2008

The performance of estimation methods for generalized linear mixed models

Damian Collins
University of Wollongong

The performance of estimation methods for generalized linear mixed models

A thesis submitted in fulfillment of the

requirements for the award of the degree

Doctor of Philosophy

from

University of Wollongong

by

Damian Collins BSc (Hons) UNSW

School of Mathematics and Applied Statistics

June 2008

THIS PAGE IS BLANK

I, Damian Paul Collins, declare that this thesis, submitted in fulfillment of the requirements for the award of Doctor of Philosophy in the School of Mathematics and Applied Statistics, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Damian Collins

June, 2008

Acknowledgments

I would first like to thank my supervisors, Ken Russell, Robin Thompson and Brian Cullis, for giving me this wonderful opportunity to improve myself, and for supporting me throughout this process. I am sorry that it was such a saga for you all, and that I didn't listen to your advice as well as I should have.

I would also like to thank Idris Barchia, Paul Nicholls and other biometricians in NSW DPI for taking care of the consulting workload during my absence.

I would also like to thank Gwenda Thompson for her friendship and camaradarie during the latter stages.

I would also like to thank the staff at SMAS at UOW for providing a supportive environment for research.

Finally, I would like to thank my parents, for always being there when I needed someone to talk to and providing a home away from home.

The NSW DPI provided me with a generous financial assistance package during my years of research.

Abstract

Generalised linear models (GLMs) are a flexible class of non-linear models for non-normally distributed response data. GLMs encompass models for discrete response data which takes one of several values rather than being measured on a continuous scale. Discrete response data is abundant in agricultural and biological research, for instance, in the mortality of animals and plants (binary/binomial data) and the scoring of disease (ordinal data).

Generalised linear mixed models (GLMMs) are an extension of GLMs which include additional random effects in the (conditional) linear predictor. Some examples of where GLMMs may be useful include the analysis of designed experiments, surveys, spatial data and longitudinal or repeated measures data.

The fundamental difficulty in using GLMMs is that no closed analytical expression for the likelihood is available. A variety of approaches have been proposed to circumvent this difficulty, including approximate likelihood approaches, such as penalized quasi-likelihood (PQL), numerical approaches, such as Gauss-Hermite quadrature (GHQ), and approaches based on the use of Monte Carlo methods, such as modern Bayesian approaches implementing Markov Chain Monte Carlo (MCMC) techniques.

Although in recent years more attention in the literature has been given to Bayesian approaches and other approaches based on Monte Carlo techniques for GLMMs, there is still widespread interest amongst practitioners in the use of approximate likelihood approaches, especially with the work of Lee & Nelder (2001, 2006). The objective of this PhD is primarily to explore the approximate likelihood approaches, as well as comparing and contrasting them with numerical and Monte Carlo approaches.

The most widely known approximate likelihood approach, PQL, is well-known to give biased estimators of the GLMM parameters for binary grouped data when the group size is small. However, the other two groups of approaches for GLMMs are not without problems. Numerical approaches such as GHQ are only suitable for GLMMs with nested random effects only, and often require very good starting values to achieve convergence. Approaches based on Monte Carlo techniques can be very computational intensive and also have convergence problems, as well as being sensitive to the choice of priors, when used within the Bayesian paradigm. The approximate likelihood approach of Lee and Nelder is claimed, by its proponents, to enjoy the computational efficiency of PQL whilst not suffering from the estimation bias issues that PQL experiences.

A background to the GLMM and inferential issues is provided in Chapter 1, with theoretical material and alternative approaches for modelling correlation in non-normal data, such as the generalized estimating equation (GEE) approach. It is argued that the GLMM is the most generally applicable model for modelling correlation and clustering in non-normal data available at present. The second chapter reviews the main estimation approaches for GLMMs, discussing in more detail the issues associated with each of the approaches already highlighted above.

Chapters 3 and 4 focus on the two most popular approximate likelihood approaches, PQL and the hierarchical GLM (HGLM) approach of Lee & Nelder (2001, 2006) respectively. Simulation studies are presented in Chapter 3 for binary and sparse Poisson data from a range of designs. These studies show that the two main factors associated with estimation biases are the group sizes and the relative magnitude of the variance components (as well as the sparcity of the Poisson data). These studies also suggest that hypothesis testing for fixed effects, against the usual null hypothesis of zero effect, can be reliably conducted using Wald tests using the estimated variance-covariance matrix of the fixed effects from PQL. Finally, they also indicate that the first order Laplace approximation may be useful for calculating approximate likelihood ratio tests for testing variance components. Chapter 4 contains discussion

of the HGLM approach of Lee and Nelder, which relies on either a first or second order approximation of the likelihood. Computational issues associated with the use of the HGLM approach are discussed in the context of a Fortran 90 implementation. Further simulation studies show that estimation biases for HGLM approaches are generally much smaller in magnitude than PQL, but the HGLM estimators can also be unstable for binary models with conditional expectations near 0 or 1. Some heuristic arguments for the relative performance of the HGLM approaches versus PQL are also presented.

Estimation biases for the PQL and the HGLM approaches are compared with Bayesian and GHQ approaches in Chapter 5 using a series of case studies. The approximate likelihood approaches performed reasonably well against Bayesian and GHQ approaches for all case studies presented, with the exception of the Rodriguez & Goldman (2001) datasets, with no finite maximum for the likelihood found using the (second order) HGLM approaches. The second order HGLM approach gave similar estimates to the Bayesian and GHQ approaches in a paired binary simulation study. Despite greater estimation biases, the PQL estimators had lower MSE than the GHQ estimators in a second paired binary (and Poisson) simulation study, in which the Bayesian estimator, with default priors, suffered estimation bias as well. PQL also performed relatively well against other approaches in a simulation study involving a randomised complete block design (RCBD) and in a simulation study involving a spatial GLMM, where PQL was compared with a much more computationally intensive Bayesian approach. These simulations also showed that the "REML-like" correction to the likelihood used by the HGLM and Bayesian approaches can give some positive estimation bias.

Whilst both approximate likelihood approaches had difficulties either in terms of estimation bias or instability, in general they perform relatively well against the other approaches and provide a useful and efficient way of fitting a wide variety of GLMMs. The use of a first or second order HGLM approach is generally preferable to PQL to achieve lower estimation biases. If PQL is employed, it is suggested that the

first order Laplace approximation be calculated for approximate testing of variance components.

Contents

List of Figures

List of Tables

THIS PAGE IS BLANK