

University of Wollongong - Research Online

Thesis Collection

Title: The performance of estimation methods for generalized linear mixed models

Author: Damian Collins

Year: 2008

Repository DOI:

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material.

Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

Research Online is the open access repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

2008

The performance of estimation methods for generalized linear mixed models

Damian Collins
University of Wollongong

Follow this and additional works at: <https://ro.uow.edu.au/theses>

University of Wollongong

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material.

Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

Recommended Citation

Collins, Damian, The performance of estimation methods for generalized linear mixed models, Doctor of Philosophy thesis, School of Mathematics & Applied Statistics - Faculty of Informatics, University of Wollongong, 2008. <https://ro.uow.edu.au/theses/1737>

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

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 camaraderie 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 sparsity 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

1	Review of basic elements of theory	1
1.1	Linear and generalized linear models and classical inferential approaches	1
1.1.1	Linear models	1
1.1.2	Generalized linear models	3
1.1.3	Maximum likelihood estimation	5
1.2	Linear mixed models	13
1.2.1	Specification	13
1.2.2	Estimation and Prediction	15
1.2.3	Usefulness of the linear mixed model	18
1.3	Further issues	18
1.3.1	Bayesian estimation	18
1.3.2	Integral approximations	21
1.4	The generalized linear mixed model	23
1.4.1	Specification	23
1.4.2	The problem of likelihood inference for GLMMs	25
1.4.3	Alternatives to GLMMs	26
1.5	Objectives of this research	29

2	Review of approaches to estimation for GLMMs	31
2.1	Approximate approaches (Laplace based)	31
2.1.1	Penalized quasi-likelihood	32
2.1.2	Hierarchical GLM approach of Lee and Nelder	41
2.2	Numerical methods – Gauss-Hermite quadrature	45
2.2.1	Quadrature for nested random effects models	46
2.2.2	Adaptive Gauss-Hermite quadrature	48
2.2.3	Implementation of Gauss-Hermite quadrature for GLMMs . . .	50
2.3	Stochastic methods (including full Bayesian MCMC)	53
2.3.1	Monte Carlo methods	53
2.3.2	Full Bayesian approaches	56
2.4	Marginal approaches and other approaches	58
2.4.1	Marginal approaches	58
2.4.2	Non-parametric GLMM – Aitkin (1999)	62
2.4.3	Modified EM approach – Steele (1996)	63
2.5	Discussion	64
3	The use of PQL for GLMMs	65
3.1	Factors affecting estimation bias	65
3.1.1	Background	65
3.1.2	Aims	67
3.1.3	Methodology	69
3.1.4	Designs with independent random effects	71
3.1.5	Designs with correlated random effects	88
3.1.6	Discussion	97

3.1.7	Case study : Beitler-Landis dataset	100
3.2	Other statistical inference using PQL	102
3.2.1	Inference concerning variance components	102
3.2.2	Inference concerning the fixed effects	105
3.3	Discussion	109
4	The HGLM approach of Lee and Nelder	114
4.1	Review of the HGLM methodology, and comparison with PQL	114
4.2	First order HGLM approaches	118
4.2.1	A Fortran 90 implementation with numerical derivatives . . .	118
4.2.2	Performance in simulation studies compared to PQL	121
4.2.3	Analytical expressions for the score equations	132
4.2.4	Adequacy of the (first order) Laplace approximation	142
4.3	Second order HGLM approaches	145
4.3.1	An expression for the second order Laplace correction term . .	146
4.3.2	Computation of the second order Laplace correction term . . .	149
4.3.3	Performance in simulation studies	149
4.4	Discussion	151
5	Case studies	156
5.1	Preliminaries	156
5.1.1	Review of alternative approaches	156
5.1.2	Software used in these case studies	158
5.2	Simple comparisons	162
5.2.1	The Beitler-Landis dataset	162

5.2.2	A paired binary simulation study	163
5.2.3	Further paired binary (and Poisson) simulation studies	166
5.2.4	The Rodriguez-Goldman datasets	169
5.2.5	Simulation study using a “typical” RCBD	173
5.2.6	The Salamander dataset	175
5.3	A simulation study using spatially correlated errors	178
5.3.1	Methods	179
5.3.2	Results	182
5.4	A “real-life” dataset with an ordinal response	184
5.4.1	Description of the dataset	184
5.4.2	Analysis of the “real-life” dataset	185
5.4.3	Simulation study	188
5.5	Discussion	191
6	Conclusions	194
	Bibliography	199
A	Appendix	216
A.1	Expressions for implicit differentiation	216
A.2	The second order Laplace approximation of an integral	217
A.2.1	Higher order Laplace approximations for a univariate integral .	217
A.2.2	The “second order” Laplace approximation	219
A.3	Delta method of calculating SEs for PQL spatial predictions	220
A.4	Laplace approximations for the ordinal (5.8) and ordinal factor analytic (5.9) models	221

List of Figures

1.1	A heuristic explanation of the first order Laplace approximation . . .	22
3.1	PQL estimation biases for the binary one-way classification model (3.1): effects of the group size and the variance parameter	73
3.2	PQL estimation biases for the binary one-way classification model (3.1): effects of the number of groups and the variance parameter . . .	74
3.3	PQL estimation biases for the binary one-way classification model (3.1): effects of the within-group fixed coefficient.	74
3.4	PQL estimation biases for the Poisson one-way classification model (3.1): the biases for the variance parameter and fixed coefficients. . . .	75
3.5	PQL estimation biases for the Poisson one-way classification model (3.1): the biases for the intercept	76
3.6	Testing Breslow's hypothesis: the marginal distributions of number of successes in grouped data using two parameter settings	79
3.7	PQL estimation biases for the binary nested two-way model (3.3) . . .	81
3.8	PQL estimation biases for the Poisson nested two-way model (3.3): the variance parameters	82
3.9	PQL estimation biases for the Poisson nested two-way model (3.3): the intercept	83
3.10	PQL estimation biases for the binary crossed two-way model (3.4) . .	84

3.11 PQL estimation biases for the Poisson crossed two-way model (3.4):	
the variance parameters	85
3.12 PQL estimation biases for the Poisson crossed two-way model (3.4):	
the intercept	86
3.13 PQL estimation biases for a crossed (3.5) binary model with many	
fixed effects	87
3.14 PQL estimation biases for a nested (3.6) binary model with many fixed	
effects	87
3.15 PQL estimation biases for the binary random coefficient model (3.7):	
the variance parameters	90
3.16 PQL estimation biases for the binary random coefficients model (3.7):	
the fixed coefficients	91
3.17 PQL estimation biases for the binary random coefficients model (3.7):	
the fixed coefficients, second plot	92
3.18 PQL estimation biases for the Poisson random coefficient model (3.7):	
the variance parameters	92
3.19 PQL estimation biases for the Poisson random coefficient model (3.7):	
the fixed coefficients	93
3.20 PQL estimation biases for the Poisson random coefficient model (3.7):	
the fixed coefficients, part two	93
3.21 PQL estimation biases for the binary AR correlated model (3.8) . . .	96
3.22 PQL estimation biases for the binary AR correlated model (3.8, part	
two)	97
3.23 PQL estimation biases for the Poisson AR correlated model (3.8) . . .	98
3.24 PQL estimation biases for the Poisson AR correlated model (3.8), part	
two)	99

3.25	Average estimated SEs vs Monte Carlo SEs for the binary one-way classification model (3.1) using PQL	107
3.26	Average estimated SEs vs Monte Carlo SEs for the Poisson one-way classification model (3.1) using PQL	108
3.27	Average estimated SEs vs Monte Carlo SEs for the binary nested two way (3.3), AR correlated (3.8) and crossed two way models (3.4) using PQL	108
3.28	Average estimated SEs vs Monte Carlo SEs for the Poisson nested two way (3.3), crossed two-way (3.4), and AR correlated (3.8) models using PQL	109
3.29	Average estimated SEs vs Monte Carlo SEs for the binary random coefficients model (3.7) using PQL	110
3.30	Average estimated SEs vs Monte Carlo SEs for the Poisson random coefficients model (3.7) using PQL	111
4.1	Estimation biases for first order HGLM approximations and PQL for the binary one-way classification model (4.5): the variance parameter.	123
4.2	Estimation biases for first order HGLM approximations and PQL in the binary one-way classification model (4.5): the intercept.	124
4.3	Estimation biases for first order HGLM approximations and PQL in the Poisson one-way classification model (4.5): the variance parameter.	125
4.4	Estimation biases for first order HGLM approximations and PQL in the Poisson one-way classification model (4.5): the intercept. (4.5).	126
4.5	Estimation biases for first order HGLM approximations and PQL for the binary one-way classification model (4.5), where the intercept $\tau_0 = 2$	129
4.6	Estimation biases for first order HGLM approximations and PQL for the binary nested two-way classification model (4.6).	131

4.7	The adequacy of the Laplace approximation for the binary one-way classification (part one)	144
4.8	The adequacy of the Laplace approximation for the binary one-way classification (part two)	144
4.9	The adequacy of the Laplace approximation for the Poisson one-way classification.	145
4.10	Estimation biases for second order HGLM approximations and PQL for the binary one-way classification model (4.5).	152
4.11	Estimation biases for second order HGLM approximations and PQL for the binary one-way classification model (4.5).	153
4.12	Estimation biases for second order HGLM approximations and PQL for the binary nested two-way classification model (4.6).	154
5.1	Profile likelihoods for the variance parameter in the Landis-Beitler model (3.9) for AGHQ, PQL and HG(0,1) approaches	164
5.2	AGHQ estimates of the variance parameter in the paired binary study (5.1) versus those from PQL, Bayesian and second order HGLM approaches.	167
5.3	Box plots of the estimates of the variance parameter for PQL and AGHQ in the second paired binary study (5.2).	169
5.4	Profile likelihoods for GHQ and HG(1,2) (no REML correction) for the first Rodriguez-Goldman dataset (5.3).	172
5.5	Diagram of sampled and predicted locations for the spatial case study (5.6) and the Matérn correlation function.	180
5.6	An illustration of the estimation errors for both PQL and Bayesian approaches for the spatial case study (5.6).	183
5.7	Design of the phytophera trial.	185

List of Tables

3.1	Values of the simulation parameters used for the one-way classification study (3.1).	71
3.2	Testing Breslow's hypothesis: comparison of the estimation bias and probabilities of low successes/failures for grouped binary data	79
3.3	Values of the simulation parameters used for the nested two-way classification study (3.3).	80
3.4	Values of the simulation parameters used for the crossed two-way classification study (3.3).	82
3.5	Values of the simulation parameters used for the crossed (3.5) and nested (3.6) binary models with many fixed effects.	86
3.6	Values of the simulation parameters used for the random coefficients model (3.7)	89
3.7	Values of the simulation parameters used for the correlated AR model (3.8)	95
3.8	The Beitler & Landis (1985) dataset used in Breslow (2003).	100
3.9	Estimates from the analysis of the Beitler & Landis (1985) dataset (3.9) using PQL, GHQ and Bayesian approaches.	101
3.10	Average parameter estimates from simulation studies based on the Beitler-Landis dataset (Table 3.8) using PQL	102

3.11	Values of the simulation parameters in model (3.10) for testing a single variance component in the one-way classification model (3.11) using PQL.	103
3.12	Type I error rates for testing a single variance component in the one-way classification model (3.11) using PQL and the Laplace approximation of the likelihood.	104
3.13	Type I error rates for testing fixed coefficients in models (3.1), (3.3), (3.4), (3.8) and (3.7) using PQL.	106
3.14	Type I error rates for testing fixed coefficients in the random coefficients model (3.7) using PQL	107
4.1	The levels of approximation using the HGLM approach of Lee & Nelder, and their corresponding likelihood expressions for fixed effects and variance parameters	117
4.2	Values of the simulation parameters used for the one-way classification study (3.1) comparing first order HGLM approximations and PQL. . .	122
4.3	Values of the simulation parameters used for the nested two-way classification study (4.6) comparing first order HGLM approximations and PQL.	130
4.4	Fortran-style pseudo-code to compute the adjustment ζ (4.8) to the mixed model equations (4.7) required for the HG(1, j) approaches ($j \geq 1$).	136
4.5	Fortran style pseudo-code required to compute the correction term (4.21) for the second order Laplace approximation.	150
4.6	Values of the simulation parameters used for the one-way classification study (3.1) comparing second order HGLM approximations and PQL.	150
5.1	Estimates from the analysis of Beitler/Landis data (table 3.8, equation (3.9)) using PQL, adaptive GHQ and Bayesian approaches.	162

5.2	Average estimates (\pm SE) for the paired binary study (5.1) using the PQL, Bugs, AGHQ and HGLM approaches.	166
5.3	Average estimates (\pm SE) from the (second) paired binary and Poisson studies (5.2) using PQL, Bayesian, AGHQ and HGLM approaches. . .	168
5.4	Estimates of variance parameters for the Rodriguez & Goldman (2001) datasets (5.3) using PQL, Bayesian, GHQ and HGLM approaches. . .	172
5.5	Average estimates from a simulation study using an RCBD design (5.4) using PQL, Bayesian, GHQ and HGLM approaches.	176
5.6	Estimates for the summer and pooled salamander datasets from model (5.4) using PQL, Bayesian and HGLM approaches.	178
5.7	Average parameter estimates, estimation and prediction errors, and true 95% confidence interval coverages for the PQL and Bayesian approaches in the spatial case study (5.6).	183
5.8	Variance component estimates from fitting models (5.7), (5.8) and (5.9) to the phytophthora dataset using PQL.	189
5.9	Estimates of the variance components associated with the factor-analytic ordinal model (5.9) to the phytophthora data using PQL.	189
5.10	Average estimates of variance components from a simulation study based on the phytophthora dataet, using the ordinal model (5.8) and the corresponding binomial model (5.10).	191
5.11	Estimated null distribution of the LRT statistic for testing (5.9) against (5.8), using the Laplace approximation and PQL.	191

THIS PAGE IS BLANK