

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Small Area Estimation
via
Generalized Linear Models.

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Statistics

At Massey University, Palmerston North, New Zealand.

Alasdair D. L. Noble

2003



CERTIFICATE OF REGULATORY COMPLIANCE

This is to certify that the research carried out in
the Doctoral Thesis entitled

Small Area Estimation via Generalized Linear Models

in the
Institute of Information Sciences and Technology
and
Statistics Research and Consulting Centre
at Massey University, New Zealand

- (a) is the original work of the candidate, except as indicated by appropriate attribution in the text and/or in the acknowledgements;
- (b) that the text, excluding appendices/annexes, does not exceed 100,000 words;
- (c) all the ethical requirements applicable to this study have been complied with as required by Massey University, other organizations and/or committees which had a particular association with this study, and relevant legislation.

Please note Ethical Authorisation code(s) were not applicable.

Candidate's Name: Alasdair Dewar Lowe Noble

Signature: 

Date: 13/12/04

Supervisor's Name: Dr Stephen Haslett

Signature: 

Date 13/12/04



Massey University
COLLEGE OF SCIENCES

INSTITUTE OF INFORMATION
SCIENCES & TECHNOLOGY
Private Bag 11 222
Palmerston North
New Zealand
T 64 6 356 9099
F 64 6 350 5750
www.massey.ac.nz
www-ist.massey.ac.nz

Integrated research and
teaching in the fields of
• Statistics
• Computer Science
• Electronics & InfoComm
Engineering

CANDIDATE'S DECLARATION

This is to certify that the research carried out for my Doctoral thesis entitled:

“Small Area Estimation via Generalized Linear Models”

in the:

Institute of Information Sciences and Technology,
and Statistics Research and Consulting Centre
Massey University,
Palmerston North,
New Zealand

is my own work and that the thesis material has not been used in part or in whole for
any other qualification.

Alasdair Dewar Lowe Noble

Signature

alnw

Date

13/12/04



SUPERVISOR'S DECLARATION

This is to certify that the research carried out for the Doctoral thesis entitled "Small Area Estimation via Generalized Linear Models" was done by Alasdair Noble in the Institute of Information Sciences and Technology, and the Statistics Research and Consulting Centre, Massey University, Palmerston North, New Zealand. The thesis material has not been used in part or in whole for any other qualification, and I confirm that the candidate has pursued the course of study in accordance with the requirements of the Massey University regulations.

Supervisor's Name

Dr. Stephen Haslett

Signature

Date

13/12/04

Abstract

Survey information is commonly collected to yield estimates of quantities for large geographic areas, for example, complete countries. However the estimates of those quantities at much smaller geographic areas are often of interest and the sample sizes in these areas are generally too small to give useful results. Small area estimation is used to make inference about those small areas with greater precision than the direct estimates, either by exploiting similarities between different small areas or by accessing additional information often from administrative records.

The majority of the traditional small area estimation methods are examples of a simple linear model Marker (1999) and this work begins by extending the model to a generalized linear model (GLM) Nelder and Wedderburn (1972) and then including structure preserving estimation (SPREE) in the classification. This had not been done previously.

SPREE had previously been fitted using the iterative proportional fitting algorithm Deming and Stephan (1940) which could be described as a “black box” approach. By expressing SPREE in terms of a GLM an alternative algorithm for fitting the method is developed which elucidates the underlying concepts. This new approach allows the method to be extended from the contingency table with categorical variables which the IPF could fit, to continuous variables and random effects models. An example including a continuous variable is given.

SPREE is a method which uses auxiliary information as well as survey data. In the past assumptions about appropriate auxiliary information have been made with little theoretical support. The new approach allows these assumptions to be considered and they are found to be wanting in some cases.

An example based on a national survey in New Zealand for unemployment statistics, is used extensively throughout the thesis. These data have characteristics that make analysis in the Bayesian paradigm appropriate. This paradigm has been applied and a conditional autoregressive error structure is considered.

Finally relative risk models are considered. It is shown that these could have been fitted using the IPF algorithm but the new approach allows combinations of other modeling techniques which are not available using IPF.

Acknowledgements

To fully acknowledge everyone who has had some input into this thesis would require far more space than is sensible to use. To have embarked on this journey at a mature age has meant that many people have supported me in many ways over a number of years, to all of them I am very appreciative. A few who deserve particular mention are listed below.

To all of you who have been supportive and though you may have felt I was being stupid attempting this, never voiced your thoughts; I thank you.

Firstly to Jeff Hunter and Dick Brook who were so encouraging when I first arrived at Massey University as an experienced teacher but a very naïve research student.

To the members of the then Department of Statistics and those now in the statistics group in the Institute of Information Sciences and Technology, you have been very patient, accepted my few strengths and unfailingly helped me in times of need (and there have been many). I think I have brought something to the group but know that I have taken far more.

To Doug Stirling who supervised my Masters thesis and in that way began my independent academic thinking.

To Greg Arnold who, as second supervisor, has helped me in many subtle ways. Your self deprecating manner belies a very thorough understanding of statistics and although the topic may not be very familiar to you you have always managed to relate it to areas that are familiar and in so doing bring a new light to the problem. This breadth has helped me often.

Finally on the academic side Steve Haslett. I am fairly sure when you took me on as a PhD student you had little idea of what you were letting yourself in for. I suspect the experience will make you more careful in your selection in future. I hope you have learnt a little about small area estimation through our work, I am sure that you have learnt some things about the variability in styles of learning. This work would not have been possible without you and I am certain that the benefit to me has been far greater than any benefit to you.

To Heather, Josie and (always last) Alex. I cannot thank you enough for your support through the past few years. It has had its ups and downs but hopefully the future will make it all worthwhile. At last I will be able to say to Josie and Alex "Yes I am a doctor now".

Finally I have tried to follow my Great Great Uncle Sir James Dewar's maxim:

"Minds are like parachutes; they only function when they are open."

I am not sure that I have succeeded all of the time.

Table of Contents

Abstract.....	i
Acknowledgements	iii
Table of Contents.....	v
Table Of Figures and Tables.....	ix
CHAPTER 1	1
Introduction	1
1.1 Small area estimation.....	1
1.2 Formulation	6
1.3 A new approach to SPREE.	9
1.4 Scope of this work	10
1.5 Additional topics.....	11
1.5.1 Bayesian approaches.....	11
1.5.2 Spatial statistics	12
1.5.3 Relative risk models.....	13
1.6 Computational aspects	14
1.7 Conclusions	16
CHAPTER 2	18
An Historical Background	18
2.1 Introduction	18
2.2 Historical methods in a linear regression framework	20
2.2.1 Demographic methods	20
2.2.2 Synthetic and related methods	25
2.2.3 Symptomatic regression.....	27
2.2.4 Structure preserving estimation (SPREE).....	31
2.2.5 Composite estimation	33
2.3 Models with area specific effects.....	36
2.4 EBLUP, EB and HB approaches	37
2.5 Concluding remarks.....	38
CHAPTER 3	40
Structure Preserving Estimation; the link with the Generalized Linear Model.	40
3.1 Introduction	40
3.2 SPREE by the iterative proportional fitting algorithm.....	42
3.3 A simple example	52
3.4 An alternative approach, the generalized linear model.	55
3.5 The new approach.....	61
3.6 Application of the GLM to our data	64
3.7 Identifying the effects and interactions which are updated by the sample survey data.....	65
3.8 A simple example	69
3.9 Another approach for binary data.	72
3.10 Fitting the models.	73
3.11 Concluding remarks.....	79
CHAPTER 4	82
An example of the Generalized Linear Model approach	82
4.1 Introduction.	82
4.2 The data	85
4.3 Application of the new algorithm	88
4.4 Brief notes on computing.....	99
4.5 Conclusions	100
CHAPTER 5	102
Quadratic and linear functions for the age variable	102
5.1 Introduction	102
5.2 A quadratic function for age.	103
5.3 Other possible models.....	109
5.4 A more realistic model.....	116

5.5	Closing comments	121
CHAPTER 6	122
	The relationship between the census and sample survey data	122
6.1	Introduction	122
6.2	The relationship between the two data sources.....	123
6.3	Practical considerations in calculating the correlations.....	128
6.4	A more detailed look at the correlations between parts of the model.....	130
6.5	Transformations of variables and the effect on correlations.....	132
6.6	Suggestions for model checking based on this.....	133
6.7	Concluding remarks.....	134
CHAPTER 7	137
	Bayesian approaches to parameter estimation.....	137
7.1	Introduction	137
7.2	Frequentist and Bayesian statistics	139
7.3	Bayesian solutions, computing approaches	143
7.4	The data	156
7.5	Choice of priors	157
7.6	Bayesian solution with a quadratic function.....	162
7.7	Variance estimation	163
7.8	Conclusions	165
CHAPTER 8	167
	Spatial models, a conditional autoregressive (CAR) approach.....	167
8.1	Introduction	167
8.2	The CAR model.....	169
8.3	Implementation in WinBUGS	170
8.4	Specification of the CAR model in WinBUGS	171
8.5	An Example	176
8.6	Edge Effects.....	183
8.7	Adjacencies other than simple geographic.....	184
8.8	Conclusions	185
CHAPTER 9	188
	A relative risk and odds ratio approach	188
9.1	Introduction	188
9.2	Relative risk models	190
9.3	A simple example	193
9.4	The data	197
9.5	Results	197
9.6	Discussion.....	201
CHAPTER 10	203
Conclusions	203
10.1	Introduction	203
10.2	The linear regression framework	204
10.3	The extension to Include SPREE	205
10.4	The wider application of the new algorithm	207
10.5	The assumptions in SPREE and the relationship between the two data sources.....	208
10.6	Bayesian approaches and variance estimation.....	210
10.7	Conditional autoregressive and relative risk models	211
10.8	Comments about the data used in this thesis and practical considerations	211
10.9	Final conclusions and suggestions for future work.....	212
Bibliography	216
Appendix A	224
Detailed calculations from Chapter 3	224
	The Iterative Proportional Fitting Algorithm Examples	224
	The Generalized Linear Model Calculations.....	228
Appendix B	237
Design Matrices Construction and Checking	237
Appendix C	243

Computer programs used in the thesis with chapter references.....	243
EG 2 Chapter 3.....	243
SPREE Equivalent model Census data and new margins.....	245
Relative risk models chapter 9.....	248
Appendix D.....	250
Examples of WinBUGS output.....	250

Table Of Figures and Tables

Figure 1.1	Generation of "Small Areas". Subdivisions of geographic regions or divisions that cut across the divisions used for sampling.	3
Table 1.1	Models fitted, computer software used and chapter references.	15
Figure 2.1	Relationship of variables in SPREE	32
Figure 3.1	Diagram showing the relationship between the two data sources and the small area estimates in a simple example.	43
Figure 3.2	Diagram showing the association structure.	46
Figure 3.3	Main effects in a two by two table	65
Figure 3.4	One two dimensional margin.	66
Figure 3.5	One single dimensional margin.	67
Figure 3.6	Two new single dimensional margins from survey data.	68
Figure 3.7	The three dimensional diagram of cell counts for the $2 \times 2 \times 2$ table presented in figure 3.3.	69
Figure 3.8	MLwiN screed for constraining parameters in a model.	77
Figure 3.9	MLwiN output screen.	78
Table 4.1	Models fitted, computer software used and chapter references.	82
Figure 4.1	Map of the Regional Authorities of New Zealand.	87
Table 4.2	Census data from Work and Income New Zealand for unemployment counts by sex and three age groups in each Region.	89
Table 4.3	Table of coefficients for the full model with categorical variables for region, sex and the two age categories. Coefficients in bold type will be carried forward.	92
Figure 4.2	Part of the MLwiN window for constraining parameters.	94
Table 4.4	Fully saturated model fitted to the sample data with constrained coefficients. The reestimated coefficients are shown in bold.	96
Table 4.5	Final estimates from the combined model.	97
Table 5.1	Table showing the different models used in this thesis, the chapters in which they are discussed, the estimation process and computer package used.	101
Figure 5.1	Graphs of unemployment counts against the three age groups by regions for males and females.	104
Figure 5.2	Matlab sparse matrix representation of the design matrix.	105
Table 5.2	Table of coefficients for the full model with categorical variables for region and sex, and linear and quadratic terms for age.	107
Table 5.3	Predicted counts for the saturated model with a quadratic term for age.	108

Table 5.4	Table of coefficients for the model with categorical variables for region and sex and a linear term for age with all interactions.	111
Table 5.5	Predicted counts and residuals for the linear model with all interactions..	112
Table 5.6	Predicted counts and residuals for the linear model with no interactions	114
Table 5.7	Table of coefficients for the model with the linear effect as the only age effect.	115
Figure 5.3	Graphs of un employment counts against the eleven age groups for males and females.	116
Table 5.8	The new unemployment margin for the five yearly age groups. The margin for sex stays as before.	117
Table 5.9	Counts for unemployment from Department of Work and Income data in five yearly intervals.	118
Table 5.10	Table of coefficients for the full model with categorical variables for region and sex and linear and quadratic terms for age.	119
Table 5.11	Predictions for unemployment in five yearly intervals.	120
Table 6.1	Relationship between "Correlation between Y_c and Y_s " and probable success of SPREE based estimation.	126
Table 7.1	Output from WinBUGS program.	150
Figure 7.1	Graphs showing the convergence of samples from a BUGS program.	151
Figure 7.2	Density curves for the five coefficients.	152
Figure 7.3	Autocorrelation plots for the iterations for the five coefficients.	153
Figure 7.4	Histograms of replicates for the margins from sample survey data.	158
Figure 7.5	Normal probability plots for the replicates of the new margins from the sample survey data.	159
Table 7.2	Means and variances for the new margins from 512 replicates of survey data.	159
Figure 7.6	Quantile-Quantile plot and histogram of probabilities reported by Anderson Darling tests for normality for the 54 cell contingency table used in the earlier example.	161
Figure 8.1	Map of regions and diagram showing adjacencies used.	177
Table 8.1	Unemployment counts by region. Census counts from NZDWI data, and estimates using the new algorithm (SPREE) and using the new algorithm including an autoregressive error structure	178

Table 8.2	Comparison of the census data with estimates from a SPREE type analysis and the new approach including a conditional autoregressive error structure, North Island regions. Prior for the precision of the CAR parameters was Gamma(0.5, 0.005)	179
Table 8.3	Comparison of the census data with estimates from a SPREE type analysis and the new approach including a conditional autoregressive error structure, South Island regions. Prior for the precision of the CAR parameters was Gamma(0.5, 0.005)	180
Figure 8.2	Graphs of estimates with and without the conditional autoregressive term	181
Figure 8.3	The first graph from figure 8.2 with large values removed to show the structure better. Regions 1 to 16 are identified	182
Figure 8.4	Map of small areas within a region (shaded) with other small areas around.	184
Figure 9.1	Survey data available for the relative risks model.	194
Figure 9.2	History graph for the constant term in the model for the census data.	195
Table 9.1	Estimates found by SPREE and the relative risks model for North Island regions.	199
Table 9.2	Estimates found by SPREE and the relative risks model for South Island regions.	200
Figure 9.3	Graph of relative risk s model vs SPREE, numbered points are noted below.	201
Figure Appendix A.1	A 2 x 2 x 2 table with new margins.	235