

A COMPARISON OF METHODS FOR ANALYSING
CORRELATED COUNT DATA.

By

Clair Alston

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

OF

THE UNIVERSITY OF NEW ENGLAND

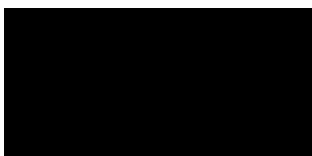
MARCH 1997

Preface

I hereby declare that this thesis describes my own original work, supervised by Dr Robert Murison (major supervisor), Dr David Smith and Dr Ian Davies.

I certify that the substance of this thesis has not already been submitted for any degree and is not currently being submitted for any other degree.

I certify that any help received in preparing this thesis, and all sources used, have been acknowledged in this thesis.



Clair Alston.

Acknowledgements

I would like to thank my supervisors, Dr Robert Murison and Dr David Smith, for their guidance and support with this work.

Graham Charles of NSW Agriculture, Narrabri, is thanked for providing the herbicide data set which I have used extensively in this thesis.

Thanks are also due to Adrian Doss who scribbled all over several “final” drafts with red ink, to Steven Harder for various computing tips, and to Dr Ian Davies, Tony and Barbara Bernardi for helpful comments on a recent draft.

Family and friends are also gratefully acknowledged for encouraging and tolerating me during the time I have spent doing this project.

Abstract

This thesis considers extensions of Generalized Linear Models (Nelder and Wedderburn, 1972) to incorporate correlated count data. Of particular interest is the Poisson random effects model which is commonly solved by approximate methods due to the complexity of calculations in maximum likelihood estimation (Diggle, Liang and Zeger, 1994, p173-5).

The methods considered fall into 4 categories;

1. quasi-likelihood techniques, (Schall, 1991), (Breslow and Clayton, 1993),
2. overdispersion models, (Van de Ven and Weber, 1995)
3. generalized estimating equations, (Liang and Zeger, 1986), and
4. Markov Chain Monte Carlo techniques, (Zeger and Karim, 1991).

These techniques are examined and compared both algebraically and through the use of a small simulation study. On this basis, some recommendations for the use of these methods in practice are made.

The variogram is used to determine which error model is appropriate to use with

a number of data sets, and use of several residual types resulting from GLMs are compared. This comparison is done so that the appropriate error model is most evident at the investigative stage of the analysis.

Contents

Preface	ii
Acknowledgements	iii
Abstract	iv
1 Introduction	1
1.1 Generalized linear models	3
1.1.1 Definition of generalized linear models	5
1.1.2 Likelihood functions	6
1.1.3 Correlation in longitudinal and temporal data	7
1.2 Models for correlation structure	8
1.3 Random effects models	15
2 Comparative analysis of herbicide experiment	19
2.1 The experimental design	19
2.2 Preliminary analysis	22
2.3 Diagnosing random effects from correlation structure	25

2.4	GLMM analysis	28
2.5	GEE analysis	33
3	Applications to count data	36
3.1	Introduction	36
3.2	Methods for analysing correlated Poisson data	41
3.2.1	Generalized estimating equations	42
3.2.2	Approximations using penalized quasi-likelihood	45
3.2.3	Approximations using marginal quasi-likelihood	47
3.2.4	Approximations using the Gibbs sampler	49
3.2.5	Dealing with overdispersion by using mixing distributions	53
4	Simulation Study	56
4.1	Outline of algorithm	56
4.2	Implementation of algorithm	58
4.3	Simulation results	59
4.4	Practical implications	61
4.5	PQL with small sample sizes	63
5	Error model diagnostics for Poisson GLMMs using the variogram	66
5.1	Introduction	66
5.2	The variogram for detecting random effects and serial correlation in correlated Poisson models.	69
5.3	Variograms from a simulation study	71

A Iteratively weighted least squares	78
B Identities	80
C Derivation of $E(\mathbf{y})$ and $\text{var}(\mathbf{y})$ in Poisson GLMM	81
D Derivation of log-likelihood and score functions in Poisson GLMM	83
E Deriving the likelihood for Poisson mixed with Gamma distribution	88
F Simulation procedure for correlated Poisson data (Exchangeable structure)	90
G Splus routine for simulation study using EQL methodology	99
H Derivation of variogram	110
References	112

List of Tables

2.1	Table of experimental treatments.	20
2.2	Treatment counts by design row.	22
4.1	Simulation results (\pm se) for exchangeable correlation structure, $\rho = 0.2$	59
4.2	Simulation results (\pm se) for exchangeable correlation structure, $\rho = 0.5$	60
4.3	Comparison of MQL and EQL results for slide example	63

List of Figures

1.1	Growth profiles from simulated data.	14
2.1	Experimental design of nutgrass trial.	21
2.2	Raw data from nutgrass experiment.	23
2.3	Contour plot of residuals from model (2.1), year 1.	24
2.4	Contour plot of theoretical correlations amongst residuals. a) Correlation between time 1 & 2, b) Correlation between time 1 & 3 and c) Correlation between time 2 & 3.	27
2.5	Comparison of estimates for nutgrass model.	30
2.6	Contour plot of residuals from MQL model, year 1.	31
2.7	Contour plot of residuals from PQL model, year 1.	32
2.8	Contour plot of residuals from GEE model, year 1.	34
3.1	Example correlation structures, a) autoregressive, b) uniform.	43
3.2	Illustration of rejection sampling concept.	51
5.1	Example variogram containing 3 error sources.	68
5.2	Comparison of two residuals in Simulation study (one set). The dotted line is 1:1.	72

5.3	Simulation study variograms for true residuals (5.9).	73
5.4	Simulation study variograms for working residuals.	74
5.5	Simulation study variograms for raw residuals.	77