsletter, Institute of Education, University of London*, 8(1):3–6, 1996.

[46] H. Goldstein and J. Rasbash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, General*, 159:505–513, 1996.

[47] C. Gourieroux and A. Monfort et al. Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118, 1993.

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

[49] X. Guyon. *Random fields on a network*. Springer-Verlag, 1995.

[50] S.J. Haberman. Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 5:815–841, 1977.

[51] D.A. Harville. Extension of the gauss-markov theorem to include the estimation of random effects. *Annals of Statistics*, 4:384–395, 1976.

[52] D.A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–339, 1977.

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

[54] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[55] P.J. Heagerty and S.L. Zeger. Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):1–26, 2000.

[56] J.J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica*, 52:271–320, 1984.

[57] C.R. Henderson. Selection Index and expected genetic advance. *Statistical Genetics and Plant Breeding*, pages 141–163, 1963.

[58] C.C. Heyde. *Quasi-likelihood and its application : a general approach to optimal parameter estimation.* Springer, New York, 1997.

[59] J.P. Hinde. *Compound Poisson regression models.* in GLIM 82: Proceedings of the International Conference on Generalized Linear Models, ed. R Gilchrist, 1982.

[60] J.P. Hinde. Random effects in generalized linear models and the EM algorithm. *Communications in Statistics - Theory and Methods*, 17:3847–3856, 1988.

[61] I. Hoeschele and B. Tier. Estimation of variance components of threshold characters by marginal posterior modes and means via gibbs sampling. *Genetics Selection Evolution*, 27:519–540, 1995.

[62] H. Hoijtink. Posterior inference in the random intercept model based on samples obtained with Markov Chain Monte Carlo methods. *Computational Statistics*, 15(3):315–336, 2000.

[63] J.L. Hopper. Variance components for statistical genetics: Applications in medical research to characteristics related to human diseases and health. *Statistical Methods in Medical Research*, 2:199–223, 1993.

[64] J. Jiang. Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, 93(442):720–729, 1998.

[65] N.L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley & Sons, Inc., 1972.

[66] M.R. Karim and S.L. Zeger. Generalized linear models with random effects - Salamander mating revisited. *Biometrics*, 48(2):631–644, 1992.

[67] J. Kiefer and J. Wolfowitz. On a theorem of Hoel and Levine on extrapolation. *The Annals of Mathematical Statistics*, 36:1627–1655, 1965.

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

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

[70] A.Y.C Kuk and C.H. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79:531–541, 1992.

[71] A.Y.C Kuk and Y.W. Cheng. The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computation and Simulation*, 59:233–250, 1997.

[72] A.Y.C Kuk and Y.W. Cheng. Pointwise and functional approximations in Monte Carlo maximum likelihood estimation. *Statistics and Computing*, 9:91–99, 1999.

[73] N.M. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

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

[75] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 57:425–437, 1995.

[76] C. Lee. Methods and techniques for variance component estimation in animal breeding - Review. *Asian-Australasian Journal of Animal Sciences*, 13(3):413–422, 2000.

[77] Y. Lee and J.A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B*, 4:619–678, 1996.

[78] E.L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer-Verlag, 1998.

[79] B.G. Leroux. Modelling spatial disease rates using maximum likelihood. *Statistics in medicine*, 19:2321–2332, 2000.

[80] E. Lesaffre and B. Spiessens. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 50:325–335, 2001.

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

[82] X. Lin and N.E. Breslow. Analysis of correlated binomial data. *Journal of Statistical Computing and Simulations*, 55:133–146, 1996.

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

[84] Q. Liu and D.A. Pierce. Heterogeneity in Mantel-Haenszel-type models. *Biometrika*, 80:543–556, 1993.

[85] T.A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 44:226–233, 1982.

[86] L.S. Magder and S.L. Zeger. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*, 91(435):1141–1151, 1996.

[87] D. Malec and J. Sedransk et al. Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92(439):815–826, 1997.

[88] P. McCullagh. Quasi-likelihood functions. *Annals of Statistics*, 11:59–67, 1983.

[89] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

[90] C.E. McCulloch. Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89(425):330–335, 1994.

[91] C.E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):162–170, 1997.

[92] C.A. McGilchrist. Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B, Methodological*, 56(1):61–69, 1994.

[93] F. Mealli and C. Rampichini. Estimating binary multilevel models through indirect inference. *Computational Statistics and Data Analysis*, 29:313–324, 1999.

[94] N. Metropolis and A.W. Rosenbluth et al. Equations of state equations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

[95] R.B. Millar and T.J. Willis. Estimating the relative density of snapper in and around a marine reserve using a log-linear mixed-effects model. *Australian and New Zealand Journal of Statistics*, 41(4):383–394, 1999.

[96] G. Molenberghs and E Goetghebeur. Simple fitting algorithms for incomplete categorical data. *Journal of the Royal Statistical Society, Series B*, 59:401–414, 1997.

[97] C. Moreno and D. Sorensen et al. On biased inferences about variance components in the binary threshold model. *Genetics Selection Evolution*, 29(2):145–160, 1997.

[98] R. Natarajan and R.E. Kass. Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95(449):227–237, 2000.

[99] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135:370–384, 1972.

[100] J.M. Neuhaus and W.W. Hauck et al. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79(4):755–762, 1992.

[101] J.M. Neuhaus and M.L. Lesperance. Estimation efficiency in a binary mixed model setting. *Biometrika*, 83(2):441–446, 1996.

[102] J.M. Neuhaus and M.R. Segal. An assessment of approximate maximum likelihood estimators in generalized linear mixed models. *Technical Report 54, Department of Epidemiology and Biostatistics, University of California San Francisco*, 1996.

[103] Y. Ochi and R.L. Prentice. Likelihood inference in a correlated probit regression model. *Biometrika*, 71:531–543, 1984.

[104] M.-S Oh and Y.B. Lim. Bayesian analysis of time series Poisson data. *Journal of Applied Statistics*, 28(2):259–271, 2001.

[105] J. Palmgren. Exponential family models and statistical genetics. *Statistical methods in Medical Research*, 9:57–72, 2000.

[106] A. Penttinen. Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method. *Jyaskyla Studies in Computer Science, Economics and Statistics*, 7:1–105, 1984.

[107] R.W. Platt, B.G. Leroux, and N.E. Breslow. Generalized linear mixed models for meta-analysis. *Statistics in Medicine*, 18:643–654, 1999.

[108] F.A. Quintana and J.S. Liu et al. Monte Carlo EM with importance reweighting and its applications in random effects models. *Computational Statistics and Data Analysis*, 29:429–444, 1999.

[109] S.W. Raudenbush and Yang M.-L et al. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics*, 9(1):141–157, 2000.

[110] B.D. Ripley. *Stochastic simulation.* John Wiley & Sons (New York; Chichester), 1987.

[111] G. Rodriguez and N. Goldman. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, General*, 158:73–89, 1995.

[112] R. Schall. Estimation in generalized linear models with random effects. *Biometrika*, 40:719–727, 1991.

[113] L. Simar. Maximum likelihood estimation of a compound poisson process. *The Annals of Statistics*, 4:1200–1209, 1976.

[114] P.J. Solomon and D.R. Cox. Nonlinear component of variance models. *Biometrika*, 79:1–11, 1992.

[115] R. Stiratell, N.M Laird, and J.H. Ware. Random-effects models for serial observations with binary responses. *Biometrics*, 40:961–971, 1984.

[116] B.C. Sutradhar and Z. Qu. On approximate likelihood inference in a Poisson mixed model. *Canadian Journal of Statistics*, 26:169–186, 1998.

[117] M. Tan and Y.S. Qu et al. A Bayesian hierarchical model for multi-level repeated ordinal data: Analysis of oral practice examinations in a large anaethesiology training program. *Statistics in Medicine*, 18(15):1983–1992, 1999.

[118] M.A. Tanner. *Tools for statistical inference : observed data and data augmentation methods.* Springer-Verlag, New York, 1991.

[119] H. Tao and M. Palta et al. An estimation method for the semiparametric mixed effects model. *Biometrics*, 55:102–110, 1999.

[120] R.J. Tempelman. Generalized linear mixed models in dairy cattle breeding. *Journal of Dairy Science*, 81:1428–1444, 1998.

[121] D.A. van Dyk. Nesting EM algorithms for computational efficiency. *Statistica Sinica*, 10(1):203–225, 2000.

[122] S.G. Walker and B.K. Mallick. Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B*, 59(4):845–860, 1997.

[123] P.M. Wang and M.L. Puterman. Analysis of longitudinal data of epileptic seizure counts - a two-state hidden Markov regression approach. *Biometrical Journal*, 43(8):941–962, 2001.

[124] Y. Wang and Y.H. Wang et al. Gibbs-Sampler approach for meta-analysis of multiple clinical trials using generalized linear model with random-effects. *Chinese Medical Journal*, 113(6):554–557, 2000.

[125] L. Watier, S. Richardson, and D. Hemon. Accounting for pregnancy dependence in epidemiologic studies of reproductive outcomes. *Epidemiology*, 8(6):629–636, 1997.

[126] D.A. Williams. Extra-binomial variation in logistic linear models. *Applied Statistics*, 31(2):144–148, 1982.

[127] R. Wolfinger and M. O'Connell. Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243, 1993.

[128] K.K.W. Yau and P.S.F. Ma. A simulation study for the binomial-logit model with correlated random effects. *Journal of Statistical Computing and Simulations*, 63:169–186, 1999.

[129] S.L. Zeger. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988.

[130] S.L. Zeger and K.-Y.Liang. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11(14):1825–1839, 1992.

[131] S.L. Zeger and M.R. Karim. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Asssociation*, 86(413):79–86, 1991.

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

[133] S.L. Zeger, K.-Y. Liang, and P.S. Albert. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060, 1988.

Appendix A

THE POLIO INCIDENCE DATA

OBS	y	OBS	y	OBS	y	OBS	y	OBS	y	OBS	y	OBS	y	OBS	y
1	0	24	1	43	1	64	1	85	1	106	4	127	0	148	1
2	1	23	1	44	0	65	0	86	1	107	2	128	0	149	0
3	0	24	5	45	1	66	0	87	0	108	3	129	0	150	1
4	0	25	0	46	0	67	1	88	1	109	3	130	1	151	0
5	1	26	3	47	1	68	2	89	1	110	0	131	0	152	2
6	3	27	1	48	0	69	0	90	0	111	0	132	1	153	0
7	9	28	0	49	1	70	0	91	2	112	2	133	1	154	0
8	2	29	1	50	0	71	1	92	1	113	7	134	0	155	1
9	3	30	4	51	1	72	2	93	3	114	8	135	0	156	2
10	5	31	0	52	0	73	0	94	1	115	2	136	0	157	0
11	3	32	0	53	1	74	3	95	2	116	4	137	0	158	1
12	5	33	1	54	0	75	1	96	4	117	1	138	0	159	0
13	2	34	6	55	1	76	1	97	0	118	1	139	1	160	0
14	2	35	14	56	0	77	0	98	0	119	2	140	2	161	0
15	0	36	1	57	1	78	2	99	0	120	4	141	0	162	1
16	1	37	1	58	0	79	0	100	1	121	0	142	2	163	2
17	0	38	0	59	0	80	4	101	0	122	1	143	0	164	1
18	1	39	0	60	2	81	0	102	1	123	1	144	0	165	0
19	3	40	1	61	0	82	2	103	0	124	3	145	0	166	1
20	3	41	1	62	1	83	1	104	2	125	3	146	1	167	3
21	2	42	1	63	0	84	1	105	2	126	0	147	0	168	6

VITA

Kerrie Nelson was born in New Zealand. She carried out her undergraduate degreee at the University of Auckland, and in 1997, came to Seattle, Washington to continue her studies. In 2002 she earned a Doctor of Philosophy in Statistics at the University of Washington.