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

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

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 adNoble

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

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

Table Of Figures and Tables

CHAPTER 1

Introduction

1.1 Small area estimation

We live in the so called knowledge economy, or the information age. This information is collected by a huge range of organizations and individuals and is then used in many different ways. Some data are collected for specific purposes, some for storage in large databases for subsequent searching, some as a by product of administrative structure and some appear to be collected simply because it is easy to do so.

Where the information required from the data is important the quality of the data collected is obviously important. This generally means that the cost of the data collection increases. The ideal data collection is, of course, to know all that you need about every object that you want to know about. This would be a census, but it is not usually possible. Instead it is common to sample some of the objects from the population, and collect and record the necessary information for them and then make a prediction about the larger group. What we will discuss in this thesis can be applied to any objects collected using a sample and to any measurements made on them.

The example used extensively in this thesis concerns counts of unemployed people. A sample of people from throughout the country are selected and they are asked whether they are in paid employment or not. This is done by Statistics New Zealand in what they call the Household Labour Force Survey

(HLFS). The survey asks more than whether the person has work or not but it is the unemployment data which we will consider. The population is known to be divided into a number of distinct groups such as males and females, age groupings and ethnicities. These are known as subpopulations in sample surveys. If membership of these subpopulations is known before the sample is selected then they may be used as strata to select the sample in such a way as to ensure that those groups are represented in the sample with given subpopulation sample sizes. Designing the sample in ways such as these makes it more complex but improves the estimates. A full discussion of the Statistics New Zealand data and how it is collected will appear in Chapter Four.

Once the sample data have been collected a value for the total number of people unemployed can be estimated. This is an estimate only as we have not asked all of the population and if a different sample were selected we would get a slightly different answer. Estimates derived directly from the data are known as direct estimates. Much of this thesis will be concerned with estimates which are derived from survey data in combination with data from other sources. These are known as indirect estimates and such estimates are based on a statistical model linking the different data sources.

Human nature, being what it is, we then commonly ask questions about smaller groups, "How do males compare with females?", "How does my town compare with the next?" and so on. It is not hard to see that as we look at smaller and smaller groups the number of people selected in the sample who come from that group also gets smaller and smaller. This causes an accuracy problem as the smaller the sample size the less accurate the estimate. For an organization such as Statistics New Zealand it is important to maintain high standards of data integrity and publishing estimates that have low

accuracy is not acceptable. Even worse it is possible that the sample size for some groups may be zero in which case it is not possible to make a direct estimate at all. Making estimates for small groups is the subject of this thesis.

In the literature there are a number of terms used to denote these smaller groups; small area, local area and small domain are common. We will use the term small area.

The small areas may be sub divisions of larger geographic regions or sub divisions which cut across the subpopulations or strata that were used in the data collection. The specific estimates required may be these individual domains or some larger aggregation of them. That is estimates may be formed for the different subpopulations or strata in each geographic area and then summed to give a total for that area.

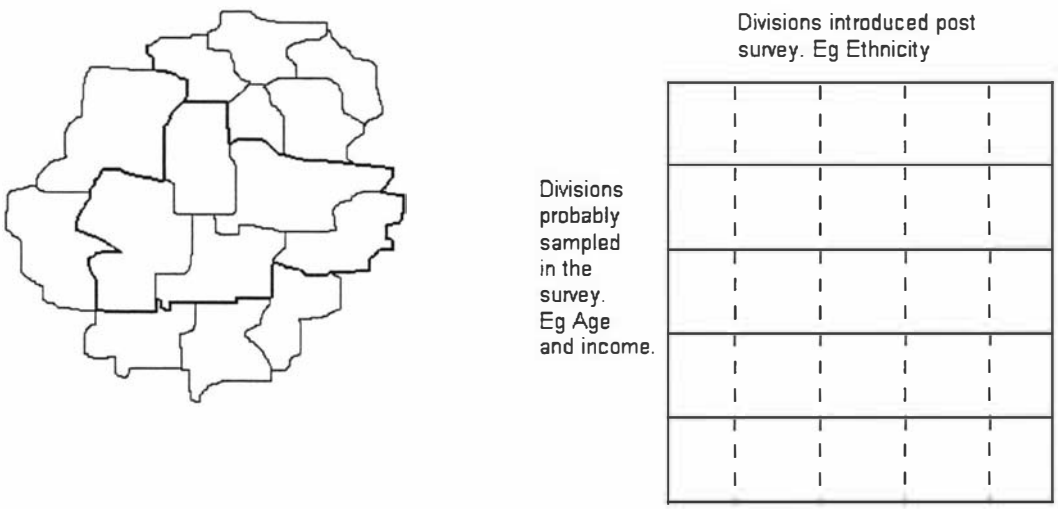


Figure 1.1 Generation of "Small Areas". Subdivisions of geographic regions or divisions that cut across the divisions used for sampling.

In the example which we will discuss a large sample survey is carried out and used to estimate counts of unemployment at a national level. Estimates for smaller administrative areas, or newly introduced strata such as industry groups, are not possible due to the small, or zero, sample sizes in many of those areas, or groups.

Nevertheless there has been an increase in demand for estimates for these small groups, particularly for appropriation or apportionment of government funds. The balance between costs involved in survey sampling and data processing has changed with the increasing costs of employing people to carry out surveys and the decreasing costs of powerful computers.

Several computationally and statistically powerful methods, with sound theoretical backgrounds, have been developed. These methods "borrow strength" from related or similar small areas through explicit or implicit statistical models that connect the small areas via supplementary data. In other words, data are available at the small area level, which can be used, via a model and past data, to estimate parameters for the unknown data at the small area level.

To return to the unemployment counts example, estimates may be possible for some subdivision or stratification of the sample survey data, for example male/female, urban/rural, age groups etc and census data could be expected to be available to give proportions of these subpopulations in each administrative area. This information can then be combined to give estimates of the unemployment count in individual administrative areas, which are the small areas in this case.

The data used to estimate the model parameters for the small area estimates are sometimes collected using complex sample surveys that are designed so

that certain predefined domains of interest are more adequately sampled. Commonly in New Zealand the sample size for Maori ethnicity will be boosted to ensure adequate estimates for this portion of the population even though they are generally less than 20% of the total population size. Despite such over sampling these surveys give good estimates for the variables of interest over the larger domains. Often the sample sizes are inadequate to give reliable estimates for the small areas, if an estimate can be found at all.

A number of small area estimation methods are reviewed in Chapter Two. Their historical use has largely been driven by the data that were available in a particular application. In the earlier methods census data were available from an earlier date and the objective was to update estimates from the census with data collected since, or by using historical relationships from more than one census.

More recently sample survey data, collected for other purposes, have been combined with different information about the small areas, from another source, to produce the estimates. The example used throughout this thesis uses unemployment data from an administrative government department as a census to be updated by survey data from a different government organization. The definitions that the two organizations use for unemployment are different but it is assumed that the relationship between the measures and other known variables is close and so that some model parameters found under one definition can be used under the other. This concept can be widened to variables that are not measuring the same thing but are believed to be closely related. The definition of "closely related" will be considered in Chapter Six and some guidelines suggested.

A number of earlier reviews of small area estimation were published for example Morrison (1971) National Research Council (1980), Purcell and Kish (1979), Zidek (1982) and McCullagh and Zidek (1987), in Platek, Rao, Sarndal and Singh (1987). More recently there have been a number of symposia/workshops, and some reviews, Rao (1986), Chaudhuri (1992) and Schaible (1992) and Ghosh and Rao (1994) presented a particularly full historical account. Ghosh and Rao (1994) was also updated in 1999, Rao (1999). Marker (1999) organises small area estimators using a generalized linear regression framework, and this has been extended to certain types of nonlinear estimation (such as apply for count data) in Noble, Haslett and Arnold (2002).

We have briefly introduced the idea of a small area estimate and the sorts of problems in which they arise. In introducing these ideas there are a number of technical issues which have also been introduced briefly. Where necessary these will be enlarged on later. The rest of this chapter will be an overview of the thesis with brief comments about the content which will be considered in the later chapters. The chapter signposts the development of the work and the topics which will be presented.

1.2 Formulation

In Chapter Two we will consider the historical solutions to the problem of small area estimation. The earlier methods are generally specific to the particular information in the data available. They were generally developed for use in estimating human populations. Several more recent techniques require a higher level of theoretical sophistication but are less dependent on the data format.

Here, and in more detail in Chapter Three, we will consider the generalized linear model, as these can be used to provide a formal structure within which almost all of the statistical procedures suggested can be placed. We will review the literature and discuss some more recent developments, and illustrate the link between the small area estimation procedures and the more general models.

The advantages of providing a framework for the different procedures are :-

1 The procedures can be compared more easily,

2 The estimation of parameters of each model is linked by the common framework,

3 The implicit assumptions in the different procedures are made more explicit by the model proposed,

4 The explicit assumptions make it easier to test them by residual analysis or other means.

We will show that the historical methods are all examples of a generalized linear model and most are examples of the simpler linear regression model which we will consider initially. Individual data sets will require their own structure in the model but the overall pattern will be shown to be the same.

The model is

$$g\left(\mathrm{E}[\mathbf{Y}]\right) = \mathbf{X}\boldsymbol{\beta} \qquad (1.1)$$

where:-

$g(\,)$ is a (possibly composite, Thompson and Baker (1981)) link function, monotonic and differentiable, that allows a wide variety of functions including the identity, ln, logit, probit, powers etc. Nelder and Wedderburn (1972).

$\mathbf{Y}$ is a vector of values for the response variable.

$\mathbf{X}$ is a known matrix constructed from values of categorical and/or continuous explanatory variables.

$\boldsymbol{\beta}$ is a vector of parameters to be estimated

The model is linear in the parameters, $\boldsymbol{\beta}$, but powers or other functions of the x variables are allowed in the $\mathbf{X}$ matrix.

The underlying distribution of the response can include any members of the exponential family of distributions which include the normal, binomial, Poisson etc. Nelder and Wedderburn (1972).

The simple models in which there is an identity link function and the response is normally distributed can be expressed as a simple linear regression model of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1.2)$$

where $\mathbf{Y}$ is a vector which is a set of observations

$\mathbf{X}$ will be a matrix, generally termed a design matrix

$\boldsymbol{\beta}$ is a vector of usually unknown parameters

$\boldsymbol{\varepsilon}$ is a vector of model errors.

Now the link function is the identity and this is the form of the framework that Marker has called generalized linear regression. In some circumstances each element of $\mathbf{Y}$ corresponds to each small area or domain. In others there is one element for each subgroup in each small area, or even for each observation in each small area. For random or mixed parameter models, $\mathbf{Y}$ can be supplemented so that the model can include random parameters, see Haslett J. and Haslett S. (2004). Generally $\mathbf{Y}$ will be defined to be l x 1, although for mixed and random coefficient models, l may not be the sample size. In particular models l will be defined explicitly. The matrix $\mathbf{X}$, vector of

parameters β, and model error vector ε will be of the appropriate order to ensure that the matrix operations can be carried out.

We have postulated a simple model which we will show can form a framework within which the various methods of small area estimation can be placed. Chapter Two will detail the models or procedures for small area estimation, that have been proposed and these will be shown to be examples of the model in equation (1.2).

Structure preserving estimation (SPREE) Purcell and Kish (1980) and its extension to a very much wider class of models will be a central theme in this thesis and will be introduced in Chapter Three. This method could not be included in the framework suggested by Marker. The generalized linear model as expressed in equation (1.1) is a wider class of models which includes the linear regression models as a subset. The SPREE method, as well as those which will be explained in Chapter Two, can be included in the generalized linear model approach. SPREE will be discussed in Chapter Three.

Linear regression models, log linear models for count data and logistic regression models will be used in various sections of the thesis. These are all examples of the GLM which, when it was introduced, unified all of these methods.

1.3 A new approach to SPREE.

As mentioned in the previous section SPREE will be properly introduced in Chapter Three. In that chapter we will look at the traditional method of fitting SPREE models using the iterative proportional fitting algorithm (IPF) of Deming and Stephan (1940). We will show that SPREE can alternatively be viewed as an application of a GLM and that this allows a different approach to obtaining the small area estimates. This new approach is more general and

so the underlying concept of SPREE can be applied to a whole new set of problems. These extensions along with examples will be considered in the remaining parts of the thesis.

The new approach is the central theme of the thesis. We believe that it has a number of advantages over the traditional method including allowing a wider range of data sources to be used and making the underlying assumptions more explicit. The detailed development of the method is in Chapter Three. An example of its use as an alternative to the earlier formulation of SPREE is in Chapter Four. The Chapter Four models could have been estimated by either the new method or the old and there would have been no difference. However the new method is then used in Chapters Five, Eight and Nine in contexts where the old method could not have been used. Chapter Six uses the new method to gain a greater understanding of some important requirements for the data used in small area estimation models, which have not previously been clear.

1.4 Scope of this work

Section 1.3 has indicated the development of this thesis. Whilst Chapter Two will consider a wide range of small area estimation methods the rest of the work will focus on a method which has certain data requirements. The method is applicable when there is a source of census data which can be used to build a model which can then be modified using sample survey data which may be more timely or, as in our example, more accurately measure the variable of interest. It is a requirement that the variable of interest, or some other variable closely related to it, be measured in both the census and survey.

1.5 Additional topics.

In the previous sections we have indicated that a new approach to SPREE will be the substance of this work. It extends the range of problems which the underlying principles of SPREE can be applied to. To illustrate the range of these extensions we will incorporate other statistical techniques. These already have considerable bodies of knowledge and there is no intention to explain them in detail. H owever they need to be briefly introduced. The important points of these additional techniques, particularly any which are central to the development of this work, will be explained at the beginning of the relevant chapter. Below are some brief notes on these topics.

This is not an exhaustive list of the possible areas into which the new approach could be extended. They are simply examples and give an indication of the range of possibilities, but do not limit them.

1.5.1 Bayesian approaches

For much of the latter part of this thesis we will use a Bayesian approach to the analysis. This will be done because the form of the data lends itself to a Bayesian approach. A secondary advantage is that we will include spatial variables and these can easily be fitted in this context.

Bayesian approaches to statistical analysis have been seen by some to be controversial. However there is now a greater acceptance partly due to the development of computer software which can estimate the distributions for the parameters in the model. In Chapter Seven we will introduce the Bayesian approach and we will show that there are good reasons for its use with the data that we have. In its earliest forms small area estimation by SPREE has assumed that the census data are exact. This assumption is not supported in the example we will use as the census data have been collected on a number

of occasions. We can thus estimate census cell counts as well as the variability in those counts. The survey data are collected as a large number of balanced repeated replicates and so again there is information about the variability in the survey data which can be exploited in the Bayesian approach. Balanced repeated replicates is not however a prerequisite to Bayesian methods.

The main advantage of the Bayesian approach is that by appropriate specification of the prior distributions of some parameters and the data, as discussed above, we have an additional set of models that can allow for that variability in the prediction of cell values for the full model, including census variation.

We will then extend the analysis in Chapter Eight to include spatial variation. This is easily done in the Bayesian paradigm and software exists to perform the analysis.

1.5.2 Spatial statistics

The basic assumptions made in most simple statistical analysis are that the data are independent and identically distributed. In many circumstances neither of these assumptions hold true. With particular relevance to the data that we will consider it is intuitively sensible to assume that data collected from some connected geographical regions defined by arbitrary political boundaries are likely to be correlated. The analysis of such correlated data has seen many advances particularly in the last 25 years or so of the last century and there are now some fairly standard approaches which can be used. These analyses assume some kind of known structure, generally a correlation structure, which is defined and parameters are estimated.

In the data that we will use there are a number of possible structures which could be investigated. We may consider that areas close to each other could

be similar or areas with similar employment opportunities. The unemployment in rural areas with a large horticulture sector may have strong seasonal effects and be similar to another area with similar industry separated by a large distance.

These spatial correlations can be included in the analysis by an appropriate covariance matrix and we will consider the inclusion of this correlation in our analysis. Many models of spatial dependence could then be fitted. However we will restrict the thesis to conditional autoregressive models (CAR) as the intent is only to show that the new algorithm proposed can be applied in a wider range of models.

1.5.3 Relative risk models

The final models which we will consider are risk, and relative risk models. In these models we will consider the risk of being unemployed, count of unemployed divided by the population size or the relative risk, the count of unemployed divided by the count of employed, rather than simply the counts of unemployed. The counts of unemployed for each cell in the table do not convey all of the information that is available. A small count of unemployed may result from a cell with few unemployed out of a large population (a "good" result) or a few unemployed out of a small population. The outliers in the count data may be caused by particularly large or small regions whereas the outliers in risk or relative risk data will be caused by regions with high, or low, unemployment rates irrespective of the region's population. By modelling the risk, or relative risk, we can include this additional information in our model.

In Chapter Nine we will consider these models in more detail, show their relationship with the models for counts and show that estimates of the

necessary parameters in the model can be found in much the same way as for the count models.

Models using relative risk data are common in applications such as epidemiology, for examples see Mollie A. in Gilks W.R. et. al. (1996) or pp73-75 Congdon P. (2001), but have not been used extensively in modelling unemployment nor in the general application of small area estimates. This may be due to the lack of appropriate data in the small area estimation setting as both the auxiliary and the survey data would require both counts of unemployed and employed, or total population counts.

1.6 Computational aspects

As the thesis develops we will use a number of different computer packages to fit the different models. The usage may seem to be complicated in places but we will try to show that where we change packages it will reproduce results already found in the previous package and that it can be used to do an analysis not supported in the other package. In some situations it may be possible to perform the new analysis in the previous package but it is more easily carried out, or the process can be made clearer, by using the new package. Inevitably, as software develops new functionality is introduced and it may become possible to easily build some models in a different package. Appendix B has listings of typical programs used for each analysis, or an explanation of the process used if the package is menu driven. These do not include every program used but are sufficient to provide templates for any other necessary programs. In many cases more than one package could be used. However we have not included the analysis in every package possible in each situation. Only where we move from one package to another have we duplicated analyses.

The table on the following page, Table 1.1, shows the relationship between the different models that are used along with the chapter in which they appear and the packages that are used.

The models that include the CAR model and the relative risk models have a number of possible combinations with other variables in the table below. A selection of the possible models has been used to illustrate the concepts.

	Chapter					
	3	4	5	7	8	9
Variability in Census				X	X	X
Categorical Variables Only	X	X		X	X	X
Linear or Quadratic Terms			X	X		
Conditional Autoregressive Error Structure					X	
Relative Risks						X
Method Used	SPREE by IPF or GLM	SPREE by IPF or GLM	GLM	GLM	GLM	GLM
Software	SAS, Splus, MLwiN	SAS, Splus, MLwiN	MLwiN BUGS	BUGS	BUGS	BUGS

Table 1.1 Models fitted, computer software used and chapter references.

The programs used are not the only ones that could have been used for each analysis. They are intended to illustrate methods and they may not even be the "best" for any particular analysis. Their purpose is to show that a new

algorithm for fitting SPREE type models works and that it has a more extensive range of applications.

1.7 Conclusions

In this chapter we have introduced the problem of small area estimation and briefly considered the range of methods developed in the latter part of the last century, and early in this one, to gain estimates for particular data structures. We have introduced a notation for general linear models and generalized linear models and we will show in Chapter Two that the methods considered all fit into the generalized linear model framework. The SPREE method has been mentioned and it will be discussed in greater detail in Chapter Three. It will be shown that SPREE too can be included in the generalized linear model framework. We will show how the concept underlying the SPREE method can be more clearly explained and then extended to data structures which are not solely categorical by the use of the generalized linear model. Some very simple examples will be used to illustrate this.

The later chapters of the thesis will be used to illustrate the concepts with more realistic examples and then to show how a wider range of models are allowed with the new algorithm.

In most methods that incorporate census data in the estimation process there is an assumption that these data can be used to give estimates without error. This is a very restrictive assumption and in many cases it is hard to justify. The implication is then that the only error term in the model is related to the survey data. We shall see in our examples later that we have a number of censuses and so there is variability associated with these census data. Estimates should reflect this error as well as that from the survey. Even if there is only one census it may be appropriate to view it as a result which if

repeated would give varied results and to try to estimate, in some way, the variability associated with the census. We will introduce Bayesian methods to easily allow this variability once known or estimated to be modeled by way of a prior distribution.

In this chapter we have introduced the idea of small area estimation and given a brief overview of the rest of the thesis. A more thorough review of the literature and discussion of the methods of small area estimation that have been used in the past is the subject of Chapter Two.

CHAPTER 2

An Historical Background

2.1 *Introduction*

As mentioned in Chapter One, Marker (1999) showed that most of the historical small area estimation methods could be placed in a framework of the linear regression model expressed in the equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2.1).$$

Most often $\mathbf{Y}$ will be the vector of estimates for the small areas. The subscript a will be used to denote the small areas $a = 1, \ldots, A$. The $\mathbf{X}$ matrix may be a matrix of data or of indicator variables, to denote membership of groups, or a mixture of the two. $\boldsymbol{\varepsilon}$ will be an error term which will generally be normally distributed with mean zero. The variance could be of a number of forms and will be discussed where necessary.

Equation (2.1) is a GLM in which the distribution of the random variable $\mathbf{Y}$ is normal and the link function is the identity.

Other notation used regularly in this chapter includes the following:

g	Subscript for groups in categorical variables	Maximum value G
h	Subscript for small domains	Maximum value H
i,j	Subscripts often for additional variables	Maximum value I, J
k	Subscript for members of a group	Maximum value K
l,d	Lengths of a vectors and sizes of matrices	
n	Sample sizes	

N	Population sizes
p	A proportion
R, r	Ratios
t	Subscript indicating time
w	Weights
θ	A parameter of interest

Notation such as $\bar{y}$ for a mean, $\hat{\beta}$ for estimates and $x_{\cdot j}$ to indicate summation over the first subscript will be adopted, diag (n_1, n_2, . . . , n_A) will be a square matrix with the elements in the bracket on the leading diagonal and zeros elsewhere. Where matrix operations are shown it will be assumed that the dimensions of matrices are such that the operation can be performed.

We will express each method presented in terms of that model and show the implied structure of the **Y** vector and **X** matrices. Where additional variables are required for a particular method they will be defined at that time.

The only major method that cannot be included in this framework is SPREE estimation which we will consider in more detail in the next chapter. We will then more fully utilize the generalized linear model as proposed by Nelder and Wedderburn (1972) and show that the methods described in this chapter as well as SPREE can be included. We will use this new framework to propose a different method of fitting SPREE estimates and show that it has advantages over the traditional methods.

This chapter will be based on Purcell and Kish (1980), Ghosh and Rao (1994), Marker (1999) and Rao (1999). The classifications that these authors use for the various methods are not identical but they are broadly similar. The

notation which they have used is also different and at times confusing. One intention of this chapter is to place the different methods into a consistent framework which would help in making comparisons.

2.2 Historical methods in a linear regression framework

2.2.1 Demographic methods

Purcell and Kish (1980) categorize these methods under the general heading of Symptomatic Accounting Techniques (SAT). These methods were all developed for estimating human population sizes and use particular available information to improve estimates from other data, possibly an earlier census. The literature is not consistent in its classification of these older techniques but they are generally extensions of a very simple model in which a previously known estimate, usually from a census, is updated by adding or subtracting births, deaths, immigration and migration counts for each small area.

This can be expressed as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ where:

- $\mathbf{Y}$ is the vector of estimates for the population of each small area,
- $\mathbf{X}$ is a matrix with columns of counts for variables used to update the previous estimate, for example the last census estimate and counts of births, deaths, immigration and migration, since the last census, each row contains the data for each small area,
- $\boldsymbol{\beta}$ is a column of 1's and –1's depending on whether the quantity gets added or subtracted from the census estimate.

For the example above in which there was a previous census and adjustments for births, deaths, immigration and migration counts the parameters in $\boldsymbol{\beta}$ would be :

$$\boldsymbol{\beta} = \begin{pmatrix} 1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}.$$

In this model there is no stochastic part as it is assumed that the values are known exactly, $\boldsymbol{\beta}$ does not need to be estimated, as it is known, and it is simply an arithmetic problem.

Administrative records do not usually detail all arrivals and departures from an area so various methods have been developed to estimate these from symptomatic variables, variables that reflect the arrivals and departures. These methods include the following which will be described below

Administrative Records (AR)	Starsinic (1974)
Housing Unit (HU)	Smith and Lewis (1980)
Vital Rates	Bogue (1950)
Composite Method	Bogue and Duncan (1959)
CMII	US Bureau of Census (1966).

Administrative Records

The AR method uses estimates for net migration from records for individuals. Births, deaths, incomes filed with taxation authorities and other such administrative data are collected and used to update census data.

Housing Units

This method expresses the current population as the number of occupied housing units multiplied by the average number of people in each housing unit plus the number in group quarters, for example, hostels or barracks. Each

of these three quantities has to be estimated. There are various methods of estimating these quantities depending again on the available data Smith and Lewis (1980).

This is again a simple linear model with no random part. The three quantities are all estimated so that the small area population can be calculated.

We can write the model as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ where:

- $\mathbf{Y}$ is the vector of estimates for the population of each small area,
- $\mathbf{X}$ is a matrix with columns of counts of housing units and numbers of people in group housing in each small area
- $\boldsymbol{\beta}$ is a column vector with two entries, the first is the average number of people per housing unit and the second is a one.

Once again the rows represent the small areas.

Vital Rates

This method is an extension of the SAT methods and in some of its later guises it fits under the heading of symptomatic regression which is discussed in the next section.

In the Vital Rates method the variable of interest is the population, **Y**, for each local area, a, Y_a. It is assumed that information on local birth rate is known for each local area from the last census, r_{a0}, and that the equivalent rate is known for the larger area for both the census year and current year R_0 and R_t. This method uses only birth data, and this is used as a symptomatic variable rather than as a component of population change. In other words the births are not added to the total but are used to generate a multiplier based on the wider area. The crude birth rates, for the local areas, are estimated by $r_{at} = r_{a0}\left(\frac{R_t}{R_0}\right)$ where r_{a0} denotes the crude birth rate for the local area, a, in the

most recent census year (t = 0). If, for the larger area in a given year t, the annual number of births is b_t then the population for the local area at year t, Y_a, is given by $Y_a - \frac{b_t}{r_{at}}$.

This method depends heavily on the validity of the assumption that

$$\frac{r_{at}}{r_{a0}} = \frac{R_t}{R_0} \qquad (2.2)$$

i.e. that the ratio, for the birth rate at the last census and now, for the local area is the same as for the larger area Marker (1983). This assumption is often questionable, a city may have suburbs with predominantly young families and others of the elderly. Hence the larger area, the city, may not represent all of its suburbs in terms of the birthrates.

The extension to a number of large areas each with its own small areas simply requires the addition of an extra subscript for the large areas. The assumption is then that the small areas within the larger area have the same ratios of rates as in equation (2.2) and that this changes for each larger area.

The method can be extended to finding another estimate of the population using the death rates in exactly the same way. The final estimate of population is then an average of the two.

We can show that this is another example of the linear model in equation (2.1) **Error! Reference source not found.**. If birth rates and numbers of births are known for each small area, a, then the size of the population can be easily calculated. Assuming numbers of births are known from administrative records the problem becomes one of estimating birth rates. The model used is:

$$r_{at} = X_{a0}\beta + \varepsilon_a \text{ where}$$

Y_{at} is the a^{th} member of the vector of unknown birth rates for the current year, t.

X_{a0} is the a^{th} member of the vector of birth rates for the last census.

β is the ratio of the current birth rate for the whole area to the birth rate for the whole area at the last census which is known and so the estimates are calculated.

If the error term is written such that $E(\varepsilon_a) = 0$ and $Var(\varepsilon_a) = \frac{\sigma^2 X_{a0}}{W_a}$

where $W_a = \frac{\text{Small Area Population at last census}}{\text{Total Population at last census}}$ then the weighted least squares estimates of β are

$$\hat{\beta} = \frac{\sum_a \left(\frac{Y_{at} X_{a0}}{X_{a0} / W_a} \right)}{\sum_a \left(\frac{X_{a0}^2}{X_{a0} / W_a} \right)}$$

which can be shown to be the ratio of the current birth rate for the whole area to the birth rate for the whole area at the last census as before. Hence $\mathbf{Y}$ can be calculated from $\mathbf{X}$ and $\boldsymbol{\beta}$. The variance assumption suggests that areas with high birth rates or smaller population sizes will be more variable.

There are a number of methods that use concepts from both Symptomatic Accounting Techniques and Symptomatic Regression. The regression method is used to estimate such accounting variables as births, deaths, immigration and migration in various categories (age-sex-race) and these estimates are used in an accounting technique to generate the estimates required. Each of the accounting variables, for example the births above, can be shown to be a

linear regression and these variables are simply summed to estimate the small area values.

Composite Method.

(This method is not what would be called composite estimation in current usage, where a weighted average of a direct and indirect estimate would be found, this will be discussed later).

The composite method is an extension of the VR method. It sums independently computed age-sex-race estimates based on births, deaths and school enrollments. Zidek (1982)

CM II

This method takes into account the net migration as well as births and deaths. The net migration is subdivided into military and civilian migration, estimates for the military migration are readily available and school enrollments are used for the civilian migration. The net migration is denoted m_{ai} and net changes due to births and deaths are as defined previously.

As noted previously these methods have been developed specifically for finding human population totals. They use symptomatic variables, for example the number of births or the number emigrating, to predict and there are a number of ways of estimating that use the characteristics of the variable in question. To extend these to other populations comparable symptomatic variables would need to be found and suitable estimators applied.

2.2.2 Synthetic and related methods

Gonzalez (1973) describes synthetic estimates as follows: "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for sub areas under the assumption that the small

areas have the same characteristics as the large area, we identify these estimates as synthetic estimates." They have become well used for a number of reasons:-

- They are simple to use
- They can be applied to general sampling designs
- There exists the potential to increase accuracy by borrowing information from similar small areas.

There are a number of related methods which will be described.

Synthetic Estimation.

Synthetic estimation uses survey data to estimate proportions of different subgroups, for example age by sex by ethnicity subgroups, and the small areas estimates are derived by taking the appropriate weighted average of these for each small area. The standard formulation for these estimates is

$$\hat{\bar{Y}}_{a..} = \frac{\sum_j N_{aj.} \bar{y}_{.j.}}{N_{a.}} \qquad \text{where}$$

N_{aj} is the size of the population of small area a and subgroup j.

N_a is the total population in area a,

$\bar{y}_{.j.}$ is the sample average for Y for subgroup j across all small areas.

The model can alternatively be written as

$$Y_{ajk} = x_{ajk}\beta_j + \varepsilon_{ajk} \qquad \text{where}$$

Y_{ajk} is the value of the variable of interest in small area a, for the jth sub group for the kth member.

$x_{ajk} = \begin{cases} 1, \text{ if } Y_{aj'k} \text{ is such that } j' = j \\ 0, \text{ otherwise} \end{cases}$ indicating membership of the jth subgroup.

ε_{ajk} an error term

The design matrix X can then be constructed and it will have $A \times J \times K$ rows and each row will be entirely zeros apart from one one which denotes the subgroup membership. By sorting this matrix so that all members of one subgroup are consecutive it can be shown that

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = diag\left(n_{.1}^{-1}, n_{.2}^{-1}, n_{.3}^{-1}, ..., n_{.j}^{-1}\right)$$

and $$\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \left(y_{.1.}, y_{.2.}, y_{.3.}, ..., y_{.j.}\right)^{\mathrm{T}}$$

The least squares estimate of $\boldsymbol{\beta}$ is then $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ and

$$\hat{\bar{Y}}_{a..} = \frac{\sum_{j} N_{aj.}\bar{y}_{.j.}}{N_{a..}}$$ as before.

2.2.3 Symptomatic regression

Two past censuses are used to define a regression relationship between the variable of interest and ratios of the symptomatic variables at each census. These ratios are then found for the last census and the present time and used to update the variable of interest.

The model used is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is the vector of the ratios of the values in the most recent census of the variable of interest in each small area a to the equivalent values in the previous census.

$\mathbf{X}$ is the (a x j) matrix of ratios for the symptomatic variables for each small area a for each of the j variables.

$\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$

Assuming $\boldsymbol{\Sigma}$ is $\sigma^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix then the least squares estimates of $\boldsymbol{\beta}$ are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

These coefficients are then used along with ratios for the symptomatic variables for the last census and the present to estimate the current values of the variable of interest at the small area level.

Ratio-Correlation

Let t = {0,1,2} denote two consecutive census years and the current year, note the three years represented may not be equally spaced. Also N_{it} and S_{ijt} are the population total and the value for the jth symptomatic variable for the a^{th} local area (a = 1, ...,A) in the year t.

Further let : $p_{at} = \dfrac{N_{at}}{\sum_a N_{at}}$ and $s_{ajt} = \dfrac{S_{ajt}}{\sum_a S_{ajt}}$ be the corresponding proportions

and write $R'_a = \dfrac{p_{a1}}{p_{a0}}$, $R_a = \dfrac{p_{a2}}{p_{a1}}$ (which is unknown as p_{it} is what we are trying to estimate)

and $r'_{aj} = \dfrac{s_{aj1}}{s_{aj0}}$, $r_{aj} = \dfrac{s_{aj2}}{s_{aj1}}$

Using the data $\left(R'_a, r'_{a1}, \ldots, r'_{aj}; a = 1, \ldots A\right)$ and multiple regression we first fit

$$R'_a = \hat{\beta}_0 + \hat{\beta}_1 r'_{a1} + \ldots. + \hat{\beta}_j r'_{aj}$$

where the $\hat{\beta}$'s are the estimated regression coefficients that link the change, R'_a, in the population proportions between the two census years to the corresponding changes, r'_{aj}, in the proportions for the symptomatic variables. Next the changes, R_a, in the postcensal period are predicted as

$$\tilde{R}_a = \hat{\beta}_0 + \hat{\beta}_1 r_{a1} + \ldots + \hat{\beta}_j r_{aj}$$

using the known changes in the symptomatic variables in the postcensal period and the estimated regression coefficients. Finally the current population counts are estimated as

$$\tilde{N}_{a2} = \tilde{R}_a p_{a1} \left(\sum_a N_{a2} \right)$$

where the total population across all local areas is found from other data.

The difference correlation method uses the same process but differences between the proportions at the two time points are used rather than their ratio.

These procedures use regression coefficients in the last intercensal period and it is assumed that these parameters are applicable to the change from the last census to the present time. Significant changes in the statistical relationship can lead to errors in the current estimates. Such changes are quite likely if the censuses are 10 years apart as the ratios are based on data that will be between 10 and 20 years old.

The methods described thus far use only census data. Use of sample survey data can help to reduce the problem of changes in the statistical relationship and give accurate current data for some small areas.

The sample-regression method

Erickson (1974) avoids the problem of out of date data by using sample estimates of R_i for those local areas where data exists. The sample regression estimators are then obtained for all areas using the known symptomatic ratios.

The sample regression method has similarities with symptomatic regression except that instead of using census data from two past censuses, data are used from a sample of the small areas and the most recent census. Thus it is

assumed that the sample of small areas is representative of all of them. The model used is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is the vector of ratios of the variable of interest for the most recent census and the survey.

$\mathbf{X}$ is the matrix of equivalent ratios for the symptomatic variables.

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$

for the small areas that are sampled,

$\boldsymbol{\Sigma} = \sigma^2\mathbf{I} + \mathbf{T}$ where $\mathbf{T} = \text{diag}\{\boldsymbol{\tau}\}$ is a diagonal matrix of the variances of the sampling errors for each small area *a*.

The weighted least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}$ and these are used as before to compute the estimates for all small areas.

Components of variance regression.

The methods considered so far have considered data at the area level but the components of variance method considers data at the element level. The error term has two components, one is random and the other is small area specific. The model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$ where

As before $\mathbf{Y}$ is the column vector of responses with elements Y_{ak}.

X is the design matrix denoting membership of each small area.

$\boldsymbol{\varepsilon}$ is the vector of errors with elements ε_{ak}.

$\boldsymbol{\varepsilon} = \mathbf{v} + \mathbf{u}$

$\mathbf{v}$ is the vector of small area effects, these will be the same for all individuals in each small area, hence $\mathbf{v}$ has elements v_a

$\mathbf{u}$ is the vector of random errors with elements u_{ak}.

$\mathbf{v} \sim N(\mathbf{0}, \Sigma_{vv})$ and $\mathbf{u} \sim N(\mathbf{0}, \Sigma_{uu})$.

This estimate is not optimal unless all of the n_a are equal Searle (1988), Harter and Fuller (1987).

The least squares estimates of $\boldsymbol{\beta}$ are then $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ The average residual for each small area can be calculated and given the assumed error structure $E(\mathbf{v}_a \mid \bar{\varepsilon}_a) = \bar{\varepsilon}_a G_a$ where $G_a = (\Sigma_{vv} + n_a^{-1}\Sigma_{uu})^{-1}\Sigma_{vv}$. So we can produce an estimate of v_a for each small area, if there is survey data for that small area.

2.2.4 Structure preserving estimation (SPREE)

Chambers and Feeney (1977) and Purcell and Kish (1980) propose structure preserving estimation (SPREE) as a generalization of synthetic estimation in the sense that it makes fuller use of reliable direct estimates. SPREE commonly uses the well known method of iterative fitting of margins in a multi-way table, Deming and Stephan (1940), where the margins are direct estimates. Although this method is based on ratios, as in the previous method, the raking procedure (IPF) guarantees that the ratio estimates are adjusted to conform to the margins.

In Figure (2.1) the data in the body of the table are generated from a census and the margins from the survey. The margins are available for the different categories but not for all of the combinations. The data in the body of the table are then adjusted to agree with the margins. This adjustment is achieved by first adjusting the rows then the columns and repeating this until convergence. The interactions between the variables are set by the census data except for any that may be defined in the new margins. The interactions that are defined by the census data remain unchanged through the iterative

process. Only the effects defined by the survey data are changed to agree with the new margins. The **S**tructure of the table is **PR**E**s**erved in the **E**stimation, hence SPREE.

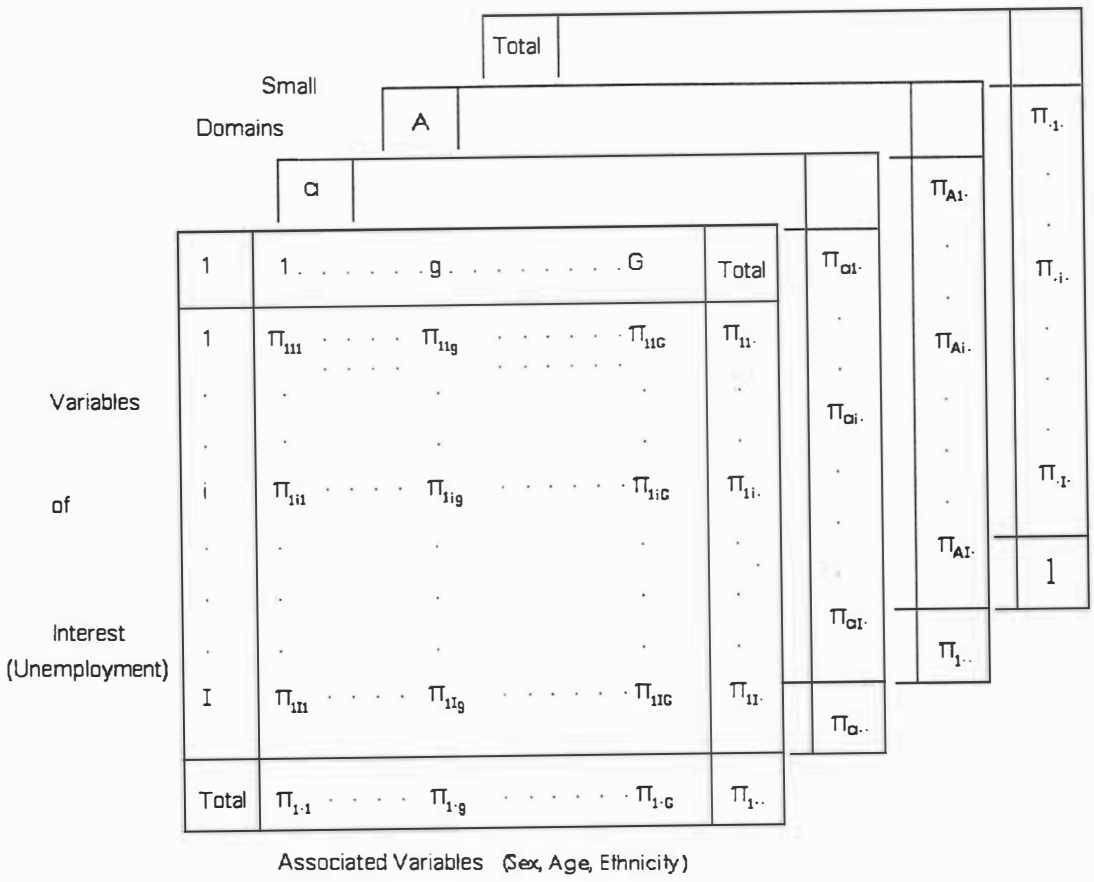


Figure 2.1 Relationship of variables in SPREE

Estimation using SPREE will be considered in more detail in the next chapter. We will describe its usual formulation using the iterative proportional fitting algorithm. We will then propose an alternative approach by setting it in the framework of the generalized linear model.

2.2.5 Composite estimation

The composite estimators cover a wide range of examples all of which can be characterised by a weighted average of two estimators. The most common simply apply a weighted average to a direct estimator and one of the estimators described above. The weights are based on the variance or mean square error of the two estimators. The composite estimator is intended to offset the bias in the indirect estimator and the variability in the direct estimator. Where few areas in the model have direct estimators, there is little advantage in the composite estimator. Because the composite estimator is a linear combination of the direct estimator and an estimator found directly from a linear model there is a linear model that includes both the direct and synthetic estimators explicitly that gives the optimal least squares estimate as the composite estimator.

Synthetic estimators tend to be biased, direct estimators may be unstable, an obvious way to reduce these two effects is to take a weighted average of the two estimates. Such estimators may be written

$$\hat{Y}_a^C = w_a \hat{Y}_{1a} + (1 - w_a)\hat{Y}_{2a} \qquad (2.3)$$

where $\hat{Y}_{1a}$ is a direct estimator and $\hat{Y}_{2a}$ is an indirect estimator, and w_a is a suitably chosen weight $(0 \le w_a \le 1)$. Using the estimators $\hat{Y}_a$ and $\hat{Y}_a^S$ for $\hat{Y}_{1a}$ and $\hat{Y}_{2a}$ respectively a number of possible choices of weights have been proposed.

- $w_{a\,(\text{opt})}$ obtained by minimising MSE($\hat{Y}_a^C$) with respect to w_a assuming cov($\hat{Y}_a$, $\hat{Y}_a^S$) $\cong 0$, the resulting weights can be very unstable. Such weighting is inversely proportional to the variances of the direct estimator. Unstable estimates of the variance cause

problems so that the weights need to be stabilised by an averaging process.

- Schaible (1978) proposes an "average" weighting scheme based on several variables. He notes that the composite estimator is quite robust to changes in the weights.
- Purcell and Kish (1979) use a common weight, w, and then minimize the average MSE. The choice of a common weight is not reasonable if the individual variances vary considerably.
- Drew, Singh and Choudhry (1982) propose the weights

$$w_a(D) = \begin{cases} 1, & \text{if } \hat{N}_a \geq \delta N_a \\ \dfrac{\hat{N}_a}{\delta N_a}, & \text{otherwise} \end{cases}$$

where $\hat{N}_a$ is the direct, unbiased estimator of the known population size N_a and δ is chosen to control the contribution of the synthetic estimator.

- Samdal and Hidiriglou (1989) propose a similar weight

$$w_a(S) = \begin{cases} 1, & \text{if } \hat{N}_a \geq N_a \\ \left(\dfrac{\hat{N}_a}{N_a}\right)^{h-1}, & \text{otherwise} \end{cases}$$

where h is subjectively chosen, they suggest, as a general purpose weight, $h = 2$.

This and the preceding weights are identical if $h = 2$ and $\delta = 1$.

These latter two weights do have some problems, notably they can fail to draw strength from the synthetic estimator when the sample size is greater than the expected sample size even though the expected sample size may not be large enough to make the direct estimator reliable. Also the weights do not take into account the size of the between area variation relative to the within

area variation for the characteristic in question, that is all characteristics would use the same weights irrespective of the differences of between area homogeneity.

BLUP

Holt, Smith and Tomberlin (1979) proposed a Best Linear Unbiased Prediction estimator of $\hat{Y}_i$ under the following model for the finite population:

$$y_{agk} = \mu_g + e_{agk} \qquad k = 1, \ldots, N_{ig};\ g = 1, \ldots, G;\ \ i = 1, \ldots, m$$

where y_{agk} is the y-value of the k^{th} unit in the cell (a,g) for small area a in domain g, μ_g's are fixed effects and the errors e_{agk} are uncorrelated with zero means and variances σ_a^2.

N_{ag} denotes the number of population elements in the large domain g which belong to the small area a. Suppose n_{ag} elements in a sample of size n fall in the cell (a,g) and let $y_{ag.}$ and $y_{.g.}$ denote the sample means for (a,g) and g respectively.

The BLUP estimator of Y_i is given by

$$\hat{Y}_a^B = \sum_g \hat{Y}_{ag}^C = \sum_g \left(\frac{n_{ag}}{N_{ag}} \hat{Y}_{ag} + \left(1 - \frac{n_{ag}}{N_{ag}} \right) \hat{Y}_{ag}^S \right)$$

where $\hat{Y}_{ag} = N_{ag} \bar{y}_{ag}$ and $\hat{Y}_{ag}^S = N_{ag} \bar{y}_{.g}$

or

$$\hat{Y}_a^B = \sum_g \left(n_{ag} y_{ag.} + \left(N_{ag} - n_{ag} \right) y_{.g.} \right)$$

which in effect sums the values in the sample from that cell and estimates the total for the rest in that cell by the mean across the g^{th} group multiplied by the number not in the sample. If the sample size for that cell is large this tends to the direct estimator, if it is small it tends to the synthetic estimator. If the sample fraction is small for all of the groups then no account is made of the

relationship between, between area and within area variation. This problem can be reduced by more sophisticated models.

We have considered the main methods considered by Purcell and Kish and Marker for small area estimation and shown that they fit into a linear regression with the exception of SPREE.

2.3 *Models with area specific effects*

Two types of models, with area specific effects, have been proposed.

In the first area-specific auxiliary data $\mathbf{x}_a = (x_{a1},\ldots,x_{al})^T$ is assumed to be known and the parameters of interest, θ_a, are assumed to be related to $\mathbf{x}_a$.

In particular we assume that $\theta_a = \mathbf{x}_a^T\beta + v_a z_a \qquad a = 1,\ldots,A$

The z_a's are known positive constants, $\boldsymbol{\beta}$ is the vector of regression parameters and the v_a's are iid random variables with $\mathrm{E}[v_a] = 0$ and $\mathrm{V}(v_a) = \sigma_a^2$.

In the second type element-specific auxiliary data $\mathbf{x}_{aj} = (x_{aj1},\ldots,x_{ajl})^T$ are available for the population elements. The variable of interest y_{aj} is assumed to be related to $\mathbf{x}_{aj}$ through a nested error regression model.

$$y_{aj} = \mathbf{x}_{aj}^T\beta + v_a + e_{aj} \qquad j = 1,\ldots,N_i;\quad a = 1,\ldots,A. \qquad (2.4)$$

where $e_{aj} = \tilde{e}_{aj}k_{aj}$ and the $\tilde{e}_{aj}$'s are iid random variables independent of the v_a's with $E(\tilde{e}_{aj}) = 0$, $V(v_a) = \sigma^2$, the k_{aj}'s being known constants and N_a being the number of elements in the a^{th} area.

For making inferences under the first model some assumptions are required that may be quite restrictive in some applications. Violation of these assumptions will introduce bias to the estimates.

Assuming that direct estimators $\hat{\theta}_a$ are available and that $\hat{\theta}_a = \theta_a + e_a$

we obtain the model $\hat{\theta}_a = x_a'\beta + v_a z_a + e_a$ which is a special case of the general mixed linear model.

In the second model we assume that the sample values also obey the model, whilst this will be true under simple random sampling it may not be appropriate under more complex sampling designs. The model can however be extended to account for such features.

Writing the model in equation (2.4) in a matrix form as

$$\mathbf{y}_i^p = \mathbf{X}_i^p\boldsymbol{\beta} + \mathbf{v}_i\mathbf{1}_i^p + \mathbf{e}_i^p$$

the model can then be written in a partitioned matrix form :

$$\mathbf{y}_a^p = \begin{bmatrix}\mathbf{y}_a \\ \mathbf{y}_a^*\end{bmatrix} = \begin{bmatrix}\mathbf{x}_a \\ \mathbf{x}_a^*\end{bmatrix}\boldsymbol{\beta} + v_a\begin{bmatrix}\mathbf{1}_a \\ \mathbf{1}_a^*\end{bmatrix} + \begin{bmatrix}\mathbf{e}_a \\ \mathbf{e}_a^*\end{bmatrix}$$ where the superscripts * denotes non-sampled elements, and the $\mathbf{x}_a$ and $\mathbf{x}_a^*$ are matrices.

The mean of each small area $\bar{Y}_a$ can be written as $\bar{Y}_a = f_a\bar{y}_a + (1-f_a)\bar{y}_a^*$ where $f_a = \frac{n_a}{N_a}$ and $\bar{y}_a$, $\bar{y}_a^*$ denote the means for sampled and non sampled elements respectively. Estimation of the overall mean can be viewed as prediction of the mean of the non sampled elements given the data.

Two forms of a model based approach are common, one for data which is at the area level and the other for data at the elemental level.

2.4 EBLUP, EB and HB approaches

Lahiri (1996) combines these approaches under the heading of Composite Estimation which was discussed in section 2.2.5. They are characterised by a weighted average of two estimates, the direct estimate and a synthetic estimate.

$$\hat{\theta}_a^C = w_a y_a + (1-w_a)\mathbf{x}_a^{\mathrm{T}}\hat{\boldsymbol{\beta}}$$

which is essentially the same as equation (2.3).

The choice of weight, w_a, is the difference between the three approaches. As these estimates are simply a linear combination of two estimates then as long as these two estimates are formed by linear models then the new estimate will also be a linear model.

These approaches are useful when direct estimates are available for a "reasonable" number of the small areas, if they are not then the weight, w_a, is equal to zero for most and the estimate is simply the synthetic estimate. They are an attempt to combine the zero bias of the direct estimate with the smaller variance of the synthetic estimate.

These approaches are all model based and so they will fit in to the linear regression framework if the model proposed also fits into this framework and the distributional properties are appropriate.

These methods are something of a transition from a concept of having available survey, and auxiliary data, and using them to estimate small area parameters and the concept of a population which can be modelled in some way. The advantages of the latter approach are that the assumptions of the model are explicitly stated, and that the model may be able to be validated from the data once it is collected.

2.5 Concluding remarks

In this chapter we have discussed many of the common methods of small area estimation. We have shown that they all fit into a linear model form as Marker (1999) had done.

Most of these models were developed for particular problems and require data in a very particular form. Although there are general principles which

can be recognized, these models may be difficult to extend to new situations when the data available are different.

All of the methods introduced are model based methods. In some of the implementations it may not be immediately clear what the model is. Each of these models includes a number of assumptions, some of which may not have been explicitly stated in the original formulation of the method. If any of these assumptions do not hold in practice then the estimates found will be biased and without consideration of the assumptions there is no way of gaining any insight into this bias. Given the explicit model, including the assumptions, some diagnostics may be possible. We will return to these issues in a number of places in the rest of this thesis.

We have mentioned SPREE briefly and stated that it does not fit into this model and in the next chapter we will show that if we use a more general model, the generalized linear model, then we can include all of the methods discussed so far as well as SPREE. We will then show that this classification allows a new method of estimating the results for SPREE which can then be applied in situations where SPREE has traditionally been inapplicable. The assumptions required for SPREE also become more explicit and the new formulation gives a better understanding of the situations under which SPREE will give good estimates and where it will not.

CHAPTER 3

Structure Preserving Estimation; the link with the Generalized Linear Model.

3.1 Introduction

In the last chapter we considered the historical methods for small area estimation and showed that they are examples of a linear model. We noted that SPREE did not fit this model and by considering a more general model we will now show that SPREE can be included in the general classification of the methods. In this chapter we will look at SPREE in detail and show how the underlying concept can be extended to new situations.

SPREE can be applied when there are two sources of data, the first will generally be a census and the second a sample survey. The census data will be available in the form of counts of membership of mutually exclusive subgroups of the population defined by categorical variables. The sample survey data are also available categorized by some of the variables but they are not able to be disaggregated to the same fine level due to the inadequate sample size.

We will initially look at the approach based on a framework of categorical data analysis due to Purcell and Kish (1980). In this method the Iterative Proportional Fitting algorithm (IPF), Deming and Stephan (1940), is used to adjust the cell values in the table of census data to agree with the new marginal totals which can be estimated from the sample survey data. Although some aspects of the structure in the table will remain the same in

this process the cell counts do not. The small area estimates may be these new cell counts or some aggregation of them.

In the general case this is called raking. IPF is used where the data are categorical. Raking is commonly used to adjust estimates so that they sum to a known total. Thus if a series of estimates for regional populations did not add up to the known national population they could be rescaled appropriately.

Marker (1999) noted that SPREE can be written as a log linear model but then went on to say that it cannot be expressed as a linear regression model.

It will be shown that raking and IPF are equivalent to fitting a log linear model to the cell values in the table of census data (which may or may not be integers) and then changing some coefficients in the model whilst keeping others the same. By extending the linear regression model to the model introduced in equation (1.1) we will show that the census data, $\mathbf{Y}_c$, can be expressed in the form

$$g\left(\mathrm{E}\left[\mathbf{Y}_c\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are partitions of a design matrix which will be defined in section 3.5. Similarly the model for the sample survey data, $\mathbf{Y}_s$, can be written as

$$g\left(\mathrm{E}\left[\mathbf{Y}_s\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + \mathbf{X}_2\boldsymbol{\beta}_{2s}\,.$$

See section 3.4 for the necessary detail on the model and possible link functions.

By a suitable choice of the partition of the $\mathbf{X}$ matrix into $\mathbf{X}_1$ and $\mathbf{X}_2$ the coefficients that change will be seen to be the components of $\boldsymbol{\beta}_1$ and those that stay the same are the components of $\boldsymbol{\beta}_2$. We will show that fitting SPREE by the IPF algorithm is equivalent to stating that $\mathbf{X}_2\boldsymbol{\beta}_{2c} = \mathbf{X}_2\boldsymbol{\beta}_{2s}$ in the two models

above. In Chapter Six this assumption will be investigated further, in conjunction with other commonly stated assumptions about the relationships between the census and survey data.

This chapter develops these ideas. In section 3.2 we look at the traditional approach to SPREE by the IPF algorithm. Then in section 3.3 we give a simple example to show the mechanics of the algorithm. Section 3.4 shows how the log linear model can be used to model the data in a contingency table, such as the census data, and then shows that this model is an example of the GLM. In section 3.5 we will formally propose the new approach. In section 3.6 we come back to the specifics of the data which is available and look at the relationships between the data sets with respect to the model that we have developed.

Section 3.7 investigates some practical considerations in the new information supplied by the survey data and looks at various configurations of new margins. Finally in section 3.8 we will work through an example of the new method using one of the configurations from section 3.7. A full example of the new method is presented in Chapter Four.

3.2 SPREE by the iterative proportional fitting algorithm

A simple example is introduced in diagrammatic form to illustrate the relationships. A census is available which includes data on employment status (employed / unemployed) categorised by ethnicity (Maori / other) and sex (male or female, as usual). This census data may be from some time ago or as in the real example that we will consider later a census in which the definition of unemployment is not the one that is required. A survey is taken which details employment status by ethnicity, either more recently or using

the approved definition, and the required estimates are the employment status by sex. This is shown in the diagram.

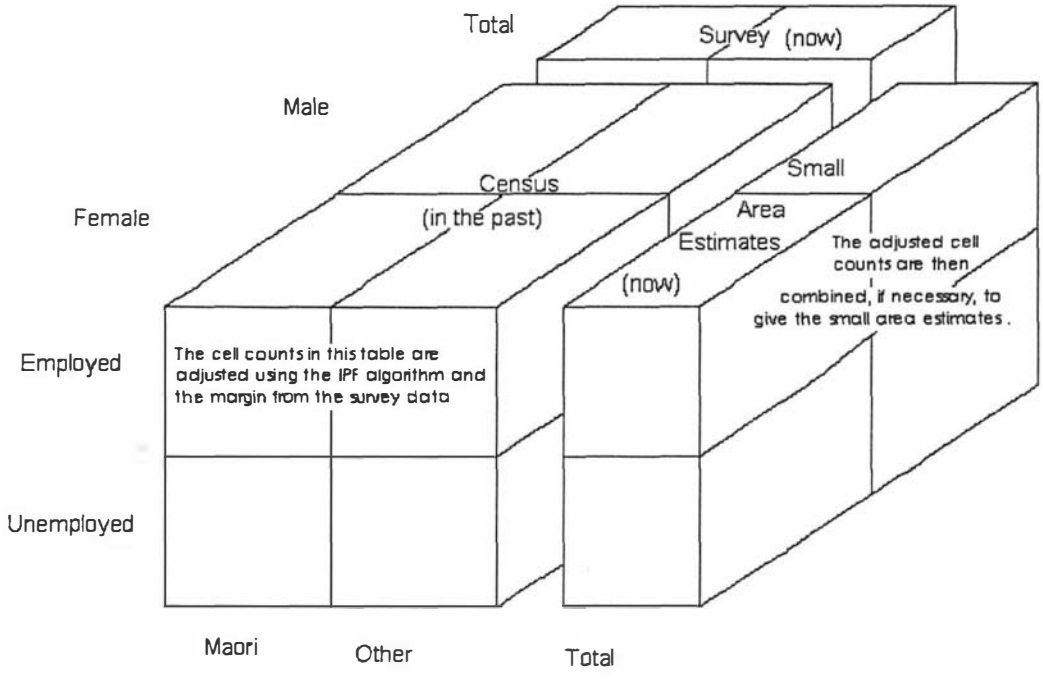


Figure 3.1 Diagram showing the relationship between the two data sources and the small area estimates in a simple example.

The iterative proportional fitting algorithm, applied to this example, would adjust the census sex by ethnicity by employment totals to agree with the survey data producing new estimates for each cell in the 2 by 2 by 2 contingency table above. These would then be summed across one variable, the ethnicities, to give the required small area estimates.

The small area estimates indicated are sums of the rows in the table and in this example are estimates of numbers of employed and unemployed by sex. There is no requirement that the final estimates should be sums and it may be that the required estimates are the body of the table, here the numbers of employed and unemployed by sex and ethnicity.

In this example we have no specific interest in ethnicity. It is an associated variable, useful because changes in ethnicity are expected to be related to the variables of interest and so improve their estimates.

We shall see how the IPF algorithm can be used to adjust the cell counts in the body of the table to conform to the new margin from the sample survey and hence produce the small area estimates. The survey provides marginal constraints to which the census counts must conform.

Two assumptions are made initially:-

- that current estimates for the variable of interest cross tabulated by appropriate associated variables are available at a large domain level
- that estimates are known for a variable, closely related to the variable of interest, cross tabulated with the same associated variables for each small domain. This variable may be the same variable at an earlier time period.

The current data will generally be available from a survey and the previous data may be from a census or some administrative data collected for some other reason.

There is a third assumption made in presenting this model. That is that there is an underlying superpopulation model governing the behaviour of the small domain frequencies over time. The concept of a superpopulation is well established and it can have a number of interpretations. Most commonly it is thought of as being a larger population from which the population that is under consideration is drawn as one of all of the possible populations that could have been drawn. Alternatively we can consider the population as being a realization of a random process for which there exists a stochastic

model Cassel, Sarndal and Wretman (1977). In the case of a contingency table this model is usually assumed to be a Poisson or multinomial model at the population level. Such models are particularly useful when the census data are from one period, and the survey estimates from another. Since the census values at the current time are unknown and must therefore be modeled.

Suppose there exists a large finite population of size N, in which each unit is labeled u where $u = 1, \ldots . N$.

Each unit is associated with a vector of responses which can contain three types of variables these determine:-

1 the small domain to which each unit belongs,

2 the variable(s) of interest in the study, in the example above sex and employment which categorise the cell for which estimates are required

3 the associated variables used in the study, in the example ethnicity.

We shall also introduce three variables:-

π the past cell relative frequencies representing the association structure. Generally found from the census data.

ξ the marginal constraints representing the allocation structure, generally found from the sample survey data.

p the desired cell estimates resulting from π and ξ.

The obvious subscripted form for each cell is suggested in which there are H small domains, I variables of interest and G associated variable categories (large domains). As was shown in *Figure 3.1* the required small area estimates may be individual cell counts or they may be sums of cells so the subscript a has not been used in the above notation to avoid any confusion. The subscript

h may refer to the small areas or it may refer to smaller divisions of them which will be summed.

Hence an indicator variable d_{higu} is defined which will be equal to 1 if unit u is in domain h and has response category i for the variable of interest and g for the associated variable, and equal to 0 otherwise. Thus a three dimensional contingency table can be constructed of order H x I x G. N_{hig} denotes the number of units in domain h, with response type i,g for the interest and associated variables. $N_{hig} = \sum_{u=1}^{N} d_{hig}$ and $p_{hig} = \frac{N_{hig}}{N}$

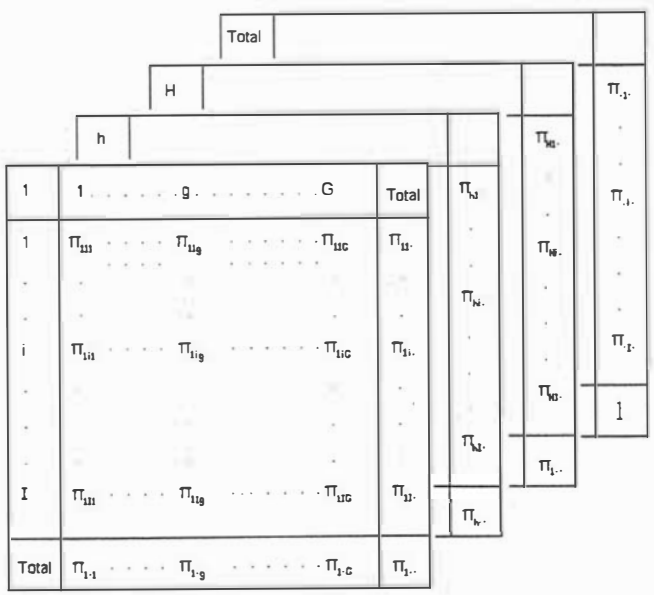


Figure 3.2 *Diagram showing the association structure.*

The associated variable categories may be combinations of a number of associated variables. For example variables for sex and three categories of age can be written as a single variable with six categories.

These probabilities can then be summarized in a three dimensional array even if the total number of variables is greater than three. This is described as a restricted canonical form as the distribution is completely specified and it is restricted by using relative frequencies which sum to a grand total of 1 over the whole table.

The association structure relates to the structure in the relative frequencies recorded at the small domain level for the variable of interest cross-tabulated with some associated variables at the time of the census or other source of that data.

Note the sample design does not place any restrictions on the methodology. The allocation structure is defined by the relative frequencies for the variable of interest cross tabulated with the associated variables found at the present time. This would generally be estimated from sample survey data. The allocation structure can also be thought of as the association structure at the current time aggregated over the small domains in each larger domain.

The object of the exercise is to disaggregate this structure to reveal the structure of the small domains and hence estimate the variables of interest over the small domains. The disaggregated estimates are denoted p_{hig} and to obtain the small domain estimates these may be summed over the associated variable for each small domain if necessary. So it is necessary to estimate p_{hig} for each cell in the table.

The marginal frequencies define the current relationship between the associated variables and the variables of interest but there is no information about the current relationship between these variables and the small domains. Only the past association structure expresses this relationship at some time in the past. It would seem reasonable to require the estimates p_{hig} to

(i) preserve the allocation structure between the associated and interest variables,

(ii) in some way carry over the association structure, observed earlier, without compromising (i) above.

If the sampling variability of the allocation structure is "not too great" these would be reasonable, but if the variability is "large" (i) may not be a good prerequisite for a model. The estimation procedure should generally ensure that $\sum_h p_{hig} = \xi_{.ig}$ ie we force the estimates to add to the known allocation structure. If we had any knowledge of change in the association structure over time then this could also be incorporated into our estimation procedure but rarely do we have such information.

In this case we have to assume that the association structure is stable over time.

It has been shown Purcell and Kish (1980) that weighted least squares, quasi maximum likelihood and information-theoretic approaches all result in the same form for the estimator

$$p_{hig} = \frac{\pi_{hig}}{\pi_{.ig}} \xi_{.ig}.$$

The assumption of a stable association structure implies that all cross product ratios not already respecified by the allocation structure are forced to be constant from the original census to the postcensal period. This is not unreasonable as we have no information on how they may have changed, if they have. However, if we believe that there may be change then this must be modeled explicitly or our estimates will not reflect this change.

In the simple example considered earlier we assume that there is no information in ξ, the allocation structure, concerning the sex effect or its two and three way interactions with the other two variables.

We can the write the following:-

$$\frac{p_{1ig}}{p_{2ig}} = \frac{\pi_{1ig}}{\pi_{2ig}}, \text{ for all } i \text{ and } g;$$ the sex effects

and $$\frac{p_{11g}p_{22g}}{p_{21g}p_{12g}} = \frac{\pi_{11g}\pi_{22g}}{\pi_{21g}\pi_{12g}}, \text{ for all } g;$$ two way interaction sex and employment at each level of ethnicity

and $$\frac{p_{1i1}p_{2i2}}{p_{2i1}p_{1i2}} = \frac{\pi_{1i1}\pi_{2i2}}{\pi_{2i1}\pi_{1i2}}, \text{ for all } i;$$ two way interaction sex and ethnicity at each level of employment

and $$\frac{p_{111}p_{221}p_{212}p_{122}}{p_{211}p_{121}p_{112}p_{222}} = \frac{\pi_{111}\pi_{221}\pi_{212}\pi_{122}}{\pi_{211}\pi_{121}\pi_{121}\pi_{221}}$$ Three way interaction sex, employment and ethnicity

The structure that is not preserved due to new information from the sample survey is:-

$$\frac{p_{h1g}}{p_{h2g}} \neq \frac{\pi_{h1g}}{\pi_{h2g}}, \text{ for all } h \text{ and } g$$ the employment effects

and $$\frac{p_{hi1}}{p_{hi2}} \neq \frac{\pi_{hi1}}{\pi_{hi2}}, \text{ for all } h \text{ and } i$$ the ethnicity effects

and $$\frac{p_{h11}p_{h22}}{p_{h21}p_{h12}} \neq \frac{\pi_{h11}\pi_{h22}}{\pi_{h21}\pi_{h12}}, \text{ for all } h$$ the two way interaction of employment and ethnicity

In other words we are assuming that the sex effect, the sex by employment interaction, the sex by ethnicity interaction and the sex by employment by ethnicity interaction all remain constant over the time since the census. The employment effect, the ethnicity effect and the employment by ethnicity interaction are allowed to change.

This can easily be mathematically extended to the general, H x I x G, case but visualising the effects and their interactions becomes more difficult. In practical applications the variables have real meaning and this helps to visualize the effects and interactions. However, for all cases the variable of interest effect and the associated variable effect and their interaction will not be preserved whilst all other effects and interactions will. Of course should any of these other effects or interactions, change over time or between the two variables of interest, if they are not the same, the estimates will be biased.

Given the data described thus far this is a reasonable solution to the problem. However there may be additional information and this should be incorporated into the analysis if possible. This additional information may come from two possible sources:-

1. independent estimates of the size of the small domains
2. current estimates of the distribution of the associated variables within the small domains.

If either of these are available then it would make sense to constrain our estimates so that they conform to this information. Hence for the first situation we could expect $\sum_i \sum_g p_{hig} = \hat{p}_{h..}$ with $\hat{p}_{h..}$ to be known and for the second $\sum_i p_{hig} = \hat{p}_{h.g}$ with $\hat{p}_{h.g}$ to be known.

In each situation if the estimates are not available from other sources it may be appropriate to use the equivalent π values. This requires additional assumptions to be made about the lack of change in the structure over time, but these may be acceptable. These new constraints are then denoted by ξ so as to be consistent with the earlier notation. This may mean that $\pi_{h.g}$ may be denoted by $\xi_{h.g}$ if they are being used as a constraint.

We now have two sets of constraints which will almost certainly not be exactly consistent. In theory an analytic solution as produced earlier is possible but in practice it is not. Deming and Stephan (1940) proposed an iterative procedure as a solution to a similar problem. Their procedure is generally referred to as iterative proportional fitting (IFP) or iterative scaling. The IPF procedure has some useful properties:-

- Provided that $p_{hig} > 0$ for all h, i, g then the procedure always converges. Fienberg (1970)
- Assuming that the underlying multinomial model is a correct model the method results in estimates that minimise the discriminant information I($p;\pi$) where $I(p;\pi) = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} Ln\left(\frac{p_{ij}}{\pi_{ij}}\right)$, where r and c are the number of rows and columns in a two way table Ireland and Kullback (1968). They also maximise the likelihood equation of the multinomial distribution, so although the estimates are not maximum likelihood they are Best Asymptotically Normal (BAN). BAN estimators maintain the properties of consistency, asymptotic normality and efficiency. Ireland and Kullback (1968), Darroch and Ratcliffe (1972) and Bishop, Fienberg and Holland (1975)
- Mosteller (1968) notes that part of the interaction structure is preserved. We have seen this above and have shown that any interactions not redefined by the new margins will remain unchanged.

For the first additional constraint the iterations follow the following:-

$${}_1p_{hig}^{(k)} = \frac{p_{hig}^{(k-1)}}{p_{.ig}^{(k-1)}}\xi_{.ig} \quad \text{and} \quad p_{hig}^{(k)} = \frac{{}_1p_{hig}^{(k)}}{{}_1p_{h..}^{(k)}}\xi_{h..} \quad \text{where } {}_1p \text{ results from the first step of}$$

the iteration. These are then repeated until a convergence criterion is met.

In the second case the first step is the same as the first set of margins are the same but the second marginal constraint results in $p_{hig}^{(k)} = \frac{{}_1p_{hig}^{(k)}}{{}_1p_{h.g}^{(k)}}\xi_{h.g}$

The same approach can be applied to tables of higher dimensions. The notation becomes clumsy as there are many subscripts and there are many more possible margins that can form the constraints. However the iterative procedure is still applicable with additional steps, one for each constraint, within each iteration.

As in the earlier case the preserved association structure of the IPF estimates is defined by the complete association structure less the structure respecified by the marginal constraints.

Structure preserving estimation is a well established method of small area estimation which uses the iterative proportional fitting algorithm to adjust the cell values of a table to new marginal totals whilst preserving some of the interaction structure of the original table. The only interactions which will change are those for which there is information in the new margins. In the next section we will look at a simple example which illustrates the process.

3.3 A simple example

As a simple example to illustrate IPF we consider a 2 x 2 table with everything known.

1	3
5	2

This is the census data as far as this example is concerned.

The new margins are also known.

	5
	8

This is the sample survey data. The survey

9	4	13

data is unable to estimate the individual cell counts but can provide counts for the margins.

The original table can also be written with its margins

1	3	4
5	2	7
6	5	11

We want to find values for the body of the table with the new margins which are consistent with the original table in terms of the odds ratio, the cross product ratio. This can be achieved using the IPF algorithm.

We multiply the row, or column, by the ratio of the margins so, in the example above, the top row is multiplied by $\frac{5}{4}$ and the second row by $\frac{8}{7}$.

This will adjust the cells in the table so that they sum to the column margin on the right hand side. This will not result in the correct row margin at the bottom.

1.25	3.75	5
5.7143	2.2857	8
6.9643	6.0357	13

However we can now multiply the columns by the equivalent ratios using the new table and the required row margin. In the example we multiply the left column by $\frac{9}{6.9643}$ and the right column by $\frac{4}{6.0357}$. The new result will have the correct row margin but the new column margin will now be incorrect. However we can iterate round this process until some convergence criterion is met. Continuing the iterations with the example above gives the following result after 10 iterations.

2.1779	2.8221	5
6.8216	1.1784	8
8.9995	4.0005	13

The detailed calculations can be found in Appendix A.

We need to set a convergence criterion but we could continue with any number of significant figures to any degree of accuracy. We will stop at this point.

The cross product ratio for the original table is $\frac{2}{15} = 0.133333$ and for the new table it is $\frac{2.1779 \times 1.1784}{6.8216 \times 2.8221} = 0.1333133$ remaining constant (within the rounding errors of the estimated cell values) at every step whilst the margins change. We will show that this is equivalent to keeping the interaction term constant in the log linear model that Marker (1999) suggested.

The extension to larger two dimensional tables simply requires the same process to be applied to the additional rows and or columns.

Tables with three or more dimensions do imply some other considerations as there are now a number of different margins which could be fitted and for each margin, or margins, different terms in the model will be kept constant. We will look at this issue in section 3.7.

We have presented an example to show the mechanics of the iterative proportional fitting algorithm. It is intended to give an understanding of the sorts of calculations necessary. The calculations become somewhat tedious but there are computer packages which will perform them simply. We will now show that the data in the contingency table can be modeled by a generalized linear model and that this then suggests a new approach that can be used to estimate the cell values.

3.4 *An alternative approach, the generalized linear model.*

Iterative Proportional Fitting is an iterative algorithm which fits new margins to counts in the cells of a contingency table. As we have seen the SPREE approach uses this algorithm to adjust counts from a census to the new margins found from sample survey data.

Now we will suggest an alternative method. We will model the census data as a log linear model which is an example of the generalized linear model. We will consider which coefficients will change when the updated margin is introduced. We will estimate the coefficients for the new model and show that the resulting estimates from the new model are the same as for the iterative proportional fit. This can be extended to several associated variables, and more than two categories for each variable, with little difficulty.

The use of the generalized linear model is a new approach. With this approach we can then relax the restriction that all variables are categorical and the concept underlying SPREE can be applied to a wider range of problems. The two assumptions that were made at the beginning of this chapter are more explicitly stated in the GLM form.

The table of census data may be modeled by a Log linear model of the form

$$Ln\left(\mathrm{E}\left[p_{ijk}\right]\right) = \mu + \alpha_i + \beta_j + \ . \ . \ + \alpha\beta_{ij} \qquad (3.1)$$

in which:-

- There may be 2 or more subscripts
- The variables are categorical

Using the development from Bishop, Fienberg and Holland (1975) we suppose that we have a table which represents proportions in two categories measured at two points in time. The basic table can be presented as

		Second Time Point	
	Category	1	2
First Time Point	1	p_{11}	p_{12}
	2	p_{21}	p_{22}

There are then two other ways in which we can present this two way table

First as the table measuring changes from the first measurement to the second,

	Category	Stays Same	Different
First Time Point	1	p_{11}	p_{12}
	2	p_{22}	p_{21}

secondly as the table measuring changes going back from the second time point to the first.

	Second Time Point	
Category	1	2
Stays Same	p_{11}	p_{22}
Different	p_{21}	p_{12}

For each of these 2 x 2 tables there is a cross product ratio, taking them in order we can write $\alpha_3 = \frac{p_{11}p_{22}}{p_{12}p_{21}}$, $\alpha_2 = \frac{p_{11}p_{21}}{p_{12}p_{22}}$ and $\alpha_1 = \frac{p_{11}p_{12}}{p_{22}p_{21}}$.

Taking logarithms of the $\{\alpha_i\}$ we get three linear contrasts :-

$$Ln(\alpha_3) = Ln(p_{11}) - Ln(p_{12}) - Ln(p_{21}) + Ln(p_{22})$$

$$Ln(\alpha_2) = Ln(p_{11}) - Ln(p_{12}) + Ln(p_{21}) - Ln(p_{22})$$

$$Ln(\alpha_1) = Ln(p_{11}) + Ln(p_{12}) - Ln(p_{21}) - Ln(p_{22})$$

Given values for these three contrasts, and remembering that $\sum p_{ij} = 1$, we have completely defined the four cell probabilities.

This formulation suggests that we should look at a model that is linear in the log scale.

The model proposed is

$$Ln\left(\mathrm{E}\left[p_{ij}\right]\right) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \quad i = 1,2;\ j = 1,2 \qquad (3.2)$$

u is the grand mean of the log(probabilities)

$u_{1(i)}$ is the deviation from the grand mean at level i of the first variable

$u_{2(j)}$ is the deviation from the grand mean at level j of the second variable

and $u_{12(ij)}$ is the deviation from $u + u_{1(i)} + u_{2(j)}$.

We can also note that $u_{1(1)} = -u_{1(2)}$, $u_{2(1)} = -u_{2(2)}$,

and $u_{12(11)} = -u_{12(12)} = -u_{12(21)} - = u_{12(22)}$

The development above is easily shown to be a special case of the generalized linear model (GLM) of Nelder and Wedderburn (1972) which is expanded by McCullagh and Nelder (1983). We briefly introduced the GLM in section 1.2. We will now show how SPREE can be expressed in terms of two GLMs which have some parameters in common.

The model is

$$g\left(\mathrm{E}[\mathbf{Y}]\right) = \mathbf{X}\boldsymbol{\beta}$$

as expressed in equation (1.1).

The model is linear in the parameters, β, but powers or other functions of the x variables are allowed in the $\mathbf{X}$ matrix.

The distributions for the underlying process $\mathbf{Y}$ can include any members of the exponential family of distributions which include the normal, binomial, Poisson etc. Nelder and Wedderburn (1972).

The simple models in which there is an identity link function and the errors are normally distributed can be expressed either as a simple linear regression or analysis of variance depending on the design matrix and include the models described by equation (1.2)**Error! Reference source not found.** and discussed in Chapter Two. If the elements of the design matrix are only indicators, variables denoting membership (or not) of a category then it is an analysis of variance, and if they are continuous variables then it is a regression. A mixture of the two results is the familiar ANCOVA.

Estimation of β for all of these cases with an identity link is simple under the assumption of independence and "constant variance", either in the different treatments in the ANOVA case or about the mean values, for regression. Then the covariance matrix is simply $\sigma^2\mathbf{I}_n$ and the maximum likelihood solution, $\hat{\boldsymbol{\beta}}$, for the parameters β is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{3.3}$$

σ^2 will also need to be estimated from the residuals as

$$\hat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{3.4}$$

The model can further be extended to link functions that are not the identity and for responses that are not normally distributed. Examples of this model include the log-linear model for Poisson distributed count data in a

contingency table. In this case the link function is a log function and the distribution is the Poisson distribution.

When the link function is not the identity the variance will not, in general, be constant. In this case a weighting is required in the iterative solution to account for the changing variance. Further the outcomes may also be correlated in which case the variance-covariance matrix has to be estimated as well as the parameters. This adds many more parameters to the model and often some covariance structure will be assumed to simplify the covariance matrix.

Although the model **Error! Reference source not found.** describes a non linear relationship between $\mathbf{Y}$ and $\boldsymbol{\beta}$, the least squares or maximum likelihood solution**Error! Reference source not found.** can be achieved by iterated generalized least squares, where at each iteration a generalized least squares algorithm is applied to a linear (rather than non linear) model under a metric which changes at each iteration and depends not only on Var($\mathbf{Y}$) but also on the partial derivatives of $\mathbf{X}\boldsymbol{\beta}$ with respect to μ=E($\mathbf{Y}$). The estimation procedure iterates between estimating the parameters $\boldsymbol{\beta}$ and estimating the variance-covariance of the residuals and $\hat{\boldsymbol{\mu}}$ until convergence. It has been shown del Pino (1989) that this results in maximum likelihood estimates of the parameters both for a generalized linear model and for the wider class of linearisable non linear models and that the dependent variable in the generalized least squares algorithm at each iteration. Initially the covariance matrix can be assumed to be $\mathbf{I}$ and the ordinary least squares estimates of the parameters are found, although other starting points are possible.

The generalized linear model can also include the range of hierarchical models for clustered survey data and split plot designs in experimental design.

In the case of the contingency table and the log linear model from equation. (3.1) the link function is a logarithm and the distribution would be Poisson. In the examples that we consider later we will not necessarily use Poisson distributions for the process but we will present evidence for the use of other distributions.

If a new marginal distribution is available (generally from a sample survey) then the elements in the table are adjusted to agree with the new margin(s). The main effects associated with the new margins will change along with any effects which are interactions only between these main effects. All other effects (main and interaction) not changed by the new margins remain as they were. In practice this will mean that at least one main effect and all interactions including that effect will not change. Thus higher order interactions predominantly remain the same. Under a model fitting regime some of the effects and the higher order interactions may be deemed to be 0; however, in general, we will fit the saturated model.

The advantage of fitting the saturated model is that it removes the need for complicated model fitting procedures which need to be justified each time. The saturated model ensures that all important effects are included in the model possibly at the expense of fitting some unnecessary terms.

Modelling a contingency table with a log linear model has been briefly introduced in this section. This has then been generalized to a GLM and the links demonstrated. In the next section we will explain the new approach and

then follow that with a number of examples of increasing complexity to show how this model can be applied to a SPREE type problem.

3.5 *The new approach*

We will now formally write the new approach and discuss its advantages

Following on from the earlier discussion and using the notation developed in Chapter One the general form of each GLM can be written as equation (1.1)

$$g\left(\mathrm{E}[\mathbf{Y}]\right) = \mathbf{X}\boldsymbol{\beta}.$$

This model can be applied to both the census and the survey data.

The parameter vector $\boldsymbol{\beta}$ can be partitioned into the parameters that are able to be estimated from the sample survey data, β_1, and those that are not, β_2. As the second partition of the parameters is not able to be estimated from the survey data the assumption is made that they will remain constant for both data sets.

The design matrix can be partitioned in the same way, subscripts c and s will denote the census data and the survey data.

We can then write $g\left(\mathrm{E}[\mathbf{Y}_c]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c}$ for the census data (3.5)

and $g\left(\mathrm{E}[\mathbf{Y}_s]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + X_2\boldsymbol{\beta}_{2s}$ for the survey data. (3.6)

In this new approach we see that the estimates of the individual cell counts are then formed by a combination of estimated parameters, some from the survey data and some from the census. Wherever possible parameter estimates from the survey data are used as they are considered to be more appropriate, either as they are more timely or measured on a more accurate variable, although this does not have to be the case.

We can now see that the SPREE model, via the IPF algorithm, is making the assumption that the second term in each of the models above is the same, that is

$$\mathbf{X}_2\boldsymbol{\beta}_{2c} = \mathbf{X}_2\boldsymbol{\beta}_{2s} \qquad (3.7)$$

We will look further at the ramifications of the assumption $\mathbf{X}_2\boldsymbol{\beta}_{2c} = \mathbf{X}_2\boldsymbol{\beta}_{2s}$ in Chapter Six of this thesis. It has been said that high correlation between the variables Y_c and Y_s indicates that SPREE is appropriate, Ambler et al (2001), Haslett, Green and Zingel (1998). We will discuss whether this is enough to ensure that the estimates are reasonable.

This expression of the method makes the assumptions used more explicit and as such is useful for the practitioner. The GLM applied to the table would be a log linear model with Poisson counts in each cell and the link function is clearly stated. The further assumptions about the association and allocation structures are simply expressed in terms of the equality of the second terms in the two models. This should make it easier to check the validity of this assumption however we will see in Chapter Six that, the very restrictions on the availability of data which make us use an estimation procedure mean that, the data is not available to check the assumption. However as we now have an explicit assumption it may be possible to collect some additional data to verify it or otherwise. This is more difficult using the IPF algorithm as the assumption is less clearly defined.

A further importance of the alternative formulation for SPREE models includes the possibility that the algorithm in this form can be applied to variables other than categorical. This introduces many new situations in which the concept underlying SPREE can be applied to other situations. The process of using data which is most likely already available to build a GLM

and then change some of the coefficients to agree with data collected in a sample survey which is not sufficient to find all of the coefficients could be applied in many ways.

Thus far we have omitted to mention anything about the possible variance of any estimates in the table. In the literature there seems to be relatively little information compared with the plethora of methods available to estimate the small area statistic. For SPREE the census data is generally considered to be measured without error and estimates of any error in the survey data are ignored in the simple application of the method. There are a number of sources of possible error in the estimates.

- The census data is unlikely to include 100% coverage and census structure may evolve over time
- The survey data will include sampling error
- The model proposed may not include all of the relevant variables

In Haslett, Green and Zingel (1998) the first two of these sources of error are considered. The data they used were unusual in that there were a number of estimates of the census data and that the survey was collected in 512 balanced replicates. These enabled appropriate estimation of the error in the final estimates. SPREE estimates were calculated for combinations of the different census data and for the survey replicates. This resulted in realistic errors for each estimate. We will consider this in more detail in Chapter Seven. Our new approach will make inclusion of this additional data somewhat easier and we will consider this later in the thesis.

The example that will be used to explain the new algorithm in this chapter is very simple in terms of having few variables and small numbers of categories for each variable, so that the underlying concepts can be explained without

other complications muddying the water. In the next chapter we will look at a more realistic example and we will find that there are some issues, such as the design matrix, which need to be considered in more detail. We will move to categorical variables with more than two categories and show how we can include variables that are not modeled by categorical variables at all.

3.6 Application of the GLM to our data

In SPREE we assume that the census data are in the form of a contingency table. This table can be modeled by a log linear model as discussed above. The sample survey data are not able to be disaggregated to the same fine level as the census data but the assumption is that there is still a table similar to the census table from which these margins have been derived.

If we consider the simple example from Section 3.3 we had the two by two table below of "census" data

1	3
5	2

which can be expressed as

$Ln\left(\mathrm{E}\left[p_{ij}\right]\right) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$ using the notation of section 3.3.

The sample survey data are shown below, the individual cells in this table are unknown.

		5
		8
9	4	13

The underlying table can also be expressed as

$Ln\left(\mathrm{E}\left[p_{ij}\right]\right) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$ but this model is not estimable from the information available.

In the next section we will discuss which parts of the model for the survey data can be estimated and which cannot. In the example given here it is intuitive that there is information in the known margins for the first three

terms in the model and none for the last, the interaction term. Birch (1963) showed that in a log linear model for a contingency table the marginal totals corresponding to unknown terms in the model are sufficient statistics given the other terms in the model. So in the case above if the interaction term $u_{12(ij)}$ is known then the margins given are sufficient statistics for the remaining terms. In the well known Chi squared test for independence the margins are used to construct a table in which the interaction is assumed to be zero we will assume that the interaction term is known from the census data and this along with the new margins, is used to estimate the other terms in the model hence estimating the individual cell counts.

3.7 *Identifying the effects and interactions which are updated by the sample survey data.*

In the simple example of SPREE which we considered in section 3.3 and have just revisited in section 3.6, we looked at a two by two table. The new margins were easy to identify and the effects that would change because of the new information contained in the sample survey data are intuitively clear. If we reconsider the two by two by two table of the form in section 3.1, and shown in *Figure 3.3* below, we will see that even with this simple extension the issue becomes less clear.

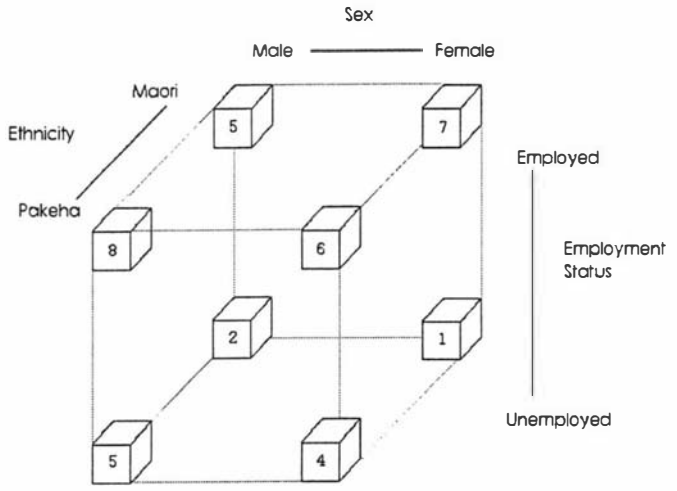


Figure 3.3 Main effects in a two by two by two table.

The three variables are Sex, Ethnicity and Employment Status and the individual cell counts represent the counts in each of the eight combinations of those categories.

The log linear model in this case has three main effects, three two way interactions and a three way interaction. Thus it can be written as

$$Ln\left(\mathrm{E}\left[p_{ij}\right]\right) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \quad (3.8)$$

We will now consider the various margins which could be found for this table. There are three two way margins an example of which is illustrated in *Figure 3.4*, a further three one way margins one of which is shown in *Figure 3.5* and a plethora of combinations of these of which a simple one is shown in *Figure 3.6*. *Figures 3.4* and *3.5* show margins constructed from the table in the diagram so that it is easy to see what they represent. *Figure 3.6* is more realistic for the problem that we have as the margins illustrated show no relationship with the given table as they are derived from the new information..

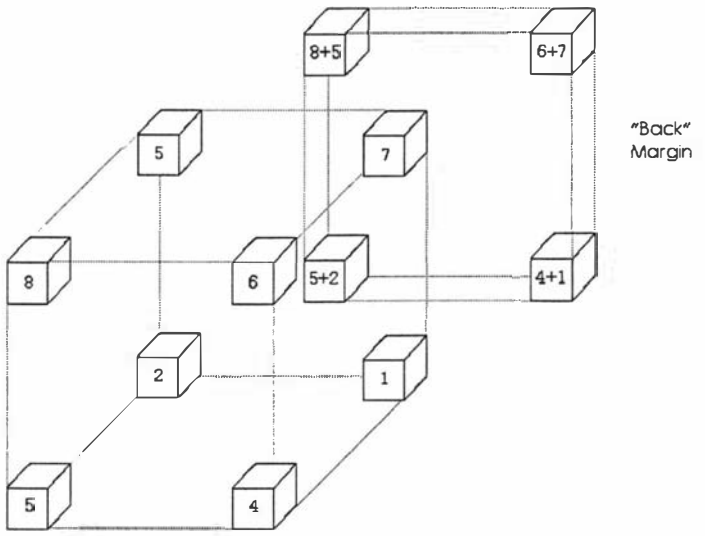


Figure 3.4 One two dimensional margin.

Similar margins can be found for the side (8 + 6, 5 + 7, 5 + 4 and 2 + 1) or for the "base", or "top", (8 + 5, 5 + 2, 6 + 4 and 7 + 1). These margins are then the counts of the two way tables, for the back margin the counts would be for Employment status by Sex.

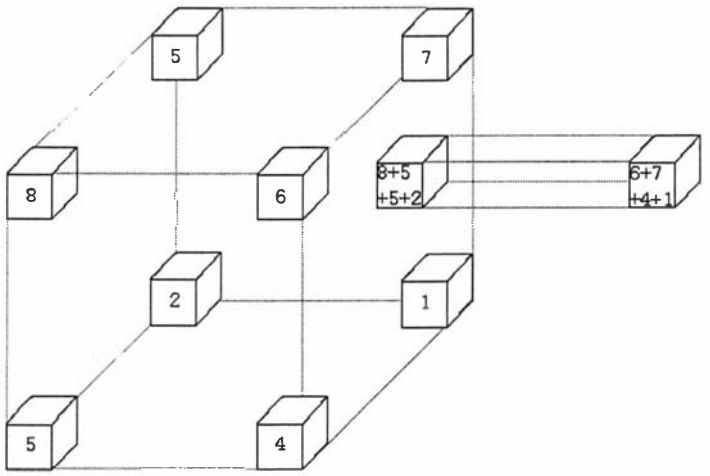


Figure 3.5 One single dimensional margin.

This margin in the diagram is the counts for Male and Female, similar margins can be found for the ethnicities, 8 + 5 + 6 + 4 and 5 + 7 + 2 + 1, and for Employment status, 8 + 5 + 6 + 7 and 5 + 2 + 4 + 1.

Alternatively we may have two new one way margins 19, 25 for the sexes and 20, 24 for the ethnicities in the example that we have shown above. Now only the main effects for sex and ethnicity are available from these new margins.

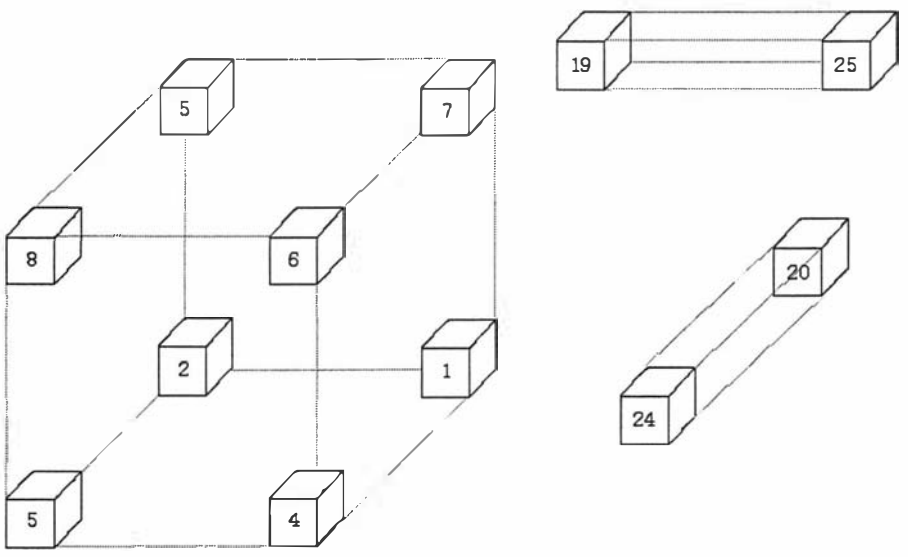


Figure 3.6 Two new single dimensional margins from survey data.

If we are trying to predict counts for Employment status then the new margins could be any combination of the examples above for the Sex and Ethnicity variables.

The details of the application of the IPF algorithm vary depending on the particular new margins which are available. In the case shown in *Figures 3.4* and *3.5* there will be convergence in one step of the IPF algorithm but the case shown in *Figure 3.6* will not, in general, converge as quickly because there are

two constraints which will not simultaneously apply but require the iterative process.

Using the model in equation (3.8) we arbitrarily denote u_1 as the sex effect, u_2 as the ethnicity effect and u_3 as the employment status effect and the interaction terms as the sensible combinations of the subscripts. The margins in *Figure 3.4* will provide estimates for u_1, u_3 and u_{13} as well as u as this is derived from the total. In *Figure 3.5* we are able to estimate u and u_1 only and in *Figure 3.6* we can estimate u, u_1 and u_2.

Later in this thesis we will consider real examples in which there are variables for two sexes, three age groups and two ethnicities as well as nine different regions. The combinations of possible marginals for a table such as this are large however the problem is ameliorated as the margins used are defined by the available data. In most cases the available data will not have a complicated structure and so the new marginals will also be reasonably simple.

3.8 A simple example

For illustrative purposes we shall look again at the simple two by two by two table used earlier in this chapter. There are three variables sex, ethnicity and employment status. These can take the values Male / female, Maori / other and Employed / Unemployed. For each cell there is a count and these can be lain out in a three dimensional diagram.

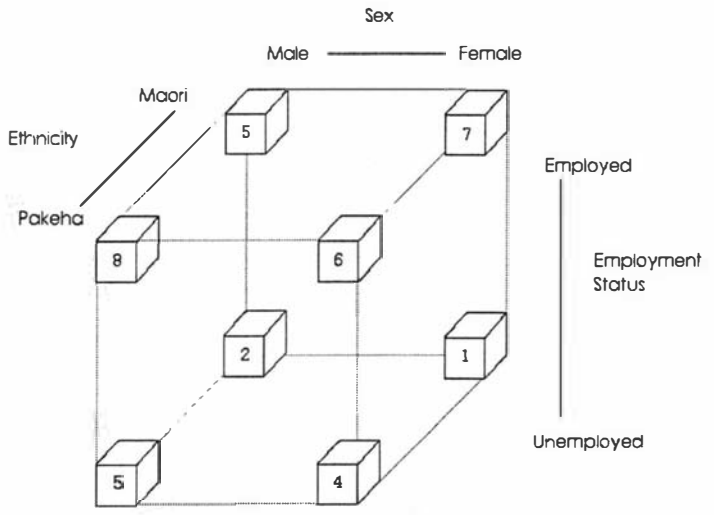


Figure 3.7 *The three dimensional diagram of cell counts for the 2 x 2 x 2 table presented in figure 3.3*

We will then fit a new Sex by Employment Status margin to this table. The margin in the table above is 13, 13, 7 and 5 and the new one that we will fit will be 14, 16, 5 and 9.

The log linear model that we have previously discussed is

$$Ln\left(E\left[\,p_{ijk}\,\right]\right) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

The terms have the following interpretations and their coefficients are given.

u	The overall mean	1.3894
$u_{1(1)} = -u_{1(2)}$	Main effect due to Sex	0.4672
$u_{2(1)} = -u_{2(2)}$	Main effect due to Employment Status	0.1084
$u_{3(1)} = -u_{3(2)}$	Main effect due to Ethnicity	0.3273
$u_{12(11)} = -u_{12(12)} = -u_{12(21)} = u_{12(22)}$	Interaction between Sex and Employment Status	-0.1206
$u_{13(11)} = -u_{13(12)} = -u_{13(21)} = u_{13(22)}$	Interaction between Sex and Ethnicity	-0.2483

$u_{23(11)} = -u_{23(12)} = -u_{23(21)} = u_{23(22)}$	Interaction between Employment Status and Ethnicity	0.0193
$u_{123(111)}$	Three way interaction	0.1368

The details of the calculation of these coefficients can be found in Appendix A The iterative proportional fitting algorithm is particularly easy to apply in this case as it converges in one step. Again the detailed calculations are shown in Apendix A. If we then fit the same model to the new values the coefficients can once again be found. Both sets of coefficients are presented in the table on the next page. It is easy to see which effects have remained the same and which have changed.

			Original Table	After IPF
Overall Mean	u		1.3894	1.5229
	$u_{1(.)}$	Sex	0.4672	0.4748
Main Effects	$u_{2(.)}$	Employment status	0.1084	0.1559
	$u_{3(.)}$	Ethnicity	0.3273	0.3273
	$u_{12(.)}$	Sex*Employment Status	-0.1206	0.0770
2 Way Interactions	$u_{13(.)}$	Sex*Ethnicity	-0.2483	-0.2483
	$u_{23(.)}$	Employment Status*Ethnicity	0.0193	0.0193
3 Way Interaction	$u_{123(.)}$	Sex*Employment Status*Ethnicity	0.1368	0.1368

The results are as expected. The parameters for the overall mean, Sex and Employment status main effects and the Sex by Employment status

interaction have all changed whilst the other main effect and interactions have stayed the same.

3.9 *Another approach for binary data.*

The variable of interest in this case is the employment status and this is a binary variable. Even with unemployment data this is not necessarily so, part time work may be classified as a third category and housepersons (wives or husbands) not seeking work could be another. The other variables may be binary, as in this case, but could also be multinomial variables for example there may be more categories of ethnicity and additional variables, eg age groups, could also be included.

Where one variable has only two levels the model can also be written as a binomial model with a logit link. The probabilities are now the proportions that are employed and unemployed in each combination of categories in the table. Previously all of the probabilities added to 1 now each column adds to 1.

The logit is defined as

$$logit(p'_{ijk}) = Log\left(\frac{p_{ijk}}{1-p_{ijk}}\right)$$

Hence

$$logit(p'_{ijk}) = Ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = Ln\left(\frac{p_{1jk}}{p_{2jk}}\right) = Ln\left(p_{1jk}\right) - Ln\left(p_{2jk}\right)$$

$$\Rightarrow Logit(\mathrm{E}\left[p'_{ijk}\right]) = u + u_{1(1)} + u_{2(j)} + u_{3(k)} + u_{12(1j)} + u_{13(1k)} + u_{23(jk)} + u_{123(1jk)} -$$

$$\left(u + u_{1(2)} + u_{2(j)} + u_{3(k)} + u_{12(2j)} + u_{13(2k)} + u_{23(jk)} + u_{123(2jk)}\right)$$

$$Logit(\mathrm{E}\left[p'_{ijk}\right]) = u + u_{1(1)} + u_{2(j)} + u_{3(k)} + u_{12(1j)} + u_{13(1k)} + u_{23(jk)} + u_{123(1jk)}$$

$$-u + u_{1(1)} - u_{2(j)} - u_{3(k)} + u_{12(1j)} + u_{13(1k)} - u_{23(jk)} + u_{123(1jk)}$$

$$= 2u_{1(1)} + 2u_{12(1j)} + 2u_{13(1k)} + 2u_{123(1jk)}$$

These effects are calculated for the table that we have been considering. The detail can be found in Appendix A. The results are given in the table below.

The overall mean	0.9344	2 x 0.4672
First main effect	0.2412	2 x 0.1206
Second main effect	-0.4966	2 x -0.2483
Interaction	0.2736	2 x 0.1368

These can be compared with the appropriate values in the first log-linear model above.

This approach is limiting as the variable of interest must have only two outcomes, as opposed to the log-linear model in which the variable of interest may have more outcomes.

In some cases the alternative approach of fitting a logistic model may be used. It has been shown that the two models are equivalent and that there is a simple relationship between parameters in the models. There are no advantages in using the logistic model at this stage so we will not consider it any further. In Chapter Nine we will revisit the model when we investigate relative risk models.

3.10 Fitting the models.

Models such as these can easily be fitted in standard statistical packages such as SAS or Splus. We have also used MLwiN which is designed for analysis of multi level models but it is able to be applied to this problem with a little effort. In this section we will show how these packages can be applied to the simple table discussed in section 3.7. We will first show how SAS and Splus

can be used and then move on to MLwiN and explain what advantages and disadvantages it has.

The model that is being fitted is equation (1.1) **Error! Reference source not found.** and **X** is a design matrix. Most statistics packages that can fit models such as these have an automatic default for this design matrix but they are not necessarily the same default. This is not a problem if the final estimates are all that matter but we wish to compare the coefficients in some models and these are dependant on the design matrix. MLwiN does not have a default design matrix as it is not specifically intended for fitting these sorts of models. For these reasons we have developed a design matrix which is used in this example for all of the packages.

There are many ways in which the problem may be parameterised, some more convenient than others. Thus far an overall mean has been found and then each effect has either been above or below this. This is the same as using +1 and –1 for each effect and the design matrix looks like the one below:-

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{pmatrix}$$

In SAS we can fit the model to the original table with the program below. The variables in the model correspond to the columns in the design matrix above.

```
proc genmod data=sec39;
model counts=u u1 u2 u3 u12 u13 u23 u123 / dist=poi noint;
run;
```

The estimates for the coefficients coincide with those calculated in section 3.7 as they should. The coefficients which will not change are then used to calculate what is often termed an offset. An offset is a term in the model for which the coefficient is known to be one and in this case represents the census value of $\mathbf{X}_2\boldsymbol{\beta}_2$. The offset in this case is the contribution of those terms to the model. It is calculated by multiplying the part of the design matrix associated with the terms that stay the same with their coefficients. In this case:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0.3273 \\ -0.2483 \\ 0.0193 \\ 0.1368 \end{bmatrix} = \begin{bmatrix} 0.2350 \\ -0.2350 \\ -0.0771 \\ 0.0771 \\ 0.4581 \\ -0.4581 \\ 0.6931 \\ \end{bmatrix}$$

The other part of the partition of the design matrix is then used with these offsets to find the new coefficients. The data for this step are the margin from the survey data. It is necessary to use the margins to build a table which only exhibits the effects to be estimated. In this case it is easy as each value can be halved and then repeated twice so the two halves of the table are the same. The SAS program is then:

```
proc genmod data=sec39;
model newcounts=u u1 u2 u12 / dist=poi noint offset= offs;
run;
```

where `newcounts` is a vector $\begin{bmatrix} 7 \\ 7 \\ 8 \\ 8 \\ 2.5 \\ 2.5 \\ 4.5 \\ 4.5 \end{bmatrix}$ and `offs` is the vector above.

The output from this program includes the following section.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
U	1	1.5227	0.1668	1.1958	1.8496	83.36	<.0001
U1	1	0.4748	0.1668	0.1479	0.8017	8.11	0.0044
U2	1	-0.1560	0.1668	-0.4829	0.1709	0.88	0.3496
U12	1	0.0770	0.1668	-0.2498	0.4039	0.21	0.6441
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

The coefficients agree with those that are calculated in Appendix A. This set of coefficients can then be used to predict the new cell values.

The data can be modeled in Splus using the glm function and a Poisson distribution. Splus includes the constant term by default so the design matrix does not need the constant term The census data is easily modeled:

census <- glm (y ~ X, family = poisson)

The matrix X is the design matrix above without the first column.

This gives the coefficients for those terms which will remain the same. They are then used to produce the offset as before. Using the sample data, a design matrix for the effects which will change and the offset the new coefficients are estimated.

sample <- glm (sampledata ~ Xa + offset(offsets), family = poisson)

The new coefficients are estimated and predicted cell counts can be calculated from the full model.

MLwiN can also be used. This program is intended for data which exhibits multiple levels, for example students in classes in schools. The table of data that we have here does not have this structure but it can all be placed as a single group within a higher level and the program will then perform appropriately. MLwiN has a very easy to use interface to define the model and the output below shows the fit for the table used earlier illustrating the interface. Not surprisingly the effects are the same as before.

$$\left.\begin{aligned} &prob_{ijk} \sim Poisson(\pi_{ijk}) \\ &prob_{ijk} = \pi_{ijk} + e_{0ijk}pcons^{*}_{j} \end{aligned}\right\}$$

$$\begin{aligned} \log(\pi_{ijk}) = {} & -2.248(1.239)cons1_{j} + 0.108(1.239)i_{ijk} + 0.467(1.239)j_{ijk} + 0.327(1.239)k_{ijk} + -0.121(1.239)i*j_{ijk} + 0.019(1.239)i*k_{ijk} + \\ & -0.248(1.239)j*k_{ijk} + 0.137(1.239)i*j*k_{ijk} \end{aligned}$$

$$pcons^{*}_{j} = pcons_{j}\pi_{ijk}^{0.5}$$

$$[e_{0ijk}] \sim (0, \Omega_{e}) : \Omega_{e} = [1.000(0.000)]$$

Extracting the effects that we know will stay the same gives

u_k = 0.327

u_{ik} = 0.019

u_{jk} = -0.248

u_{ijk} = 0.137

Note: These estimates are available to greater accuracy.

To estimate the other effects we need to use the new margin and the known effects.

In MLwiN the offset does not have to be calculated as it can estimate the new coefficients with constraints which can be that some coefficients are predetermined.

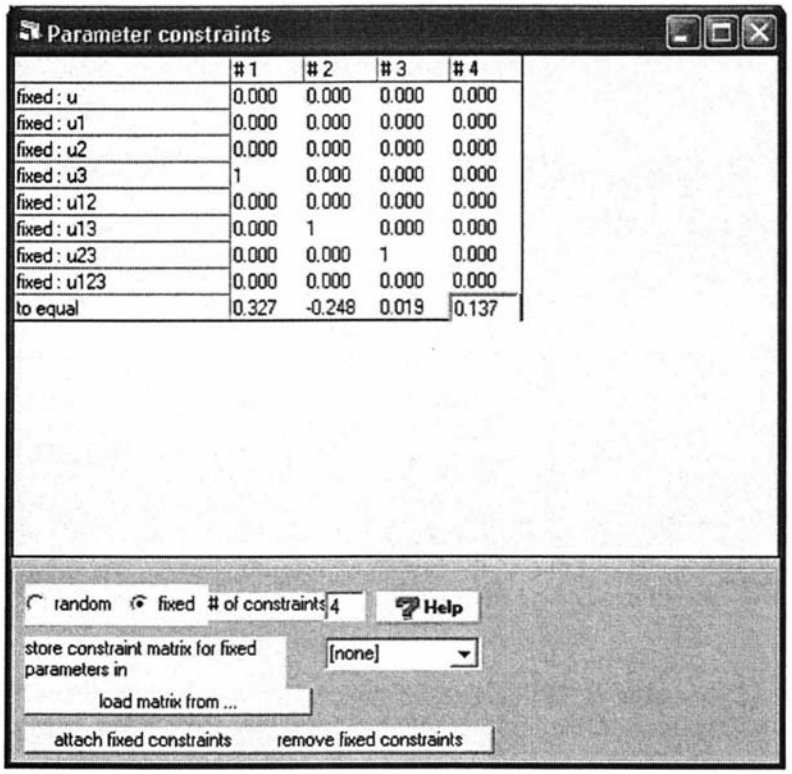


Figure 3.8 MLwiN screen for constraining parameters in a model.

If we now fit the model to the same new data vector as before the results are as expected below.

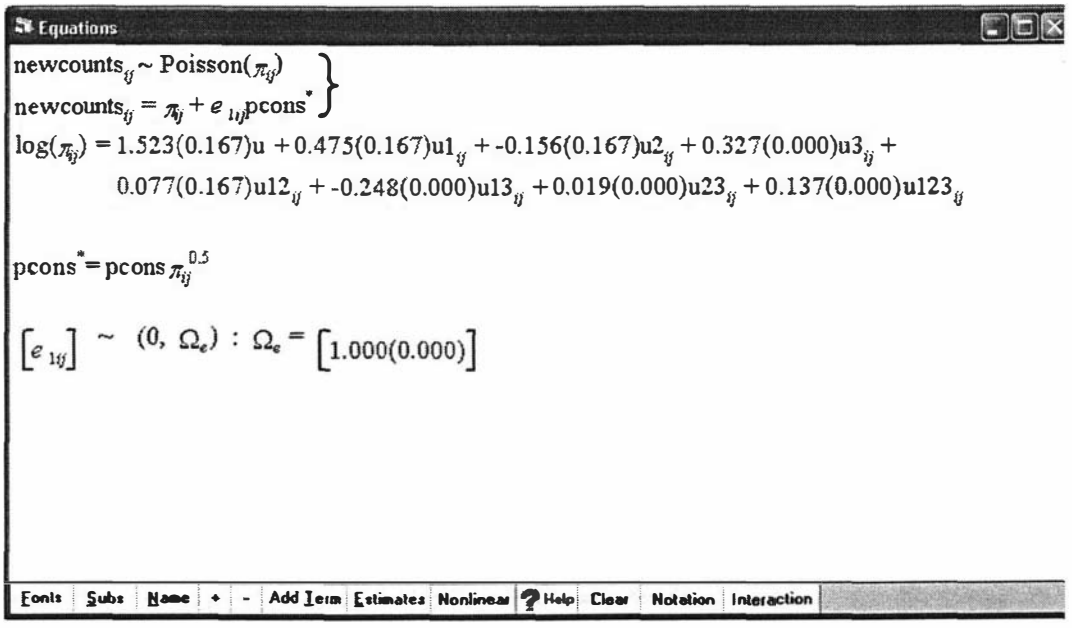


Figure 3.9 MLwiN output screen.

Each of these packages has a number of algorithms available for fitting these models. In the examples above the iterative generalized least squares algorithm has been used. See del Pino (1989) for further details.

We have now shown that the SPREE method can be fitted using a generalized linear model approach and that standard commercial software packages can be used for the fitting. The GLM is fitted to the census data and the parameters that are reestimated by the survey data are found using the survey data and the coefficients from the census model for those parameters that remain the same.

3.11 Concluding remarks

In this chapter we have looked at the SPREE method for small area estimation using IPF. We have shown that it is equivalent to fitting a generalized linear model to the census data and then refitting the model to the margins of the

sample survey data whilst holding some of the parameters fixed from the census model. This approach makes the assumptions that are made in the IPF approach, much more transparent and it allows a wider range of models to be fitted to the data including relaxing the requirement that all variables must be categorical. In Chapter Five we look at an example in which age is modeled with a quadratic form and the approach includes the possibility of including continuous variables in the model. In Chapter Eight a conditional autoregressive error structure is included which adds a continuous random variable to the model. These are extensions that SPREE is unable to include.

In looking at some of these other models later in this thesis we will see that this extension also allows us to include some variation in the data for both the census and the sample survey data. This is possible with SPREE but in the example that we consider later there are 8 replicates of the census data and 512 replicates of the survey. This would mean doing the IPF 520 times to reflect all of the variability in the data. Haslett, Green and Zingel (1998) have shown that the 520 replicates will result in adequate estimates as opposed to the 512 timeds 8 or 4096 replicates which would appear to be required.

Whilst the extended approach makes many new modelling scenarios possible using the concept that underlies SPREE there do remain practical difficulties which may affect a particular situation. The nature of the data collected will constrain the range of models possible. Although it has been stated that increases in computing power make many of the small area estimation methods possible it still needs to be remembered that there are limitations of memory and processor speed which may prevent some estimation methods on some models being used with some data sets. As an example the simple construction and checking of design matrices can become a significant activity and it is clearly crucial that the design matrix is correct. Even in the smallest

practical example used later in the thesis the design matrix is 54 by 54, for the largest it is 384 by 384. In a matrix of these dimensions ensuring that every element is correct is not simple. Some approaches to these problems will be mentioned later in the thesis when real data sets are used.

In the next chapter we will discuss the data which will be used for the examples in this thesis. It will become clear that the data is suitable for SPREE type estimation. In Chapter Two we considered a range of small area estimation methods and it should be made clear that there may be other approaches which could be used for the data that will be described.

It is our intention to show that SPREE can be applied in a wider range of situations than has previously been suggested and to show that the new approach has some advantages. Whether this is the "best", however that may be measured, estimation method for any particular application or even for the application which we present is not the main thrust of this work.

CHAPTER 4

An example of the Generalized Linear Model approach

4.1 Introduction.

In the previous chapter we showed an alternative approach to SPREE estimation using a generalized linear model approach. We wrote the model for the census data and expressed the survey data using a model of the same form. We estimated the coefficients in the model for the census data and up dated some of those using the information from the survey data. Some of the coefficients estimated using the census data were used in the final model along with the ones estimated using the survey data. The new cell values were then predicted from this model.

In this chapter we will apply this new approach to an example for unemployment counts in the nine regions of the North Island of New Zealand classified in three age groups and two sexes. We shall see that the new approach will give identical results to the SPREE approach using the iterative proportional fitting algorithm for this example. As only categorical variables are used either approach is applicable.

This new algorithm will be shown, in later chapters, to have much wider applicability than the SPREE method using IPF. In the simple example in Chapter Three the coefficients in the model could be evaluated easily but in any practical applications we require a computer package to estimate the coefficients. In the previous chapter we showed how some packages could be used for the simple example. We will continue to use these packages for the

examples in this thesis and also introduce another, WinBUGS, as it has features which will be shown to be useful.

The table below, Table 1.1 from Chapter One, shows the various models which we will introduce in the rest of this thesis, the method that will be used and the software packages involved. It also shows by omission that there are a large number of other combinations of the models which we have not used.

	Chapter					
	3	4	5	7	8	9
Variability in Census				X	X	X
Categorical Variables Only	X	X		X	X	X
Linear or Quadratic Terms			X	X		
Conditional Autoregressive Error Structure					X	
Relative Risks						X
Method Used	SPREE by IPF or GLM	SPREE by IPF or GLM	GLM	GLM	GLM	GLM
Software	SAS, Splus, MLwiN	SAS, Splus, MLwiN	MLwiN BUGS	BUGS	BUGS	BUGS

Table 4.1 Models fitted, computer software used and chapter references

In Chapter Three we introduced the SPREE model using iterative proportional fitting and showed an alternative formulation using GLMs. In this chapter we will show that the two approaches will produce the same results for the example data. The range of models will be extended to linear

and quadratic models for one of the categorical variables, age in this case, in Chapter Five and then the other models in the table will be introduced in the subsequent chapters. Some of the possible combinations are not given as examples as this would become very repetitive and add little to the overall development. The models used are intended to show some possibilities which the new algorithm allows. Many other possibilities are left unmentioned but the breadth of application should become clear. The table includes all of the models which we will consider. We will look later at the conditional autoregressive (CAR) and relative risk (RR) models. These will not be combined with all of the other possibilities. The table shows the chapters in which the models that have been considered are described.

In most cases there are a number of computer packages that could be used to estimate the coefficients in the different models. This thesis is not intended to compare the different packages and in each case we will only perform the analysis with one package in detail but will try to indicate which alternate packages could be used for some of the models. The packages that have been used include SAS, S-Plus, MLwiN and WinBUGS. The latter two are not main stream statistical packages like the other two, they have been used as they have an emphasis which has made them useful in the later parts of this work. Their special features will be explored at the appropriate time. The list of computer packages is in no way exhaustive and is only intended to show that a number of commonly available, "off the shelf", packages can be used for the analysis using this algorithm. We will see that some of the design matrices become quite large and so the speed of computing does become an issue. The majority of the computing was completed on a Pentium I machine with a cpu speed of 166MHz and 64 Mb of RAM. More recently a 2GHz Pentium 4 machine with 1Gb RAM has been available and this has improved the speed

of the computing considerably. The relative risk models were analysed using the latter hardware.

The analyses in the thesis were constrained by the available data. Some of the models will be of little practical value for this example data, however it is important to remember that there may be other situations in which those models could be appropriate. This thesis intends to indicate the possible extensions to the method which become possible by using the generalized linear model approach and which may be applied to different data structures.

4.2 The data

In New Zealand, Statistics New Zealand carries out a Household Labour Force Survey (HLFS) quarterly. This survey has a sample of 15,000 households and approximately 30,000 individuals in the civilian, non institutionalised, usually resident population aged 15 years and over. The groups that are excluded are:

- those living in non-private dwellings
- long term residents of old peoples homes, hospitals and psychiatric institutions
- inmates of penal institutions
- members of the permanent armed forces
- members of the non-New Zealand armed forces
- overseas diplomats
- overseas visitors who expect to be resident in New Zealand for less than 12 months
- those aged less than 15 years of age

- people living on offshore islands except Waiheke island.

Young males tend to be under-represented, their high mobility making them difficult to include in the sample.

There is a questionnaire for each household and an individual questionnaire for each eligible person in the household.

To be employed the person only needs to work for one hour in the week, or to work, even unpaid, in a family business. To be unemployed the person must be available to start a job and to be actively seeking work, not just looking in the paper. Statistics New Zealand uses the International Labour Organisation (ILO) definition of unemployment

It must be made clear that whilst the data collected in this survey in the March quarter of 1996 is the basis of the data used in this thesis some manipulation of the data has occurred so that it is able to illustrate the types of analysis that we have used. In fact some categories have been combined whilst others have been disaggregated to produce tables with the structure that we want. It should be noted that the data in this form are not of a sufficient standard that Statistics New Zealand would publish the data. Were the techniques to be used by an organisation such as Statistics New Zealand they could generate the necessary data structure and figures directly from the survey but it would not be identical to the example that we have used as they would be able to correctly identify the categories whilst we have only been able to make intelligent guesses. However these considerations in no way affect the reliability or validity of the statistical methods developed here to analyse the data.

The Department of Work and Income in New Zealand (WINZ) collects monthly data on registered unemployed people through its national network

of offices. The aims of the Department are to distribute unemployment benefits to those that are entitled to them and to help all unemployed people, who seek it, to find work. The registered unemployed are eligible for a benefit which is paid by central government. The benefit is limited to certain categories of people and is distributed with some restrictions to encourage the beneficiary to find work. The Department of Work and Income have a more restricted administrative definition of unemployment than the ILO definition.

While the definition of unemployed as used by the two organizations is different, the two unemployment variables are highly correlated and are said to allow the association structure generated by The Department of Work and Income data to be used for a SPREE approach to estimating unemployment rates under the Statistics New Zealand definition. Green, Haslett and Zingel (1998). Whether this is in fact a correct statement will be considered further in Chapter Six. Statistics New Zealand currently publish a national unemployment rate and estimates at the Regional level (16 nationally). They would like to publish estimates at the Territorial Authority (74) level. The sample sizes in the HLFS are too small in many Territorial Authorities for Statistics New Zealand to publish these direct estimates. Even some Regional Authorities have larger errors than Statistics New Zealand would like. The "census" data provided by WINZ is used to find the majority of the parameters and the survey data, only supplying a margin for the three age groups and one for the two sexes, and implicitly a new total, is used to reestimate the three main effects and the constant term.

This has been a very brief introduction to the data that will be used for the practical examples in this thesis. The exact nature of the sampling, adjustments for non response and other details are not given. We will assume that the SNZ and WINZ data are appropriate for the analysis. Clearly this is

important in any real analysis but we only want to show that the new approach is practicable.

For administrative purposes New Zealand is divided into 16 Regional Authorities. There are also 74 smaller Territorial Authorities which predominantly are subdivisions of the Regions with some minor exceptions. A map showing the Regional Authority boundaries is given below.

Figure 4.1 Map of the Regional Authorities of New Zealand

4.3 Application of the new algorithm

For the example below the data are restricted to the North Island which includes nine of the sixteen national Regional Authorities. The

unemployment counts are divided into three age groups 15 to 24, 25 to 49 and 50 or over and two sexes, male and female.

The data from the Department of Work and Income is used to form the census association structure for a saturated log linear model and the fit for that model is shown below. The data are for March 1996. Some values are non integer because the boundaries used by the two organizations are not identical. Where boundaries do not coincide the WINZ data have been apportioned to the SNZ regions and hence some fractions have been introduced.

Region		Age Groups		
		15 to 24	25 to 49	Over 50
Northland	Male	1794.4	4386.6	632.2
	Female	1060.3	2555.0	465.0
Auckland	Male	7233.1	16542.7	2324.7
	Female	5040.4	8104.5	1796.9
Waikato	Male	2924.9	4955.0	792.0
	Female	2047.8	2849.2	654.0
Bay of Plenty	Male	2552.5	5169.5	755.0
	Female	1862.4	1681.2	601.0
Gisborne	Male	680.0	1511.0	203.0
	Female	438.0	1011.0	120.0
Hawkes Bay	Male	1819.9	3470.9	550.0
	Female	1137.0	1109.1	377.0
Taranaki	Male	1295.9	1982.1	332.2
	Female	896.0	1530.7	277.1
Manawatu-	Male	2552.7	4287.6	752.0
Wanganui	Female	1693.3	2220.2	538.5
Wellington	Male	4073.4	6911.3	1145.7
	Female	2555.7	2395.0	753.3

Table 4.2 *Census data from Work and Income New Zealand for unemployment counts by sex and three age groups in each region.*

There are a number of ways to parameterise the model and the choice of these is arbitrary. As has already been discussed some software packages will generate the design matrix given the categorical variables and a statement as to which terms and interactions to include, for example SAS and S-Plus. Other

packages require a design matrix to be explicitly given either as a series of column vectors, MLwiN, or as a complete matrix, WinBUGS. If a design matrix is generated then SAS and S-Plus are able to accommodate that matrix. MLwiN can easily accept data from Excel so it can be used to generate the column vectors. Alternatively MATLAB is easy to use to generate the design matrix and can be used to save the matrix in a comma delimited form which can then be copied directly into WinBUGS, SAS or S-Plus or copied into Excel for use in MLwiN.

The parameterization which is used consistently in this thesis has already been discussed. It takes a constant term and then looks at deviations from that constant. Thus if the variable has two levels the effect will either be added or subtracted, the design matrix then consists of a column of 1's or -1's. If there are more than two possibilities there will be one fewer columns than the number of possibilities and these will be 1's, 0's or -1 and all columns will sum to 0.

Appendix B includes the design matrix for the example given here with instructions needed to form it and some comments on checking that it is correct. It is very easy to check the matrix used in Chapter Three as it is only 8 by 8. The one for the example in this chapter is 54 by 54 and can be printed on a single page with each row on one line so again it is not too hard to check visually. We can easily see patterns which repeat themselves throughout the matrix. The entire matrix only contains 1's 0's or -1's and in Excel it is easy to colour these numbers differently, a macro can be written to do this in a large matrix, and then the patterns are obvious. Later we will consider the problem with 16 regions, 2 sexes, 3 age groups and 2 ethnicities which requires a 192 by 192 design matrix. Checking that this one is correct becomes more difficult

and the extension to larger matrices is easy but they become much more difficult to check.

The details of the model parameterization are:

A constant term;

Sex is +1 for males and –1 for females;

Two terms Age1 and Age2 for the three age groups

for ages 15 to 24 Age1 is +1and Age2 is 0,

for ages 25 to 49 Age1 is 0 and Age2 is +1,

for ages 50 and over both Age1 and Age2 are -1.

Similarly for the nine regions there are eight terms which define the nine regions in terms of 1's, 0's and -1's.

As a consequence figures quoted are differences from the mean.

We will now work through the example detailing each step in the algorithm.

Step A Estimate the log linear model for the census data. The table of coefficients is given on the next page. The main effects are in the upper part of the table with the interaction lower down. The table is split into two parts. The first two lines are the coefficients which will be estimated again when the sample survey data is introduced. The rest of the table gives the values of the coefficients which will be carried forward into the model for the small areas from the census data.

β_1 Effects	Constant	Sex	Age1	Age2
Estimated from Census Data	7.300	0.240	0.223	0.723

β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.

Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
-0.086	**1.262**	**0.238**	**0.097**	**-1.142**	**-0.308**	**-0.580**	**0.077**

Two way interactions	Sex by Age1	**-0.034**	Sex by Age2	**0.118**

Two way interactions, Regions with Sex

-0.012	**-0.018**	**-0.057**	**0.037**	**-0.013**	**0.091**	**-0.106**	**-0.007**

Two way interactions, Regions with Age1

-0.207	**-0.079**	**0.042**	**0.068**	**-0.078**	**0.057**	**0.040**	**0.040**

Two way interactions, Regions with Age2

0.179	**0.072**	**-0.029**	**-0.130**	**0.239**	**-0.133**	**0.020**	**-0.065**

Three way interactions, Regions with Sex and Age1

0.068	**-0.007**	**0.029**	**-0.086**	**0.026**	**-0.062**	**0.083**	**0.006**

Three way interactions, Regions with Sex and Age2

-0.077	**0.017**	**-0.025**	**0.166**	**-0.145**	**0.121**	**-0.123**	**-0.023**

Table 4.3 *Table of coefficients for the full model with categorical variables for region and sex, and for the two age categories. Coefficients in bold type will be carried forward.*

Step B Select those coefficients which remain the same, and those which will be changed.

The coefficients that are held the same are indicated in the table above in bold type. These terms have been selected because the new margins only give information about the two age effects and the sex effect, as well as a different population size. Hence these effects and the constant term will change in the refitted model. There are four independent new pieces of information and so four coefficients will change in the model. (The survey data might also have been used to refit the Sex by Age1 and Sex by Age2 effects if sufficiently accurate survey estimates were available for these subpopulations.)

Step C Constrain the coefficients that stay constant to be the same as they were under the census based log linear model. In MLwiN this is done by filling out the window as shown below.

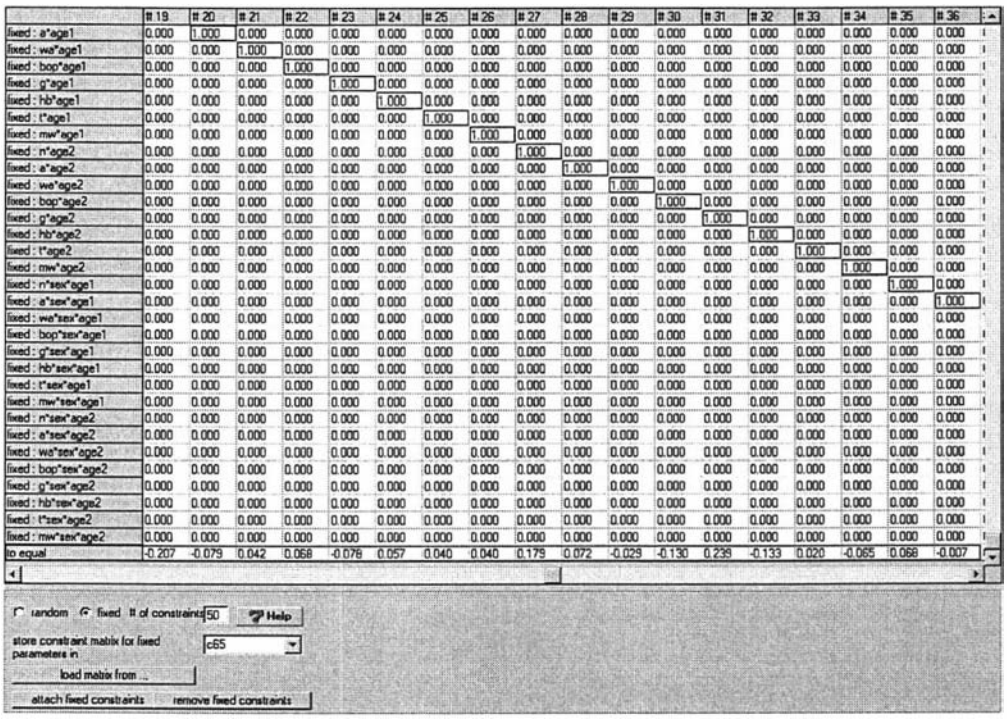


Figure 4.2 Part of the MLwiN window for constraining parameters.

It can be seen that in each column there is a 1 in the row for that particular term and the value of the coefficient in the row at the bottom labelled "to equal". So, for example, in the column number 22 in the window above we see a 1 in the bop*age1 row and the number 0.068 in the row at the bottom. This relates to the coefficient in the table 4.1 in the Bay of Plenty column in the two way interactions Region with Age1 row which is seen to be 0.068.

Step D Refit the model using the new sample based margins. To do this we need to build a new table from marginal data using only the effects that will change, i.e in this case construct the independent 3 x 2 table for the new age and sex margins. This table then includes the information about the effects that we wish to

change but contains no information about the other effects which will come from the coefficients in the table from the census data.

19078	22503	5534	47116	Female
25156	29671	7296	62125	Male
44235	52175	12831	109241	
15 to 24	25 to 49	Over 50		

Then divide these by 9 to give the nine regions.

2119.856	2500.361	614.895
2795.144	3296.861	810.772

Refit the model to this table with each entry repeated 9 times for each region, constrain the chosen coefficients and the new coefficients will be estimated. The table with the estimated coefficients for the final model is shown on the next page.

β_1 Effects	Constant	Sex	Age1	Age2
Estimated from Census Data	7.300	0.240	0.223	0.723
Estimated from Survey Data	**7.226**	**0.105**	**0.374**	**0.505**

β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.

Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
-0.086	1.262	0.238	0.097	-1.142	-0.308	-0.580	0.077

Two way interactions	Sex by Age1	-0.034	Sex by Age2	0.118

Two way interactions, Regions with Sex

-0.012	-0.018	-0.057	0.037	-0.013	0.091	-0.106	-0.007

Two way interactions, Regions with Age1

-0.207	-0.079	0.042	0.068	-0.078	0.057	0.040	0.040

Two way interactions, Regions with Age2

0.179	0.072	-0.029	-0.130	0.239	-0.133	0.020	-0.065

Three way interactions, Regions with Sex and Age1

0.068	-0.007	0.029	-0.086	0.026	-0.062	0.083	0.006

Three way interactions, Regions with Sex and Age2

-0.077	0.017	-0.025	0.166	-0.145	0.121	-0.123	-0.023

Table 4.4 Fully saturated model fitted to the sample data with constrained coefficients. The reestimated coefficients are shown in bold.

Step E Finally, predict the new values for the table from the new model which is the combination of the coefficients from the two data sources.

Region		Age Groups		
		15 to 24	25 to 49	Over 50
Northland	Male	1693.7	2861.9	547.8
	Female	1312.3	2185.7	528.4
Auckland	Male	6827.2	10792.7	2014.4
	Female	6238.3	6933.1	2041.7
Waikato	Male	2760.8	3232.7	686.3
	Female	2534.5	2437.4	743.1
Bay of	Male	2409.2	3372.6	654.2
Plenty	Female	2305.0	1438.2	682.9
Gisborne	Male	641.8	985.8	175.9
	Female	542.1	864.8	136.4
Hawkes Bay	Male	1717.8	2264.4	476.6
	Female	1407.2	948.8	428.3
Taranaki	Male	1223.2	1293.1	287.9
	Female	1108.9	1309.4	314.9
Manawatu-	Male	2409.4	2797.3	651.6
Wanganui	Female	2095.7	1899.3	611.9
Wellington	Male	3844.8	4509.0	992.8
	Female	3163.1	2048.8	855.9

Table 4.5 Final estimates from the combined model

An iterative proportional fit for the same original table and the new margins yields the same result but without estimating the log linear model coefficients.

One advantage of the explicit use of a generalized linear model is that the algorithm is not then restricted categorical data. Continuous variables may be incorporated and, via a careful respecification of the model detailed for example in Goldstein (1995), random effects as well as fixed effects may be fitted using the IGLS algorithm.

4.4 Brief notes on computing

The comments in section 3.9 are all relevant to these models and to the rest of this thesis. We have explained briefly the parameterization used for the example here and it follows the same principles as the parameterization used in section 3.9. MLwiN has been used for the example here but SAS or S plus could equally easily been used or a number of other packages. The model fitting follows the method used in section 3.9 and only differs in the size of the data set and the complexity of the design matrix.

Some difficulty in using MLwiN for larger tables has been experienced. The program reserves some columns in its worksheet for internal use and the design matrix may use those same columns. Under those circumstances data in the design matrix has been lost and the program was unable to fit the model. A workaround to not use the columns used internally was partially successful but correctly identifying the columns used and avoiding them was somewhat tedious. The columns used internally begin at column 90 and about the next 10 columns are used. As the design matrix used for the above example is only 54 columns wide there is plenty of space for data columns as well and the problem is avoided. MLwiN intends to avoid these problems in future versions.

4.5 Conclusions

In this chapter we have seen how the algorithm proposed in Chapter Three can be applied in a number of standard computer packages to a real data set. As has been noted the same results could have been obtained using the traditional IPF algorithm in this case. The direct estimates, at the Regional Authority level, are not available for comparison with the se estimates. The survey data were only able to provide the margins used in the example.

The extension beyond this to larger tables is simple in theory as more categories can be added to each variable or more variables can be added to the problem. It becomes harder to display the tables and to visualise the array of data but the log-linear model is easily extended with more terms and higher order interactions. The computing power required increases with the size of the table. Most of the examples used in this thesis have been modified so that the computing did not become a serious issue. They are representative of real problems but may not be as complex as they could be. The full analysis of the data available would involve 74 Territorial Authorities, two sexes, three age groups and four ethnicities. This then requires a 74 by 2 by 3 by 4 contingency table with a 1776 by 1776 design matrix. The computing power available as this work was begun was not adequate for this problem, the latest Pentium 4 type processors will cope although they are still slow.

The table 4.1 at the beginning of this chapter foreshadowed the development in the rest of the thesis and we will look at the areas in which the new algorithm can be applied but where IPF could not. We shall consider a much wider range of models and give a number of examples. In the next chapter we will consider modelling the age structure in the table either with a linear or a quadratic function. We will show that there is some evidence of curvature in the plots of age and so a quadratic function would appear to be more

appropriate although the model with a linear age variable is more parsimonious.

CHAPTER 5

Quadratic and linear functions for the age variable

5.1 Introduction

In the first three chapters we developed a new algorithm for small area estimation based on the concepts associated with structure preserving estimation, and in Chapter Four we presented an example. We suggested that this new algorithm allows different models to be fitted which are not possible with SPREE. The rest of the thesis will concentrate on investigating some of the options available to us and showing that the new algorithm can be applied in these cases. Below is part of the table, from Chapter Four, showing the models that we are proposing and the way that they are interlinked. We have already considered the SPREE model and the model for categorical variables using the generalized linear model in Chapter Three. We have shown that they are in fact the same.

Thesis Chapter	Categorical x Linear or quadratic ✔	variability in census ✔	CAR	RR	Method Used	Software Package
5	✔				GLM	MLwiN, BUGS
7	x	✔			GLM	BUGS
7	✔	✔			GLM	BUGS
8	x	✔	✔		GLM	BUGS
8	✔	✔	✔		GLM	BUGS
9	✔	✔		✔	GLM	BUGS

Table 5.1 *Table showing the different models used in this thesis, the chapter in which they are discussed, the estimation process and computer package used.*

The table shows this model allowing no variability in the census data. In other words the parameters that are carried forward from the census data model to the sample survey data model are taken to be fixed. This restriction will be relaxed in Chapter Seven.

The initial extension will be to investigate quadratic and linear terms for the age variable rather than using a categorical approach. To demonstrate the approach we will begin by using the three age groups that have been used previously although it may be more useful to use more age groups in a practical model. In section 5.4 we will extend the idea to 11 age groups which is a more meaningful model.

Using the three age groups the results should be the same as for the three categories as the quadratic function should model the three groups exactly. In effect we are simply reparameterizing the same model. If we include all of the interaction terms then we again have a saturated model and the results should be the same as for SPREE using categorical variables evaluated either using IPF or the new approach. We will consider two other models a linear one with all of the interactions and a quadratic model that only includes a quadratic main effect and no interactions. There are many other possible models that could be tried but this is not an exercise in model fitting more an illustration of the possibilities.

5.2 A quadratic function for age.

As has been mentioned earlier an advantage of recognising that SPREE can be included in the generalized linear model framework is that it can then be extended to include variables that are not categorical, which is not possible with SPREE using the IPF algorithm. Essentially the concept that makes SPREE work is that census data is available which is used to estimate the coefficients in a log linear model which defines the association structure. Sample survey data is used to reestimate the effects and interactions which

are able to be found from the survey data. These are combined to build a model which is used to predict the small area estimates.

The same concept can apply to data which is not entirely categorical, if there is a way of estimating the coefficients in a proposed underlying model. In SPREE the coefficients do not need to be estimated as iterative proportional fitting can be used to change the values in the body of the table to be consistent with the new margins. Here we are using iterative generalized least squares to estimate the model coefficients for the census data and then reestimating the coefficients which will change given the survey data. This does not require all of the variables, or in fact any of the variables, to be categorical, although in many applications they will be. There may also be a mixture of categorical and continuous. This has the potential to improve the small area estimates. Instead of collapsing a continuous variable into a few categories, which requires some assumptions about that variable, the variable may be able to be modelled more appropriately, even with quite simple functions for example linear or quadratic.

It is noted here that the data that we have is not able to be analysed with truly continuous variables. We do not have elementwise data for the census and privacy laws make it difficult to operate on the unit record data from the survey. However we can show the effect of modelling the age category by either a quadratic or a linear form. Currently there are three categories of age and so a quadratic form will fit these exactly.

Looking at the two graphs in *Figure 5.1* below we can see that there appears to be some curvature in the data when plotted against the age variable and would suspect that a quadratic form would fit somewhat better than a linear function. We will fit the linear function later in this chapter.

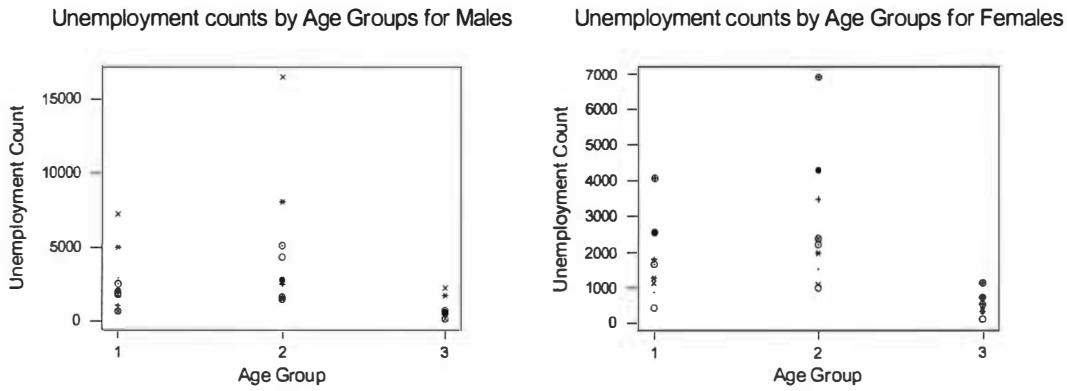


Figure 5.1 *Graphs of unemployment counts against the three age groups by regions for males and females.*

The analysis that is reported below has been carried out in MLwiN using the iterative generalized least squares estimation option. The design matrix is a modification of the one in Appendix B, and used in Chapter Four, in which the Age 1 variable uses values of 20, 37.5 and 57,5 for the three ages. These are mid points in the age categories used. The Age 2 variable is simply Age 1 squared. The interaction terms are then found using these values. The design matrices are predominantly zeros and so the patterns of numbers other than zeros can be used to check the matrix as in Appendix B. The Matlab sparse matrix for this design matrix is shown below and can be seen to be the same as the one for the categorical variables as no account is taken for the numbers that are now not 1 or –1 but –17.5 and 20 or 306.25 or 400.

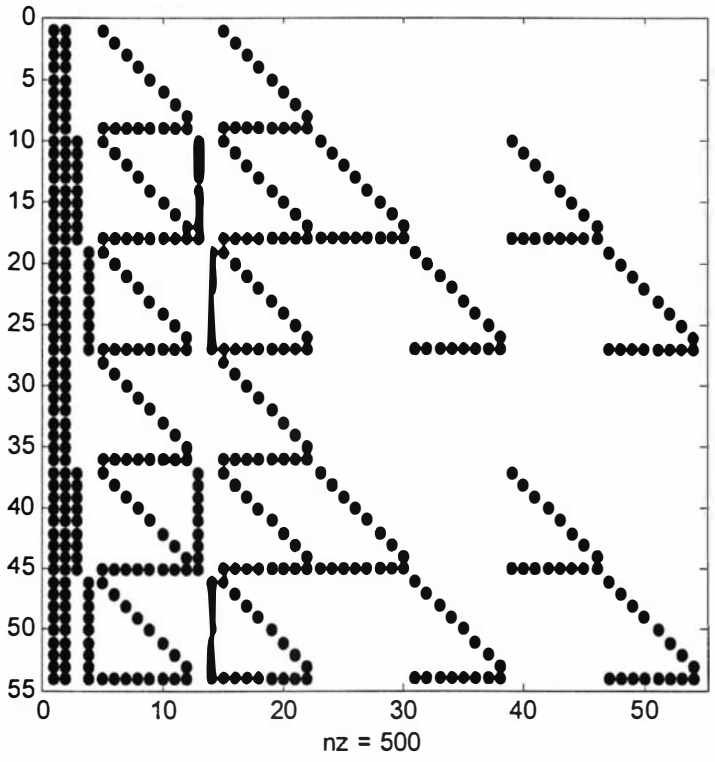


Figure 5.2 Matlab sparse matrix representation of the design matrix.

When fitted to the census data we get the coefficients reported in Table 5.2 on the next page. The coefficients are split into those that change between the census and survey data and those that are held constant, in other words β_1 and β_2. there are two values for each of β_1, the values estimated from the census and those from the survey. These are in the top section of the table and the rest are in the lower section. The complete set of coefficients β_1 from the survey data and β_2 from the census are then used to predict the cell counts which are given in Table 5.3.

As both models are saturated and there are three age categories in the example in Chapter Four and a quadratic term in this case the results should be the same. Comparing the two tables we find that this is so except for some small differences (in the 4th or 5th significant figure) which will be due to rounding errors in entering the coefficients which remain constant between the census data model and the sample survey data model.

β_1 Effects	Constant	Sex	Age	Age Squared
Estimated from Census Data	8.0225	0.3583	-0.0237	-0.0030
Estimated from Survey Data	7.73090	0.22287	-0.02831	-0.00205

β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.

Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
0.09358	1.33439	0.20891	-0.03374	-0.90292	-0.44082	-0.55976	0.01189
Two way interactions		Sex by Age	-0.00007	Sex by Age Squared	-0.00050		
Two way interactions, Regions with Sex							
-0.08803	-0.00153	-0.08162	0.20331	-0.15742	0.21216	-0.22924	-0.02917
Two way interactions, Regions with Age							
0.00823	0.00306	-0.00181	-0.00156	0.00031	-0.00094	-0.00247	-0.00111
Two way interactions, Regions with Age Squared							
-0.00079	-0.00032	0.00013	0.00056	-0.00102	0.00057	-0.00008	0.00028
Three way interactions, Regions with Sex and Age							
-0.00243	0.00012	-0.00115	0.00194	0.00094	0.00139	-0.00252	0.00005
Three way interactions, Regions with Sex and Age Squared							
0.00033	-0.00007	0.00011	-0.00071	0.00061	-0.00052	0.00053	0.00009

Table 5.2 Table of coefficients for the full model with categorical variables for region and sex, and linear and quadratic terms for age

Region		Age Groups		
		15 to 24	25 to 49	Over 50
Northland	Male	1693.4	2862.1	547.8
	Female	1311.9	2185.6	528.4
Auckland	Male	6827.4	10792.7	2015.1
	Female	6237.9	6932.3	2042.2
Waikato	Male	2761.1	3232.6	686.5
	Female	2534.8	2437.1	743.2
Bay of Plenty	Male	2408.9	3372.3	654.4
	Female	2304.6	1438.0	683.0
Gisborne	Male	641.9	985.8	176.0
	Female	542.1	864.8	136.4
Hawkes Bay	Male	1718.0	2264.5	476.7
	Female	1407.3	948.7	428.5
Taranaki	Male	1223.3	1293.1	287.8
	Female	1109.0	1309.6	314.8
Manawatu-Wanganui	Male	2409.8	2797.5	651.8
	Female	2095.4	1899.0	612.5
Wellington	Male	3844.8	4508.8	992.3
	Female	3163.6	2048.7	855.8

Table 5.3 *Predicted counts for the saturated model with a quadratic term for age.*

We can see that the full quadratic model makes the same predictions as the saturated model with the categorical variables and as the original SPREE algorithm as they are all saturated models with no error term. We can use these predictions as a base level to compare other models. We will use the usual deviance statistic:-

$$G^2 = \sum_i x_i Ln \frac{\hat{m}_i}{x_i} \qquad (5.1)$$

where x_i are the data, in our case the predicted cell counts from the saturated model

$\hat{m}_i$ are the predicted cell counts for the model

to make the comparisons. We need to be aware that with large values in the cell counts and the large number of cells quite small residuals in each cell will sum to a large value overall so it is likely that most alternate models will be significantly different from the saturated one and hence would be rejected under the usual criteria. We should also note that given the complex survey design and the estimation using both sample survey and census data the distribution of the cell counts would be unlikely to be Poisson. Use of G^2 in a formal hypothesis test would be inappropriate but large changes would still indicate an improvement or worsening of the model.

This section is important as it represents the first example in which the new approach, suggested in this thesis, has been applied to a model which could not be constructed using SPREE and the IPF algorithm.

5.3 Other possible models.

A linear function for age or a quadratic main effect with no interactions.

We have just shown that a quadratic function can be fitted to the age categories and as there are three categories in this example we get the same results. We could also see whether a linear term in age would give an adequate fit to the data with a reduction in the number of parameters in the model. Table 5.4 below shows the estimated coefficients for this model, the terms with XXX in the coefficient column are not included in the model. Table 5.5 then shows the predicted cell values along with the residual calculated from the predicted values for the saturated model, i.e either the SPREE model

or the new algorithm with categorical age variables or the quadratic function, as they all give the same results.

β_1 Effects	Constant	Sex	Age	Age Squared
Estimated from Census Data	7.4881	0.2838	-0.0206	XXX
Estimated from Survey Data	7.2451	0.1338	-0.0255	XXX

β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.

Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
-0.0291	1.2813	0.2216	0.0597	-1.0404	-0.3528	-0.5768	0.0451
Two way interactions		Sex by Age	0.0009	Sex by Age Squared	XXX		
Two way interactions, Regions with Sex							
-0.0302	-0.0016	-0.0605	0.0889	-0.0719	0.1278	-0.1460	-0.0137
Two way interactions, Regions with Age							
0.0060	0.0025	-0.0013	-0.0020	0.0024	-0.0015	-0.0013	-0.0012
Two way interactions, Regions with Age Squared							
XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX
Three way interactions, Regions with Sex and Age							
-0.0020	4.93E-5	-0.0009	0.0026	-0.0008	0.0018	-0.0026	-0.0002
Three way interactions, Regions with Sex and Age Squared							
XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX

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*

		Age Groups					
		15 to 24		25 to 49		Over 50	
Region	Gender	Count	Residual	Count	Residual	Count	Residual
Northland	Male	2392.1	-698.7	1668.6	1193.5	1105.5	-557.7
	Female	1871.1	-559.2	1356.1	829.5	938.6	-410.2
Auckland	Male	9358.4	-2531.0	6365.8	4426.8	4098.3	-2083.2
	Female	7436.7	-1198.8	4886.3	2046.0	3023.6	-981.4
Waikato	Male	3319.4	-558.3	2079.8	1152.8	1219.0	-532.5
	Female	2872.1	-337.4	1796.0	641.1	1050.2	-307.0
Bay of Plenty	Male	3123.8	-714.9	2054.2	1318.1	1272.3	-617.9
	Female	2263.4	41.2	1315.6	122.3	707.7	-24.7
Gisborne	Male	870.1	-228.2	582.1	403.7	367.7	-191.8
	Female	773.8	-231.7	514.3	350.5	322.4	-186.1
Hawkes Bay	Male	2164.2	-446.2	1413.7	850.8	868.9	-392.2
	Female	1410.4	-3.2	837.7	110.9	461.9	-33.4
Taranaki	Male	1414.3	-191.0	859.3	433.7	486.3	-198.5
	Female	1368.0	-259.1	880.5	429.1	532.2	-217.4
Manawatu-Wanganui	Male	2874.5	-464.6	1827.0	970.5	1088.4	-436.7
	Female	2323.1	-227.7	1436.6	462.4	829.4	-216.9
Wellington	Male	4616.6	-771.8	2914.1	1594.6	1722.5	-729.2
	Female	3161.7	1.9	1800.0	248.7	945.5	-89.8

Table 5.5 *Predicted counts and residuals for the linear model with all interactions.*

From a simple look at the residuals it is immediately clear that the quadratic model is likely to fit better. The residuals for the first age group are almost all negative, the two positive ones are quite small. The second age group has

positive residuals throughout and the third age group has negative residuals in all cells. This would clearly suggest curvature in the data.

The deviance was calculated and found to be $G^2 = 15041.36$ with 36 degrees of freedom and the reduction in parameters in the model is 18. The increase in G^2 from the saturated model is far too large to make this a suitable model.

Another model was considered in which a quadratic term for age was included but there were no interaction terms. This suggests that the same quadratic effect is present for all sex by region groups. The estimated coefficients for this model are given in Table 5.6 and the predictions with residuals are shown in Table 5.6 below.

For this final model the deviance is $G^2 = 4078.6$ with 20 degrees of freedom. This is a considerable reduction in deviance from the linear model and with fewer parameters. However the deviance is too big to imply that this model is as good as the saturated model

		Age Groups					
		15 to 24		25 to 49		Over 50	
Region	Gender	Count	Residual	Count	Residual	Count	Residual
Northland	Male	2392.1	-698.7	1668.6	1193.5	1105.5	-557.7
	Female	1871.1	-559.2	1356.1	829.5	938.6	-410.2
Auckland	Male	9358.4	-2531.0	6365.8	4426.8	4098.3	-2083.2
	Female	7436.7	-1198.8	4886.3	2046.0	3023.6	-981.4
Waikato	Male	3319.4	-558.3	2079.8	1152.8	1219.0	-532.5
	Female	2872.1	-337.4	1796.0	641.1	1050.2	-307.0
Bay of	Male	3123.8	-714.9	2054.2	1318.1	1272.3	-617.9
Plenty	Female	2263.4	41.2	1315.6	122.3	707.7	-24.7
Gisborne	Male	870.1	-228.2	582.1	403.7	367.7	-191.8
	Female	773.8	-231.7	514.3	350.5	322.4	-186.1
Hawkes Bay	Male	2164.2	-446.2	1413.7	850.8	868.9	-392.2
	Female	1410.4	-3.2	837.7	110.9	461.9	-33.4
Taranaki	Male	1414.3	-191.0	859.3	433.7	486.3	-198.5
	Female	1368.0	-259.1	880.5	429.1	532.2	-217.4
Manawatu-	Male	2874.5	-464.6	1827.0	970.5	1088.4	-436.7
Wanganui	Female	2323.1	-227.7	1436.6	462.4	829.4	-216.9
Wellington	Male	4616.6	-771.8	2914.1	1594.6	1722.5	-729.2
	Female	3161.7	1.9	1800.0	248.7	945.5	-89.8

Table 5.6 *Predicted counts and residuals for the linear model with all interactions.*

β_1 Effects	Constant	Sex	Age	Age Squared
Estimated from Census Data	8.05820	0.27960	-0.02307	-0.00314
Estimated from Survey Data	7.77586	0.12859	-0.02769	-0.00212

β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.

Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
-0.04934	1.27124	0.22523	0.06765	-1.05010	-0.34719	-0.57259	0.04849

Two way interactions	Sex by Age	XXX	Sex by Age Squared	XXX

Two way interactions, Regions with Sex

-0.02323	-0.00067	-0.05654	0.07819	-0.06833	0.12069	-0.13511	-0.01266

Two way interactions, Regions with Age

XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX

Two way interactions, Regions with Age Squared

XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX

Three way interactions, Regions with Sex and Age

XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX

Three way interactions, Regions with Sex and Age Squared

XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX

Table 5.7 *Table of coefficients for the model with the quadratic main effect as the only age effect.*

Two more models have been proposed and fitted using the new approach. The linear model has flaws which are obvious with the most basic diagnostics. Simply by looking at the residuals we have seen that there is a pattern which would be better modeled by a quadratic term. The quadratic model with no interaction term is also inadequate and looking at the two graphs in Figure 5.1 it is not hard to see why. The quadratic form is quite different for some regions and at least some interaction terms should be included.

The models introduce in this section are models that SPREE in its traditional form could not fit. We have used the GLM approach and shown that where the two approaches could both be applied, in the saturated model with a quadratic effect, the estimates are the same. We have then applied the GLM approach to models that are not able to be fitted with SPREE.

5.4 A more realistic model

Eleven age groups and a quadratic age effect.

The most detailed data that Statistics New Zealand collects is in five yearly categories which makes eleven age categories for each region by sex. With the new algorithm we are able to model this in a more parsimonious way using a quadratic term for the age groups. In general the counts follow a curve which could be approximated by a quadratic.

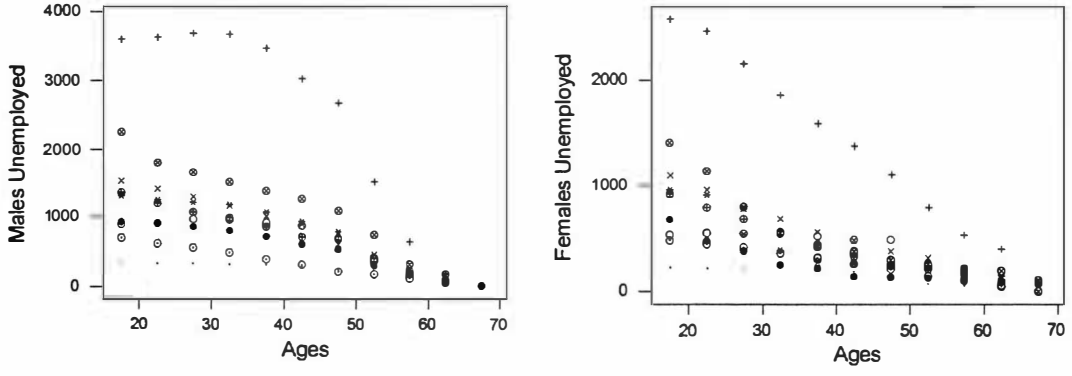


Figure 5.3 Graphs of unemployment counts against the eleven age groups for males and females.

Additional polynomial or other terms could be included but the purpose of this paper is not to find the best model, but simply to demonstrate that the new algorithm opens up a much wider range of possible small area models than available in regression and SPREE.

Using the same algorithm as in Section 5.2 and the data in Tables 5.8 and 5.9 we estimate the coefficients in the model, in Table 5.10, and hence the predicted values in Table 5.11 below.

Age	17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5
Margin	24760	19475	14254	11325	9965	9067	7564	5214	3654	3269	694

Table 5.8 *The new unemployment margin for the five yearly age groups. The margin for sex stays as before.*

From Table 5.8 the independent table can be generated in the same way as Table 5.2, to use as data for Step D

Census data in 5 yearly age groups. Column headings are the centre of the age range for that group.

Region		17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5
North	M	882	912	968	954	918	869	678	402	180	50	0
	F	525	535	540	535	511	485	484	235	178	52	0
Auck	M	3603	3630	3690	3666	3476	3032	2679	1520	631	174	0
	F	2572	2468	2148	1863	1598	1386	1110	789	523	396	89
Waikato	M	1523	1402	1298	1156	1008	893	600	459	263	68	2
	F	1096	952	774	683	551	476	365	307	215	132	0
B o P	M	1302	1251	1219	1176	1077	927	771	451	231	73	0
	F	953	909	534	382	270	261	234	220	186	143	52
Gisborne	M	347	333	328	327	312	298	246	121	69	13	0
	F	221	217	220	224	218	187	162	64	47	9	0
Hawkes	M	924	896	843	791	703	601	533	306	195	49	0
Bay	F	675	462	376	249	213	143	128	123	106	83	65
Taranaki	M	690	606	554	485	397	326	220	176	116	40	0
	F	470	426	406	355	308	256	206	142	93	39	3
Man	M	1357	1196	1073	986	861	704	664	362	225	165	0
Wang	F	911	782	683	555	427	316	239	201	158	100	80
W'ton	M	2263	1810	1664	1513	1380	1268	1086	732	321	93	0
	F	1413	1143	796	546	405	365	283	251	204	189	109

Table 5.9 *Counts for unemployment from Department of Work and Income data in five yearly intervals*

β_1 Effects	Constant	Sex	Age	Age Squared
Estimated from Census Data	6.20135	-0.41529	0.05591	-0.00131
Estimated from Survey Data	7.50247	-0.54423	-0.02632	-0.00029

β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.

Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
-1.54162	0.72169	0.41673	0.69281	-2.42406	0.53471	-0.53682	0.68335

Two way interactions	Sex by Age	0.04182	Sex by Age Squared	-0.00055

Two way interactions, Regions with Sex

0.53136	-0.16852	0.17345	-0.76242	0.69340	-0.72056	0.69263	0.04976

Two way interactions, Regions with Age

0.07943	0.02992	-0.00755	-0.03641	0.07922	-0.05184	0.00277	-0.03603

Two way interactions, Regions with Age Squared

-0.00091	-0.00035	0.00006	0.00046	-0.00100	0.00066	-0.00009	0.00045

Three way interactions, Regions with Sex and Age

-0.03027	0.00867	-0.01273	0.04885	-0.04633	0.04922	-0.04676	-0.00479

Three way interactions, Regions with Sex and Age Squared

0.00035	-0.00011	0.00015	-0.00062	0.00061	-0.00062	0.00057	0.00008

Table 5.10 *Table of coefficients for the full model with categorical variables for region and sex, and linear and quadratic terms for age*

Predictions in 5 yearly age groups. Column headings are the centre of the age range for that group.

Region		17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5
North	M	775	810	789	718	608	481	355	244	157	94	52
	F	598	602	576	524	454	374	293	218	154	108	66
Auck	M	3177	3219	3057	2723	2274	1781	1308	900	581	352	200
	F	3368	2675	2126	1691	1347	1074	857	684	547	437	350
Waikato	M	1439	1239	1033	835	654	496	365	260	179	120	78
	F	1393	1052	800	614	475	371	292	232	185	150	122
B o P	M	1182	1115	1001	855	696	539	397	278	186	118	71
	F	1371	830	537	371	274	216	182	164	158	162	179
Gisborne	M	299	299	281	248	206	161	119	82	54	33	19
	F	149	254	241	214	177	138	100	68	43	25	14
Hawkes	M	858	781	683	575	464	361	269	193	133	88	56
Bay	F	906	528	332	225	165	131	111	103	102	110	127
Taranaki	M	669	540	429	334	255	192	142	103	73	51	35
	F	583	490	404	326	258	200	152	114	83	59	42
Man	M	1281	1062	868	698	553	432	332	251	187	138	100
Wang	F	1256	867	618	455	346	271	219	183	158	141	129
W'ton	M	1976	1694	1415	1151	912	704	529	388	277	192	130
	F	1976	1166	738	501	365	286	240	216	208	216	240

Table 5.11 *Predictions for unemployment in five yearly intervals*

We could compare these predictions with the predictions for the saturated model either from SPREE using iterative proportional fitting, or via the generalized linear model algorithm outlined in Chapter Four.

The new approach to SPREE has been extended to a model with 11 age groups and a quadratic function fitted to this data. This cannot be done in the traditional use of SPREE.

5.5 Closing comments.

In this chapter we have shown that the new algorithm can be used in situations where SPREE will not work. The change in the modelling approach to using linear and quadratic terms or to dropping some of the interactions allows more sophisticated, more parsimonious, or both, models to be considered. It should now be clear that this approach can be used with other data structures and as long as the model can be expressed as a generalized linear model, and that there is both census and sample survey data available, we can use the algorithm to estimate values of the response variable. The algorithm can in fact be used in applications other than those related to small area estimation in that if there are important explanatory variables in a model that are not measured in the sample data we may be able to estimate the coefficients for those terms from other data.

Statistics New Zealand collects its data for the HLFS in five yearly intervals of which there are eleven. It is not currently known how detailed WINZ are able to report their data however it would seem that a polynomial function, either quadratic or cubic should be adequate, would be a useful approach to modelling that many age intervals. As there is no intention to extrapolate to other ages polynomial models may well be adequate for the data.

By interpolating data from the current example we have been able to fit a quadratic term to the eleven age groups thus showing that the new algorithm can be used in this way.

CHAPTER 6

The relationship between the census and sample survey data

6.1 Introduction

In Chapter Three we developed the new algorithm for a generalized linear model approach to small area estimation. It was noted then and reiterated in Chapter Four that the response variables for the census data and the sample survey need not be exactly the same. In the example used throughout this thesis the two measures of unemployment are not the same but Haslett, Green and Zingel (1998) stated that they are highly correlated. Ambler, Caplan, Chambers, Kovacevic and Wang (2001) also state that the strong relationship between the two variables that they use is adequate to justify the use of the method, they even draw the scatter plots to demonstrate the strong relationship between the two data sources. We will consider this point in more detail in this chapter and suggest some general principles under which the new approach will give good estimates. These principles will equally apply to the classical use of the SPREE method as the new approach is identical to SPREE when a saturated model is used with categorical variables. We can also use these principles in situations where the new approach is applicable but SPREE is not.

An extension of the analysis can be used to generalize the results and investigate which parts of the model are driving the system.

It should be noted that the correlation here is a measure of linear association between the counts in equivalent cells in two tables of the same dimensions and not the more usual correlation between pairs of continuous variables.

6.2 *The relationship between the two data sources*

In describing the new algorithm in Chapter Three we wrote that "a generalized linear model will be fitted to census data and then some of the lower order parameters in the model will be adjusted in line with sample survey data". It was also noted that "we can partition the design matrix and the vector of parameters into two, the parameters that are estimated by the survey data and those that are not". Hence we wrote the two models as equations (3.5)**Error! Reference source not found.** and (3.6) **Error! Reference source not found.** which we reiterate here along with the implicit assumption in equation **Error! Reference source not found.**:

$g\left(\mathrm{E}\left[\mathbf{Y}_c\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c}$ for the census data

Error! Reference source not found.

and

$g\left(\mathrm{E}\left[\mathbf{Y}_s\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + X_2\boldsymbol{\beta}_{2s}$ for the survey data.

Error! Reference source not found.

However the survey data is not sufficiently detailed to estimate $\mathbf{X}_2\boldsymbol{\beta}_{2s}$ so in SPREE we assume that $\mathbf{X}_2\boldsymbol{\beta}_{2c} = \mathbf{X}_2\boldsymbol{\beta}_{2s}$

Error! Reference source not found.

Expressing the algorithm in this way allows us to investigate the correlation between the census and survey data. To simplify the notation initially we will define two univariate vectors $\mathbf{u}$ and $\mathbf{v}$ (each of length l) which can be written in the form $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ and $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$.

Then $\rho(\mathbf{u},\mathbf{v}) = \rho\left(\left(\mathbf{u}_1 + \mathbf{u}_2\right),\left(\mathbf{v}_1 + \mathbf{v}_2\right)\right)$ (6.1)

$$= \frac{E\left(\left(\mathbf{u}_1 + \mathbf{u}_2\right)\left(\mathbf{v}_1 + \mathbf{v}_2\right)\right) - E\left(\mathbf{u}_1 + \mathbf{u}_2\right)E\left(\mathbf{v}_1 + \mathbf{v}_2\right)}{\sqrt{Var\,(\mathbf{u})}\sqrt{Var\,(\mathbf{v})}}$$

=

$$\frac{E(\mathbf{u}_1\mathbf{v}_1) - E(\mathbf{u}_1)E(\mathbf{v}_1) + E(\mathbf{u}_1\mathbf{v}_2) - E(\mathbf{u}_1)E(\mathbf{v}_2) + E(\mathbf{u}_2\mathbf{v}_1) - E(\mathbf{u}_2)E(\mathbf{v}_1) + E(\mathbf{u}_2\mathbf{v}_2) - E(\mathbf{u}_2)E(\mathbf{v}_2)}{\sqrt{Var(\mathbf{u})}\sqrt{Var(\mathbf{v})}}$$

$$= \left(\frac{Var(\mathbf{u}_1)Var(\mathbf{v}_1)}{Var(\mathbf{u})Var(\mathbf{v})}\right)^{\frac{1}{2}} \rho(\mathbf{u}_1, \mathbf{v}_1) + \left(\frac{Var(\mathbf{u}_1)Var(\mathbf{v}_2)}{Var(\mathbf{u})Var(\mathbf{v})}\right)^{\frac{1}{2}} \rho(\mathbf{u}_1, \mathbf{v}_2)$$

$$+ \left(\frac{Var(\mathbf{u}_2)Var(\mathbf{v}_1)}{Var(\mathbf{u})Var(\mathbf{v})}\right)^{\frac{1}{2}} \rho(\mathbf{u}_2, \mathbf{v}_1) + \left(\frac{Var(\mathbf{u}_2)Var(\mathbf{v}_2)}{Var(\mathbf{u})Var(\mathbf{v})}\right)^{\frac{1}{2}} \rho(\mathbf{u}_2, \mathbf{v}_2)$$

$$= w_{11}\,\rho(\mathbf{u}_1, \mathbf{v}_1) + w_{12}\,\rho(\mathbf{u}_1, \mathbf{v}_2) + w_{21}\,\rho(\mathbf{u}_2, \mathbf{v}_1) + w_{22}\,\rho(\mathbf{u}_2, \mathbf{v}_2)$$

where $w_{ij} = \left(\frac{Var(\mathbf{u}_i)Var(\mathbf{v}_j)}{Var(\mathbf{u})Var(\mathbf{v})}\right)^{\frac{1}{2}}$.

It should be noted that as $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ and $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ Var(**u**) = Var(**u**1) + Var(**u**2) and similarly for **v**, provided (**u**1, **u**2) and (**v**1, **v**2) are both independent sets. Hence we can show that $\sum_{i=1}^{2}\sum_{j=1}^{2}(w_{ij})^2 = 1$

We will now define **C**, a centering matrix (about the mean), such that $\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J}_n$, **I** is an *n* by *n* identity matrix and $\mathbf{J}_n$ is an *n* by *n* matrix of ones. Since $(\mathbf{u} - \bar{\mathbf{u}})^T (\mathbf{u} - \bar{\mathbf{u}}) = \mathbf{u}^T\mathbf{C}\,\mathbf{u}$ and as *u* is a univariate vector quantity, we can write Var(**u**) as $\frac{1}{n}\mathbf{u}^T\mathbf{C}\,\mathbf{u}$

And $w_{ij} = \left(\frac{\left(\mathbf{u}_i{}^T\mathbf{C}\,\mathbf{u}_i\right)\left(\mathbf{v}_j{}^T\mathbf{C}\,\mathbf{v}_j\right)}{\left(\mathbf{u}^T\mathbf{C}\,\mathbf{u}\right)\left(\mathbf{v}^T\mathbf{C}\,\mathbf{v}\right)}\right)^{\frac{1}{2}}$

So for the census and survey data we can write:-

$$\rho(\mathbf{Y}_c, \mathbf{Y}_s) = w_{11}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_1\beta_{1s}) + w_{12}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_2\beta_{2s}) + w_{21}\rho(\mathbf{X}_2\beta_{2c}, \mathbf{X}_1\beta_{1s}) + w_{22}\rho(\mathbf{X}_2\beta_{2c}, \mathbf{X}_2\beta_{2s})$$

$$\text{with } w_{ij} = \left[\frac{\left[(\mathbf{X}_i\beta_{ic})^T \mathbf{C} (\mathbf{X}_i\beta_{ic})\right]\left[(\mathbf{X}_j\beta_{js})^T \mathbf{C} (\mathbf{X}_j\beta_{js})\right]}{\left[\mathbf{Y}_c^T \mathbf{C} \mathbf{Y}_c\right]\left[\mathbf{Y}_s^T \mathbf{C} \mathbf{Y}_s\right]} \right]^{\frac{1}{2}} \quad i = 1,2 \quad j = 1,2 \quad (6.2)$$

We can use equation (6.2) to give us insight into the relationship between the census and survey data.

However equation (6.2) is a relationship between the two sets of data, the GLM approach expresses a relationship between the expected values and the linear part of the model and also includes a link function.

To a first approximation we can consider the correlation between the transformed data from the two sources rather than the expected values. This is only true where the transformed value exists and for some of the distributions the transformation cannot be applied to reasonable data points. For example zero counts cannot be transformed by the logarithmic transformation. However, the effect of adding a small amount to those values so that the transformation can be applied will be small when the correlation is considered. The literature seems to suggest that the correlation under consideration has not been transformed at all. This will have an effect on the correlation however again with the typical link functions that are used the changes will not be great.

Let us begin by considering the categorical case. By selecting a suitable parameterization the design matrix, $\mathbf{X}$ in this case, can always be made orthogonal and so the two matrices formed by partitioning $\mathbf{X}$, $\mathbf{X}_1$ and $\mathbf{X}_2$, would also be orthogonal. An example of an orthogonal design matrix for the nine North Island regions is given in Appendix B. In most cases in this thesis a different parameterization has been used as it gives a more intuitive meaning to the coefficients, however it is possible to reparameterize the

design matrix so that it is orthogonal. If so it can be seen that $\rho(\mathbf{X}_1\boldsymbol{\beta}_{1c}, \mathbf{X}_2\boldsymbol{\beta}_{2s})$ and $\rho(\mathbf{X}_2\boldsymbol{\beta}_{2c}, \mathbf{X}_1\boldsymbol{\beta}_{1s})$ will be zero as they are uncorrelated.

Hence we can simplify the correlation to

$$\rho(\mathbf{Y}_c, \mathbf{Y}_s) = w_{11}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_1\beta_{1s}) + w_{22}\rho(\mathbf{X}_2\beta_{2c}, \mathbf{X}_2\beta_{2s}) \qquad (6.3)$$

given a suitable parameterization.

We have shown that equations (3.5)**Error! Reference source not found.**, (3.6)**Error! Reference source not found.** and (3.7)**Error! Reference source not found.** are the model that is required to hold for the method to work well. It can now be seen that if equation **Error! Reference source not found.** holds then $\rho(\mathbf{Y}_c, \mathbf{Y}_s) = w_{11}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_1\beta_{1s}) + w_{22}$ so high correlations between the two data sources may be due to the second term if w_{22} is large. However considering (6.3) we can draw a table relating how well the SPREE based methods will work with the correlation between the two data sources.

		Correlation between $\mathbf{Y}_c$ and $\mathbf{Y}_s$	
		High	Low
SPREE based Estimation	Works Well	$w_{22}\rho(\mathbf{X}_2\beta_{2c}, \mathbf{X}_2\beta_{2s})$ large. $w_{11}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_1\beta_{1s})$ unimportant.	$w_{22}\rho(\mathbf{X}_2\beta_{2c}, \mathbf{X}_2\beta_{2s})$ large compared to other term. $w_{11}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_1\beta_{1s})$ small.
	Works Poorly	$w_{22}\rho(\mathbf{X}_2\beta_{2c}, \mathbf{X}_2\beta_{2s})$ small. $w_{11}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_1\beta_{1s})$ large	$w_{22}\rho(\mathbf{X}_2\beta_{2c}, \mathbf{X}_2\beta_{2s})$ small. $w_{11}\rho(\mathbf{X}_1\beta_{1c}, \mathbf{X}_1\beta_{1s})$ large compared to other term.

Table 6.1 Relationship between "Correlation between $\mathbf{Y}_c$ and $\mathbf{Y}_s$" and probable success of SPREE based estimation

We can see that equation **Error! Reference source not found.** could be true even if $\mathbf{Y}_c$ and $\mathbf{Y}_s$ are not highly correlated. Of more concern even if $\mathbf{Y}_c$ and $\mathbf{Y}_s$ are highly correlated it is still possible for equation **Error! Reference source not found.** to not hold should $\mathbf{X}_1\boldsymbol{\beta}_{1c}$ and $\mathbf{X}_1\boldsymbol{\beta}_{1s}$ also be

highly correlated. In this situation although the two sets of data are highly correlated the effects of the second parts of the two models would be different and the estimates may not be good. It should also be noted that even if the $\mathbf{X}_2\boldsymbol{\beta}_{2c}$ and $\mathbf{X}_2\boldsymbol{\beta}_{2s}$ are highly correlated this does not imply equality and so the criterion needed to ensure good estimates is not necessarily met even if they are highly correlated.

The problem that we are considering is not limited to the correlation of the different parts of the two models. The weights w_{ij} are measures of the amount of the variability that is accounted for by each part of the model. If w_{22} is small, then it is of little importance whether $\mathbf{X}_2\boldsymbol{\beta}_{2c}$ and $\mathbf{X}_2\boldsymbol{\beta}_{2s}$ are highly correlated or not as they account for little of the overall correlation. However if it is large it becomes much more important that they are highly correlated.

Now we will consider the more usual situation in which the design matrix is not orthogonal. The correlation has four terms in it and we still require equation **Error! Reference source not found.** to hold. We can redraw Table 6.1 above and split the correlation into the last term and the other three. Much the same comments can be made for this as were made earlier. High correlations between the data in the two data sources does not guarantee that the actual assumption in equation **Error! Reference source not found.** holds and the assumption may hold even if the correlations are quite low.

Finally the concepts above can be extended to models in which there are continuous variables. The relationship between the two data sets are the same in this case and the partition of the design matrix can be carried out in the same way. The conclusions about the correlations between different parts of the models are then identical for continuous variables.

6.3 *Practical considerations in calculating the correlations.*

To pick up on a point made earlier, there is a linked problem that the two data sets are not necessarily easily tested for correlation. In the example that we have the census data consists of cell counts for the complete table but the sample survey data only gives the new margins. We can check the correlation of the counts in the two margins, by summing appropriate cells in the census data, but not more than that since information at the finer level is not available from the (secondary) survey data.

We are able to check the correlation of $\mathbf{X}_1\boldsymbol{\beta}_{1c}$ and $\mathbf{X}_1\boldsymbol{\beta}_{1s}$ by estimating each of them. $\mathbf{X}_1\boldsymbol{\beta}_{1c}$ can be found by fitting the saturated model to the census data and then using the coefficients $\boldsymbol{\beta}_{1c}$ along with the partition $\mathbf{X}_1$ to generate the counts in the table. $\mathbf{X}_1\boldsymbol{\beta}_{1s}$ can be found by using the standard method to produce an independent table from the new margins from the survey data.

If this is lower than the correlation between $\mathbf{Y}_c$ and $\mathbf{Y}_s$ then the correlation between $\mathbf{X}_2\boldsymbol{\beta}_{2c}$ and $\mathbf{X}_2\boldsymbol{\beta}_{2s}$ must be higher than the correlation between $\mathbf{Y}_c$ and $\mathbf{Y}_s$ and the algorithm should produce useful estimates if $\mathbf{Y}_c$ and $\mathbf{Y}_s$ are highly correlated.

The example that we have used has considered unemployment measured against two different standards, the ILO definition and the WINZ administrative definition. Alternatively a variable may have been measured in a census and later a survey is carried out to update the information. The argument will still apply. As the census data are from an earlier time period there are no guarantees that an identical relationship holds in the current time period. Using the generalized linear model approach we can see what parts of the model are assumed to remain constant for the two data sets irrespective of whether the same or a different variable is being measured. Use of the explicit model allows the possibility of seeing directly which effects and interactions

are being carried forward from the census data and allows more explicit justification of such action.

However checking the correlation between Y_c and Y_s is not simple either as is shown in the following example which uses a simple two by two table discussed in Chapter 2.

1	3	With the margins	1	3	4	Given a new:			5
5	2	added this gives	5	2	7	set of margins			8
			6	5			9	4	

We can find $\mathbf{X}_1\boldsymbol{\beta}_{1_c}$ and $\mathbf{X}_1\boldsymbol{\beta}_{1_s}$ by calculating the values for the independence model from the margins when we assume that the single interaction term is zero. This will result in the two tables shown below.

2.182	1.818	4	and	3.462	1.538	5
3.818	3.182	7		5.538	2.462	8
6	5			9	4	

We can thus fill out the table below, using the notation that we have used above.

$\mathbf{X}_1\boldsymbol{\beta}_{1c}$	$\mathbf{X}_2\boldsymbol{\beta}_{2c}$	$\mathbf{Y}_c$	$\mathbf{X}_1\boldsymbol{\beta}_{1s}$	$\mathbf{X}_2\boldsymbol{\beta}_{2s}$	$\mathbf{Y}_s$
2.182	-1.182	1	3.462		
1.818	1.182	3	1.538		
3.818	1.182	5	5.538		
3.182	-1.182	2	2.462		

However we are unable to fill in anything in the last two columns. In SPREE and in the new algorithm we assume that $\mathbf{X}_2\boldsymbol{\beta}_{2_s}$ is equal to $\mathbf{X}_2\boldsymbol{\beta}_{2_c}$ and with that information we can calculate our predicted values for $\mathbf{Y}_s$.

We can at least check whether $\rho\left(\mathbf{X}_1\boldsymbol{\beta}_{1_c}, \mathbf{X}_1\boldsymbol{\beta}_{1_s}\right)$ is small as all of the information required is available. In the example above the correlation is 0.759 which

appears to be quite high. The general conclusion is that model diagnostics are not always available when secondary sources, such as the WINZ data in our example, are used for small area estimation.

Of course the survey data are available and direct estimates of all cell values may exist. These are not acceptable estimates for the small areas as the sample sizes in many areas will be small, and in some cases zero, so the estimates are unacceptable considered in isolation. However it may be that they are useable for finding the correlations that we are interested in as we could reasonably expect those areas with zeros or small sample sizes to be areas in which the population was also small. If the sampling was clustered we could still use the estimates for those areas represented in the clustered sample to derive the correlations.

6.4 *A more detailed look at the correlations between parts of the model.*

We can further generalize the comments made above by considering other partitions of the design matrix. Although the method that we propose partitions the matrix into two parts this does not have to be the case. There could be a number of sources of data from which parameters are estimated and used in the final model. Even if there are still only two data sources it may be helpful to consider more partitions of the design matrix. If there are d coefficients in the model and hence d columns in the design matrix it can, at most, be partitioned into d (l by 1) column vectors, each of which is associated with a parameter in the model. In the models which we are considering $d = l$ as the models are saturated. By generalizing the approach above we can consider all of the pairwise correlations which may give us more insight into the relative importance of the individual terms in the model. Initially we could look only at the census data and identify those variables in the model which contribute most to the variability in the census. It would be best if the

coefficients for these variables were reestimated by the sample survey data. Then if we are able, the correlations between the same variables from the two data sources should be considered. Coefficients for variables, for which there is a high correlation, can be estimated from the census data but those for which the correlation is low should be estimated by the survey data. This is not a firm rule as has been shown, and often the choice of which parameters can be reestimated by the survey data is made before data collection begins as they are the variables or interactions for which sample sizes, in the various levels of the variable, are too small. However if the survey is part of an ongoing programme of data collection it may be possible to change the survey design or collect additional information if this would allow current uses of the survey to continue and improve the estimation of small area statistics.

Defining $\mathbf{X}_1$ as the vector formed by the first column of the design matrix, $\mathbf{X}_2$ as the second and so on until $\mathbf{X}_d$ and similarly partitioning the vector β of coefficients we can rewrite equations (3.5) and (3.6) in the form

$$g\left(\mathrm{E}\left[\mathbf{Y}_c\right]\right) = \mathbf{X}_1\beta_{1c} + \mathbf{X}_2\beta_{2c} + \ . \ . + \mathbf{X}_i\beta_{ic} \ . \ + \mathbf{X}_d\beta_{dc} \text{ for the census data}$$

and $$g\left(\mathrm{E}\left[\mathbf{Y}_s\right]\right) = \mathbf{X}_1\beta_{1s} + \mathbf{X}_2\beta_{2s} + \ . \ . + \mathbf{X}_i\beta_{is} \ . \ + \mathbf{X}_d\beta_{ds} \text{ for the survey data.}$$

Now we can consider the correlation between the census and survey data again.

$$\rho\left(\mathbf{Y}_c, \mathbf{Y}_s\right) \approx \rho\left(\mathbf{X}_1\beta_{1c} + \mathbf{X}_2\beta_{2c} + .. + \mathbf{X}_i\beta_{ic} .. + \mathbf{X}_d\beta_{dc}, \mathbf{X}_1\beta_{1s} + \mathbf{X}_2\beta_{2s} + .. + \mathbf{X}_i\beta_{is} . + \mathbf{X}_d\beta_{ds}\right)$$

The form of the expansion is exactly the same as in the earlier case and so the general term in the expansion can be written

$$\left[\frac{\left[\left(\mathbf{X}_i\beta_{ic}\right)^T \mathbf{C}\left(\mathbf{X}_i\beta_{ic}\right)\right]\left[\left(\mathbf{X}_j\beta_{js}\right)^T \mathbf{C}\left(\mathbf{X}_j\beta_{js}\right)\right]}{\left[\mathbf{Y}_c{}^T\mathbf{C}\ \mathbf{Y}_c\right]\left[\mathbf{Y}_s{}^T\mathbf{C}\ \mathbf{Y}_s\right]}\right]^{\frac{1}{2}} \rho\left(\mathbf{X}_i\beta_{ic}, \mathbf{X}_j\beta_{js}\right)$$

Hence the correlation can be written as $\rho(Y_c, Y_s) = \sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}\ \rho\left(\mathbf{X}_i\beta_{ic}, \mathbf{X}_j\beta_{js}\right)$

where the weights w_{ij} are as defined earlier and the β's denote individual coefficients rather than vectors.

Once again we can begin by considering the case when the design matrix can be written in an orthogonal form. The correlation simplifies considerably and the only terms that would appear in the correlation are those for which $i = j$. We then need to consider the parts of the model which will remain the same between the census and survey data and those that change. The terms which are highly correlated are the best to carry through from the census data and those that have a low correlation are best reestimated.

If the design matrix is not orthogonal then all of the pairs of correlations are included and with 54 terms in the model in Chapter Four there would be 2916 correlations to consider. This is clearly not going to be very easy to use. However by considering firstly the variables themselves and then their first order interactions and so on we can see which variables in the two models are highly correlated and which are not. It may be possible to ensure that those pairs of variables that are highly correlated are placed in the second part of the partition and those that are not are included in the first. Being realistic this is unlikely to be possible as we do not have sufficient data and because the variables in each part of the partition is controlled by the available margins not the whim of the analyst.

Finally as we have already stated even if terms are highly correlated it does not mean that they are equal, which is the assumption that is required.

6.5 Transformations of variables and the effect on correlations.

A final point which needs to be mentioned is that the correlations discussed generally use the cell counts from the census and survey as data. The correlations which we have discussed in this chapter have all been on the log

transformed scale. In Chapter Three when we introduced the generalized linear model we used a notation which meant that the data, **Y**, was on the transformed scale. In the literature, when correlations are discussed as a justification for the use of SPREE estimation, it is implied that the correlation is between the variables on the untransformed scale. The correlations are not usually affected by a log transformation, if there are no extreme points, and it should not cause too many problems.

When applying a log transformation to data a zero cell count would intuitively be a cause for concern as the log of 0 is undefined. In large contingency tables a few zero cell counts do not in fact cause a problem in modeling the table Feinberg (1980). However these could be a concern when investigating correlations and need to be considered. If the correlations are on the count scale then they are again of no concern, if they are on the transformed scale an adjustment for those counts will be required.

Whilst this may not be a major concern we should acknowledge that this situation exists and note that in the case where a few small areas may be much larger in their counts than others, those areas may affect the correlation unduly.

6.6 Suggestions for model checking based on this.

We have shown in this chapter that correlation between the two data sources is not the necessary requirement for SPREE based estimation to be appropriate. We have also noted that the data is not available to check the correlations in the different parts of the model which would give some support for the use of the method. This does not give any ideas as to a better method to decide whether the approach is sensible in a given situation. However, there are some insights that can be gained from this chapter which may help in the design of the data collection which can then aid some checking.

If the survey is ongoing then the same design does not need to be used on every occasion and it may make sense to check different parameters in subsequent surveys or to use different amounts of resource on different occasions so that more model checking may be carried out. This introduces some difficulty in practice and the advantages and disadvantages need to be balanced carefully. Survey data is now being collected for which model based small area estimation is a known objective, more work needs to be done on survey design when this is one of the objectives.

In the example which we have used throughout this thesis there is an important opportunity to do some sophisticated model checking. In New Zealand every five years a full census is carried out. Recent censuses have included questions on employment status. This data must create an opportunity to check the estimates from other model based methods using data from the same period. It would also be possible to consider some of the ideas that have been developed in this chapter with respect to the census, WINZ and HLFS data.

6.7 Concluding remarks

Haslett et al (1998) commented that as the two variables for unemployment that they were considering were highly correlated they could be used in a SPREE approach. This assertion seems reasonable and has been used by researchers in the field to justify the method. Ambler, et al (2001) talk about a strong relationship between the variables" and show scatter plots to support this. Longford (2002), in discussing auxiliary information, defines auxiliary information as information that is closely related to "strongly correlated with" the variable of interest. It should be noted that Longford is discussing a range of models for small area estimation for which correlation would be a useful measure of the appropriateness of the auxiliary information but as we have

shown in this chapter for some models correlation may not be a good criterion.

The use of the new algorithm, and the notation introduced, make the conditions required, for the method to work well, more transparent. It has been shown that simple correlation between the variable of interest and some other variable is not necessarily a good criterion to use to assess the auxiliary data being used. It is possible both for the data sources to be highly correlated and for the method to give poor results and for the data sources to have low correlation but the method to still work well. The relationship is more complicated than was suggested and that it is not possible to confirm, or otherwise, the applicability of the method simply from the data available.

We have also shown that the statement that the two variables are correlated cannot be confirmed from the data directly. This suggests that any investigation of variables that may be used for census data needs to be more rigorous than simply considering the correlation.

There may well be auxiliary data available that allows the statement to be checked or the relationship between the variables may be able to be investigated in some other way. For the example that we have considered the variables are ILO unemployment and unemployment as defined by WINZ. It is not hard to confirm that the majority of those unemployed qualify under both definitions and hence we would expect the counts to be highly correlated and approximately in some fixed ratio. It may be worth investigating those individuals who do not qualify under both definitions in particular to ensure that they are not correlated with the regional variables. Currently this data is not available from SNZ, because of privacy legislation in the Statistics Act, and so the checks have not been made. For SNZ and WINZ this may be less of a problem.

All this does not seem to be particularly useful. We have a simple method to check whether the method is likely to give good results. However the data that is available is inadequate to apply it. We now have clearer ideas about what is important if the method is to work well. Particularly in an ongoing survey process such as the HLFS that SNZ carries out the sample may be selected differently on different occasions so that particular aspects of the model may be checked.

As a final comment we should acknowledge that it is very likely that some of the researchers mentioned in this chapter may not have been using the term correlation under quite the strict statistical sense that we have. It may be that the implication that the structure of the dependence of the two response variables (one from the census and one from the survey) on the explanatory variables may be similar and that this is inferred by common sense. It would seem sensible to assume that the way WINZ data is affected by region and interactions between regions and other variables would be similar to the SNZ data. Hence SPREE and our new approach should work in this situation. There is however, no statistical evidence to support this proposition.

CHAPTER 7

Bayesian approaches to parameter estimation

7.1 Introduction

As explained in Chapter Four the WINZ data which we are using as the census data are actually collected at monthly intervals and there are a number of sets of census data available for each analysis. We therefore have information not only for the mean value for each cell in the table for the census but also on the variability for each of these cells. The number of sets of census data that should be used to estimate the variability is a balance between using the most up to date data and using sufficient censuses to ensure accurate estimation of the census variability. Should there be a trend or any cyclic variation these would increase the apparent variability. We have chosen to use the previous eight months of census data, because this was all that was available to us. This is an arbitrary decision and more careful consideration of the WINZ data would make sense in a real analysis. In particular, having several years of monthly data, once they are available, would allow some assessment of seasonality.

The sample survey that SNZ carries out for the HLFS is designed in 512 balanced repeated replicates so 512 independent estimates of the mean are available for the new margins and so estimates of the variance can be made from these.

As we have information about the mean and variance for both sets of data it would make sense to formally use this information and so we have chosen to use a Bayesian analysis using MCMC for most of the analyses from this point onwards in the thesis. We do not intend to give a detailed explanation of

Bayesian data analysis nor to try to justify the use of Bayesian methods in general but we feel that this approach is appropriate for these data because, as mentioned above, we have very good estimates of priors for both the mean and variability. This will be considered in more detail later in the chapter.

In this chapter we will briefly introduce the Bayesian approach to statistics and show why it is appropriate in this context. We will then demonstrate that these techniques will result in the same point estimates for the examples that we have already considered under assumptions which seem reasonable. Finally we will show how they can give more complete estimates of the true error in the estimated counts.

In the wider context of small area estimation it has usually been assumed that the census data is correct and has no error associated with it even when the small area estimation is being used to estimate census undercount. If for example we acknowledge that there is often census undercount then it becomes hard to justify the stance that the census values are correct and without error. We have already noted that the census data that we have is collected regularly and that seasonal patterns are highly likely so variability in the census data would be reasonable. Whether in individual cases we are able to quantify the variability in the census data is another problem but with the data that we are considering there is no difficulty, providing the monthly censuses available are essentially measuring the same thing. Further for the WINZ data there is not appreciable error in the sense that the data indicate the recipients of the unemployment benefits.

It is also worth noting that the method proposed can be used, as could SPREE, when we do not have a complete census but survey data that is sufficiently detailed to estimate the coefficients in the model and then a further, less detailed, survey which can be used to update some of those coefficients. In this case there would clearly be variability inherent in the "census" data and

this should be acknowledged in the analysis. SPREE is able to be used to include the variability in the census and the sample survey in this data set as we can simply repeat the IPF for each combination of census data and survey replication. The estimates so produced can then be used to find the mean and variance. A rather more computationally efficient modification of this is effectively the approach taken by Haslett et al (1998).

It is a concern that most small area estimation techniques do not take this census variability into account at all and so the small area estimates may be estimated to be more accurate than is really true.

7.2 Frequentist and Bayesian statistics

There has been considerable discussion over the latter half of the last century between the traditional, frequentist, statisticians and those who follow a Bayesian philosophy. Below is a very brief introduction to the two approaches, there are many books which give a complete exposition for example Lindley (1985) and Bernardo and Smith (1994). Many of these books have been written by Bayesian exponents who tend to expound the advantages of their philosophy and illustrate the difficulties of the frequentist ideas. The debate became quite heated for a time but there appears, now, to be a broader acceptance of the Bayesian method particularly as advances in computer software and computer power have made their application to real problems possible.

The frequentist evaluates procedures based on repeated sampling, imagining an infinite replication of the same inferential problem and evaluating properties over this repeated sampling framework for fixed values of unknown parameters. Good procedures perform well over a broad range of parameter values in this imagined sampling experiment. In reality of course

only one sample is generally taken and this is analysed as if it was one of many.

The Bayesian requires a sampling model and, in addition, a joint prior distribution for the parameters. Unknown parameters are considered random and all inferences are based on their distribution conditional on observed data (the posterior distribution). The Bayesian evaluates the posterior distribution for a given data set and hence estimates the parameters of interest.

There has been much written about the appropriateness of the two schools of thought. The frequentist approach gained much popularity in the first half of the twentieth century and became the established approach to statistical analysis, The Bayesian approach, though it has roots in Bayes theorem from the eighteenth century, has been developed in detail more recently and therefore had to be shown to be an improvement over other methods. A lot of effort and time was spent in discrediting frequentist theories which was not necessarily the best way to win the hearts and minds of the majority of statisticians. The situation now is somewhat less combative and many practicing statisticians will accept the two paradigms and choose an appropriate method for each particular situation. Much of the criticism of the Bayesian approach is aimed at a number of points.

- It may be hard to accept that the parameters to be estimated should be considered as random variables.
- There is a reluctance to formally include subjectively chosen priors.
- It is often difficult to perform the analysis as the calculations generally include the evaluation of difficult integrals.

On the other hand the Bayesians would retort

- The concept of a fixed population makes little sense in most real situations and so parameters related to the populations should rightly be considered as random variables.
- Almost all experiments and surveys use subjective information. It would be best to formally acknowledge this.
- The interpretation of hypothesis tests and confidence intervals is much more sensible in a Bayesian analysis.

We will sidestep these issues by showing that for the data that we have there are good reasons for accepting the Bayesian paradigm. As far as the priors are concerned we will show that our data has well defined and easily justified prior distributions which we can use. The concept of a population parameter as a random variable and not a fixed value, in our context, is sensible. Unemployment, whether measured on the International Labour Organisation basis or via administrative records in a governmental department, is difficult to define and there are always individuals who do not fit the definition easily. Thus the actual count of unemployed individuals will depend on subjective decisions made by individuals at a particular time and a random variable may describe this best. There may also be benefit in modeling change in values over time but for the purposes of this thesis we will choose not to as the example data is only available for eight months.

The data that we are using has two features which are somewhat unusual in the small area estimation context.

- Firstly, and most significantly, there exist a number of sets of census data for the time period of the survey each of which results in different parameter. Also the survey is designed so that estimates of variance can be found, in this case it is collected in balanced repeated replicates. Thus the population parameters certainly vary

and we can use the data to suggest appropriate distributions for these parameters.

- Secondly we can estimate parameters for the prior distributions directly from past data.

The model that will be used is still that in equations (3.5) **Error! Reference source not found.** and (3.6) **Error! Reference source not found.** along with the assumption in equation (3.7)**Error! Reference source not found..** Hence we write $g\left(\mathrm{E}\left[\mathbf{Y}_c\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c}$ for the census data and $g\left(\mathrm{E}\left[\mathbf{Y}_s\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + X_2\boldsymbol{\beta}_{2s}$ for the survey, as before. The WINZ data forms $\mathbf{Y}_c$ and we will use eight separate sets of census data. The SNZ survey data provides margins for $\mathbf{Y}_s$ and the survey is formed in balanced replicates hence we have the 512 estimates for these margins.

From a purely pragmatic viewpoint the analysis will be carried out using the WinBUGS software (v 1.3 April 2000). In Chapter Eight we will investigate fitting conditional autoregressive (CAR) error structures in our model and WinBUGS has a built in function for estimating the CAR model which we will consider then. A definition of CAR models is left till Chapter Eight. MLwiN is able to do the estimation using the Bayesian paradigm but partly because of the difficulties that we discussed in section 4.4 we have not used it for this part of the work. MLwiN does not, at present, offer the CAR models. WinBUGS is specifically designed for analysis of data in a Bayesian setting.

Finally we will see that WinBUGS allows a somewhat neater way of modelling the survey data as we do not need to predict the values in the table assuming independence and generate the new data: we can simply tell the program what the new margins are and how the cell values generate these margins. The program then takes care of the rest.

We have discussed the basic differences between the Bayesian and frequentist approaches to statistical analysis and have suggested reasons for using the Bayesian paradigm for the data that we have.

7.3 Bayesian solutions, computing approaches

One of the problems for Bayesian approaches until recently has been the difficulty in applying them in real situations. Finding analytic solutions to most problems is difficult. With the advent of more powerful computing at an affordable price computer software has been developed to handle Bayesian approaches to statistical analysis using simulation based methods most notably Markov Chain Monte Carlo (MCMC) methods. MCMC methods are not restricted to Bayesian statistics and can equally be applied to frequentist analyses. They are based on simulation methods in which a Markov chain converges to a stationary distribution and this stationary distribution is sampled so that sample statistics can be calculated. Because the sample is drawn from the distribution of that parameter, interval estimates can be found simultaneously. The common sampling algorithms such as Gibbs and Metropolis-Hastings, do not give independent samples but are correlated. This effect can be reduced by systematically sampling from the simulations. If every tenth or twentieth simulation is saved then the autocorrelation is reduced as the autocorrelation may not extend to these lags. Sometimes much larger lags are required. This is not generally a problem as the cost is in the computing time to generate the additional simulations and this is not very high. This process of systematically selecting some of the sample is termed thinning. This autocorrelation will not affect the value of the estimate but will cause underestimation of the variance. Much has been written on MCMC methods see Besag (2001) or Gilks, Richardson and Spiegelhalter (1996) for more detail and more references.

One advantage of choosing the Bayesian approach using MCMC methods is that it is trivial to gain estimates of functions of the parameters in the model, hence we can find estimates for the small area counts if they are sums of cells in the table and the algorithm will yield accurate estimates of the variance of these estimates with little additional effort.

As mentioned in the last section we have used one such package WinBUGS for much of the analysis from this point on.

We shall first show that we can repeat the analysis that we have already done both with the purely categorical variables and then with the quadratic model for age. It will be shown that there is no need to have one program to find the parameters β_2 for the census data and then transfer them to the program for the survey data as it can all be done in one program. Initially we will reproduce the analysis of Chapter Three except we have used data for all sixteen Regional Authorities and the data is also subdivided into two ethnic groups. This allows a more useful analysis to be carried out and as we are no longer using MLwiN the restriction on the number of columns used in the worksheet is not relevant. The design matrix is now 192 by 192 and is constructed in the same way as for the earlier example. However it is not given anywhere in this thesis as it takes up a large amount of paper and is unlikely that anyone would want to look at it.

The model used is the one described in equations (3.5) **Error! Reference source not found.**, (3.6)**Error! Reference source not found.** and (3.7)**Error! Reference source not found.**. The algorithm for this first example estimates the model for the census data, (3.5) **Error! Reference source not found.** and the coefficients are recorded and presented as data in the model for the survey data (3.6) **Error! Reference source not found.**. In a Bayesian model the definition of the priors is required as well as the relationship between the data and the

parameters of interest. For the first part of the analysis, finding the coefficients for the saturated model for the census data the model is

$$Ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{Y}_c \sim N(\boldsymbol{\mu}, \boldsymbol{\tau})$$

$$\beta_i \sim N(\text{beta}_i, \text{taubeta}_i)$$

beta_i and taubeta_i are subjectively estimated from prior information. We need to make decisions about the values for $\boldsymbol{\tau}$ and the taubeta_i. These are the estimates of the variance (Note WinBUGS actually uses the precision, the inverse of the variance, rather than the variance) of the data and the coefficients. A detailed discussion of the values used for these is given in section 7.5. The programs below use somewhat arbitrary values.

The program below has been set out in five sections to illustrate the general form of a Bayesian analysis in WinBUGS. The three sections in the "model" part are each within a loop beginning "for(i in 1 : N)". The first part defines the data and a distribution, along with definitions of the parameters for that distribution. In this case it is proposed that each cell value is given by the appropriate value in the data vector $\mathbf{Y}$ and that the cells are normally distributed with a mean and the precision is the reciprocal of twice the mean value. This is chosen arbitrarily at the moment and a more detailed discussion of the prior distributions is included later. WinBUGS uses the precision, the inverse of the variance to define the normal distribution. Small precisions indicate large variance hence a diffuse prior.

Next we have some statements which define the model that is being used to link the data to the unknown coefficients. In this case it is simply stating that $Ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$. WinBUGS does not do matrix multiplication by default instead it multiplies the individual values and the summation has to be explicitly carried out. Next the distribution for the parameters in the linear part of the

model is given. In this case little is known and so as is common we use a very diffuse normal distribution with mean zero.

The final parts are the data and the initial estimates of the unknown values. The data are fairly obvious and as we have no idea about the unknown values initial values of 0 are sensible. We can have more than one set of initial values. It is good practice to have at least a second set that are quite different from the first so that, assuming they both converge to the same estimates, it is reasonable to say that the initial values do not bias the estimates.

```
model;  {

#Part 1 Prior distribution for data
for( i in 1:N ) { prec[i] <- 1/(2*mu[i])  Y[i] ~ dnorm(mu[i] , prec[i]) }


#Part 2 model
for( i in 1 : N ) { for (j in 1:N) {    mew[i,j] <- X[ i, j]*beta[j]        }
            log(mu[i]) <- sum(mew[i, ])          }


#Part 3 Prior distribution for the unknown parameters
for (i in 1:N) {  beta[i] ~ dnorm(0, 0.00000001)  }
          }

#Part 4 the data
  list(N = 192,
X = structure(.Data= c(
Large design matrix edited out of the
program at this point, 36864 1's 0's or -1's
comma separated.
),
.Dim = c(192,192)),
 Y=c(
192 data values edited out. A count for
each Region by sex by age by ethnicity.
), )
#Part 5 The initial values for the coefficients beta.
  list( beta= c(
192 0's edited out
```

The prior distribution for the cell counts is normally distributed with a mean mu and variance twice the mean.

The model is (3.5) $g\left(\mathrm{E}\left[\mathbf{Y}_c\right]\right) = \mathbf{X}\boldsymbol{\beta}_c$ where the link function is the logarithm.

The coefficients are given relatively uninformative priors

The data part of the program includes both the X, design matrix, and the census cell values for the table

The program also requires initial values

))	for each coefficient to be estimated. These are set at 0.

This program will fit the loglinear model to the census data and estimate the coefficients as we have done in MlWin or S-Plus previously. We can then manually transfer the coefficients which remain constant, β_{2c}, to the program to reestimate the coefficients for the new margins from the survey data as in the program below.

The prior distribution for the data is derived from the census data. For this illustrative program the proposition is that the data are normally distributed with a variance which is set to 100, the precision is the inverse of the variance hence 0.01. In fact there are 8 sets of census data and two approaches could be applied. Either we could suggest that the most recent is the best to predict conditions now and the previous seven can be used to estimate the variability. Or the distribution can be assessed from all eight sets of data and hence a mean and variance calculated. The latter approach is not generally recommended but we will discuss the data that we have and how it can best be used in section 7.5 and consider in more detail the appropriate priors to use.

The second part of the program then takes the estimates for the model from the first part and carries them through to the model for the survey data and reestimates those coefficients which will change, as before. The model in this case is

$$Ln(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$$

$$\mathbf{Y}_S \sim N(\boldsymbol{\mu}, \boldsymbol{\tau})$$

$$\beta_{1i} \sim N(\text{beta}_{1i}, \text{taubeta}_{1i})$$

$\beta_{2i} \sim N(\text{beta}_{2i}, \text{taubeta}_{2i})$ These are estimated in the first program

τ and taubeta$_{1i}$ are estimated as in the first program.

Note that in this program there is no need to calculate the values in the contingency table assuming independence. We can simply give a prior distribution for the margins and then define the margins in terms of the appropriate sums of the individual cells in the table. To simplify this process the sum for each Sex by Age by Ethnicity combination is found and then these are summed to give the margins.

```
model; {
#Part 1
#Statements about the survey data, marg.
for( i in 1 : 7 ) {     marg[i] ~ dnorm(margins[i], 0.01)        }


for ( i in 1:12) {    SAE[i] <- sum(mu[(i*16)-15 : i*16]) }
                margins[1] <- sum(SAE[1:6])
                margins[2] <- sum(SAE[7:12])
for (i in 3:5) { margins[i] <- SAE[i-2] + SAE[i+1] + SAE[i+4] + SAE[i+7] }
                margins[6] <- sum(SAE[1:3]) + sum(SAE[7:9])
                margins[7] <- sum(SAE[4:6]) + sum(SAE[10:12])



#Part 2 This includes calculation of the offset and its inclusion in the
model
for (i in 1:N) {   for (j in 1:187){     offs[i,j] <- X[i, j+5]*betab[ j ]   }     }
for (i in 1:N) {   for (j in 1:5)   {     mewa[i,j] <- X[i , j]*betaa[j]     }     }
for (i in 1:N) {   mu[ i ] <- exp(sum(mewa[i , 1:5]) + sum(offs[i , 1:187])) }



# Part 3
for (i in 1:5) {   betaa[i] ~ dnorm( 0.0,1.0E-8)          }
          }
          #Part 4
            list(N = 192,
          X = structure(.Data= c( Large design matrix edited out
of the program at this point, 36864 1's 0's or -1's comma
separated. ),
.Dim = c(192,192)),
marg=c(
62125, 47118,
```

The prior is defined for the survey data which is some margins of the table

The relationship between the cell values and the margins is defined. The SAE variable is the sum of each Sex by Age by Ethnicity class across the regions. This is only found for convenience.

The log linear model is defined as before, this time including an offset for the coefficients that are being carried forward from the census model. These are betab[] in the program

Prior for the new coefficients to be estimated from the survey data.

```
44236, 52176, 12831,
75752, 33491   ),
betab=c(  187 values for the coefficients transferred from
the earlier results   )
)
#Part 5 Two sets of initial values included in this program to check
convergence.
list(  beta = c( 5, .1, .5, .5, .4)  )
list(  beta = c( 3, 0, 0, 0,0.1)   )
```

Data as in the previous program.

Two sets of initial values to check convergence.

The values given for betab[i], β_{2c} in the model, in this program are those values for the coefficients in the model which have been carried forward from the first program acting on the assumption in equation **Error! Reference source not found.**.

The full results from these two programs can be found in Appendix D. The results of a Bayesian analysis using MCMC simulations is a vector of values from each simulation. In a program with a number of variables and a large number of iterations these vectors become large and require computer memory to be stored. Storing all values slows the computer down so WinBUGS saves only the values for variables that are selected to be monitored. WinBUGS can then report a number of results and other information for checking convergence of the estimates and other diagnostics. Some of these are shown below for the coefficients that are re-estimated from the sample survey data. These were called betaa[] in the program and are β_{1s} in equation (3.6)**Error! Reference source not found.**. There are only five coefficients which makes the output manageable. The first is the coefficient for the overall mean the second is the difference in the means for the sex effect and the others are the age effects and ethnicity.

For the estimates of the individual cell values there are 192 values and the output is excessive for this document.

First we can see the estimates of the centre of the distribution. These are calculated from all of the stored values of each variable.

The mean and median are given along with standard deviation of the estimate and an MC error which is calculated as the standard deviation of means of batches of estimates see p50 Gilks et al (1996) for more details. The MC error should become very small for large samples if the estimates have converged. 2.5 and 97.5 percentiles are also given, these estimate a Bayesian 95% credible interval which a frequentist would probably loosely interpret as a confidence interval. It may not be symmetric.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
betaa[1]	5.502	0.00433	2.375E-4	5.493	5.502	5.509	4001	6000
betaa[2]	0.3124	0.004483	1.421E-4	0.3039	0.3123	0.3214	4001	6000
betaa[3]	0.3588	0.006734	2.968E-4	0.3457	0.3587	0.3718	4001	6000
betaa[4]	0.5592	0.006373	2.578E-4	0.547	0.5592	0.5719	4001	6000
betaa[5]	0.5491	0.004753	1.893E-4	0.5397	0.5492	0.5583	4001	6000

Table 7.1 Output from WinBUGS program.

If these estimates are to have any credibility it is important that the chain has converged, to assess this we can look at a graph of the estimates in order. These are shown below and after a short initial period getting to a stable value they deviate very little. We would assume that these have converged, the estimates derived from the chain may be the mean, median or mode. It will not matter which is used if the distribution is unimodal and symmetric. In this work we expect most of the estimates to be normally distributed and so all three should be the same.

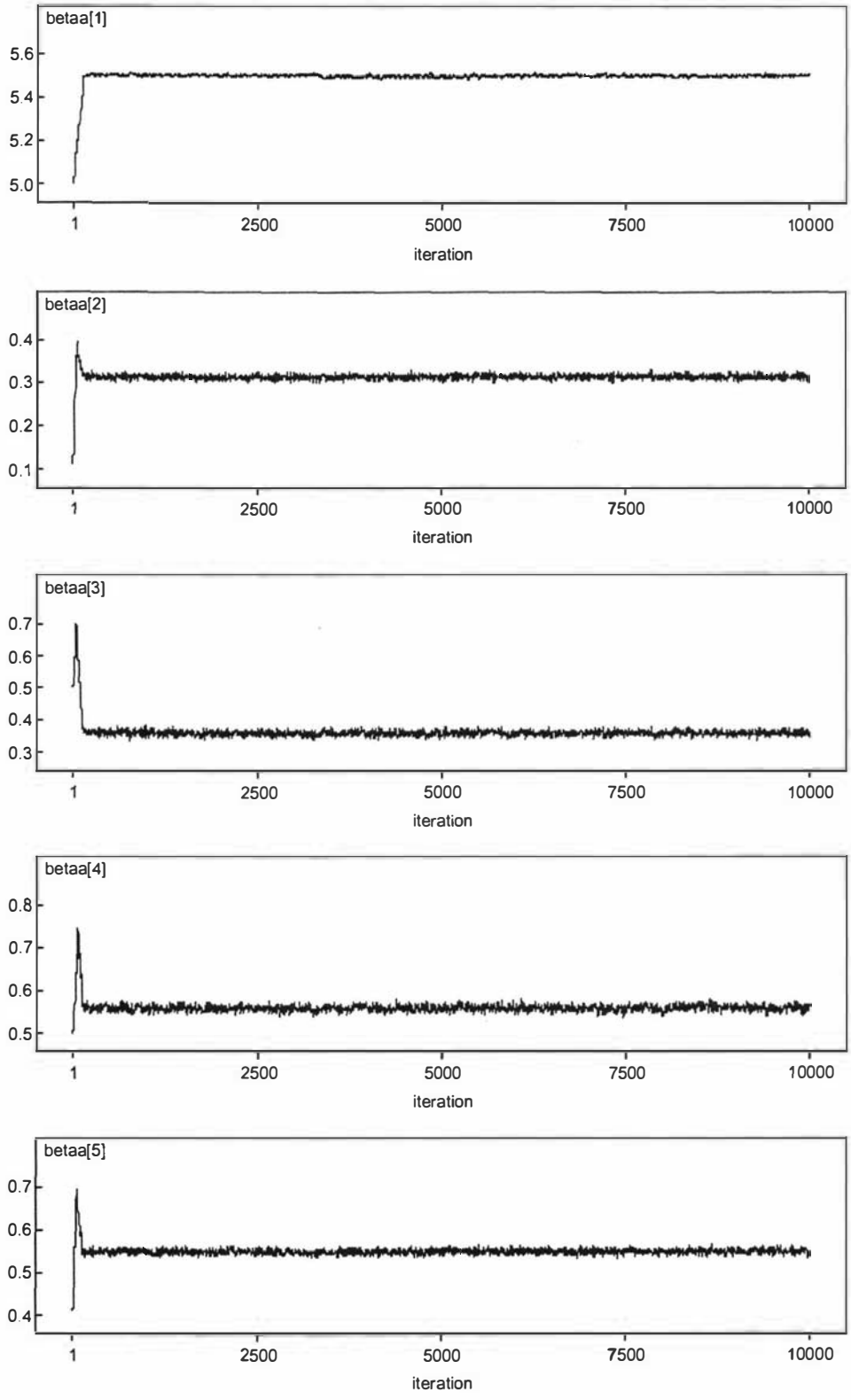


Figure 7.1 Graphs showing the convergence of samples from a BUGS program

Below are also density graphs of the samples and the autocorrelation functions. These can also be used for diagnostic purposes and these show a well behaved chain with relatively little autocorrelation. A less well behaved chain is shown in Appendix D.

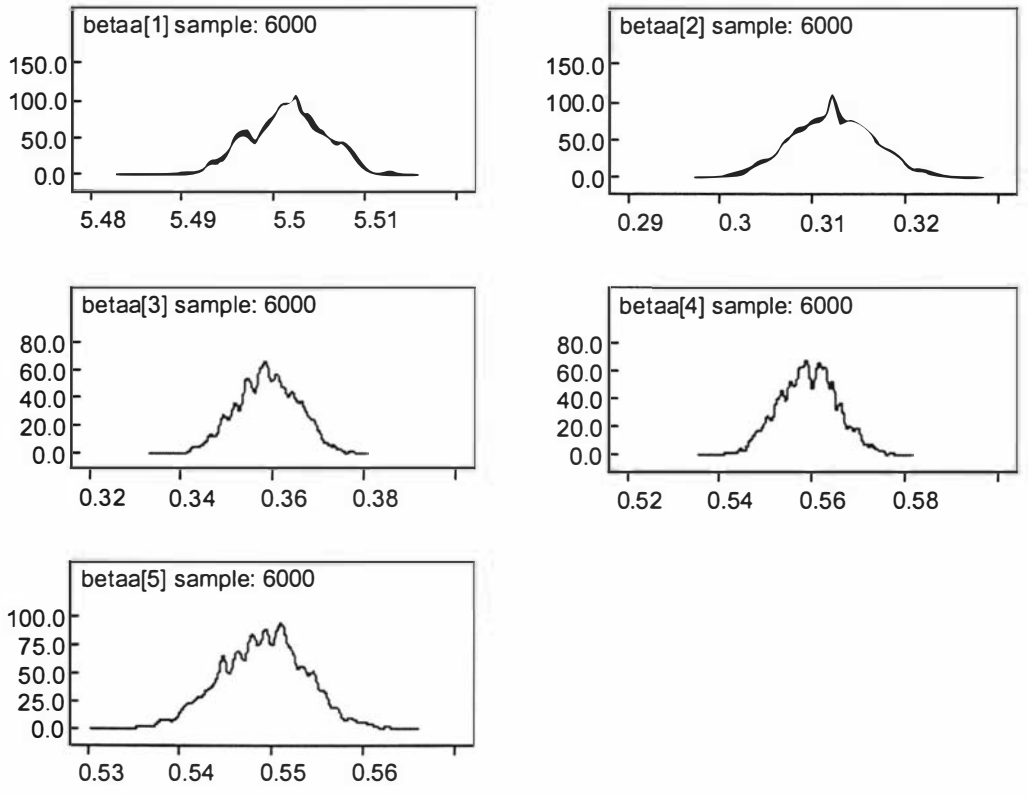


Figure 7.2 Density curves for the five coefficients

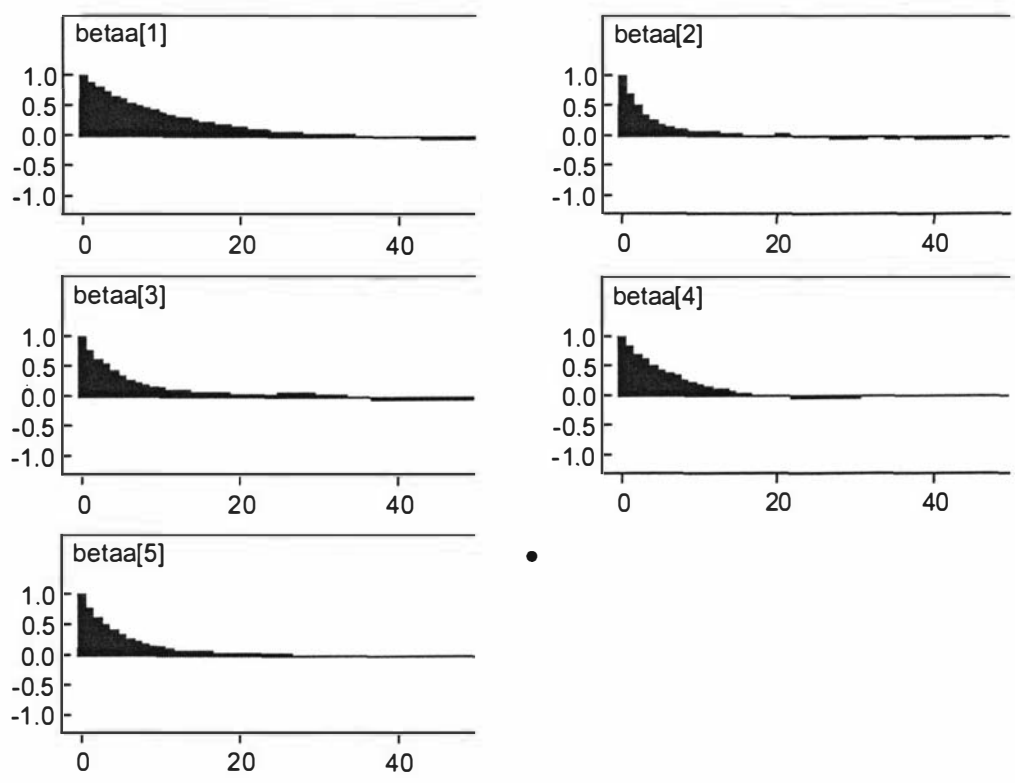


Figure 7.3 Autocorrelation plots for the iterations for the five coefficients

There was no thinning of the simulations in this output and so there is some autocorrelation shown in the graphs above. It does not extend much beyond lag 20 in any of the graphs and by letting the chain run for longer and thinning the results this could quite easily be reduced.

The two programs described above mimic the analysis performed in MLwiN in Chapter Four. However the MCMC approach allows us to combine the two programs into one by getting the program to estimate the coefficients for the census data and for the survey data, at each iteration. This program is given below.

The model now is a combination of the previous two

$$Ln(\boldsymbol{\mu}_c) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c}$$

$$Ln(\boldsymbol{\mu}_s) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + \mathbf{X}_1\boldsymbol{\beta}_{1s}$$

$$\mathbf{Y}_c \sim N(\boldsymbol{\mu}_c, \boldsymbol{\tau}_c)$$

$$\mathbf{Y}_s \sim N(\boldsymbol{\mu}_s, \boldsymbol{\tau}_s)$$

$$\boldsymbol{\beta}_{1c} \sim N(\text{beta}_{1c}, \text{taubeta}_{1c})$$

$$\boldsymbol{\beta}_{2c} \sim N(\text{beta}_{2c}, \text{taubeta}_{2c})$$

$$\boldsymbol{\beta}_{1s} \sim N(\text{beta}_{1s}, \text{taubeta}_{1s})$$

```
model;  {
for( i in 1 : N )  Y[i] ~ dnorm(muc[i] , precc[i])              }
for( i in 1 : N ) {       for (j in 1:N) {   mewc[i,j] <- X[ i, j]*beta[j] }
                                          log(muc[i]) <- sum(mewc[i, ])  }
for (i in 1:N) {  beta[i] ~ dnorm(0.0, 0.000001) }
  for( i in 1 : 7 )  marg[i] ~ dnorm(margins[i] , 0.00000001 )     }

                margins[1] <- sum(SAE[1:6])
                margins[2] <- sum(SAE[7:12])
for (i in 3:5) {  margins[i] <- SAE[i-2] + SAE[i+1] + SAE[i+4] + SAE[i+7] }
                margins[6] <- sum(SAE[1:3]) + sum(SAE[7:9])
                margins[7] <- sum(SAE[4:6]) + sum(SAE[10:12])

for (i in 1:12)  { SAE[i] <- sum(mu[(i*16)-15 : i*16]) }
for (i in 1:16)  { region[i] <-
mu[i]+mu[i+16]+mu[i+32]+mu[i+48]+mu[i+64]+mu[i+80]+mu[i+96]+mu[i+
112]+mu[i+128]+mu[i+144]
+mu[i+160]+mu[i+176] }
# for (i in 1:N) {   for (j in 1:187){     offs[i, j] <- X[i, j+5]*beta[ j+5 ]  }   }

for (i in 1:N) {   for (j in 1:5) {     mewa[i,j]  <- X[i , j]*betaa[j]   }      }
for (i in 1:N){  Xbeta[i] <- sum(mewa[i , 1:5])   offset[i] <-  sum(offs[i ,
1:187])   log(mu[ i ]) <- Xbeta[i] + offset[i]   }
betaa[1] ~ dnorm( 5.0, 0.1)
for (I in 2:5) {   betaa[i] ~ dnorm( 0.0, 0.1)   }

}
#data
 list(N = 192,      #Again the design matrix is
removed.
```

This program is a combination of the two previous ones. There is little difference in the program statements but functionally there are some important differences which are explained below.

```
Y=c(  Data values for the original table removed.
),
precc=c(    values edited out     ),
marg=c( 62777, 45501, 41652, 54583, 12042, 74725, 33553 )
)
list(   betaa= c( 5, 0.1, 0.01, 0.001, 0.4)   ,
beta = c(
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0),
 b = c(
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0) )
```

There is a significant difference in this last program. The coefficients that are carried forward from the model for the census data are no longer constant values but they are simulated, at each iteration of the program and hence the distribution of the final estimates includes variability not only from the new margins but also from the original census data through these varying coefficients. The variance of these final estimates should be closer to reflecting the true situation assuming that the model is appropriate.

The simple model proposed above still only uses categorical variables. We could also repeat the analysis using a quadratic function as in Chapter Five or using the five yearly age groups also discussed in Chapter Five. These require small changes to the program above but the main steps are the same. The design matrix **X** needs to be changed in the same way as in Chapters Five. This will be considered in more detail in section 7.6.

7.4 The data

In a true Bayesian analysis it would be discouraged to use the data itself to estimate priors. The priors are estimates of the state of knowledge before the data is collected and they should be adjusted by the data when it is known. If both the priors and data are found from the data collection process then the analysis will be biased as greater weight is given to these particular data.

For the data that we have we can use one of the replicates for the data and the rest to estimate the prior. For the WINZ data it is entirely reasonable to use the most recent data set for the new margin and some previous data sets to estimate the variability for the priors. With the sample survey data it is less clear as the data is collected in 512 balanced replicates but we could envisage one of these being the data and the rest being used to estimate the variability. The alternative approach would be to use the mean and variance of all replicates for data and for the prior. This approach is not recommended in Bayesian analyses as the same data is being used twice and the assumption is made that the priors and the data are independent. The effect will be to downward bias estimates of variability. In the case of the 512 replicates for the sample survey data there would be little difference in practice however when only using 8 sets of WINZ data there will be a greater difference here.

To make the results comparable with the results used earlier in this thesis the latter, less rigorous method of estimation for the data and the priors has been used. In any practical use of the new approach using a Bayesian analysis decisions would need to be made as to the best use of the data available with respect to the data and the priors but reusing the data is not recommended. For the WINZ data it could be argued that a mean of the three months that occur whilst SNZ HLFS data is being collected is best for the data and using some previous months would give an estimate of the variability for the prior. As far as the HLFS data is concerned, as has been suggested above, it is less

clear how it should be used but the difference that would be expected would be smaller. Previous quarters of data could be combined to estimate the variability in the replicates. What has been developed below puts aside these issues in order to illustrate the technique given only eight periods of data. Work should be carried out to ensure that the best use of the data is being achieved. In the rest of this chapter we will use averages and variances from all replicates.

7.5 Choice of priors

The programs above have used prior distributions for various parameters and for the data. In any Bayesian analysis these priors should be justified. In general if there is no information about a distribution then a prior with a fairly flat wide distribution is used hence the unknown parameters in the generalized linear model may be given a normal distribution with large variance. WinBUGS uses the precision instead of the variance, precision is the inverse of variance, so this would look like, beta[i] ~ dnorm(0, 0.00000001), in the program. The exact precision is arbitrary as long as it is small enough to give a large variance. The choice of precision can affect the speed of convergence without having a marked effect on the estimates, so it may need some fine tuning.

The prior distributions for the new margins are given by the 512 balanced replicates in the sample survey. Each replicate gives an estimate for the new margin and these distributions are shown in the graphs on the next page for the two sexes, three age groups and two ethnicities.

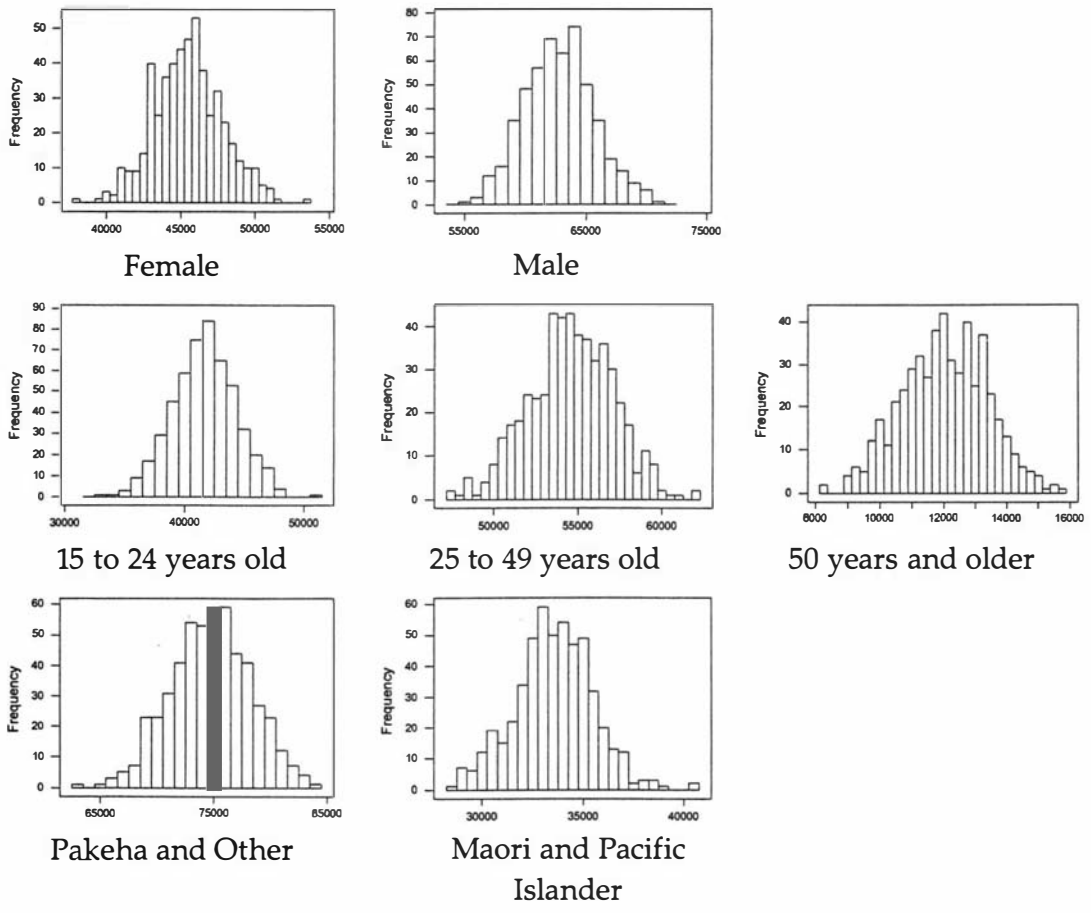


Figure 7.4 Histograms of replicates for margins from sample survey data

From the histograms above and the normal probability plots below there is no evidence to suggest that the distribution of the estimates in the replicates is not normal. The p values for an Anderson and Darling test of normality are given and the smallest is p = 0.131 which again suggests that normal distributions would be appropriate Anderson and Darling (1954).

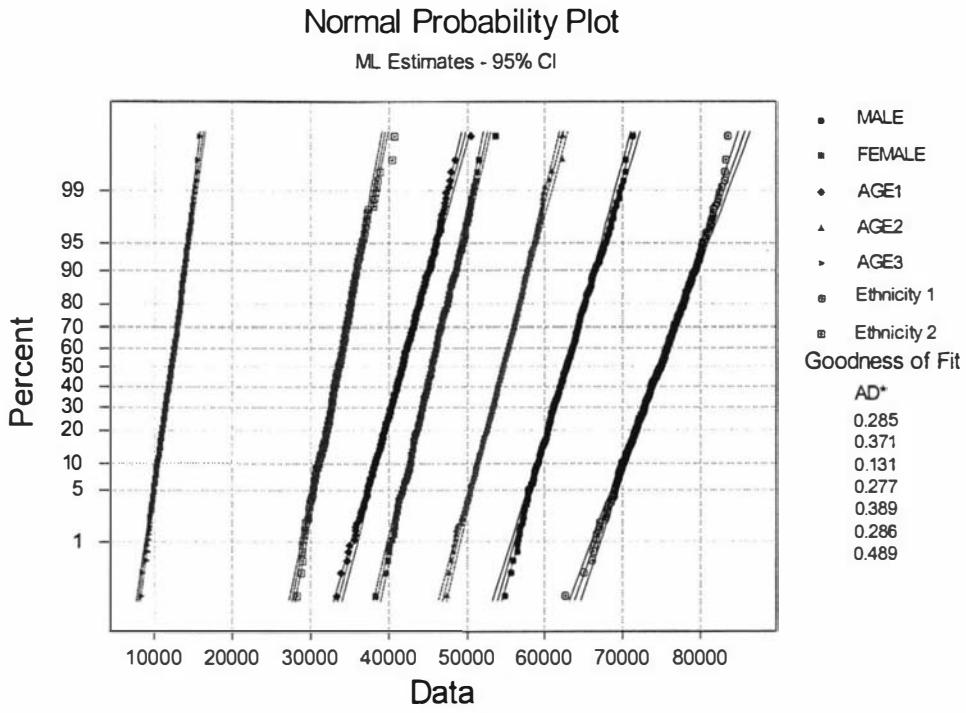


Figure 7.5 *Normal probability plots for the replicates of the new margins from the sample survey data.*

New Margins	Mean	Standard Deviation	Variance	$\frac{\text{Variance}}{\text{Mean}}$	Coefficient of Variation
Male	62777.1	2875.5	8268500	131.7	0.0458
Female	45501.3	2288.2	5235859	115.1	0.0503
Age 15 to 24	41652.4	2640.8	6973825	167.4	0.0634
Age 25 to 49	54583.6	2508.2	6291067	115.3	0.0460
Age 50 and over	12042.4	1336.6	1786500	148.4	0.1110
Pakeha and Other	74725.0	3523.0	12411529	166.1	0.0471
Maori and Pacific Islander	33553.0	1914.0	3663396	109.2	0.0570

Table 7.2 *Means and variances for new margins from 512 replicates of survey data*

The ratios of variance to mean and the coefficients of variation give some indication of a constant ratio. We could set the prior for each value in the margins with the mean and variance as calculated or use the mean and use an average coefficient of variation of around 130 as an estimate of the ratio. We have used the actual estimates of both the mean and variance.

Table 7.2 clearly indicates that the new margins should not be treated as Poisson variables. The variances are much too large to be considered equal to the mean. There are a number of mechanisms which can lead to over-dispersion such as this. They are often related to large inter-cluster variability. The complex cluster design of the HLFS may well introduce this over-dispersion. It is common in generalized linear models to assume that the variance is linearly related to the mean and not equal

$$Var(Y_i) = \sigma^2 \mathrm{E}[Y_i]$$

This is one solution to the problem and we could assume that $\sigma^2 \approx 130$ and continue the analysis. The generalized linear model is very flexible and the link function is not dependent on the distribution. The logarithmic link function is the canonical link for the Poisson distribution and hence has some desirable properties but it is not restricted to that distribution. The suggestion that a log linear model is appropriate for data in a contingency table still stands, we would expect the effects to be multiplicative and so the logarithmic transformation is sensible. It appears from the discussion above that the margins from the survey data are normally distributed and we will see below that this assumption is not unreasonable for the cell counts in the complete table of WINZ data.

In the Bayesian setting using WinBUGS we are also able to set each margin, or cell count, with its own mean and precision so we no longer need to assume some relationship between mean and variance.

The cell counts in the census data present rather more problems. Here there are only 8 replicates and there are up to 1776 cells if we consider 74 Territorial Authorities, two sexes, three age groups and four ethnicities. If we look at the nine North Island Regions broken down by the two sexes and three age groups then there are 54 cells in the table. One approach is to calculate the p values for the test of normality for each set of eight values and then to look at the distribution of these p values. Under the null hypothesis that the cell counts are each normally distributed, across the eight censuses, the p values would be uniformly distributed. The results for the table with 54 cells are given below.

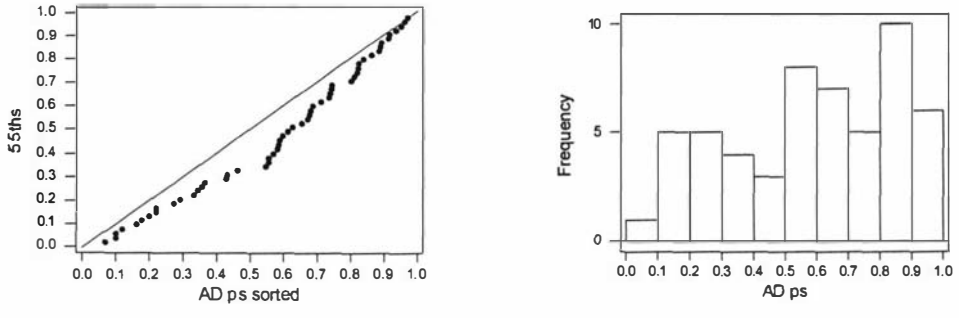


Figure 7.6 *Quantile-Quantile plot and histogram of probabilities reported by Anderson Darling tests for normality for the 54 cell contingency table used in the earlier example.*

We can see that the p values are skew down and the quantile quantile plot suggests this as well. The test used to generate these p values was Anderson and Darling which tests whether the data comes from a specified distribution in this case the normal distribution Anderson and Darling (1954). The pattern of p values was the same for the Kolmogorov Smirnov statistic and Cramer-von Mises.

None of this is evidence that the cell counts are not normally distributed and so we have used normal priors for the cell counts. It should be made clear that

the possibility of seasonal changes in the unemployment have not been investigated. With only eight months of data these would be difficult to identify but with more monthly data this should certainly be considered. However the method used for estimating the small area counts would be the same but the priors and data would be affected. Thus again we should say that the data analyses in this thesis are intended to illustrate the new approach and not to find the best model.

As discussed above in making this decision we have moved away from the Poisson or multinomial count models that are the normal assumption in contingency tables of count data. The development of the log linear model for a contingency table presented in Chapter Three made no distributional assumptions in suggesting that an additive model on the log scale was appropriate. We could appeal to the central limit theorem and with large sample sizes in most cells of the table we could expect the Poisson cell counts to be approximately normal. Or we could simply use a model which proposed that the cell counts were normally distributed and that the model was linear in the log scale.

7.6 Bayesian solution with a quadratic function.

The same program as in section 7.3 can be used for the model with a quadratic function, we only need to change the design matrix so that the appropriate column vectors, for both the main effects and interactions, have the age and the age squared in them as opposed to a one. When using three age groups the quadratic model and the earlier model with two age effects give the same results as we have seen in Chapter Five. This is as expected as the quadratic model has two parameters for age as does the model with categorical variables. The model for the nine North Island regions with data for 11 five-yearly age groups between 15 and 65 has also been fitted as in Chapter Five.

The results for these two models are the same as for those presented in Chapters Four and Five.

7.7 Variance estimation

In this thesis we have said little about the estimation of the variance of cell counts. In SPREE the usual measure of accuracy of the estimates is an estimate of the mean square error with respect to the survey design Purcell (1979) and Ghosh and Rao (1994).

Remembering the definition of variables from Chapter Three

- π the past cell relative frequencies representing the association structure. Generally found from the census data.
- ξ the marginal constraints representing the allocation structure, generally found from the sample survey data.
- $\mathbf{p}$ the cell estimates resulting from applying the IPF algorithm to π and ξ.

Along with the following subscripts

- h small domains
- i variables of interest
- g associated variable categories

We can express the estimated (est) means square error (mse) as

$$\text{est}[\text{mse}(\mathbf{p}\mid\boldsymbol{\pi})] = \hat{\mathrm{V}}(\mathbf{p}\mid\boldsymbol{\pi}) + \left(\hat{\mathrm{E}}(\mathbf{p}\mid\boldsymbol{\pi}) - \hat{\mathbf{P}}\right)\left(\hat{\mathrm{E}}(\mathbf{p}\mid\boldsymbol{\pi}) - \hat{\mathbf{P}}\right)^{\mathrm{T}} \quad (7.1)$$

where $\hat{\mathbf{P}}$ denotes another estimate of the cell counts at the time of the survey based on long term averages.

Haslett, Green and Zingel (1998) developed an estimator for the variance which could be used for the data available in the present example. Haslett (2003) has provided the following derivation.

It is better to consider the true cell counts $\mathbf{P}$ to be a superpopulation parameter rather than a parameter fixed with respect to the survey design, as

P depends on a distribution that depends on the stochastic properties of π as well as $\mathbf{p}$ rather than only on the survey design itself.

Letting ε denote expectation with respect to the ς superpopulation distribution, and E denote the usual expectation with respect to the survey design, with υ_ς and V_d denoting the corresponding variances we then have the following alternative interpretation of $\text{est}[\text{mse}(\mathbf{p}\mid\pi)]$ as an approximately unbiased estimate of the combined superpopulation and survey design variance of $\mathbf{p}$.

Taking the joint expectation of equation (7.1) first over the design (for which π is fixed) then over the superpopulation and assuming that the variance term is design unbiased yields:

$$\varepsilon\mathrm{E}\left[\text{est}[\text{mse}(\mathbf{p}\mid\pi)]\right]=\varepsilon\left[\mathrm{V_d}(\mathbf{p}\mid\pi)\right]+\varepsilon\mathrm{E}\left[\left(\hat{\mathrm{E}}(\mathbf{p}\mid\pi)-\hat{\mathbf{P}}\right)\left(\hat{\mathrm{E}}(\mathbf{p}\mid\pi)-\hat{\mathbf{P}}\right)^T\right]$$

Replacing $\hat{\mathrm{E}}(\mathbf{p}\mid\pi)$ in the bias term by $\mathrm{E}(\mathbf{p}\mid\pi)$ of which it is an estimate and approximating $\left(\mathrm{E}(\mathbf{p}\mid\pi)-\hat{\mathbf{P}}\right)$ by $\left(\mathrm{E}(\mathbf{p}\mid\pi)-\mathbf{P}\right)$ (assuming that the long term averages are close to the true values) we can write:

$$\begin{aligned}
&\approx\varepsilon\left[\mathrm{V_d}(\mathbf{p}\mid\pi)\right]+\varepsilon\mathrm{E}\left[\left(\mathrm{E}(\mathbf{p}\mid\pi)-\mathbf{P}\right)\left(\mathrm{E}(\mathbf{p}\mid\pi)-\mathbf{P}\right)^T\right]\\
&=\varepsilon\left[\mathrm{V_d}(\mathbf{p}\mid\pi)\right]+\varepsilon\left[\left(\mathrm{E}(\mathbf{p}\mid\pi)-\mathbf{P}\right)\left(\mathrm{E}(\mathbf{p}\mid\pi)-\mathbf{P}\right)^T\right]\\
&=\varepsilon\left[\mathrm{V_d}(\mathbf{p}\mid\pi)\right]+\upsilon_\varsigma\left(\mathrm{E}(\mathbf{p}\mid\pi)\right) \qquad (7.2)
\end{aligned}$$

from the usual formula for conditional variance (see for example C.R. Rao (1973))

$$=\mathrm{E}\varepsilon\left[\left(\mathbf{p}-\mathrm{E}\varepsilon(\mathbf{p})\right)\left(\mathbf{p}-\mathrm{E}\varepsilon(\mathbf{p})\right)^T\right]$$

which is the unconditional combined variance of $\mathbf{p}$ under the superpopulation and over the design.

If the small area estimates are sums of cells in the table, the variances will be sums of the appropriate variances.

The first term on the right hand side of equation (7.2) is the estimated survey variance of $\mathbf{p}$ given the most recent census association structure π which can be estimated via balanced repeated replicates sub sampling Wolter(1985). This is achieved by varying the margins of the table used to generate the SPREE estimates while leaving the underlying body of the table fixed. The second term is the estimated model variance of $E(\mathbf{p} \mid \boldsymbol{\pi})$ gained from varying the association structure for the given survey margins. It is convenient that both of these terms can be estimated using the same underlying algorithm as SPREE estimation itself. Hence they are computationally simple and efficient relative to considering all combinations of internal structure and margins.

In the Bayesian context the estimation of the variance is somewhat simpler as the distribution of all parameters of interest are generated by the simulations. If the priors include appropriate estimates of the precision for both the margins found from the sample survey data and for the cell values from the census data, then the distributions of the small area estimates will reflect these and the standard deviation reported by WinBUGS should be close to those calculated by the method above.

7.8 Conclusions

The Bayesian approach to statistics has very briefly been introduced and we have shown why the data that we are using is appropriate for this approach. Using the same model as in the analysis in Chapter Four we have shown that the Bayesian approach and BUGS software package result in the same estimates as those from the other packages that we have used and have been shown to be the same as the SPREE estimates using the IPF algorithm.

The nature of the data has made the choice of priors quite simple. The estimates formed by the MCMC approach to the Bayesian model allow the variability in the estimates to be found with ease. Even in the initial, two step, program where the coefficients in the GLM for the census data are estimated

and then these are used as part of the data to estimate the other coefficients for the survey data the variability in the estimated coefficients is carried through to the second step as part of the prior for those variables. In the final version of the program the two sets of data are fitted in the same program and the variability in the estimates of the coefficients for the census data is carried through to the estimation of the coefficients for the survey data by default. Hence the estimates of the variability in the predicted counts should reflect all of the variability inherent in the data. This is not true in most small area estimation procedures since neither MCMC nor the method of Haslett(2003), as in section 7.7, is used.

We now have the structure and the software to extend the models that we can use in new directions. The generalized linear model is very flexible in its application and by extending it to random effects as well as fixed effects an even wider range of possibilities emerge. We will illustrate some of these possibilities by introducing a conditional autoregressive (CAR) model to allow for spatial correlation between cells in the table. The CAR model will be discussed in detail in the next chapter.

In Chapter Nine we will also use the Bayesian approach in extending the tables to include counts of employed and we will show that this is the same as modeling the relative risk of unemployment.

These two extensions are in no way exhaustive of the opportunities, they simply show two possible extensions to the model that can be applied when using the new approach that we have proposed.

CHAPTER 8

Spatial models, a conditional autoregressive (CAR) approach

8.1 Introduction

Much of the background material in this chapter is covered in "Statistics for Spatial Data" by Cressie N. (1993) Wiley. Some sections will be specifically referenced where further detail may be required.

Data collected in a wide range of situations have long been recognised as having spatial variation. Initially most research was focused on reducing the effect of spatial dependence in the results. R.A. Fisher at Rothamstead Experimental Station in England in the 1920's and 30's went to great lengths to reduce the effects of spatial variability in field trials largely by data collection design. Ideas such as replication, randomization and blocking were introduced. Since that time the emphasis has moved to explicitly modeling the spatial variation both in the types of agricultural experiments that Fisher was involved with and in a wider range of problems. This wider range of problems includes applications in geostatistics, particularly mining, atmospheric science, soil science, epidemiology, public health and analysis of data from satellites. Ghosh, Natarajan, Stroud and Bradley (1998) and Rao (2003) describe spatial approaches to small area estimation including the use of models with spatial autocorrelation though they do not use auxiliary data. This range of applications continues to expand rapidly.

Most data analysis is performed on the assumption that the data collected are independent observations, in a spatial analysis this assumption is relaxed to allow for sites "close" to each other to be more "similar" than sites widely spaced. Measures of similarity and closeness will depend on the application.

These measures then need to be translated into the value of a variable. We will explain these terms as they relate to our data later in this chapter.

In relaxing the assumption of independence other less restrictive assumptions are introduced. We will assume that there is a correlation between adjacent sites. This correlation may change as "distance" between site centres changes. This distance may be Euclidean. Alternatively the correlation could be fixed for neighbours so that it is nonzero between areas with common boundaries. This thesis will define adjacency to mean sites with common boundaries. This is one of many possible definitions of spatial interdependence but it will suffice to show that the new approach that we are proposing can include models of this type. A brief discussion of some other possible adjacencies is included in section 8.7.

Spatial data are commonly categorised into a number of types.

Three commonly used are:

- data with a continuous spatial index such as geo-statistical data where variables change continuously,
- those with a lattice index where a regular or irregular grid is drawn on the area and observations are based on the grid "squares",
- spatial point processes where the variable of interest is the location of "events" which are clustered by some mechanism.

Lattice data can be either regular or irregular depending on whether the space is divided into a regular pattern, such as a grid placed on top of a city map, or an irregular pattern, often imposed by geographical boundaries. As the Regional Authority boundaries in New Zealand are based on watersheds, and as these are not regular, we will consider irregular lattices.

This chapter will consider correlation based on geographical adjacency, hence regions next to each other will be considered similar, but there are other

similarities which could be exploited. Individual cells in the table are defined by sex, age and ethnicity variables as well as geographic regions. Members of an ethnic group could be considered similar as could males or females. Even some age groups may be related. These possibilities raise many modeling issues. The intention in this chapter is only to show that the models can be fitted and the parameters estimated. We will look briefly at some model comparison but that is not the focus of this work.

In spatial models there is often a problem of what to do with areas which are on the edge of the lattice. Like the surface tension on the surface of a liquid estimates are pulled in towards the areas within the lattice with no effect from outside. We will suggest that this is of no concern in our case because the boundaries of the set of regions considered is not arbitrary but the boundary problem may need to be considered in the more general setting.

8.2 The CAR model

Gaussian conditional autoregressive (CAR) processes have been used in spatial statistics to describe the association between random variables observed at fixed sites in some Euclidean space Cressie (1993). An additional term is added to the model which is zero unless the sites are adjacent. In the case of adjacent sites the additional term is normally distributed. For examples of recent work see Besag and Kooperberg (1995) and Pettit, Weir and Hart (2002)

The spatial dependence is modeled by the covariance matrix which includes terms for adjacent areas. This matrix must be symmetric (if i is adjacent to j then j must also be adjacent to i).

$$\mathbf{Y} \sim N\left(\boldsymbol{\mu}, (\mathbf{I}-\mathbf{C})^{-1}\mathbf{M}\right) \tag{8.1}$$

where

Y is the vector of data values subscripted for census or survey as appropriate

μ is a vector of means of the distributions of the

C is the (l x l) matrix of c_{ij} where $c_{ii} = 0$, $c_{ik} = 0$ where i and k are not adjacent areas and $c_{ij}\tau_i^2 = c_{ji}\tau_i^2$ τ_i^2 is the variance of y_i conditional on the neighbours

M is the diagonal (l x l) matrix diag$\left(\tau_1^2,\cdots,\tau_l^2\right)$

Note that in the case where adjacency of areas is used in the model C is not an adjacency matrix, but a function of the adjacency matrix, the number of neighbours and possibly unequal variances depending on M.

8.3 Implementation in WinBUGS

The models that we have fitted in WinBUGS are of the form

$$Ln\left(mu_i\right) = \mathbf{x}_{1i}^{\mathrm{T}}\mathrm{b}_1 + \mathbf{x}_{2i}^{\mathrm{T}}\mathrm{b}_2 + \alpha + b_i$$

with appropriate subscripts for census and survey data as discussed in Chapter Three.

$Ln\left(mu_i\right)$ is the data i.e. y_i

$\mathbf{x}_{1i}^{\mathrm{T}}$ and $\mathbf{x}_{2i}^{\mathrm{T}}$ are the i^{th} row of $\mathbf{X}_1$ and $\mathbf{X}_2$ as defined in Chapter Three

$\alpha + b_i$ is the conditional autoregressive component

ν is defined in the literature as a vector of pseudo errors

$$\boldsymbol{\nu} = \left(\mathbf{I} - \mathbf{C}\right)\left(\mathbf{Z} - \boldsymbol{\mu}\right)$$

or equivalently

$$y_i - \mu_i \equiv \sum_{j=1}^{n} c_{ij}\left(y_i - \mu_i\right) + \nu_i$$

For the implementation of WinBUGS

$y_i - \mu_i = \alpha + b_i$ and we can write

$$\alpha + b_i = \sum_{j=1}^{n} c_{ij}\left(\alpha + b_j\right) + \nu_i$$

Define $c_{ij}^{'} \equiv \begin{cases} -1 & if\ i = j \\ c_{ij} & if\ i \neq j \end{cases}$ and $\mathbf{c}_i = (c_{i1}, ...c_{in})^T$, $\mathbf{c}_i^{'} = (c_{i1}, .., c_{i,i-1}, -1, c_{i,i+1}, ..c_{in})^T$

Then $$\alpha\left(1 - \sum_{j=1}^{n} c_{ij}\right) = \sum_{j=1}^{n} c_{ij}' b_j + \nu_i$$

$$\nu_i = \alpha\left(1 - \sum_{j=1}^{n} c_{ij}\right) - \sum_{j=1}^{n} c_{ij}' b_j$$

Then $$\nu_i = \alpha\left(1 - \mathbf{c}_i^T \mathbf{1}\right) - \left(\mathbf{c}_i^{'}\right)^T \mathbf{b} \quad = -\left(\mathbf{c}_i^{'}\right)^T \left(\alpha \mathbf{1} + \mathbf{b}\right)$$

Hence $$\boldsymbol{\nu} = -\left(\mathbf{I} - \mathbf{C}\right)\left(\alpha \mathbf{1} + \mathbf{b}\right) \qquad (8.2)$$

where $\mathbf{1}$ is an l x 1 vector of 1's

and $\mathbf{b}$ is the l x 1 vector of $b's$

Noble and Haslett (2001), Noble, Haslett and Arnold (2001).

So it can be seen that there is a deterministic relationship between ν (as defined in the literature) and α and $\mathbf{b}$ as defined in WinBUGS.

8.4 Specification of the CAR model in WinBUGS

The specification of the CAR model in WinBUGS requires the adjacencies to be defined. The four parameters of the car.normal distribution are the adjacency, the weights, the number and the precision. Thus the statement reads

```
b[1:192] ~ car.normal(adj[],weights[],num[],tau)
```

Where

- adj[] This is a string of comma separated numbers denoting the cells adjacent to a given cell. The first part of the string in the example used in this chapter is: 2, 1,3, 2,4,6,7,8, 3,5,6, 4,6, 3,4,5,8, 3,8, 3,6,7,9, 8, 11,12,13,14, 10,12, 10,11,14, 10,14,15,16, 10,12,13,15, 13,14,16, 13,15 The spaces have been added to read it more clearly. The string indicates that cell 1 is adjacent to 2, cell 2 is adjacent to cells 1 and 3 etc. In this example this string is repeated 12 times with 16 added to

each number each time so that each sex by age by ethnicity combination is paired with adjacent regions.

- weights[] Are weights which may be applied to each adjacency, these could be related to distance etc. In the example here they will all be 1 to simplify matters.
- num[] This is a houskeeping variable for the program to interpret the adj[] variable. It gives the number of neighbours for each area which the program uses internally.
- tau This is the precision of the spatial variable.

One point to note is that the car.normal distribution is parameterized to include a sum to zero constraint on the random effects. This means that a separate intercept term must be included in the model. In WinBUGS it is recommended that this be assigned an improper uniform prior which is defined by the dflat() statement in the program. The dflat() distribution corresponds to an improper prior on the whole real line, being in effect a uniform prior of infinite width. This is an improper prior as it does not integrate to one. Besag and Kooperberg (1995) show that this parameterization is equivalent to the unconstrained parameterization with no separate intercept.

A full example of a program, in fact the one used for the data analysis in the next section, is included in Appendix C.

The model that is being applied is the same as has been used previously for the census and survey data but the model for the survey data has the additional CAR terms.

$$Ln(\boldsymbol{\mu}_c) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c}$$

$$Ln(\boldsymbol{\mu}_s) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + \mathbf{X}_1\boldsymbol{\beta}_{1s} + \alpha + \mathbf{b}$$

$$\mathbf{Y}_c \sim N(\boldsymbol{\mu}_c, \boldsymbol{\tau}_c)$$

$$\mathbf{Y}_s \sim N(\boldsymbol{\mu}_s, \boldsymbol{\tau}_s)$$

$$\boldsymbol{\beta}_{1c} \sim N(\text{beta}_{1c}, \text{taubeta}_{1c})$$

$$\boldsymbol{\beta}_{2c} \sim N(\text{beta}_{2c}, \text{taubeta}_{2c})$$

$$\boldsymbol{\beta}_{1s} \sim N(\text{beta}_{1s}, \text{taubeta}_{1s})$$

$\alpha \sim U(-\infty, \infty)$ Note this replaces the constant term in $\boldsymbol{\beta}_{1s}$, $\mathbf{X}_1$ no longer includes a column of ones.

$\mathbf{b} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ as reviewed in Besag and Kooperberg (1995).

Note α and $\mathbf{b}$ are defined as in section 8.3.

Once more there are a number of modeling considerations which should be addressed if this model, or one related to it, is to be used in a practical situation. The CAR part of the model has only been added to the sample part of the model, alternatively it could be added to the census part or both. If we believe that the census data accurately models the regional effects then there is no advantage in modeling the correlation between the regions. If on the other hand we believe that the census data does not reflect the regional differences in the variable of interest then modeling the correlation between adjacent regions may improve the estimates. The sampling variability of the survey data will be smoothed by borrowing strength from the adjacent regions and so the estimates should be more stable with the CAR model.

The program is essentially the same as the one in Chapter Seven with the additional section defining the CAR model as below.

for (i in 2:5) { betaa[i] ~ dnorm(0.0,1.0E-8) }	Uninformative priors for four of the new coefficients as in the previous chapter.
]	
betaa[1] ~ dflat()	The intercept term beta[1] is given the dflat() prior as described above.
b[1:192] ~ car.normal(adj[], weights[], num[], tau)	CAR normal prior
tau ~ dgamma(.001, .001) sigma<-1/sqrt(tau) }	Prior distribution for tau for the car.normal distribution

Choice of a suitable prior for the precision, tau, of the spatial variable is less clear than the priors for the other parameters that we have discussed. Active research continues in the area of specification of priors for variance parameters at present Browne and Draper (2001). A gamma distribution is commonly used but the choice of parameters of the distribution is largely subjective. Two equal parameters give a mean of the distribution of one and this is common, but the variance is the second parameter over the first squared and this will change the probability of selecting values close to zero or far from it. If the spatial effect is in fact negligible values close to zero are desirable Kelsall and Wakefield (1999).

The choice of the gamma distribution may historically have been for convenience as it is a conjugate prior and analytically this was advantageous. Modern software makes this less important hence there may be better priors to use. Such distributions as uniform within a defined range and half normals are being used and this aspect should be considered in more detail. There has been some discussion recently, on the BUGS email discussion group, on the

exact meaning of an uninformative prior and it is erroneous to believe that simply because the distribution is fairly flat it will be uninformative. The gamma distributions used for the precision parameter of the CAR term have low density across a wide range of the positive real numbers, but they do have a maximum and the position of this maximum along with the probability that the precision could be close to zero can affect the outcome. Greater weight at some point may affect the results even if it is only slightly greater. An assessment of the robustness of the estimates to varying priors should always be carried out. If the prior is supposed to be uninformative then using other uninformative priors should give similar final estimates.

We have chosen values which gave clear changes in the estimates but which were not closely dependent on the priors chosen. We would reiterate that the intention of this thesis is to show how these more sophisticated models become possible using our approach whereas under SPREE they are not.

If the new approach was to be used in practice and a spatial variable such as this was to be included then the choice of priors is certainly important. With data such as ours, which is collected regularly, past data can be used and over time the spatial component would not be expected to change quickly. Hence we can try different priors with the past data and find ones which, over time, smooth the estimates in an appropriate manner. It is possible that these priors may change over time but we would not expect them to change suddenly unless there were administrative, or legislative, changes which introduced a sudden change to the data. In a "one off" analysis it would be more difficult to justify the choice of a prior which was at all informative as with only a single set of data there would be no information to use. An uninformative prior with some weight close to zero may be the best solution in this case.

8.5 An Example

We shall apply the conditional autoregressive model to the sixteen Regional Authorities data for both islands of New Zealand. We will subdivide the population into two sexes, three age groups and two ethnic groupings.

The census data, from the NZDWI is used to estimate the parameters in the saturated loglinear model for the two by three by two by sixteen contingency table for sex, age, ethnicity and Regional Authority categories. The survey data gives new margins for the sex, age and ethnicity variables and the overall mean. In the results presented in Table 8.1 the original census data is given along with the estimates for counts using the new algorithm with categorical variables, the latter being equivalent to SPREE. No spatial dependence is included. Finally estimates are given using the new algorithm and a CAR model to include spatial dependence. In the results below these are called "Census", "SPREE" and "CAR" respectively.

The adjacency matrix, which defines "similar" cells, is constructed by defining regions that have a mutual boundary as being similar. This is then applied to sex by age by ethnicity sub groups within each region. That is Northland male Pakehas aged 15 to 24 are considered to be similar to Auckland male Pakehas aged 15 to 24 etc. The matrix is symmetric so the reverse of the above is also true. In this analysis we have assumed that the North Island is separate from the South Island, that is, there is no adjacency between any North and South Island regions. In New Zealand the southern region of the North Island, Wellington, and the northern region of the South Island, Marlborough, are very different in terms of their population and work related activities. The diagram below shows these adjacencies.

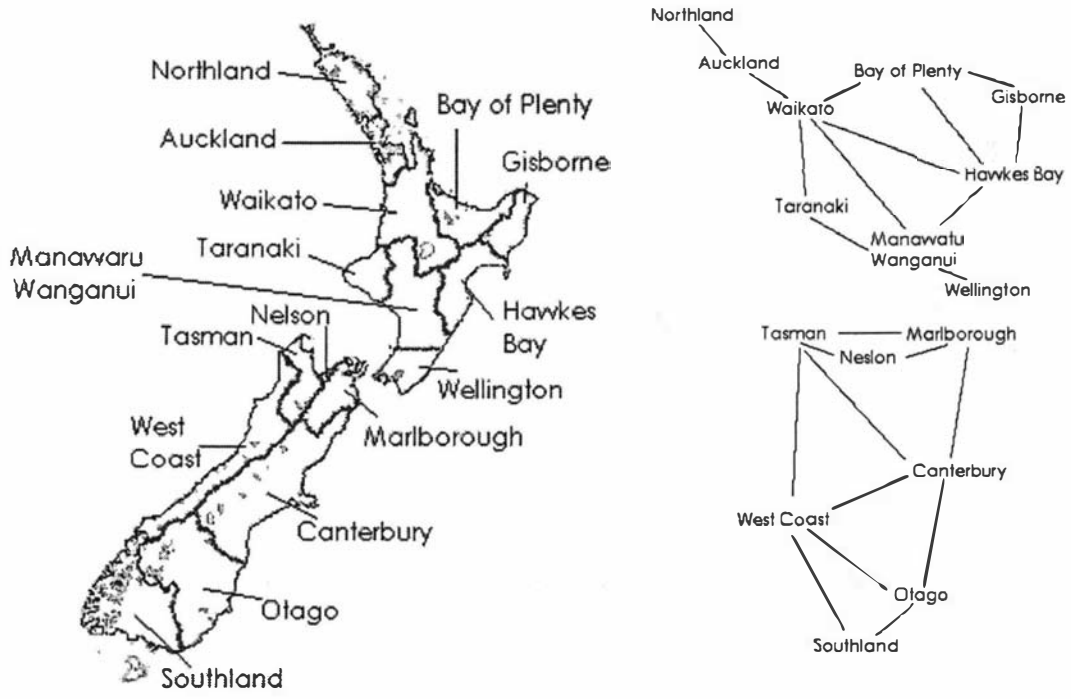


Figure 8.1 Map of regions and diagram showing the adjacencies used.

The prior for the precision of the spatial dependence is as discussed above with a gamma (0.001, 0.001) distribution. This has a mean of one and a variance of 1000.

North Island				South Island			
Region	Census	SPREE	CAR	Region	Census	SPREE	CAR
Northland	9600	5857	5969	Tasman	1614	1139	1086
Auckland	39699	26057	26544	Nelson	1859	1337	1274
Waikato	14278	9383	9530	Marlborough	1789	1249	1187
Bay of Plenty	12606	7874	8003	West Coast	2043	1431	1354
Gisborne	3722	2157	2178	Canterbury	19354	14058	13378
Hawkes Bay	8018	5087	5170	Otago	8260	6001	5702
Taranaki	5627	3867	3924	Southland	4690	3288	3119
Manawatu-Wanganui	11112	7437	7556				
Wellington	17909	12051	12278				

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.*

Statistics New Zealand report standard errors for counts in a region for the direct estimates from the survey data (not given here) in the range 4% to 10%, for the SPREE estimates they are 0.3% and for the model including the CAR error term they are in the range 0.6% to 1.4%.

Tables 8.2 and 8.3 shows counts for each method subdivided by age group ethnicity and sex for the Regions in the North and South Islands. in

		North Island											
		Male						Female					
	Ethnicity	Pakeha and Other			Maori and Pacific Island			Pakeha and Other			Maori and Pacific Island		
	Age	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over
Northland	SPREE	474	1070	238	616	1069	106	419	623	153	444	574	73
	CAR	398	1057	270	595	1183	139	529	640	98	401	485	61
Auckland	SPREE	2555	6105	930	2028	2563	553	2454	4229	744	1974	1681	241
	CAR	2144	6039	1060	1962	2839	730	3097	4335	479	1786	1422	201
Waikato	SPREE	1177	1706	349	792	1000	306	1266	1125	286	676	626	73
	CAR	988	1685	398	768	1107	404	1598	1153	185	612	530	61
Bay of Plenty	SPREE	755	1227	287	802	1228	308	825	820	245	65	635	77
	CAR	634	1212	328	774	1360	406	1042	843	158	601	536	64
Gisborne	SPREE	146	243	53	279	487	90	143	147	26	217	289	38
	CAR	123	241	59	270	538	120	181	150	17	196	244	31
Hawkes Bay	SPREE	595	961	210	483	673	197	557	506	134	369	382	46
	CAR	498	925	239	467	746	259	705	519	87	334	323	39
Taranaki	SPREE	625	770	169	240	274	140	603	567	119	176	463	22
	CAR	524	761	192	232	304	185	761	580	76	159	138	18
Manawatu	SPREE	1060	1504	316	543	663	267	1042	964	233	411	387	47
Wanganui	CAR	889	1486	358	526	734	353	1316	987	149	372	326	39
Wellington	SPREE	1814	2697	521	852	983	369	1597	1577	340	682	540	79
	CAR	1520	2664	592	823	1088	488	2019	1614	219	617	456	66

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)*

		South Island												Male	F
	Ethnicity	Pakeha and Other			Maori and Pacific Island			Pakeha and Other			Maori and Pacific Island				
	Age	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over		
Tasman	SPREE	155	350	69	23	38	50	158	215	40	17	20	3		
	CAR	156	334	60	20	45	59	161	219	32	15	13	2		
Nelson	SPREE	221	334	72	37	50	49	236	238	49	28	21	1		
	CAR	221	319	63	33	59	59	240	242	40	25	14	1		
Marlborough	SPREE	184	322	76	47	57	57	181	206	50	35	31	3		
	CAR	184	308	66	40	67	68	184	209	42	32	21	3		
West Coast	SPREE	217	765	86	30	44	64	182	235	52	26	28	2		
	CAR	218	444	76	26	52	76	184	240	42	22	19	1		
Canterbury	SPREE	2426	3388	688	400	441	484	2671	2422	511	352	251	24		
	CAR	2438	3234	598	344	520	578	2704	2466	417	297	172	19		
Otago	SPREE	1095	1578	336	126	136	231	1051	1056	201	98	81	12		
	CAR	1100	1505	293	108	160	276	1066	1073	165	83	56	10		
Southland	SPREE	516	756	187	126	174	148	545	525	116	101	85	9		
	CAR	518	722	163	108	205	177	552	534	95	85	58	8		

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)*

Figure 8.1 gives a clearer comparison of the SPREE and CAR estimates with different gamma priors for the precision of the CAR parameters.

Parameters for Tau

Prior
Gamma(0.01, 0.01)

Mean 1

Variance 100

sigma = 0.02949

Prior
Gamma(0.1, 0.01)

Mean 10

Variance 1000

sigma = 0.009335

Prior
Gamma(0.5, 0.0005)

Mean 100

Variance 2000000

sigma = 0.006429

Graph of Estimates including a CAR structure vs Estimates without.

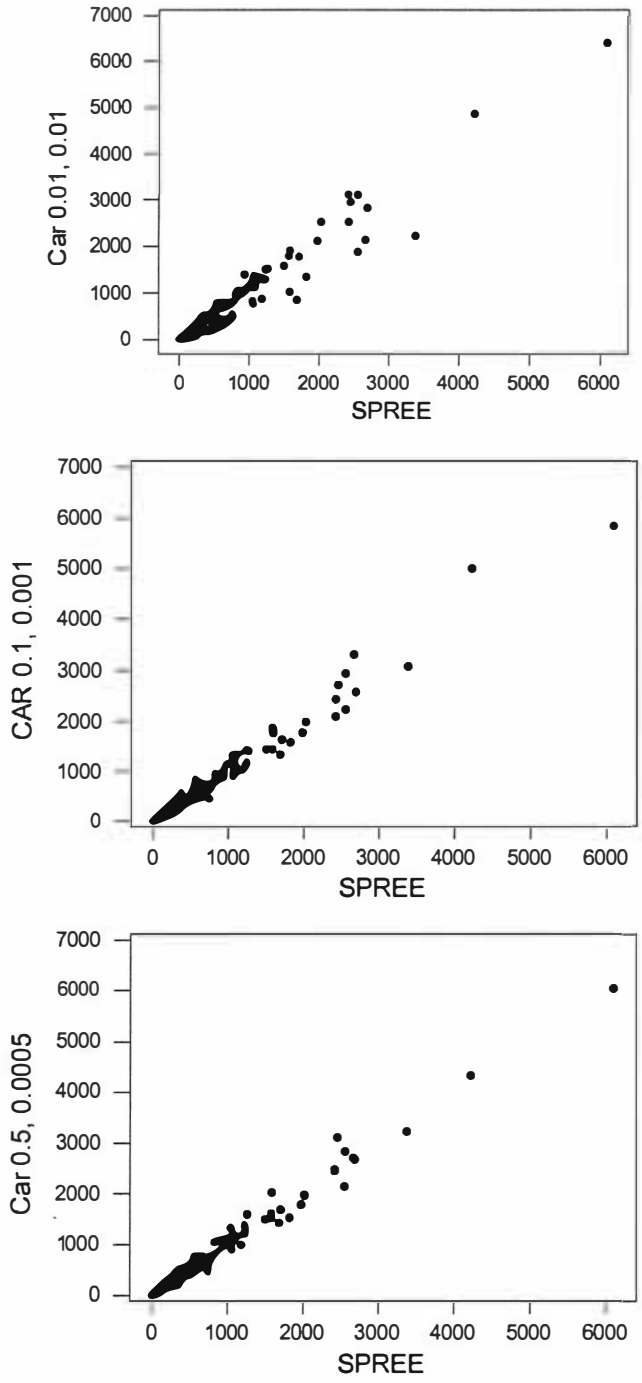


Figure 8.2 *Graphs of estimates with and without the conditional autoregressive term.*

The graph below shows the estimates with and without the CAR part in the model for the gamma 0.01, 0.01 prior of the precision parameter without the values greater than 750. When the larger values are included the scales on the axes are such that the detail is lost. The larger values are removed and so the fanning out of the estimates is more obvious. This is not unexpected given that the linear model is on a logarithmic scale.

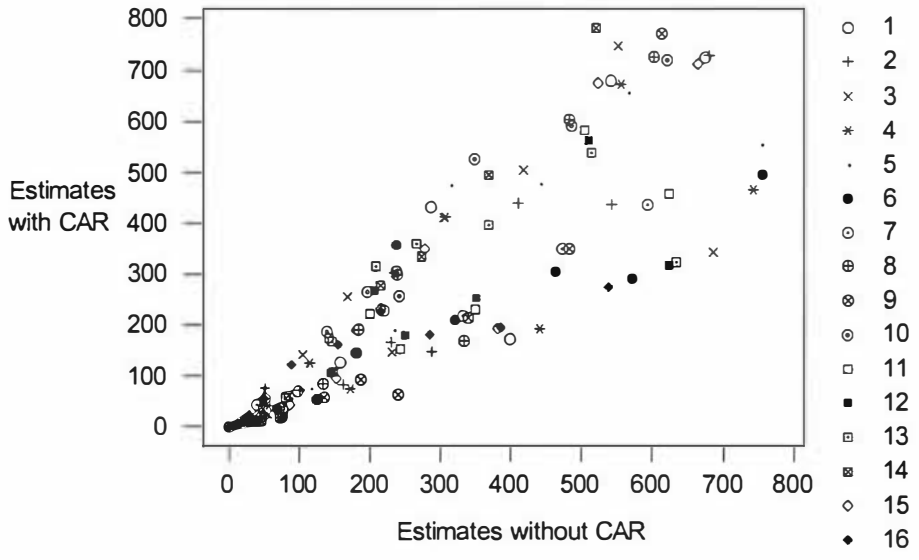


Figure 8.3 The first graph from figure 8.2 with large values removed to show the structure better. Regions 1 to 16 are identified.

In general the estimates for the North Island Regions change by a similar proportion in each sex by age by ethnicity combination compared with the SPREE estimates without the CAR term. The South Island estimate for the same sex by age by ethnicity combination changes in the opposite direction. The CAR terms form blocks of nine and seven almost identical values for each sex by age by ethnicity combination, for the two Islands. This is due to not having any adjacency between the two Islands and, I suspect, the high degree of interlinking between the small number of regions. Modelling the North Island alone did not change the pattern for that island in the CAR terms.

In section 8.7 we will discuss correlations between regions which may not be strictly spatial, in these cases there may be adjacencies which cross between the islands.

The model can be fitted but the results do not seem to be particularly useful in this example. Including the 74 Territorial Authorities may give more room for sensible correlations between the smaller areas.

The choice of the prior for the variance if the CAR term is important and much more work would be needed to ensure that it was appropriate. In the graphs in figure 8.1 we have demonstrated that the choice of prior does affect the estimates in the example. I don't believe that we are in a position to make firm decisions on the appropriate prior for this model at this stage. We have mentioned earlier that we only had eight months of data for the census and it would be more useful to use more months of census data to look at the seasonal patterns which we would expect to be found in this sort of data and then to consider the CAR terms in more detail.

The method proposed in this thesis can be used to fit CAR terms to the model.

8.6 Edge Effects

As mentioned in the introduction to this chapter edge effects can cause concern in spatial modeling. The problem arises when the area under study is part of a larger area. In this case there are data for points outside the area under consideration which are probably unknown. If in the diagram below we are interested in the larger area with small areas one, two, three, four and five then if we are considering small area one it will be it will be influenced by two, three and four but not five if we are considering shared boundaries. However small area one is also likely to be influenced to some extent by areas six, seven and eight for which there may be no data. The influence may also be somewhat different if different political structures or administrative

organization exists in the larger areas which have common boundaries with our small area one.

Figure 8.4 Map of small areas within a region (shaded) with other small areas around.

For the unemployment data in this thesis this is not a problem as the areas are two islands which are not influenced by the areas around them since there are none, only sea. Should we be considering other variables such as mean temperature then the effect of the oceans bounding some areas would need to be considered.

The choice of adjacencies between the two islands may be a matter of some debate. We could assume that the two islands were independent of each other as a simple assumption, in which case the CAR part of the model should be applied to each island separately.

8.7 Adjacencies other than simple geographic

This chapter has concentrated on spatial correlation between neighbouring regions with a common boundary. With the data that we have for unemployment it is easy to surmise that there may be other correlations, and hence other adjacencies, such as those within a gender or ethnicity. It is

perfectly reasonable to suggest that counts of unemployed females may be similar in different regions and similar for ethnic groups. Also there may be correlations within age groups. These effects may well be stronger than the relatively crude geographic adjacency that we have considered here. Even within the geographic adjacencies there may be stronger reasons for considering regions with similar industries to be adjacent even though they may be geographically separated.

The governmental regions discussed here are defined with other variables in mind, not unemployment, so the adjacency may not be well modeled using those boundaries.

The approach that we have taken in this chapter can easily be adapted for other kinds of correlations as they only depend on the definition of adjacency and the construction of the adjacency matrix.

If we are considering correlations other than simple spatial effects then the division between the islands may not be a factor. For example if we were to hypothesize that unemployment rates amongst young Maori women were similar throughout the country then we may consider them to be adjacent, in this case the physical separation of the two islands may not be relevant.

8.8 Conclusions

The new approach that we have proposed has been used to fit a CAR model to adjust the cell counts so that a predefined correlation structure is included in the model. We have shown that models such as this are possible with the new approach.

The choice of correlation structure is subjective. It is important that if models such as these are used in practice then a great deal of work is needed to justify the use of the model and to define a suitable correlation structure. In small area estimation boundaries between geographic areas and divisions between

levels of categories, such as age groups, are subjectively chosen. These both most often represent an arbitrary choice and in fact the underlying variable is continuous. In those circumstances an autoregressive structure can smooth the effect.

The analyst would hope that using these sorts of models would have the effect of smoothing the estimates and over time they should appear more stable than the direct estimates or the ordinary SPREE estimates. This would enable some checking of the estimates as the direct estimates from the survey data are unbiased but more variable. If the bias in the synthetic estimates is acceptable then they should be close to the mean of the direct estimates but smoother.

This chapter has demonstrated that the new approach offers more modeling possibilities than SPREE using the IPF algorithm. It allows both the use of auxiliary data in the form of administrative records and to gain strength from close neighbours to improve the estimates. Whilst it is feasible to do the equivalent analysis first fitting the SPREE model and then fitting a CAR model to the SPREE estimates the ease with which it can be done simultaneously in the new approach and, using the Bayesian paradigm the ease with which estimates of the variance can also be found, are useful improvements. The variance estimates are found as a byproduct of the sampling as discussed in Chapter Seven and include by default all of the sources of variation if they are included in the priors of the model.

As far as the example presented is concerned there seem to be some drawbacks in the way that the CAR model has been applied. I think that there are too many links between adjacent regions, and the division between the two islands, to make sensible adjustments. The result is the similar adjustments to blocks of cells and the compensatory increase and decrease in the two islands. This does not seem to be reasonable for this data. The regions

are large and all include both urban and rural areas, the variables in the model may capture the important differences between these regions and so there is little to be gained by including the CAR term.

It is more likely that the CAR model would prove useful if we consider the 74 smaller Territorial Authorities and define "adjacencies" between the larger urban centres, and between some of the predominantly rural areas even though they are not contiguous.

CHAPTER 9

A relative risk and odds ratio approach

9.1 Introduction

An alternative approach to modeling unemployment is to consider the risk of being unemployed in different regions. The risk is defined as the ratio of unemployed over the employable population. An alternative, the relative risk is often used in medical and epidemiological studies. It is defined as the ratio of two associated risks. In the medical scenario these are commonly the risk of contracting a disease under two different scenarios. For the data that we are modeling we could imagine a relative risk of unemployment in Auckland and Wellington or any other pair of regions. The comparison does not need to be regional, there is a relative risk for males and females and for different age groups or ethnicities.

The statistical benefit of doing this is that the relative risks are much more similar across regions than the number of unemployed which is directly related to regional population size.

We will define two possible ratios (one risk measured in two ways). The first is the ratio of unemployed to the total population for each Region by sex by age by ethnicity combination and the second is the ratio of unemployed to employed in each combination. The latter is termed an odds ratio as the two groups are disjoint but their union forms the population. These ratios could be applied at any level of categorisation but it makes most sense to apply them at the lowest level of agregation in the table. We will consider both of these in this chapter and show that they are closely related. This will enable us

to apply the new algorithm which we have proposed to the relative risks approach and show that it is a simple extension of the method.

As previously the algorithm that we have proposed is not restricted to categorical variables and any of the analyses such as the quadratic model for age, or the CAR model could be applied to the relative risk data. We will only consider the simpler situation where we apply the algorithm to the categorical data and find a relative risk for each cell in the table but the idea is readily extendible. The immediate effect of finding the relative risk is to smooth the data, as regions with high counts of unemployment are generally regions with high counts of employed and those with low counts of unemployment have small populations and hence small counts of employed. Those regions that are truly unusual will be better identified in this way as the effect of different sized populations is ameliorated. Counts can still be recovered from the estimated relative risk given the appropriate multipliers if needed.
There are a number of equivalent ways of expressing the model and we will discuss those in the next section along with how to estimate the parameters.

The census data that we have available from WINZ does not have any information about the employed people in a region as this organisation is not concerned with those in the employed group. We will use national census data of population in each region subdivided by the sexes, age and ethnic categories that we have previously used as margin totals. The alternative (and equivalent) structure is to increase the size of the original table so it contains both employed and unemployed counts and fit an equivalent overall log linear rather than logistic type model. See for example Bishop Fienberg and Holland (1975)

The SNZ HLFS data includes both employment and unemployment counts so the true relative risk can be calculated in this case.

9.2 *Relative risk models*

We shall begin by considering the risk model in which the count of unemployment (N_u) in a cell is divided by the population for that cell (N_p) rather than the number of employed (N_e). The subscripts u, e and p will be used in this sense throughout this chapter. Of course there is a simple relationship between these three variables and one could be dispensed with, but the exposition below will be easier to follow if we use all three. Using the notation from earlier chapters we have written our model:-

$$g\left(\mathrm{E}\left[\mathbf{Y}_c\right]\right) = X_1\boldsymbol{\beta}_{1c} + X_2\boldsymbol{\beta}_{2c} \qquad \text{for the census data}$$

$$\text{and} \quad g\left(\mathrm{E}\left[\mathbf{Y}_s\right]\right) = X_1\boldsymbol{\beta}_{1s} + X_2\boldsymbol{\beta}_{2s} \qquad \text{for the survey data.}$$

To make the notation simpler wi will explicitly state the link functions in each model. So for the log linear model so we can write:-

$$Ln\left(\mathrm{E}\left[\mathbf{N}_{uc}\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c} \qquad \text{for the census data}$$

$$\text{and} \quad Ln\left(\mathrm{E}\left[\mathbf{N}_{us}\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + \mathbf{X}_2\boldsymbol{\beta}_{2s} \qquad \text{for the survey data.}$$

The method developed does not require that it be a log linear model and there may be instances when some other form of the GLM may be appropriate.

In this section we are interested in a model for both employed and unemployed, and at this stage the distinction between the census and sample survey data is not important. However the estimation procedure will be the new algorithm that we have proposed using both the census and sample survey data. We will make comment where the different sources of data are relevant. We will also modify the partition of the $\mathbf{X}$ matrix and the vector of parameters $\boldsymbol{\beta}$ to explicitly show the constant term. We can then write the model as:

$$Ln\left(E\left[\mathbf{N}_u\right]\right) = \mathbf{1}\mu + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 \qquad (9.1)$$

We now wish to extend the model to incorporate both employment and unemployment counts, continuing to consider log linear models.

Let $\mathbf{N}_r$ be the vector of the ratios with i^{th} element $\left(\frac{N_{ui}}{N_{pi}}\right)$.

One extension gives: $Ln\left(\mathrm{E}\left[\mathbf{N}_r\right]\right) = \mathbf{1}\mu_{(e)} + \mathbf{X}_1\boldsymbol{\beta}_{1(e)} + \mathbf{X}_2\boldsymbol{\beta}_{2(e)}$ (9.2)

Equivalently $Ln\left(\mathrm{E}\left[\mathbf{N}_u\right]\right) = \boldsymbol{\mu}_{(e)} + \mathbf{X}_1\boldsymbol{\beta}_{1(e)} + \mathbf{X}_2\boldsymbol{\beta}_{2(e)}$ (9.3)

where $\boldsymbol{\mu}_{(e)} = \mathbf{1}\mu_{(e)} + Ln\left(\mathbf{N}_p\right)$ (9.4)

As mentioned previously the intuitive advantage of this model is that it models something more stable ie departures from the average rates, for example of employment or unemployment, over small areas or cells, rather than the conditional probability of belonging to some cell given that a person is employed, or unemployed as in model 9.1.

If $\mathbf{N}_p = \mathbf{1}N_p$ (i.e. equal population sizes in all cells) then models (9.1) and (9.3) are identical provided the underlying processes for $\mathbf{N}_u$ and $\mathbf{N}_p$ have the same statistical distributions. Note that model (9.2) (or equivalently (9.3)) treats $\mathbf{N}_p$ as a vector of fixed constants. Extension to the case where $\mathbf{N}_p$ (like $\mathbf{N}_u$) is treated as random provides a wider set of models. These are discussed below.

Before leaving model (9.3) however, note that $\mathbf{N}_p$ might be the sum of employed and unemployed in each cell or where only one of employed or unemployed numbers is available for all cells, it may be the population count (employed + unemployed + other) in each cell. From this point of view it provides an extension over model (9.1) even when some information on employed and unemployed numbers is unavailable, provided some reasonable measure of cell population size is available. Because model (9.3) uses information not used in model (9.1), if population size information is known and reliable, estimates for small areas (for example of unemployed) will generally be more accurate under model (9.3). Note too that changes in $\mathbf{N}_p$ between earlier census and current survey can also be incorporated into

model (9.3). Immigration and emigration between cells can be explicitly accounted for by incorporating the new estimate of the population from the survey data $\mathbf{N}_{ps}$, into the estimates of the parameters. The same principle applies when the two estimates are of closely related variables as in the data that has been used in this thesis

When $\mathbf{N}_p$ is treated as random rather than fixed, equation (9.2) still applies but it cannot be rewritten in the form of equation (9.3) without introducing a random quantity into the right hand side.

Let $\mathbf{N}_{r2}$ be the vector with ith element $\dfrac{N_{ui}}{N_{pi}-N_{ui}}$.

An alternative is then to reparameterise (9.2) as

$$Ln\left(\mathrm{E}\left[\mathbf{N}_{r2}\right]\right)=\mathbf{1}\mu_{(r2)}+X_1\boldsymbol{\beta}_{1(r2)}+X_2\boldsymbol{\beta}_{2(r2)} \qquad (9.5)$$

Because $\dfrac{N_{ui}}{N_{pi}-N_{ui}}=\dfrac{p_{i0}}{1-p_{i0}}$ where $p_{i0}=\dfrac{N_{ui}}{N_{pi}}$, equation (9.5) can be rewritten as

$$Logit\left(\mathrm{E}\left[\mathbf{p}_0\right]\right)=\mathbf{1}\mu_{(r)}+X_1\boldsymbol{\beta}_{1(r)}+X_2\boldsymbol{\beta}_{2(r)} \qquad (9.6)$$

where $Logit\left(\mathbf{p}_0\right)$ has ith element $Ln\left(\dfrac{p_{i0}}{1-p_{i0}}\right)$

So equation (9.5) is a logistic regression model (with continuous and/or categorical explanatory variables). It is also a relative risk model because in each "cell" it measures the relative risk of being unemployed.

Note however, that models (9.3) and (9.5) differ in another respect namely that $\mathbf{N}_u$ and $\mathbf{p}_0$ have different distributions. In the simplest case $\mathbf{N}_u$ is related to the Poisson and $\mathbf{p}_0$ to the binomial, but where replicate sample survey, or census, information is available in each "cell" the distributions may, for example, be normal with different variances in each cell. Replicate observations give a direct estimate of the variance instead of relying on the variance function of the Poisson or binomial distributions. The actual distribution may be close enough to a Normal distribution in both cases but

unless the replicate sample is large the direct estimate may not be better than the variance function.

Where both $\mathbf{N}_u$ and $\mathbf{N}_e$ are known in each "cell" then the relative risk model is preferable. If the total for each cell includes an "other" category then the relative risk of being unemployed (relative to employed) is not estimable, unless the total population is also known, and model (9.3) is likely to be preferable. If all categories, employed, unemployed and other are available in each cell, a log linear model (or a model involving a transform of the generalizable linear model type) is best, because it makes use of all of the available information.

In all cases immigration and emigration from cells (or their analogue, changes in the values of explanatory variables, for continuous explanatory variables) can be allowed for in estimates by incorporating $\mathbf{N}_{ps}$, or its affine transform, as part of the appropriate model estimates. In the case of the relative risks model estimation of $\mathbf{N}_u$ proceeds from the model estimates of $\mathbf{p}_0$ via $N_{ui} = p_{i0}N_{pi}$ for each cell *i*. In other cases $\mathbf{N}_u$ is explicitly part of the model.

In fact although the logistic model for the risks is intuitively attractive it is easier to program the count model of the expanded table with both employed and unemployed. The two approaches are equivalent (see above, Chapter 3 and Bishop Fienberg and Holland(1975) for further detail) so it makes no difference which is used, given the marginal totals.

9.3 A simple example

To demonstrate the differences, and hopefully the advantages, in using a relative risks approach we will begin by considering the two by two by two table with four cells of "unemployed" and four cells of "employed". We will assume that we have the best data available from the survey, that is the new margins for the unemployed along with the total counts of employed plus

unemployed for all four cells. This is shown in *Figure 9.1* below. The census data would be known for all cells in the table as before. With this data available the margins for the employed group can be found by simple arithmetic, these are used in the program to estimate the parameters in the model.

For the two by two by two table the model can be written as

$$Ln\left(\mathrm{E}\left[n_{ijk}\right]\right) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$$

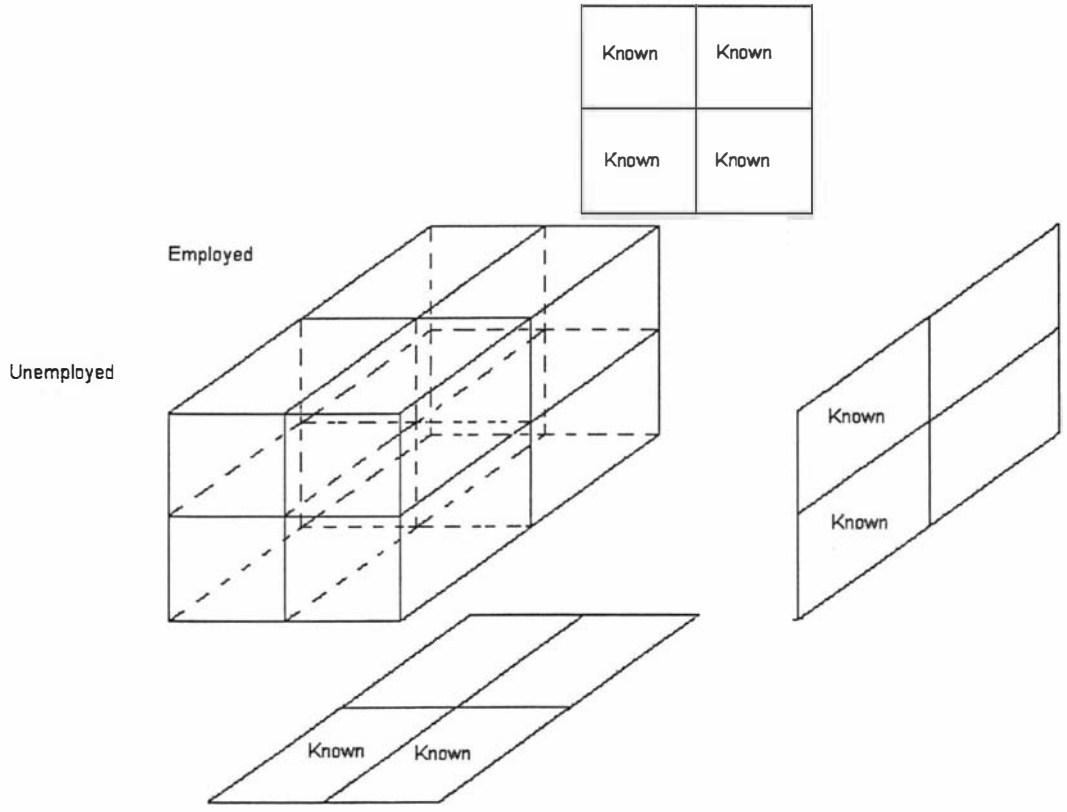


Figure 9.1 Survey data available for the relative risks model.

For a simple example let the data in the original table be 8, 6, 5, 4 for the unemployed and 75, 58, 40 and 25. The new margins from the survey are 19, 25, 20 (and 24 which is unnecessary as 19 + 25 has to be the same as 20 + 24), 90, 60, 50 and 30. The program for this analysis is in the Appendix C and the results are below. The analysis was repeated with three sets of initial values and it can be seen in the "History" graph for beta[1] below that one of the

chains took some time to converge. One of the sets of initial values was some way away from the final estimates and so the chain wandered around the space before converging. This is a positive result as it suggests that the chain is moving widely around the space and so is less likely to have converged at a local maximum. As all three sets of starting values converged to the same estimates these estimates are robust to the starting values. The coefficient beta[1] is the constant term in the model for the census data. The graphs for the other parameters monitored were similar in form to this.

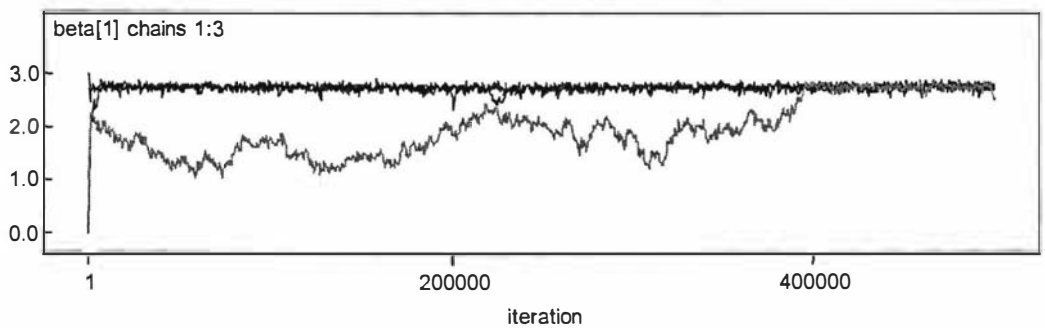


Figure 9.2 History graph for the constant term in the model for the census data

Whilst it may be a concern that convergence was slow for one of the chains note that the initial values for the coefficients in the two models were unlikely and so the fact that all three did eventually converge to the same values is encouraging.

The complete results of the simulations are given below. The tables of the model coefficients beta[ ] and beta[ ] are for the census and survey data respectively. There is only one coefficient carried forward from the census data to the new model and seven coefficients are found from the new margins, in this case.

start	400000	**thin**	500

sample	600					
node	**mean**	**sd**	**MC error**	**2.5%**	**median**	**97.5%**
beta[1]	2.746	0.05834	0.002992	2.626	2.752	2.849
beta[2]	0.1652	0.05582	0.00321	0.05399	0.1642	0.282
beta[3]	0.303	0.05402	0.002742	0.1971	0.3002	0.4186
beta[4]	-1.087	0.05732	0.002971	-1.202	-1.082	-0.9873
beta[5]	-0.02215	0.05694	0.002958	-0.1336	-0.022	0.08804
beta[6]	-0.03017	0.05588	0.003197	-0.1382	-0.02841	0.08644
beta[7]	-0.05326	0.05314	0.002722	-0.155	-0.05791	0.06111
beta[8]	0.04435	0.05719	0.002931	-0.06731	0.04429	0.1567

start	400000		thin	500		
sample	600					
node	**mean**	**sd**	**MC error**	**2.5%**	**median**	**97.5%**
Betaa[1]	3.031	0.01476	7.329E-4	3.001	3.031	3.06
Betaa[2]	0.1237	0.02225	8.982E-4	0.08041	0.1236	0.1676
Betaa[3]	0.1624	0.01838	7.865E-4	0.1267	0.1616	0.2007
betaa[4]	0.6562	0.02465	0.001077	-0.7011	-0.6568	-0.6094
Betaa[5]	0.04741	0.035	0.00156	-0.1129	-0.04868	0.01997
betaa[6]	0.2151	0.03839	0.001536	-0.2897	-0.2144	-0.1367
betaa[7]	0.3026	0.02318	8.824E-4	-0.3496	-0.3006	-0.2567

Finally the new cell counts are given below.

start	400000		thin	500		
sample	600					
node	**mean**	**sd**	**MC error**	**2.5%**	**median**	**97.5%**
mu[1]	8.57	1.126	0.05029	6.604	8.522	10.95
mu[2]	10.35	1.318	0.05333	7.781	10.27	13.2
mu[3]	11.37	1.13	0.05494	9.136	11.4	13.57
mu[4]	13.56	1.24	0.06703	11.3	13.56	16.13
mu[5]	81.37	1.487	0.06195	78.4	81.41	84.26
mu[6]	49.67	1.61	0.06527	46.56	49.79	52.81
mu[7]	38.59	1.457	0.0646	35.7	38.56	41.46
Mu[8]	16.38	1.575	0.07464	13.21	16.44	19.38

The analysis given is no different in practice to the analysis of a two by two by two table in which the third dimension is another variable such as ethnicity not employment status. *Figure 9.1* shows the simple situation and the principle can be extended to any table. The addition of the "employed" category simply adds an additional dimension to the original table. We have shown, in section 9.2, that this is a relative risks model if the third dimension is the employment status. The data for the employed group may not be available in the same administrative register for the census data and the new margins may require additional data but these issues are dependent on the application and can be dealt with on an individual basis.

9.4 The data

Looking again at the New Zealand unemployment data that we have considered in other sections of this thesis we would need estimates of the employed in each cell of the table as well as unemployed to be able to carry out the analysis using a relative risk approach. Currently census data is not available for the employed group parallel to the unemployed data that WINZ can provide. Statistics New Zealand hope that soon they will have data available from the Inland Revenue Department, the collectors of personal tax, from which estimates could be made. For the purposes of this exercise we have used estimates of the population numbers from the census and from the HLFS estimates of the proportion or numbers of employed. Using the census data estimates of the employed can be found by simple subtraction. Similarly for the new margins the estimates of the employed are found by subtraction from the estimates of the total population. In this case we decided to only include some of the values for the margins as in the example in the previous section. If the numbers of males and females is known then the total is also known so the count for one of the age groups is unnecessary as is the count for one of the ethnicities. The overall total is known either from the two sexes or the two categories of ethnicity so WinBUGS is able to estimate all of the parameters.

The table is now 384 by 384 and the design matrix is constructed in the same way as previously. Other than these differences there are no changes from the previous programs that we have used. The program is given in Appendix C

9.5 Results

The results are given in the tables below. For each region the SPREE estimate and relative risks results are given. The SPREE estimate was found using the new algorithm. The tables are split into the regions on the two Islands.

		North Island											
		Male						Female					
	Ethnicity	Pakeha and Other			Maori and Pacific Island			Pakeha and Other			Maori and Pacific Island		
	Age	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over
Northland	SPREE	516	675	172	534	777	256	658	556	162	258	460	72
	Rel Risk	468	1055	297	663	1152	143	401	593	185	461	598	97
Auckland	SPREE	3511	4081	605	2373	2976	519	4713	3469	586	1384	1815	174
	Rel Risk	2519	6025	1156	2185	2766	433	2342	4028	895	2057	1750	317
Waikato	SPREE	1097	1382	376	714	1087	292	1128	1007	328	498	614	76
	Rel Risk	1161	1684	435	854	1077	132	1207	1072	347	704	652	94
Bay of Plenty	SPREE	961	1096	272	866	549	452	932	749	164	658	543	90
	Rel Risk	744	1212	358	864	1325	175	787	783	295	693	662	101
Gisborne	SPREE	183	234	38	248	275	157	156	133	23	236	290	28
	Rel Risk	143	241	65	301	524	76	136	140	33	227	302	50
Hawkes Bay	SPREE	590	866	160	435	493	324	566	592	119	301	508	27
	Rel Risk	588	924	262	520	725	93	531	483	161	384	398	61
Taranaki	SPREE	529	769	183	227	236	142	515	521	122	163	200	16
	Rel Risk	616	761	208	257	295	38	575	540	142	183	171	29
Manawatu	SPREE	993	1262	330	547	591	276	957	860	256	337	420	40
Wanganui	Rel Risk	1043	1486	393	585	715	92	992	921	179	429	402	64
Wellington	SPREE	1595	2573	543	788	703	486	1610	1520	197	467	583	86
	Rel Risk	1788	2661	647	921	1059	160	1524	1505	407	712	562	103

Table 9.1 *Estimates found by SPREE and the relative risks model for North Island regions.*

South Island

	Ethnicity	Pakeha and Other			Maori and Pacific Island			Pakeha and Other			Maori and Pacific Island		
	Age	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over	15 - 24	25 - 49	50 and over
Tasman	SPREE	205	254	81	48	33	19	172	130	41	22	26	2
	Rel Risk	154	346	87	25	41	5	151	204	49	19	20	4
Nelson	SPREE	222	459	96	50	37	21	147	177	43	26	30	2
	Rel Risk	218	331	90	40	54	4	225	228	59	30	22	2
Marlborough	SPREE	212	355	72	57	64	26	154	182	42	35	48	2
	Rel Risk	183	319	97	50	61	8	172	197	61	35	32	5
West Coast	SPREE	293	435	91	48	37	27	169	176	42	27	23	2
	Rel Risk	214	459	109	32	48	8	173	224	64	28	29	4
Canterbury	SPREE	2108	4131	1071	370	415	258	1633	1919	474	333	339	29
	Rel Risk	2392	3343	855	433	474	72	2544	2310	615	369	264	33
Otago	SPREE	1239	1659	463	159	122	94	744	723	187	122	108	11
	Rel Risk	1081	1555	418	135	147	23	1001	1009	240	103	84	16
Southland	SPREE	115	1807	54	17	2098	437	458	3306	106	221	6	8
	Rel Risk	509	747	233	135	187	39	519	501	140	106	89	12

Male F

Table 9.2 *Estimates found by SPREE and the relative risks model for South Island regions.*

It is hard to make much sense of these results in the table form above but a graph of the relative risks estimates against the SPREE estimates suggests that the majority change little but that a few have quite large changes.

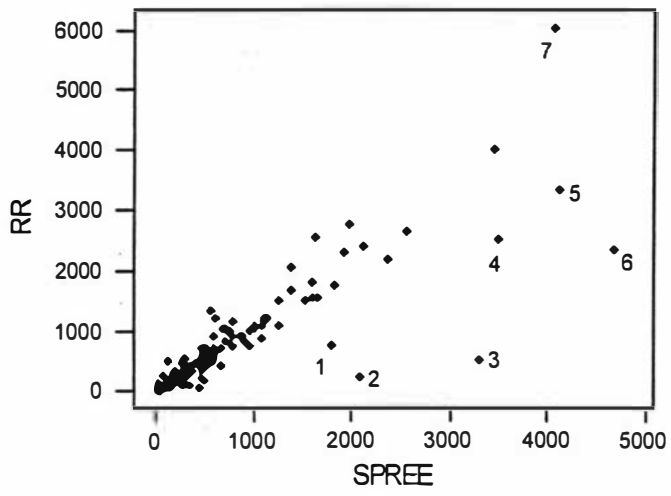


Figure 9.3 Graph of relative risks model vs SPREE, numbered points are noted below.

The points on the graph that have been labeled one to seven appeared, by inspection, to be points where the relative risks model and SPREE do not predict similar values.

Point	Relative Risks Estimate	SPREE Estimate	Description
1	747	1807	Southland, Male, 25 to 49, Pakeha and other
2	187	2098	Southland, Male, 25 to 49, Maori and PI
3	501	3306	Southland, Female, 25 to 49, Pakeha and other
4	2519	3511	Auckland, Male, 15 to 24, Pakeha and other
5	3343	4131	Canterbury, Male, 25 to 49, Pakeha and other
6	2342	4713	Auckland, Female, 15 to 24, Maori and PI
7	6025	4081	Auckland, Male, 25 to 49, Pakeha and other

Auckland and Southland are represented three times each. The Auckland and Canterbury results reflect the suggestion made earlier that areas which have particularly large or small populations may not be well estimated if the total population is not taken into account. This may also explain the Southland results where the changes are even greater in percentage terms. SPREE seems to be overestimating them by a large amount.

There are probably a number of other points which deserve careful consideration if the intention is to justify a model. Repeating comments made earlier if the method was to be used to produce estimates of unemployment for public release more work is required to validate the model for this data.

We did not expect to change everything by a large amount because the estimates must still conform to the fixed numbers in each margin. Adding in the information about the total population should improve the estimates, and it is where the total population is very different that large changes have occurred.

9.6 Discussion

In this chapter we have introduced the concept of relative risk as an alternative approach to modeling the unemployment data. We have shown that the methods that we have introduced in this thesis can to be applied to this model as effectively as to the other models that have been used.

SPREE estimation using the IPF could be used for fitting relative risk models as we have shown that the relative risks model can be fitted simply by modeling the enlarged table including both employed and unemployed. The Bayesian approach that we have used in this chapter has some advantages that we have discussed earlier. The approaches in Chapters Five and Eight are examples of possible extensions. There is the potential to use combinations of the techniques eg to combine the relative risks model with the spatial

autoregressive type errors or to use continuous variables in the relative risk model.

The combination of models gives the analyst a much wider range of tools to apply to a particular situation. We have not used any of these more complicated models but there are no reasons in principle why these approaches cannot be included along with the relative risks in an analysis, without any further extension of the theory being necessary.

Relative risk approaches have commonly been used in epidemiology and in small area estimation in disease mapping Rao (2003) but they do not seem to have had much use with unemployment data. This seems a little surprising as there does not seem to be a conceptual difference between the risk of contracting a disease the risk of becoming unemployed.

The real difference between modeling the counts and the relative risk is in the definition of the extreme values. With the count data large centres have large counts simply because the population is large. With the risk data large values are those areas with high unemployment relative to their size.

CHAPTER 10

Conclusions

10.1 Introduction

In this thesis we have achieved a number of things.

- We have repeated the work of Marker (1999) in placing the common methods of small area estimation into a single model formulation, that of a linear regression model. We have simplified his notation.
- By extending the model to a GLM we have included SPREE with the other methods which had not previously been done.
- By expressing SPREE in the form of a GLM we have shown that there is an alternative algorithm which can be used to estimate the cell values in the contingency table formed by the categorical variables which are typical of the SPREE approach.
- This new approach
  - allows the underlying concept of SPREE to be applied in a wider range of situations as it is not restricted to categorical variables;
  - the underlying process is clearer and this helps to understand the IPF algorithm;
  - it makes the assumptions inherent in SPREE more transparent and hence enables practitioners to assess their validity in any particular analysis;
  - it has facilitated an investigation of the relationship between the two data sources and shown that commonly held beliefs,

relating to correlation between data from different sources, are not necessarily accurate;

- Although we have fitted saturated models, as is typical in SPREE the new approach is better able to identify the important variables in the model.

- We have used a Bayesian approach to the estimation which has allowed us to estimate the errors in the estimated cell counts.

- Conditional Autoregressive and Relative Risk models have been applied to illustrate the wider range of models that can be used with the new approach.

- The work has been extensively illustrated with an example using data from Statistics New Zealand and the Department of Work and Income in New Zealand.

We shall cover each of these in a separate section and then make a few general comments and make suggestions of areas which require further work.

10.2 The linear regression framework

In Chapter Two we introduced the simple linear model and showed that most of the traditional small area estimation methods can be written in this form. This work had previously been done by Marker but by expressing the model in a simple form and then using this form throughout the chapter we have simplified the notation. Some of this is a little contrived as the original method may have been expressed in a simple way and to show that it is in fact an example of a linear regression model requires some algebraic manipulation which loses some of the original simplicity. The Vital Rates method in section 2.2.1 is an example of this. The use of the relative birthrates in different areas to indicate population change is a simple concept but the expression of this method in the framework is less simple.

Most of the early methods of small area estimation were devised with a particular application in mind and for specific data structures. By expressing these methods in terms of one model it makes it easier to apply the principles to new situations and new data sets. These methods are not widely used now but were part of the development of more modern approaches. All of the methods improve on the direct estimates by utilizing auxiliary information. Broadly speaking there are two sources for this auxiliary information. It may arise from other areas within the same data set or from additional data most commonly administrative records.

Most of these traditional methods use administrative data either from a census or from other administrative records or both. Commonly they use more recent related data to update an earlier census. The available auxiliary data has determined the form of the method used and most are quite specific to their application.

10.3 The extension to Include SPREE

The main advantage of writing the traditional methods in the form expressed in Chapter Two is that it is easily extendable to the GLM. This allows SPREE to be included in the general form. This had not previously been done.

We have expressed SPREE in terms of equations **Error! Reference source not found.**, **Error! Reference source not found.** and the assumption in **Error! Reference source not found.** that is:

We would write $g\left(\mathrm{E}\left[\mathbf{Y}_c\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1c} + \mathbf{X}_2\boldsymbol{\beta}_{2c}$ for the census data

Error! Reference source not found.

and $g\left(\mathrm{E}\left[\mathbf{Y}_s\right]\right) = \mathbf{X}_1\boldsymbol{\beta}_{1s} + X_2\boldsymbol{\beta}_{2s}$ for the survey data.

Error! Reference source not found.

We see that the SPREE model, via the IPF algorithm, is making the assumption that the second term in each of the models above is the same, that is

$$\mathbf{X}_2\boldsymbol{\beta}_{2c} = \mathbf{X}_2\boldsymbol{\beta}_{2s}$$

Error! Reference source not found.

SPREE is restricted to categorical variables which define cells in a contingency table. The new approach removes this restriction to categorical variables and hence allows a much wider range of models. We have fitted a quadratic term to the age variable in our example in Chapter Five and considered CAR error structures in Chapter Eight neither of which could be fitted using the IPF algorithm. Thus we have shown that this approach allows the broader the range of models that we suggested. These are only two of a much larger group of possible models that could be used with this new approach.

An additional advantage of this new approach is the way it reveals the underlying process in SPREE. A senior SNZ statistician saw a draft version of some of this work and immediately declared that for the first time he understood what SPREE, using the IPF algorithm, was doing. It is not easy to see which cross product ratios will remain constant in the IPF algorithm and what the interpretation of those cross product ratios is. It is much easier to see which terms in the model can be re-estimated from the survey data and so see which are being estimated from the census data. This allows us to at least intuitively critique the assumption even if, as has been shown in Chapter Six, it is not easy to check it rigorously.

An important aim of statistical methods is to understand the underlying process. Explaining the results of sophisticated techniques to users is often difficult so a method which makes this clear has advantages. It is of note that in Nelder and Wedderburn (1972) an entire section is written expounding the advantages of the generalized linear model as a tool for teaching statistics.

Whilst not suggesting that this thesis should be compared in any way with that paper I support the idea that understanding the processes involved in an analysis is important. If these processes are easily understood then there are advantages to all users. Education can be considered in its broadest sense and if the method improves the understanding of the analyst then it should be encouraged. The generalized linear model approach to SPREE estimation makes the procedure much clearer and thus should help analysts to understand their work and make it easier to explain to others.

10.4 The wider application of the new algorithm

This new algorithm is more general than SPREE using the IPF algorithm. SPREE is confined to contingency tables and so all variables must be categorical. The concept underlying SPREE is to model the census data and then use some of the structure from that model in a new model for the survey data. Structure that is found in the survey data is used, where there is insufficient data the structure from the census data is used. This is achieved by using the IPF algorithm to adjust the cell counts in the table of census data to fit the new margins from the survey.

By explicitly modeling the data using the log linear model we have shown that the new margins are sufficient statistics for the terms in the model associated with those margins when the other terms are estimated from the census data. This process could be applied to a much wider range of problems than just small area estimation. Models could be developed using a small amount of data collected and then estimating terms from a model for similar data.

The new approach has allowed new modeling strategies to be employed and a few examples of these have been presented. Whilst it may be possible to use the IPF algorithm as part of a modeling process which includes some of the

extensions which we have suggested the application is very simple using the new approach. The examples in Chapters Five, Eight and Nine are possibilities which hopefully signpost the sorts of models which could be applied.

The Bayesian approaches to estimation which we have used in the latter part of this thesis are not easily applied to traditional SPREE estimation. They do have some advantages which we have illustrated by the ease of fitting the CAR models in the Bayesian paradigm and also the ease with which variance estimation is achieved. The potential also exists to combine information from more than two data sources in the Bayesian approach which could well be useful as there are moves by many statistical agencies to collect data from a range of administrative registers. There is also the possibility of being able to use more than one survey instrument. These advances will no doubt generate their own problems but this approach at least suggests a direction to begin.

10.5 The assumptions in SPREE and the relationship between the two data sources

The assumption made when SPREE, using IPF, is used is that there is some structure in the table of census values which is assumed to be the same in the equivalent table of values that could be found from the survey if there was sufficient data. This structure is determined by various ratios and cross product ratios of cell counts in the table. We have shown that this is equivalent to some terms and interactions in the log linear model for the survey data being assumed to be equal to the equivalent terms and interactions in the model for the census data.

Our approach using the log linear model has shown that the structure in the table can be expressed as coefficients in a model and that some of these coefficients are assumed to be the same for the census and survey data.

We have found that it has sometimes been assumed that the correlation between the cell counts in the table for the census data and the equivalent, but unknown, cell counts in the survey data is a useful indicator of the effectiveness of the method. This correlation may be found by considering past results. By writing the method in terms of the log linear models it has been shown that this is not a good measure of how well the method will work.

The new approach casts more light on the relationship between the two sources of data. We have shown that correlation between the census and survey data does not necessarily mean that the method would work well. There are circumstances in which the method can work well when the correlation is low and conversely the method may work poorly when the correlation is high.

We have shown that the assumption being made is $\mathbf{X}_2\boldsymbol{\beta}_{2c} = \mathbf{X}_2\boldsymbol{\beta}_{2s}$ and that given the data available it is not possible to check this. The ease with which this assumption can be written when the algorithm is used in this form is an advance. It enables a closer inspection of the assumption being made and there may be situations in which it can be partly assessed for accuracy.

The part of the model expressed in $\mathbf{X}_2\boldsymbol{\beta}_2$ is largely made up of higher order interaction terms. Depending on the data sources it may be possible to compare some or all of these higher order interactions to gain a better understanding of the efficacy of the method.

Using the IPF algorithm the assumption is defined in terms of cross product ratios in a large contingency table. Assessing how well the assumption holds is not easy firstly because the data is not available as has been explained but also because the interpretation of the cross product ratios in a large table are not intuitive.

Finally we would note that we suspect that some authors have probably used the term correlation to suggest that the internal structure in the two tables is similar rather than in the strict statistical sense of the term. This has not always been made explicit in the literature and we would suggest that if this is the case it should be made clearer.

10.6 Bayesian approaches and variance estimation

A substantial part of this thesis has used a Bayesian approach to the analysis. In Chapter Seven we explained why we felt that a Bayesian approach was sensible for the data that was available. The use of Bayesian methods has increased rapidly in recent years with increasing computer power which has allowed MCMC methods of analysis to be used. There is an increasing acceptance of these approaches in many application areas as they may be simpler to apply.

The form of the data that has been collected suggests plausible prior distributions for the data and model parameters. The repeated census data is unusual and gives estimates of variability directly. The design of the sample survey is such that variance estimates are also readily available.

One advantage for us was that the conditional autoregressive models which we wanted to investigate were easily fitted in WinBUGS. Whilst this may have been the initial attraction to using a Bayesian analysis it was clear from the start that the data leant themselves to this approach.

Another advantage, which was not part of the decision to move to a Bayesian approach, is that by using MCMC techniques the estimation of the variance of the estimates is an automatic byproduct of the process. The simulation, once it has converged, results in a large number of draws from the posterior distribution for each parameter in the model and these can be stored if necessary. Thus there is a large sample from the distribution and it can be

used to estimate the variance by simple calculation from the sample. This is true not only for the parameters in the model but any defined functions of them. In this way the individual cell counts can be simulated along with sums of them in our case the regional totals. By including these in the simulations variance estimates of these quantities are immediately available. We had not initially considered the variance estimation problem and by taking the Bayesian approach this was neatly solved.

10.7 Conditional autoregressive and relative risk models

In Chapters Eight and Nine two extensions to the basic model have been illustrated. These are two of a wide range of possible models, and combinations of models, which could be applied using the new approach. The applicability of these or other model stop particular data sets will depend on the situation but we have shown how they can be fitted in principle.

10.8 Comments about the data used in this thesis and practical considerations

It has been said a number of times through this work that the example used to illustrate the ideas is intended only to show that the techniques can be used on a real data set and that we have not followed through the analysis as we would if the objective was to produce small area estimates for unemployment in the regions.

The data that we have used in our examples is not ideal for the purpose of estimating unemployment counts as we only had eight months of census data available. We would expect unemployment figures to be seasonal, particularly in some parts of the country where seasonal crops are grown, it would be an improvement to have at least one year of data so that a sensible model could be fitted to the census data. More than one year of data would be desirable to assess the seasonal model but this was not available to us.

However this does not negate the work that we have done as this will simply change the coefficients in the model for the census data which are carried forward to the model for the survey data and improve estimates by reducing variability among census "replicates".

The approach which has been described in this thesis is a method of obtaining synthetic estimates which will be biased if the coefficients which are carried forward from the model for the census data are not exactly correct. It is hoped that the bias will be small and these estimates would be useful to use along with the unbiased direct estimates from the survey, where they exist, to form a composite estimate. Much has been written about these forms of estimates see Longford (2002) for a detailed exposition.

Composite estimates can be formed directly from the survey and auxiliary data however for these to be useful the bias in the estimates from the auxiliary data must be small. The method that we propose does not require the bias to be small as long as the bias is related to the terms in the model that are re-estimated from the survey data. For example it is easy to imagine auxiliary data that consistently over or under estimates the counts by the same amount. This would simply change the value of the constant term in the model and hence would be of no concern as this will always be estimated from the survey data.

10.9 Final conclusions and suggestions for future work

As this piece of work reaches its conclusion it becomes ever clearer that the end is simply an arbitrary point in time and that there are many areas of investigation that should be followed up. The thesis has proposed a new approach to SPREE estimation and shown that this approach has some advantages over the earlier algorithm. A few examples have been given

however these have simply been indicative of the possibilities available with the new approach.

There has been no attempt to actually answer the question "Can we use this method to produce good estimates of unemployment at the small area level?" As has been explained above the data which we had was not conducive to answering this question in an effective way.

There are a number of specific areas of further work which should be carried out. These are all applications of the new approach and so do not represent any new work. A number of suggestions are included below.

- A full analysis of the 74 Territorial Authorities is now possible. We had problems earlier due to computing power but with more recent machines this problem can be attempted. Haslett, Green and Zingel (1998) found estimates of the variance of the SPREE estimates as discussed in Chapter Seven but using slightly different data. The Bayesian estimates of variance should be checked on the same datasets to see if they agree with their estimates
- Access to raw data for the survey to look at correlations and parameter estimates in the survey data using the same model as for the census. Although they will not be estimated very accurately they will be unbiased and so should be able to be compared at least qualitatively.
- Use of census data to assess the accuracy of the estimates using various methods. Data has been collected in the recent census in New Zealand which would make it possible to compare the estimates with an accurate figure.
- Considerations in the survey design which may make checking the models better. The work in Chapter Six on checking the assumptions was inconclusive in strategies to use for checking the assumptions. It

does enable the analyst to investigate the important variables in the census data and to adjust the survey design to check that these variables are behaving correctly in the survey data.

- In Chapter Eight we considered a simple model for spatial adjacency. There are other definitions of proximity such as similar ethnicities, or more subjectively regions with similar employment profiles, which may yield more useful results when incorporated in a CAR error term.
- The relative risk models have only been considered as far as ensuring that it is possible to use the method to fit them. The data used was not very sophisticated and it was not expected that they would yield useful results from a practical point of view. However the approach could be expected to be valuable given good estimates of the counts of employed or good estimates of the totals of employed and unemployed for each combination of the other categorical variables. These models also have the potential to estimate a number of categories of employment such as part time employed, there is no restriction on the table being divided in two by the categorization so there can be a number of possibilities which better represent the true employment situation. Of course to achieve this information is needed on all of these categories. This may require the combination of data from a number of sources, IRD the income tax department in New Zealand, WINZ, SNZ and others. This brings with it many problems such as data matching and variable reliability of the different data sources, but in principle there is nothing to prevent the approach being used. The Bayesian methods which we have used are particularly appropriate when the sources of data are of different levels of reliability. The priors can be adjusted to take this into account.

- Composite estimation using these estimates and the raw survey estimates this would ensure that in the few areas where there were large sample sizes the estimates were close to the survey estimates and that where there were small sample sizes the SPREE type estimates would dominate. One problem of this is that totals may not agree with national totals but the small area estimates could be rescaled to allow for this.

Bibliography

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M. and Wang, S. (2001). Combining unemployment benefits data and LFS data to estimate ILO unemployment for small areas: An application of a modified Fay-Herriot method. *Proceedings of the International Association of Survey Statisticians, Meeting of the International Statistical Institute*, Seoul, August 2001.

Anderson, T.W. and Darling D.A.(1954). A Test of Goodness of Fit. *Journal of the American Statistical Association* **49**, 765 - 769.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28 - 36.

Bernardo, J.M. and Smith, A.F.M. (1994). Bayesian Theory. John Wiley and Sons.

Besag, J. (2001). Markov Chain Monte Carlo for Statistical Inference. Center for Statistics and the Social Sciences Working Paper No. 9, *University of Washington*

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 4, 733 - 746

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society Series B.* **25** 220 - 233

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). Discrete Multivariate Analysis. MIT Press, Cambridge, Massachusetts.

Bogue , D.J. (1950). A technique for making extensive postcensal estimates. *Journal of the American Statistical Association* **45,** 149 - 163.

Bogue, D.J. and Duncan, B.D. (1959). A composite method of estimating postcensal population of small areas by age, sex and colour. Vital Statistics –Special Report 47, No. 6, National Office of Vital Statistics, Washington, DC.

Browne, W.J. and Draper, D. (2001) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Technical Report, Institute of Education, London.

Cassel, C-M, Samdal, C-E, and Wretman, J. H. (1977). Foundations Of Inference In Survey Sampling. John Wiley, New York.

Chambers, R.L. and Feeney, G.A. (1977). Log linear models for small area estimation. Unpublished paper, Australian Bureau of Statistcs.

Chaudhuri, A. (1992). Small domain statistics: a review. Technical report ASC/92/2, Indian Statistical Institute, Calcutta.

Congdon, P. (2001). Bayesian Statistical Modelling, John Wiley and Sons, Ltd.

Cressie, N. (1993). Statistics for Spatial Data, John Wiley and Sons, Ltd.

Darroch, J.N. and Ratcliffe, D. (1972). Generalised iterative scaling for loglinear models. *Annals of Mathematical Statistics.* **43**, 1470 - 1480

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the exact expected marginals are known. *Annals of Mathematical Statistics.* **11**, 427 - 444

Drew, D., Singh, M.P. and Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology* **18**, 17 - 47

Ericksen, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association* **69**, 867 - 875.

Fienberg, S.E. (1970). The Analysis of Cross-Classified Categorical Data. The MIT Press.

Fienberg, S.E. (1980). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*. **41**, 907 - 917.

Fuller, W.A and. Harter, R.A. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics* (Platek R., Rao J.N.K., Samdal C-E. and Singh M.P. eds) Wiley, New York

Ghosh, M. and Rao, J.N.k. (1994). Small area estimation: an appraisal. *Statistical Sciences* **9**, 55 - 93.

Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B. (1998). Generalized Linear Models for Small-Area Estimation. *Journal of the American Statistical Association* **93**, 273 - 282.

Gilks, W.R., Richardson S. and Spiegelhalter D.J. (1996). Markov Chain Monte Carlo in Practice. Chapman and Hall.

Goldstein, H. (1995). Multilevel Statistical Models. Arnold.

Gonzalez, M.E. (1973). Use and Evaluation of synthetic estimators. *Proceedings of the Social Statistics Section* 33 - 36. American Statistical Association, Washington, DC.

Haslett, S. (2003). Personal Communication.

Haslett J. and Haslett S. (2004). *Linear Algebra and its Applications.*

Haslett, S., Green, A., and Zingel, C. (1998). Small Area Estimation Given Regular Updates of Census Auxiliary Variables. *Proceedings of the New Techniques and Technologies for Statistics Conference. November, Sorrento, Italy,* 206 - 211.

Holt, D., Smith, T.M.F. and Tomberlin (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association* **74**, 405 - 410.

Ireland, C.T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*. **55**, 179 - 188.

Kelsall, J.E. and Wakefield, J.C. (1999). Discussion of "Bayesian models for spatially correlated disease and exposure data", by Best et al. In *Bayesian Statistics 6* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith ed.), 151, Oxford University Press. Quoted in "Bayesian approaches to disease mapping" Wakefield, J.C., Best, N.G. and Waller, L. in *Spatial Epidemiology Methods and Applications* P. Elliot, J.C. Wakefield, N.G. Best and D.J. Briggs (2000), Oxford University Press.

Lahiri, P. (1996). Small-Area Estimation with an Application. Lecture notes for a Presentation at Statistics New Zealand, Wellington NZ.

Lindley, D. V. (1985). Making Decisions. John Wiley New York.

Longford, N. T. (2002). Small-area estimation in national surveys. Short Course at the RSS Conference, Plymouth

Marker, D. A. (1983). Organization of small area estimators. *Proceedings of Survey Research Methods Section.* 409 - 414. American Statistical Association, Washington DC.

Marker, D. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics.* **15**, 1 - 24.

McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman and Hall, London.

McCullagh, P. and Zidek, J. (1987). Regression methods and performance criteria for small area population estimation. In *Small Area Statistics* (Platek, R., Rao, J.N.K., Samdal, C. E. and Singh, M. P., eds) 62 - 74. Wiley New York.

Mollie, A. (1996). Bayesian Mapping of Disease. In *Markov Chain Monte Carlo in Practice.* (Gilks W.R. , Richardson, S and Spiegelhalter, D. J. eds) 359-380 Chapman and Hall, London.

Morrison, P. (1971). Demographic information for cities: A manual for estimating and projecting local population characteristics. RAND report R-618-HUD

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association* **63**, 1 - 28.

National Research Council (1980). *Panel on Small-Area Estimation of Population and Income. Estimating Population and Income of Small Areas.* National Academy Press, Washington, DC.

Nelder, J. A. and Wedderburn, (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A.* 135, part 3, 370 – 384.

Noble, A. D, L., and Haslett, S. J. (2001) A Generalized Linear Model Approach to Small Area Estimation using Census and Survey Data and Conditional Autoregressive Models. Poster presented to A Conference in Celebration of Wayne Fuller's Seventieth Birthday, Iowa, June 2001.

Noble, A.D.L., Haslett S.J. and Arnold G. (2001). Extension of Small Area Estimation using Generalized Linear Models. To Spatially Correlated Data. Paper presented to the Royal Statistical Society Spatial Modelling theme conference, Glasgow, July 2001.

Noble, A.D.L., Haslett S.J. and Arnold G. (2002). Small Area Estimation via Generalized Linear Models. *Journal of Official Statistics*

Pettit, A. N., Weir, I. S. and Hart, A. G. (2002). A Conditional Autoregressive Gaussian Process for Irregularly Spaced Multivariate Data with Application to Modelling Large Sets of Binary Data. *Statistics and Computing*.**12**, 353 – 367.

del Pino, G. (1989). The Unifying Role of Iterative Generalized Least Squares in Statistical Algorithms. *Statistical Science.* 4, 394-408.

Platek R., Rao J.N.K., Samdal C-E. and Singh M.P. (eds) *Small Area Statistics* Wiley, New York)

Purcell, N.J. (1979) Efficient estimation for small domains: a Categorical data analysis approach. Unpublished PhD Thesis. University of Michigan.

Purcell, N.J and Kish, L. (1980). Postcensal estimates for local areas (domains). *International Statistical Review* 48, 3 - 18.

Rao, C. R. (1965). Linear Statistical Inference and Its Applications. Second Edition. John Wiley and Sons, New York.

Rao, J.N.K. (1986), Synthetic estimators, SPREE and best model based predictors. In *Proceedings of the Conference on Survey Research Methods in Agriculture* 1 - 16. U.S. Dept. Agriculture, Washington, DC.

Rao, J.N.K. (1999). Some Recent Advances in Model-Based Small Area Estimation. *Survey Methodology* **25**, 175 - 186.

Rao, J.N.K. (2003). Small Area Estimation. Wiley Series in Survey Methodology. Wiley-Interscience, John Wiley and Sons.

Sarndal, C.E. and Hidiriglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association* **84**, 266 - 275.

Schaible, W.L. (1978). Choosing weights for composite estimators for small area statistics. In *Proceedings of the Survey Research Methods Section* 741 - 746. American Statistical Association, Washington DC.

Schaible, W.L. (1992). Use of small area statistics in U.S. Federal Programs. In *Small Area Statistics and Survey Designs.* (G.Kalton, J. Kordos and R.Platek, eds.) **1** 95 - 114. Central Statistics Office, Warsaw.

Searle, S. R. (1988). Best Linear Unbiased Estimation in Mixed Models of the Analysis of Variance. In *Probability and Statistics, Essays in Honor of Franklin A. Graybill.* (J.N. Srivastava Editor). Elsevier Science Publishers B. V. (North-Holland).

Smith, S.K. and Lewis, B.B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography* **17** 323 - 340

Starsinic, D.E. (1974). Development of population estimates for revenue sharing areas. Census Tract Papers, Ser.GE40, No.10 U.S. Government Printing Office, Washington, DC.

Thompson, R. and Baker, R. J. (1981) Composite Link Functions in Generalized Linear Models. *Applied Statistics*, **30**, 125 - 131

US Bureau of the Census (1966). Methods of population estimation: Part I, Illustrative procedures of the Bureau's Component Method II. *Current Population Reports, Series P.25, No. 339.* U.S. Government Printing Office, Washington DC.

Wolter, K. (1985). Introduction to variance estimation. Springer-Verlag, New York.

Zidek, J.V. (1982). A review of methods for estimating the populations of local areas. Technical Report 82 - 4, Univ. British Columbia, Vancouver.

Appendix A

Detailed calculations from Chapter 3

The Iterative Proportional Fitting Algorithm Examples

From Section 3.3 The simple example on SPREE by the iterative proportional fitting algorithm.

Page 50

There follows a simple example to illustrate IPF. In the simplest case with a 2 x 2 table we start with everything known viz:-

1	3	With the margins	1	3	4	Given a new:		5
5	2	added this gives	5	2	7	set of margins		8
			6	5			9	4	13

First Iteration	1.25	3.75	5
	5.7143	2.2857	8
	6.9643	6.0357	13

Continuing the iterations with the example above gives the following results.

1.6154	2.4852	4.1006	1.9697	3.0303	5
7.3846	1.5148	8.8994	6.6383	1.3617	8
9	4	13	8.6080	4.3920	13

2.0594	2.7598	4.8192	2.1367	2.8633	5
6.9406	1.2402	8.1808	6.7872	1.2128	8
9	4	13	8.9239	4.0761	13

2.1549	2.8098	4.9647	2.1702	2.8298	5
6.8451	1.1902	8.0352	6.8151	1.1849	8
9	4	13	8.9853	4.0147	13

2.1738	2.8194	4.9932
6.8262	1.1806	8.0066
9	4	13

2.1768	2.8232	5
6.8206	1.1794	8
8.9974	4.0026	13

2.1774	2.8214	4.9988
6.8226	1.1786	8.0012
9	4	13

2.1779	2.8221	5
6.8216	1.1784	8
8.9995	4.0005	13

From Section 3.7
Page 75
IPF fitting for a 2 x 2 x 2 table

Original					19	25		
Table					20	18		
			5	7				
			2	1			15	24
	8	6						
	5	4					23	20
Iteration 1								
					19	25		
					22.5043	21.4956		
			8	11.2				
			3.2	1.6			24	24
	6.9565	5.2173						
	4.3478	3.4782					20	20
Iteration 2								
					19	25		
					19	25		
			6.75425	13.0258				
			2.70170	1.86084			24.3426	24
	5.8732	6.0679						
	3.6707	4.0453					19.6573	20
Iteration 3								
					19	25		
					19.0332	24.9667		
			6.65916	12.8425				
			2.66366	1.83464			24	24
	5.9756	6.1737						
	3.7347	4.1158					20	20
Iteration 4								
					19	25		
					19	25		
			6.64753	12.8596				
			2.65901	1.83709			24.0032	24
	5.9652	6.1819						
	3.7282	4.1213					19.9967	20
Iteration 5								
					19	25		
					19.0003	24.9996		
			6.64662	12.8578				
			2.65865	1.83684			24	24
	5.9661	6.1829						
	3.7288	4.1219					20	20

Iteration 6

				19	25		
				19	25		
		6.64651	12.8580				
		2.65860	1.83686			24.0000	24
5.9660	6.1830						
3.7287	4.1220					19.9999	20

The Generalized Linear Model Calculations

From Page 76

To estimate the log linear model for this table we first need to calculate the log probabilities as shown below.

Log of the Counts

		1.60943	1.94591
		0.69315	0
2.07944	1.79176		
1.60943	1.38629		

The log linear model can be expressed as:-

$$Ln\left(\mathrm{E}\left[p_{ijk}\right]\right) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

Where

u	The overall mean	1.3894
$u_{1(1)} = -u_{1(2)}$	Main effect due to Employment Status	0.108438

= half of the difference of the averages for the left and right sides of the table. ie

½((2.07944 + 1.60943 + 1.60943 + 0.693147)¼ - (1.79176 + 1.38629 + 1.94591 + 0)¼)

$u_{2(1)} = -u_{2(2)}$	Main effect due to Sex	0.467209
$u_{3(1)} = -u_{3(2)}$	Main effect due to Ethnicity	0.327305

The interactions are calculated as the average deviance from the two associated main effects

$u_{12(11)} = -u_{12(12)} = -u_{12(21)} = u_{12(22)}$	Interaction between Sex and Employment Status	-0.120635
$u_{13(11)} = -u_{13(12)} = -u_{13(21)} = u_{13(22)}$	Interaction between Sex and Ethnicity	0.0192688
$u_{23(11)} = -u_{23(12)} = -u_{23(21)} = u_{23(22)}$	Interaction between Employment Status and Ethnicity	-0.248342

Finally the three way interaction can be seen to be the remaining deviance given all of the other effects.

$u_{123(111)}$	Three way interaction	0.136770

From Section 3.8
Page 78
Logistic Model

			Male	Female	Male	Female
Counts	Maori	Employed			5	7
		Unemployed			2	1
	Other	Employed	8	6	**7**	8
		Unemployed	5	4	Column	totals bold
			13	**10**		
Probabilities					0.714286	0.875
(p_{ijk})					0.285714	0.125
			0.615385	0.6		
			0.384615	0.4		

The Logits of the top row can be found

Logits $Ln\left(\frac{0.615385}{1-.615385}\right)$ = 0.47001 Similarly. 405465 Similarly 0.916292 Similarly 1.945910

The interpretation of the effects in this model is different from that in the log linear model, however the four effects can be calculated in a similar way in each.

u	The overall mean	0.9344192
$u_{1(1)} = -u_{1(2)}$	First main effect	0.2412682
$u_{2(1)} = -u_{2(2)}$	Second main effect	0.4966817
$u_{12(11)} = -u_{12(12)} = -u_{12(21)} = u_{12(22)}$	Interaction	0.2735407

These can be compared with the appropriate values in the first log-linear model above.

0.9344192 = 2 x 0.467209

0.2412682 = 2 x 0.120635

0.4966817 = 2 x 0.248342

and 0.2735407 = 2 x 0.136770

Calculation of of new coefficients for a new margin when the IPF algotrithm would converge in one step.

Looking at the new table we will have the following:

New Margin

14	16
5	9

Total = 44

Probabilities

0.318	0.364
0.114	0.204

a_{112}	a_{122}
a_{212}	a_{222}

a_{111}	a_{121}
a_{211}	a_{221}

Hence it can be seen that $a_{111} + a_{112} = 0.318$ etc

From the model $Ln(a_{111}) = \hat{u} + \hat{u}_i + \hat{u}_j + \hat{u}_k + \hat{u}_{ij} + \hat{u}_{ik} + \hat{u}_{jk} + \hat{u}_{ijk}$

and $Ln(a_{112}) = \hat{u} + \hat{u}_i + \hat{u}_j - \hat{u}_k + \hat{u}_{ij} - \hat{u}_{ik} - \hat{u}_{jk} - \hat{u}_{ijk}$

Note: The exact form of these depends on the parameterisation. In this case i,j and k are either –1 or 1 hence only the signs change. With other parameterisations the form will change but the solution follows the same process.

Combining the three equations above yields:

$$Ln(0.318) = \hat{u} + \hat{u}_i + \hat{u}_j + \hat{u}_k + \hat{u}_{ij} + \hat{u}_{ik} + \hat{u}_{jk} + \hat{u}_{ijk} + \hat{u} + \hat{u}_i + \hat{u}_j - \hat{u}_k + \hat{u}_{ij} - \hat{u}_{ik} - \hat{u}_{jk} - \hat{u}_{ijk} \text{ or}$$

$$0318 = e^{\hat{u}} e^{\hat{u}_i} e^{\hat{u}_j} e^{\hat{u}_k} e^{\hat{u}_{ij}} e^{\hat{u}_{ik}} e^{\hat{u}_{jk}} e^{\hat{u}_{ijk}} + e^{\hat{u}} e^{\hat{u}_i} e^{\hat{u}_j} e^{-\hat{u}_k} e^{\hat{u}_{ij}} e^{-\hat{u}_{ik}} e^{-\hat{u}_{jk}} e^{-\hat{u}_{ijk}}$$

$$\Rightarrow \quad 0.318 = e^{\left(\hat{u}+\hat{u}_i+\hat{u}_j+\hat{u}_{ij}\right)} \left(e^{\left(\hat{u}_k+\hat{u}_{ik}+\hat{u}_{jk}+\hat{u}_{ijk}\right)} + e^{-\left(\hat{u}_k+\hat{u}_{ik}+\hat{u}_{jk}+\hat{u}_{ijk}\right)} \right)$$

Note: the first factor has only the unknown effects and the second factor has only the known effects. So by substitution the second factor will be a constant hence

$$\Rightarrow \quad \frac{0.318}{e^{(\hat{u}_k+\hat{u}_{ik}+\hat{u}_{jk}+\hat{u}_{ijk})}+e^{-(\hat{u}_k+\hat{u}_{ik}+\hat{u}_{jk}+\hat{u}_{ijk})}} = e^{(\hat{u}+\hat{u}_i+\hat{u}_j+\hat{u}_{ij})}$$

or $\quad c_1 = \hat{u}+\hat{u}_i+\hat{u}_j+\hat{u}_{ij}$

Taking the other three values from the new margins three more linear equations will result and hence the four equations in four unknowns can be solved simultaneously to give the unknown effects.

These effects can be combined with those that stay constant to constitute the new model and hence the estimates of the values for the new table found.

For the example above:-

$$\Rightarrow \quad \frac{0.318}{e^{(0.327+0.019-0.248+0.137)}+e^{-(0.327+0.019-0.248+0.137)}} = e^{(\hat{u}+\hat{u}_i^*+\hat{u}_j+\hat{u}_{ij})}$$

$$\Rightarrow \quad \frac{0.318}{e^{0.235}+e^{-0.235}} = e^{(\hat{u}+\hat{u}_i+\hat{u}_j+\hat{u}_{ij})}$$

and hence $\quad \hat{u}+\hat{u}_i+\hat{u}_j+\hat{u}_{ij} = 0.1547$

Similarly $\quad \hat{u}-\hat{u}_i+\hat{u}_j-\hat{u}_{ij} = 0.1815$

$$\hat{u}+\hat{u}_i-\hat{u}_j-\hat{u}_{ij} = 0.0515$$

$$\hat{u}-\hat{u}_i-\hat{u}_j+\hat{u}_{ij} = 0.0816$$

$$\Rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{u}_i \\ \hat{u}_j \\ \hat{u}_{ij} \end{pmatrix} = \begin{pmatrix} 0.1547 \\ 0.1815 \\ 0.0515 \\ 0.0816 \end{pmatrix}$$

and $\hat{u} = -2.261 \quad \hat{u}_i = -0.156 \quad \hat{u}_j = 0.475 \quad \hat{u}_{ij} = 0.077$

The full set of parameters for the new table is then

$\hat{u}$	$\hat{u}_i$	$\hat{u}_j$	$\hat{u}_k$	$\hat{u}_{ij}$	$\hat{u}_{ik}$	$\hat{u}_{jk}$	$\hat{u}_{ijk}$
-2.261	-0.156	0.475	0.327	0.077	0.019	-0.248	0.137

Using these and the design matrix allows the new table to be easily constructed.

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} -2.261 \\ -0.156 \\ 0.475 \\ 0.327 \\ 0.077 \\ 0.019 \\ -0.248 \\ 0.137 \end{pmatrix}$$

$$= \begin{pmatrix} -1.63 \\ -1.784 \\ -2.512 \\ -1.81 \\ -2.1 \\ -1.63 \\ -3.426 \\ -3.196 \end{pmatrix}$$ Exponentiate these and the result is $$\begin{pmatrix} 0.1959 \\ 0.1680 \\ 0.0811 \\ 0.1637 \\ 0.1225 \\ 0.1959 \\ 0.0325 \\ 0.0409 \end{pmatrix}$$

and multiplying by 44, the grand total, gives $$\begin{pmatrix} 8.621 \\ 7.390 \\ 3.569 \\ 7.201 \\ 5.388 \\ 8.621 \\ 1.431 \\ 1.801 \end{pmatrix}$$

These can be compared with the results of the iterative proportional fitting on page 75 and we can see that within the rounding errors in only using 3 decimal places for the coefficients in the GLM the results are the same.

It should be noted that there was an analytic solution to the GLM in this case but this will not always be the case. When there is an analytic solution the IPF algorithm converges in one step. More generally the IPF algorithm will not converge in one step and in those cases there will not be an analytic solution to the GLM problem. In that case an iterative solution to find the coefficients

for the GLM will be required such as the iterative generalized least squares algorithm which is used in MLwiN.

The matrix used in calculating the unknown parameters comes easily from the design matrix when it is partitioned into the known and unknown parameters.

$\hat{u}$	$\hat{u}_i$	$\hat{u}_j$	$\hat{u}_k$	$\hat{u}_{ij}$	$\hat{u}_{ik}$	$\hat{u}_{jk}$	$\hat{u}_{ijk}$
1	1	1	1	1	1	1	1
1	-1	1	-1	1	-1	1	-1
1	1	-1	-1	1	1	-1	-1
1	-1	-1	1	1	-1	-1	1
1	1	1	1	-1	-1	-1	-1
1	-1	1	-1	-1	1	-1	1
1	1	-1	-1	-1	-1	1	1
1	-1	-1	1	-1	1	1	-1

The example above is particularly easy as the unknown parameters form four linear equations in four unknowns and are thus easily solved. In the more general case the constraints formed by the fixed parameters may leave some non linear equations to be solved. These may have a closed form and can thus be solved however, it is easier to fix the "known" parameters and to then estimate the model for a table which reflects the new margins but does not have any of the internal structure defined by the known parameters. This will then estimate the unknown parameters in the model.

For example if we take the 2 by 2 by 2 example above but the new known margins are now the margins for employment status and sex with no interaction as in the diagram below:-

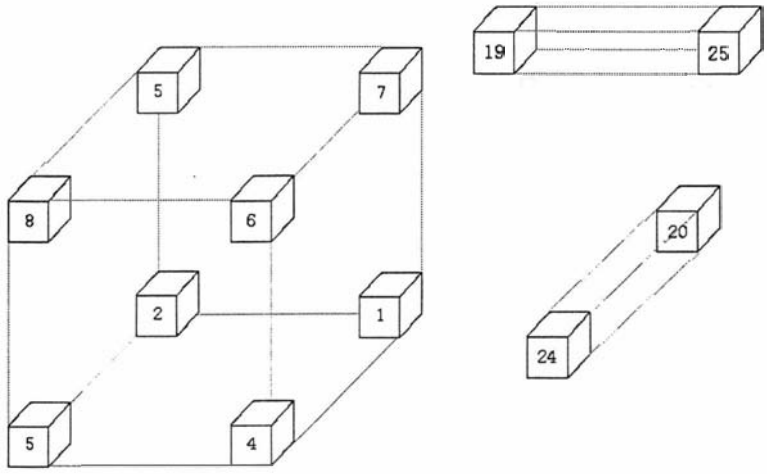


Figure Appendix A.1 A 2 x 2 x 2 table with new margins.

Using the notation above the parameters that will remain the same in this case will be u_k, u_{ij}, u_{ik}, u_{jk}, and u_{ijk} and those to be estimated are u , u_i , and u_j.

$$\frac{19}{44} = e^{\hat{u}} e^{\hat{u}_i} e^{\hat{u}_j} e^{\hat{u}_k} e^{\hat{u}_{ij}} e^{\hat{u}_{ik}} e^{\hat{u}_{jk}} e^{\hat{u}_{ijk}} + e^{\hat{u}} e^{\hat{u}_i} e^{\hat{u}_j} e^{-\hat{u}_k} e^{\hat{u}_{ij}} e^{-\hat{u}_{ik}} e^{-\hat{u}_{jk}} e^{-\hat{u}_{ijk}}$$

$$+ e^{\hat{u}} e^{\hat{u}_i} e^{-\hat{u}_j} e^{\hat{u}_k} e^{-\hat{u}_{ij}} e^{\hat{u}_{ik}} e^{-\hat{u}_{jk}} e^{-\hat{u}_{ijk}} + e^{\hat{u}} e^{\hat{u}_i} e^{-\hat{u}_j} e^{-\hat{u}_k} e^{\hat{u}_{ij}} e^{-\hat{u}_{ik}} e^{\hat{u}_{jk}} e^{\hat{u}_{ijk}}$$

$$= e^{\hat{u}} e^{\hat{u}_i} (e^{\hat{u}_j} e^{\hat{u}_{ij}} (e^{\hat{u}_k} e^{\hat{u}_{ik}} e^{\hat{u}_{jk}} e^{\hat{u}_{ijk}} + e^{-\hat{u}_k} e^{-\hat{u}_{ik}} e^{-\hat{u}_{jk}} e^{-\hat{u}_{ijk}}) + e^{-\hat{u}_j} e^{-\hat{u}_{ij}} (e^{\hat{u}_k} e^{\hat{u}_{ik}} e^{-\hat{u}_{jk}} e^{-\hat{u}_{ijk}} + e^{-\hat{u}_k} e^{-\hat{u}_{ik}} e^{\hat{u}^*_{jk}} e^{\hat{u}_{ijk}}))$$

$$= e^{\hat{u}} e^{\hat{u}_i} (e^{\hat{u}_j} k_1 + e^{-\hat{u}_j} k_2) \text{ as everything in } k_1 \text{ and } k_2 \text{ are known.}$$

This is nonlinear and four other similar nonlinear relations can be formed. These are not independent as the sums of the two margins are equal so there are five equations with the constraint that the cell values add to the correct margins. Their solution is not possible analytically so we would have to resort to other methods. This should not come as a surprise as the alternatives are iterative procedures and if there was an analytic solution to the problem in general then there would be no need for the iterative solutions.

If we consider the loglinear model for the original table

$$Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} -2.248 \\ 0.108 \\ 0.467 \\ 0.327 \\ -0.121 \\ 0.019 \\ -0.248 \\ 0.137 \end{pmatrix}$$

We now need to reestimate the coefficients which are changed for the new margins by building a 2 by 2 by 2 table with the given margins and no interactions and then estimate the three unknown effects given that all other effects are the same as before. This can be achieved by taking the new margins and generating the cell values in the table assuming independence. In this case the two new margins will generate the table:

19	**25**	
8.636	11.364	**20**
10.364	13.636	**24**

Which then need to be halved to produce the top and bottom "layers" and apply the log transformation.

The model above can be rewritten with the design matrix and the vector of coefficients partitioned into two parts. The first part is related to the parameters that are updated by the new data and the second is those parameters that stay the same. For the original table:

$$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} -2.248 \\ 0.108 \\ 0.327 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 0.467 \\ -0.121 \\ 0.019 \\ -0.248 \\ 0.137 \end{pmatrix}$$

The new coefficients can be calculated from the reconstructed table, which become the new data **Y**, and values calculated for the second term above which is commonly called an offset.

Appendix B

Design Matrices Construction and Checking.

Many of the design matrices used in this thesis have been quite large but they all have structure which has made it easier to construct them. An example is given below of the construction of a 192 by 192 design matrix for the 16 Regional Authorities with one column for gender, three age groups and two ethnicities. This one has been used extensively in the later chapters.

I have found it best to begin by looking at that structure and an Excel spreadsheet numbering each column of the matrix and stating which effect that column defines has been helpful.

1	Constant	6	Region1 (R1)	21	SA1	26	SR1	41	A1R15	56	A2R1
2	Gender (S)	7	R2	22	SA2	27	SR2	42	A1R2	57	A2R2
3	Age 1 (A1)	8	R3	23	SE	28	SR3	43	A1R3	58	A2R3
4	Age 2 (A2)	9	R4	24	A1E	29	SR4	44	A1R4	59	A2R4
5	Ethnicity (E)	10		25	A2E	30		45		60	
		11				31		46		61	
		12				32		47		62	
		13				33		48		63	
		14				34		49		64	
		15				35		50		65	
		16				36		51		66	
		17				37		52		67	
		18				38		53		68	
		19				39		54		69	
		20	R15			40	SR15	55	A1R15	70	A2R15

71	ER1	86	SA1E	88	SA1R1	103	SA2R1	118	A1ER1	133	A2ER1
72	ER2	87	SA2E	89	SA1R2	104	SA2R2	119	A1ER2	134	A2ER2
73	ER3			90	SA1R3	105	SA2R3	120	A1ER3	135	A2ER3
74	ER4			91	SA1R4	106	SA2R4	121	A1ER4	136	A2ER4
75				92		107		122		137	
76				93		108		123		138	
77				94		109		124		139	
78				95		110		125		140	
79				96		111		126		141	
80				97		112		127		142	
81				98		113		128		143	

82			99		114		129		144	
83			100		115		130		145	
84			101		116		131		146	
85	ER15		102	SA1R15	117	SA2R15	132	A1ER15	147	A2ER15

148	SER1	163	SA1ER1	178	SA2ER1
149	SER2	164	SA1ER2	179	SA2ER2
150	SER3	165	SA1ER3	180	SA2ER3
151	SER4	166	SA1ER4	181	SA2ER4
152		167		182	
153		168		183	
154		169		184	
155		170		185	
156		171		186	
157		172		187	
158		173		188	
159		174		189	
160		175		190	
161		176		191	
162	SER15	177	SA1ER15	192	SA2ER15

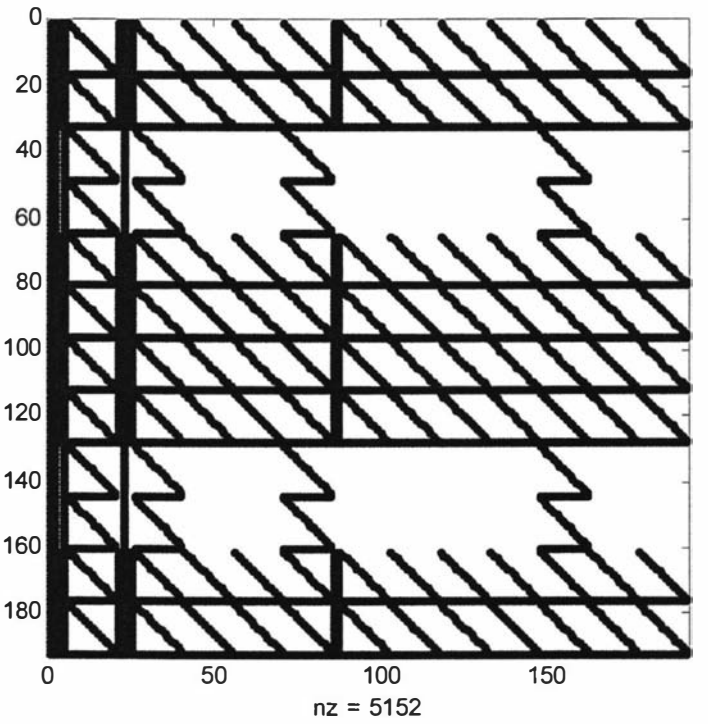
With this information it is easy to generate the design matrix after entering the first 20 columns in the above example. Matlab and Excel have been used for the generation. In Matlab the first columns are generated by using the repetitive pattern and in Excel the first entries can be typed in and the copying and pasting used. In Excel the rest of the matrix can be constructed by multiplying columns together. In Matlab the multiplication can be done in blocks. A series of Matlab commands to generate a design matrix are given below.

Matlab statements for the 192 x 192 design matrix.

```
A=ones(192,192);
A=-9*A;
A(:,1)=[ones(192,1)];
A(:,2)=[ones(96,1); -1*ones(96,1)]
A(:,3)=[ones(32,1);zeros(32,1);-1*ones(32,1);ones(32,1);zeros(32,1);-1*ones(32,1)];
A(:,4)=[ zeros(32,1);ones(32,1);-1*ones(32,1); zeros(32,1);ones(32,1);-1*ones(32,1)];
A(:,5) = [ones(16,1);-1*ones(16,1);ones(16,1);-1*ones(16,1);ones(16,1);-1*ones(16,1);
ones(16,1);-1*ones(16,1);ones(16,1);-1*ones(16,1);ones(16,1);-1*ones(16,1)];
for i=1:15 T=[zeros(i-1,1);1;zeros(15-i,1);-1]  A(:,i+5)=[T;T;T;T; T;T;T;T; T;T;T;T]  end
A(:,21)=A(:,2).*A(:,3);
A(:,22)=A(:,2).*A(:,4);
A(:,23)=A(:,2).*A(:,5);
A(:,24)=A(:,3).*A(:,5);
```

```
A(:,25)=A(:,4).*A(:,5);
for i=1:15 A(:,i+25)=A(:,2).*A(:,i+5); end
for i=1:15 A(:,i+40)=A(:,3).*A(:,i+5); end
for i=1:15 A(:,i+55)=A(:,4).*A(:,i+5); end
for i=1:15 A(:,i+70)=A(:,5).*A(:,i+5); end
A(:,86)=A(:,2).*A(:,3).*A(:,5);
A(:,87)=A(:,2).*A(:,4).*A(:,5);
for i=1:15 A(:,i+87)=A(:,2).*A(:,3).*A(:,i+5); end
for i=1:15 A(:,i+102)=A(:,2).*A(:,4).*A(:,i+5); end
for i=1:15 A(:,i+117)=A(:,3).*A(:,5).*A(:,i+5); end
for i=1:15 A(:,i+132)=A(:,4).*A(:,5).*A(:,i+5); end
for i=1:15 A(:,i+147)=A(:,2).*A(:,5).*A(:,i+5); end
for i=1:15 A(:,i+162)=A(:,2).*A(:,3).*A(:,5).*A(:,i+5); end
for i=1:15 A(:,i+177)=A(:,2).*A(:,4).*A(:,5).*A(:,i+5); end
SA=sparse(A);
spy(SA)
```

The final; two commands write the matrix in a "sparse format" which can then be looked at with the spy() command.



This only shows entries that are non zero and so there is no distinction between 1's and -1's. However, as a start at checking that the matrix is correct it is useful particularly with these larger ones.

There are more sophisticated ways of producing these design matrices and larger loops to generate multiple blocks of columns are possible however this seemed to be a good compromise between saving time and being sure it was correct.

An example of an orthogonal design matrix.

This is relevant to Chapter Six where orthogonal design matrices are briefly discussed.

c	s	a1	a2	r1	r2	r3	r4	r5	r6	r7	r8	sa 1	sa 2	sr 1	sr 2	sr 3	sr 4	sr 5	sr 6	sr 7	sr 8	a1 r1	a1 r2	a1 r3	a1 r4	a1 r5	a1 r6	a1 r7	a1 r8	a2 r1	a2 r2	a2 r3	a2 r4	a2 r5	a2 r6	a2 r7	a2 r8	sa 1r 1	sa 1r 2	sa 1r 3	sa 1r 4	sa 1r 5	sa 1r 6	sa 1r 7	sa 1r 8	sa 2r 1	sa 2r 2	sa 2r 3	sa 2r 4	sa 2r 5	sa 2r 6	sa 2r 7	sa 2r 8
1	1	1	-1	0	0	0	0	0	0	0	8	1	-1	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	-8
1	1	1	-1	0	0	0	0	0	0	7	-1	1	-1	0	0	0	0	0	0	7	-1	0	0	0	0	0	0	7	-1	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	7	-1	0	0	0	0	0	0	-7	1
1	1	1	-1	0	0	0	0	0	6	-1	-1	1	-1	0	0	0	0	0	6	-1	-1	0	0	0	0	0	6	-1	-1	0	0	0	0	0	-6	1	1	0	0	0	0	0	6	-1	-1	0	0	0	0	0	-6	1	1
1	1	1	-1	0	0	0	0	5	-1	-1	-1	1	-1	0	0	0	0	5	-1	-1	-1	0	0	0	0	5	-1	-1	-1	0	0	0	0	-5	1	1	1	0	0	0	0	5	-1	-1	-1	0	0	0	0	-5	1	1	1
1	1	1	-1	0	0	0	4	-1	-1	-1	-1	1	-1	0	0	0	4	-1	-1	-1	-1	0	0	0	4	-1	-1	-1	-1	0	0	0	-4	1	1	1	1	0	0	0	4	-1	-1	-1	-1	0	0	0	-4	1	1	1	1
1	1	1	-1	0	0	3	-1	-1	-1	-1	-1	1	-1	0	0	3	-1	-1	-1	-1	-1	0	0	3	-1	-1	-1	-1	-1	0	0	-3	1	1	1	1	1	0	0	3	-1	-1	-1	-1	-1	0	0	-3	1	1	1	1	1
1	1	1	-1	0	2	-1	-1	-1	-1	-1	-1	1	-1	0	2	-1	-1	-1	-1	-1	-1	0	2	-1	-1	-1	-1	-1	-1	0	-2	1	1	1	1	1	1	0	2	-1	-1	-1	-1	-1	-1	0	-2	1	1	1	1	1	1
1	1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
1	1	0	2	0	0	0	0	0	0	0	8	0	2	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16
1	1	0	2	0	0	0	0	0	0	7	-1	0	2	0	0	0	0	0	0	7	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	-2
1	1	0	2	0	0	0	0	0	6	-1	-1	0	2	0	0	0	0	0	6	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	12	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	12	-2	-2
1	1	0	2	0	0	0	0	5	-1	-1	-1	0	2	0	0	0	0	5	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	10	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	10	-2	-2	-2
1	1	0	2	0	0	0	4	-1	-1	-1	-1	0	2	0	0	0	4	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	8	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	8	-2	-2	-2
1	1	0	2	0	0	3	-1	-1	-1	-1	-1	0	2	0	0	3	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	6	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	6	-2	-2	-2	-2	-2
1	1	0	2	0	2	-1	-1	-1	-1	-1	-1	0	2	0	2	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	4	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	4	-2	-2	-2	-2	-2
1	1	0	2	1	-1	-1	-1	-1	-1	-1	-1	0	2	1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	2	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	2	-2	-2	-2	-2	-2
1	1	0	2	-1	-1	-1	-1	-1	-1	-1	-1	0	2	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	-2	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	-2	-2	-2	-2	-2	-2
1	1	-1	-1	0	0	0	0	0	0	0	8	-1	-1	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-8	-8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-8	-8
1	1	-1	-1	0	0	0	0	0	0	7	-1	-1	-1	0	0	0	0	0	0	7	-1	0	0	0	0	0	0	0	-7	1	0	0	0	0	0	-7	1	0	0	0	0	0	0	0	-7	1	0	0	0	0	0	-7	1
1	1	-1	-1	0	0	0	0	0	6	-1	-1	-1	-1	0	0	0	0	0	6	-1	-1	0	0	0	0	0	0	-6	1	1	0	0	0	0	0	-6	1	1	0	0	0	0	0	0	-6	1	1	0	0	0	0	-6	1	1
1	1	-1	-1	0	0	0	0	5	-1	-1	-1	-1	-1	0	0	0	0	5	-1	-1	-1	0	0	0	0	-5	1	1	1	0	0	0	0	-5	1	1	1	0	0	0	0	0	-5	1	1	1	0	0	0	0	-5	1	1	1
1	1	-1	-1	0	0	0	4	-1	-1	-1	-1	-1	-1	0	0	0	4	-1	-1	-1	-1	0	0	0	-4	1	1	1	1	0	0	0	-4	1	1	1	1	0	0	0	-4	1	1	1	1	0	0	0	-4	1	1	1	1

1	1	-1	-1	0	0	3	-1	-1	-1	-1	-1	-1	-1	0	0	3	-1	-1	-1	-1	-1	0	0	-3	1	1	1	1	1	0	0	-3	1	1	1	1	1	0	0	-3	1	1	1	1	1	0	0	-3	1	1	1	1	1
1	1	-1	-1	0	2	-1	-1	-1	-1	-1	-1	-1	-1	0	2	-1	-1	-1	-1	-1	-1	0	-2	1	1	1	1	1	1	0	-2	1	1	1	1	1	1	0	-2	1	1	1	1	1	1	0	-2	1	1	1	1	1	1
1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1
1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	-1	1	-1	0	0	0	0	0	0	0	8	-1	1	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	8
1	-1	1	-1	0	0	0	0	0	0	7	-1	-1	1	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	7	-1	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	7	-1
1	-1	1	-1	0	0	0	0	0	6	-1	-1	-1	1	0	0	0	0	0	-6	1	1	0	0	0	0	0	6	-1	-1	0	0	0	0	0	-6	1	1	0	0	0	0	0	-6	1	1	0	0	0	0	0	6	-1	-1
1	-1	1	-1	0	0	0	0	5	-1	-1	-1	-1	1	0	0	0	0	-5	1	1	1	0	0	0	0	5	-1	-1	-1	0	0	0	0	-5	1	1	1	0	0	0	0	-5	1	1	1	0	0	0	0	5	-1	-1	-1
1	-1	1	-1	0	0	0	4	-1	-1	-1	-1	-1	1	0	0	0	-4	1	1	1	1	0	0	0	4	-1	-1	-1	-1	0	0	0	-4	1	1	1	1	0	0	0	-4	1	1	1	1	0	0	0	4	-1	-1	-1	-1
1	-1	1	-1	0	0	3	-1	-1	-1	-1	-1	-1	1	0	0	-3	1	1	1	1	1	0	0	3	-1	-1	-1	-1	-1	0	0	-3	1	1	1	1	1	0	0	-3	1	1	1	1	1	0	0	3	-1	-1	-1	-1	-1
1	-1	1	-1	0	2	-1	-1	-1	-1	-1	-1	-1	1	0	-2	1	1	1	1	1	1	0	2	-1	-1	-1	-1	-1	-1	0	-2	1	1	1	1	1	1	0	-2	1	1	1	1	1	1	0	2	-1	-1	-1	-1	-1	-1
1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1
1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	-1	0	2	0	0	0	0	0	0	0	8	0	-2	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-16
1	-1	0	2	0	0	0	0	0	0	7	-1	0	-2	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-14	2
1	-1	0	2	0	0	0	0	0	6	-1	-1	0	-2	0	0	0	0	0	-6	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	12	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	-12	2	2
1	-1	0	2	0	0	0	0	5	-1	-1	-1	0	-2	0	0	0	0	-5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	10	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	-10	2	2	2
1	-1	0	2	0	0	0	4	-1	-1	-1	-1	0	-2	0	0	0	-4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	8	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	-8	2	2	2	2
1	-1	0	2	0	0	3	-1	-1	-1	-1	-1	0	-2	0	0	-3	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	6	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	-6	2	2	2	2	2
1	-1	0	2	0	2	-1	-1	-1	-1	-1	-1	0	-2	0	-2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	4	-2	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	-4	2	2	2	2	2	2
1	-1	0	2	1	-1	-1	-1	-1	-1	-1	-1	0	-2	-1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	2	-2	-2	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	-2	2	2	2	2	2	2	2
1	-1	0	2	-1	-1	-1	-1	-1	-1	-1	-1	0	-2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	-2	-2	-2	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2
1	-1	-1	-1	0	0	0	0	0	0	0	8	1	1	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	-8	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	8
1	-1	-1	-1	0	0	0	0	0	0	7	-1	1	1	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	-7	1	0	0	0	0	0	0	7	-1	0	0	0	0	0	0	7	-1
1	-1	-1	-1	0	0	0	0	0	6	-1	-1	1	1	0	0	0	0	0	-6	1	1	0	0	0	0	0	-6	1	1	0	0	0	0	0	-6	1	1	0	0	0	0	0	6	-1	-1	0	0	0	0	0	6	-1	-1
1	-1	-1	-1	0	0	0	0	5	-1	-1	-1	1	1	0	0	0	0	-5	1	1	1	0	0	0	0	-5	1	1	1	0	0	0	0	-5	1	1	1	0	0	0	0	5	-1	-1	-1	0	0	0	0	5	-1	-1	-1
1	-1	-1	-1	0	0	0	4	-1	-1	-1	-1	1	1	0	0	0	-4	1	1	1	1	0	0	0	-4	1	1	1	1	0	0	0	-4	1	1	1	1	0	0	0	4	-1	-1	-1	-1	0	0	0	4	-1	-1	-1	-1
1	-1	-1	-1	0	0	3	-1	-1	-1	-1	-1	1	1	0	0	-3	1	1	1	1	1	0	0	-3	1	1	1	1	1	0	0	-3	1	1	1	1	1	0	0	3	-1	-1	-1	-1	-1	0	0	3	-1	-1	-1	-1	-1
1	-1	-1	-1	0	2	-1	-1	-1	-1	-1	-1	1	1	0	-2	1	1	1	1	1	1	0	-2	1	1	1	1	1	1	0	-2	1	1	1	1	1	1	0	2	-1	-1	-1	-1	-1	-1	0	2	-1	-1	-1	-1	-1	-1
1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1

Appendix C

Computer programs used in the thesis with chapter references.

EG 2 Chapter 3.

```
model;  {

for( i in 1 : N ) {     precc[i] <- muc[i]
                          Yc[i] ~ dnorm(muc[i] , 1000)
                    }

# Multiply X with beta. This is elementwise multiplication.
  for( i in 1 : N ) {
      for (j in 1:N) {
         mewc[i,j] <- X[ i, j]*beta[j]
                    }
#Model log(mu) = X.beta  sum(**) sums the elementwise values from previous
log(muc[i]) <- sum(mewc[i, ])
                  }
# Priors for beta
for (i in 1:N) {  beta[i] ~ dnorm(0.0, 0.000001)  }

#The next block sums appropriate blocks of the mu's to give the margins.
for (i in 1:4) { margin[i] ~ dnorm(margins[i],1000)  }

            margins[1] <- mu[1] + mu[3] + mu[5] + mu[7]
            margins[2] <- mu[2] + mu[4] + mu[6] + mu[8]
            margins[3] <- mu[1] + mu[2] + mu[3] + mu[4]
            margins[4] <- mu[5] + mu[6] + mu[7] + mu[8]

#Generate the "offsets"
for (i in 1:N) {     offs[i,4]<- X[i,3]*beta[3]
   for (j in 5:8){     offs[i, j] <- X[i, j]*beta[ j ]     }
               }

#mewa are the contributions for the 3 coefficients for each row, they are also summed later
for (i in 1:N) {   mewa[i,1]  <- X[i , 1]*betaa[1]
                                                    mewa[i,2]  <- X[i , 2]*betaa[2]
                                                    mewa[i,3]  <- X[i , 4]*betaa[3]      }

#model log (mus) as the sum of the two parts, mewa from the new coefficients, Xbeta[i] and offs from the known
betas offset[i].
for (i in 1:N){
Xbeta[i] <- sum(mewa[i , 1:3])
offset[i] <-  sum(offs[i , 4:8])
log(mu[ i ]) <- Xbeta[i] + offset[i]
}

# Define distribution of unknown coefficients betaa[i]
                betaa[1] ~ dnorm(0, 0.00001)
                betaa[2] ~ dnorm(0, 0.00001)
                betaa[3] ~ dnorm(0, 0.00001)
                betaa[4] ~ dnorm(0, 0.00001)

}

#data

 list(N = 8,

X = structure(.Data=c(
```

```
 1, 1, 1, 1, 1, 1, 1, 1,
 1,-1, 1, 1,-1,-1, 1,-1,
 1, 1,-1, 1,-1, 1,-1,-1,
 1,-1,-1, 1, 1,-1,-1, 1,
 1, 1, 1,-1, 1,-1,-1,-1,
 1,-1, 1,-1,-1, 1,-1, 1,
 1, 1,-1,-1,-1,-1, 1, 1,
 1,-1,-1,-1, 1, 1, 1,-1      ),
.Dim = c(8,8)),
Yc=c(8,6,5,4,5,7,2,1),
margin = c(19,25,20,24)



)
)


#iNITS

  list(


 betaa= c(0, 0, 0, 0)    ,
  beta = c( 0,0,0,0,0,0,0,0)

)
```

SPREE Equivalent model Census data and new margins.

```
model;  {
for( i in 1 : N ) {    precc[i] <- 2/muc[i]                Y[i] ~ dnorm(muc[i],precc[i] )              }

# Multiply X with beta. This is elementwise multiplication.
  for( i in 1 : N ) {       for (j in 1:N) {          mewc[i,j] <- X[ i, j]*beta[j]                    }
#Model log(mu) = X.beta  sum(**) sums the elementwise values from previous
log(muc[i]) <- sum(mewc[i, ])
              , }
# Priors for beta
          for (i in 2:N) { beta[i] ~ dnorm(0.0, 0.1) }
              beta[1] ~ dnorm(6,0.1)
for (i in 1:5) { precmarg[i] <- 1/margins[i]                          marg[i] ~ dnorm(margins[i],precmarg[i])    }

for (i in 1:12) { SAE[i] <- sum(mu[(i*16)-15:i*16])  }

                                        margins[1] <- sum(SAE[1:6])
                                        margins[2] <- sum(SAE[7:12])
                                        margins[3] <- sum(SAE[1:2]) + sum(SAE[7:8])
                                        margins[4] <- sum(SAE[3:4]) + sum(SAE[9:10])
                                        margins[5] <- SAE[1] + SAE[3] + SAE[5] + SAE[7] + SAE[9] + SAE[11]


for (i in 1:16) { region[i] <- mu[i] + mu[i+16] +  mu[i+32] +  mu[i+48] +  mu[i+64] +  mu[i+80] +  mu[i+96] +  mu[i+112] +
mu[i+128] +
                                                                                    mu[i+144] +  mu[i+160] +
mu[i+176]  }

for (i in 1:N)  { for (j in 1:5)    { mewa[i,j] <- X[i,j]*betaa[j]   }    }
for (i in 1:N)  { for (j in 1:187)  { offs[i,j]   <- X[i,j+5]*beta[j+5]    }    }

for (i in 1:N) {    Xbeta[i] <- sum(mewa[i,1:5])
                                          offset[i] <- sum(offs[i,1:187])
                                        log(mu[i]) <- Xbeta[i] + offset[i]      }

for (i in 2:5)  { betaa[i] ~ dnorm(0.0, 0.1)  }
                  betaa[1] ~ dnorm(5,0.1)
}

#data
  list(N = 192,


 Y=c(605.05, 3264.94, 1504.49, 965.14, 187.13, 759.77, 799.17, 1354.90, 2317.78,
198.52,
281.85, 235.58, 277.50, 3100.51, 1399.44, 659.61,
1155.93, 3808.40, 1488.47, 1505.38, 522.88, 907.04, 449.77, 1020.15, 1599.36, 44.06,

70.31, 87.52, 56.63, 750.26, 236.34, 236.25,
1687.13, 9626.46, 2690.56, 1934.85, 383.25, 1476.16, 1214.62, 2371.69, 4253.15,
552.44,
527.03, 507.89, 733.62, 5342.05, 2487.80, 1192.55,
2475.79, 5937.87, 2316.52, 2845.12, 1127.13, 1558.87, 634.11, 1536.22, 2278.25, 87.28,
116.49, 131.11, 102.37, 1022.42, 315.94, 402.77,
352.81, 1380.09, 518.31, 426.15, 77.88, 311.12, 249.88, 468.21, 772.32, 102.93,
107.26, 113.05, 128.13, 1019.67, 497.90, 277.69,
230.28, 1204.80, 665.95, 670.93, 196.13, 428.74, 304.41, 582.22, 805.03, 107.88,
107.57, 124.00, 138.88, 1054.59, 503.53, 323.18,
407.48, 2384.98, 1231.03, 802.06, 138.63, 541.51, 586.26, 1012.84, 1552.72, 153.86,

229.74, 175.77, 176.50, 2596.15, 1021.82, 529.64,
633.55, 2819.39, 965.82, 950.19, 309.88, 527.43, 251.58, 587.11, 973.53, 24.80,
40.02, 49.34, 36.63, 503.11, 139.95, 143.65,
747.37, 5072.62, 1349.43, 983.98, 176.00, 606.68, 679.87, 1156.04, 1891.02, 257.88,

285.69, 247.31, 282.13, 2905.59, 1267.01, 629.52,
1010.85, 2961.76, 1102.48, 1118.64, 509.75, 673.18, 287.43, 682.61, 951.68, 34.63,

36.34, 54.68, 48.75, 442.48, 143.15, 149.59,
172.05, 839.01, 323.17, 275.89, 29.50, 151.48, 133.94, 262.47, 384.00, 45.27,
54.95, 56.88, 59.00, 576.79, 226.53, 130.70,
121.63, 399.08, 121.76, 127.89, 63.38, 75.99, 36.38, 77.81, 130.56, 4.63,
```

```
2.13, 5.80, 2.88, 40.08, 20.37, 15.25
),

marg=c(
62777, 45501,
41652, 54583,
74725
),
)


#iNITS

 list(

beta = c(
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0,0,0,0,0,0,0,0,0,  0,0),

betaa = c(0,0,0,0,0)


 )



****************************************************

model;  {
#Ys is the sample data ~ distributed Poison with means mus
  for( i in 1 : 7 ) {     margf[i]<-marg[i]
                          marg[i] ~ dpois(margr[i])          }
                         w2 <- sqrt(2)
                         w1 <- 1 - w2
for (i in 1:7) { margr[i]<- max((margins[i] - w1*margf[i])/w2,1)     }
                              margins[1]<-sum(SAE[1:6])
                              margins[2]<-sum(SAE[7:12])
                              margins[3]<-SAE[1] + SAE[2] + SAE[7] + SAE[8]
                              margins[4]<-SAE[3] + SAE[4] + SAE[9] + SAE[10]
                              margins[5]<-SAE[5] + SAE[6] + SAE[11] + SAE[12]
                              margins[6]<-SAE[1] + SAE[3] + SAE[5] + SAE[7] + SAE[9] + SAE[11]
                              margins[7]<-SAE[2] + SAE[4] + SAE[6] + SAE[8] + SAE[10] + SAE[12]
for (i in 1:12) { SAE[i]<-sum(mu[(i*16)-15:i*16])  }
#Generate the "offsets" (offs are a 192 by 187 matrix, each row is summed later to give the offset)
for (i in 1:N) {   for (j in 1:187){      offs[i,j] <- X[i, j+5]*betab[j ]           }          }
#mewa are the contributions for the 5 coefficients for each row, they are also summed later
for (i in 1:N) {   for (j in 1:5) {      mewa[i,j]  <- X[i , j]*betaa[j]           }          }
#model log (mus) as the sum of the two parts, mewa from the new coefficients and offs from the offset
for (i in 1:N){    log(mu[ i ]) <- sum(mewa[i , 1:5]) + sum(offs[i , 1:187]) + b[i]  }
# Define distribution of unknown coefficients betaa
                     betaa[1] ~ dflat()
for (i in 2:5) {   betaa[i] ~ dnorm( 0.0, 0.1)        }
                     b[1:192] ~ car.normal(adj[], weights[], num[], tau)
tau ~ dgamma(1, 1)  sigma<-1/sqrt(tau)
}
#data
 list(N = 192,


.Dim = c(192,192)),

marg=c( 62777, 45501, 41652, 54584, 12042, 74725, 33553),

betab=c(0.407130684, 1.868427822, 0.894371834, 0.76101946, -0.613398189, 0.304888339, -0.075503581,
        0.624522371, 1.07824282, -1.60030457, -1.471540467, -1.341729148, -1.41281209, 0.932581453,
        0.005315636, -0.250450775, -0.054169847, -0.167789806, -0.038335777, 0.049000249, -0.110442518, -
        0.152903928, -0.095928472, -0.086768054, -0.074426372, -0.027391611, -0.031511732, -0.035318339, -
        0.024173933, 0.091439541, 0.145637528, 0.068423154, 0.154050947, 0.039586646, 0.075964469, -
        0.087744076, -0.014571071, 0.097763495, 0.011771352, 0.007253497, 0.036067675, 0.088523206,
        0.073767051, 0.099922218, -0.229302179, 0.066255636, -0.068252504, -0.138489828, 0.092563412, -
        0.020823439, 0.215314478, 0.155597869, -0.002457967, 0.024529188, 0.135922252, 0.014737804, -
        0.062780314, -0.02534937, -0.028429568, -0.002131589, -0.068252504, -0.07626669, 0.029120218, -
```

```
0.110167982, -0.157168395, -0.440192352, -0.225224391, -0.245908888, -0.424534694, -0.828666685, -
0.351808644, -0.020318805, -0.124450436, -0.091288707, 0.427926831, 0.472779671, 0.252758495,
0.462037338, 0.412225804, 0.510564624, 0.114557775, 0.134489178, 0.161020196, 0.108006214,
0.05562827, 0.049492633, 0.081634324, 0.049015239, 0.055625612, 0.047601348, 0.049849354, -
0.082633207, -0.152229683, -0.050504064, -0.130590633, -0.093882168, -0.064928737, 0.143076998,
0.0920967, 0.059207753, 0.094322374, 0.07249116, 0.064753571, -0.020469713, 0.022894599,
0.050179869, -0.064740132, -0.096191848, -0.064746375, -0.124536469, -0.072823753, -0.104228353, -
0.150440139, -0.173558378, -0.009215927, -0.047425184, 0.05222705, -0.004356085, 0.057070355,
0.013261893, -0.017709379, 0.086249033, -0.007230574, -0.006060762, 0.009938126, 0.034285347,
0.112629852, -0.136981332, 0.0747397, -0.071934666, -0.109719408, -0.112672403, -0.093617847, -
0.007754004, -0.041066491, 0.013218078, 0.129528142, 0.014286, 0.057290311, 0.063433323,
0.065838193, 0.144671922, 0.15048028, 0.113725453, 0.04186245, 0.021534296, 0.134679125,
0.070744323, 0.002548924, 0.018800163, 0.061247361, -0.066332735, -0.179279514, -0.066436161, -
0.071355907, -0.094785788, -0.061083354, -0.14874883, -0.057152911, -0.046611367, -0.037064495, -
0.137240048, -0.068391335, -0.017108143, -0.030948976, -0.031972477, 0.039599282, 0.142734563,
0.049603694, 0.128795126, 0.092499333, 0.061940529, -0.137565706, -0.094151998, -0.021671023, -
0.052561686, -0.105204978, -0.025070143, -0.021987832, -0.008638045, -0.043544363, 0.058995735,
0.074446091, 0.061007695, 0.158090215, 0.070946572, 0.065153907)

adj=c(2,1,3,2,4,6,7,8,3,5,6,4,6,3,4,5,8,3,8,3,6,7,9,8,11,12,13,14,10,12,10,11,14,10,14,15,16,10,12,13,15,13,14,16,13
,15,18,17,19,18,20,22,23,24,19,21,22,20,22,19,20,21,24,19,24,19,22,23,25,24,27,28,29,30,26,28,26,27,30
,26,30,31,32,26,28,29,31,29,30,32,29,31,34,33,35,34,36,38,39,40,35,37,38,36,38,35,36,37,40,35,40,35,38
,39,41,40,43,44,45,46,42,44,42,43,46,42,46,47,48,42,44,45,47,45,46,48,45,47,50,49,51,50,52,54,55,56,51,
53,54,52,54,51,52,53,56,51,56,51,54,55,57,56,59,60,61,62,58,60,58,59,62,58,62,63,64,58,60,61,63,61,62,64,61,63,66
,65,67,66,68,70,71,72,67,69,70,68,70,67,68,69,72,67,72,67,70,71,73,72,75,76,77,78,74,76,74,75,78,74,78,79,80,74,7
6,77,79,77,78,80,77,79,82,81,83,82,84,86,87,88,83,85,86,84,86,83,84,85,88,83,88,83,86,87,89,88,91,92,93,94,90,92,
90,91,94,90,94,95,96,90,92,93,95,93,94,96,93,95,98,97,99,98,100,102,103,104,99,101,102,100,102,99,100,101,104,9
9,104,99,102,103,105,104,107,108,109,110,106,108,106,107,110,106,110,111,112,106,108,109,111,109,110,112,109
,111,114,113,115,114,116,118,119,120,115,117,118,116,118,115,116,117,120,115,120,115,118,119,121,120,123,124
,125,126,122,124,122,123,126,122,126,127,128,122,124,125,127,125,126,128,125,127,130,129,131,130,132,134,135
,136,131,133,134,132,134,131,132,133,136,131,136,131,134,135,137,136,139,140,141,142,138,140,138,139,142,138
,142,143,144,138,140,141,143,141,142,144,141,143,146,145,147,146,148,150,151,152,147,149,150,148,150,147,148
,149,152,147,152,147,150,151,153,152,155,156,157,158,154,156,154,155,158,154,158,159,160,154,156,157,159,157
,158,160,157,159,162,161,163,162,164,166,167,168,163,165,166,164,166,163,164,165,168,163,168,163,166,167,169
,168,171,172,173,174,170,172,170,171,174,170,174,175,176,170,172,173,175,173,174,176,173,175,178,177,179,178
,180,182,183,184,179,181,182,180,182,179,180,181,184,179,184,179,182,183,185,184,187,188,189,190,186,188,186
,187,190,186,190,191,192,186,188,189,191,189,190,192,189,191),
weights = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,),
num = c(1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4,
2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3,
2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4,
3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4,
1, 4, 2, 3, 4, 4, 3, 2,  1, 2, 5, 3, 2, 4, 2, 4, 1, 4, 2, 3, 4, 4, 3, 2 )

)


#iNITS
list(  betaa= c( 5, .1, .5, .5, .4))

list(betaa=c(10, 10, 10, 10,10))
```

Relative risk models chapter 9

```
model;  {
for( i in 1 : N ) {  Y[i] ~ dnorm(muc[i],precc[i] )              }

for( i in 1 : N )  {   for (j in 1:N)   {  mewc[i,j] <- X[ i, j]*beta[j]   }   log(muc[i]) <- sum(mewc[i, ])          }
beta[1] ~ dnorm(7,0.001)
beta[2] ~ dnorm(-1,0.001)
for (i in 3:N) {  beta[i] ~ dnorm(0.0, 0.000001)  }
for (i in 1:10) {    marg[i] ~ dnorm(margins[i],precmarg[i])    }
for (i in 1:24)  { SAEES[i] <- sum(mu[(i*16)-15:i*16])   }
                        margins[1] <- sum(SAEES[1:6])
                        margins[2] <- sum(SAEES[7:12])
                        margins[3] <- sum(SAEES[1:2]) + sum(SAEES[7:8])
                        margins[4] <- sum(SAEES[3:4]) + sum(SAEES[9:10])
                        margins[5] <- SAEES[1] + SAEES[3] + SAEES[5] + SAEES[7] + SAEES[9] + SAEES[11]
                        margins[6] <- sum(SAEES[13:18])
                        margins[7] <- sum(SAEES[19:24])
                        margins[8] <- sum(SAEES[13:14]) + sum(SAEES[19:20])
                        margins[9] <- sum(SAEES[15:16]) + sum(SAEES[21:22])
                        margins[10] <- SAEES[13] + SAEES[15] + SAEES[17] + SAEES[19] + SAEES[21] + SAEES[23]
for (i in 1:16) { region[i] <- mu[i] + mu[i+16] +  mu[i+32] +  mu[i+48] +  mu[i+64] +  mu[i+80] +  mu[i+96] +  mu[i+112] +
                               mu[i+128] + mu[i+144] +  mu[i+160] +  mu[i+176]  }
for (i in 1:N)  { for (j in 1:10)    {  mewa[i,j] <- X[i,a[j]]*betaa[j]   }     }
for (i in 1:N)  { for (j in 1:374)   {  offs[i,j]   <- X[i,b[j]]*beta[b[j]]     }     }
for (i in 1:N)  {         Xbeta[i] <- sum(mewa[i,1:10])
                          offset[i] <- sum(offs[i,1:374])
                          log(mu[i]) <- Xbeta[i] + offset[i]       }
betaa[1] ~ dnorm(7, 0.001)
betaa[2] ~ dnorm(-1,0.001)
for (i in 3:10)  {        betaa[i] ~ dnorm(0.0, 0.00001)  }
}
#data
  list(N = 384,

Y=c(605.05,3264.94,1504.49,965.14,187.13,759.77,799.17,1354.90,2317.78,198.52,281.85,235.58,277.50,3100.51,1399.44,659
        61,1155.93,3808.40,1488.47,1505.38,522.88,907.04,449.77,1020.15,1599.36,44.06,70.31,87.52,56.63,750.26,236.34,
        236.25,1687.13,9626.46,2690.56,1934.85,383.25,1476.16,1214.62,2371.69,4253.15,552.44,527.03,507.89,733.62,53
        42.05,2487.80,1192.55,2475.79,5937.87,2316.52,2845.12,1127.13,1558.87,634.11,1536.22,2278.25,87.28,116.49,13
        1.11,102.37,1022.42,315.94,402.77,352.81,1380.09,518.31,426.15,77.88,311.12,249.88,468.21,772.32,102.93,107.26
        ,113.05,128.13,1019.67,497.90,277.69,230.28,691.53,211.91,280.65,122.38,149.00,61.90,148.19,256.03,7.52,6.48,12
        .72,12.13,117.17,37.76,63.49,407.48,2384.98,1231.03,802.06,138.63,541.51,586.26,1012.84,1552.72,153.86,229.74,
        175.77,176.50,2596.15,1021.82,529.64,633.55,2819.39,965.82,950.19,309.88,527.43,251.58,587.11,973.53,24.80,40.
        02,49.34,36.63,503.11,139.95,143.65,747.37,5072.62,1349.43,983.98,176.00,606.68,679.87,1156.04,1891.02,257.88,
        285.69,247.31,282.13,2905.59,1267.01,629.52,1010.85,2961.76,1102.48,1118.64,509.75,673.18,287.43,682.61,951.6
        8,34.63,36.34,54.68,48.75,442.48,143.15,149.59,172.05,839.01,323.17,275.89,29.50,151.48,133.94,262.47,384.00,45
        .27,54.95,56.88,59.00,576.79,226.53,130.70,121.63,399.08,121.76,127.89,63.38,75.99,36.38,77.81,130.56,4.63,2.13,
        5.80,2.88,40.08,20.37,15.25,3743.95,49623.06,16875.51,7356.86,1046.88,5236.23,4559.83,12217.10,19037.22,1898.
        48,2064.15,1814.42,1403.50,27773.49,12475.56,4875.39,2039.07,18627.60,5956.53,4144.62,1143.13,2156.96,1075.
        23,3172.85,5738.64,216.94,281.69,313.48,207.37,3301.74,1395.66,865.75,13246.87,141844.54,44465.44,24239.15,3
        897.75,16672.84,14507.38,28560.31,59051.85,5822.56,6029.97,5314.11,4717.38,71239.95,26919.20,14454.45,3851.
        21,33769.13,9748.48,7233.88,1909.88,3599.13,1782.89,5176.78,10154.75,366.72,476.51,551.89,363.63,5553.58,188
        8.06,1478.23,13831.19,93833.91,33866.69,23525.85,3384.13,14706.88,11397.12,23140.79,39661.68,4581.07,4658.7
        4,4956.95,3808.87,53907.33,20839.10,10383.31,2557.72,10534.47,3858.09,3459.35,1159.63,1667.00,791.10,1977.8
        1,3352.97,138.48,125.52,238.28,157.88,1768.83,607.24,749.51,3683.52,50572.02,16307.97,7126.94,1025.38,5364.4
        9,4605.74,12034.16,20795.28,1706.14,2059.26,1587.23,1378.50,27523.85,13813.18,4535.36,2733.45,21339.61,6567
        .18,5013.81,1353.13,2602.57,1148.42,3727.89,6767.47,212.20,304.98,295.66,210.37,3399.89,1461.05,851.35,15106.
        63,154434.38,47372.57,26903.02,4100.00,18439.32,15392.13,30978.96,64200.98,6205.12,6624.31,5681.69,4942.87,
        76479.41,29011.99,14963.48,6236.15,41968.24,12119.52,10254.36,2978.25,5203.82,2219.57,6487.39,12715.32,424.
        37,541.66,658.32,452.25,5759.52,1847.85,1666.41,14483.95,109621.99,37555.83,26574.11,4008.50,17381.52,13470
        .06,27509.53,46138.00,4841.73,5720.05,5444.12,3973.00,64825.21,25051.47,12038.30,3019.37,13035.92,4291.24,3
        964.11,1407.63,2058.01 ,814.62 ,2221.19 ,3778.44 ,170.38 ,154.88 ,225.20 ,139.12 ,1709.92 ,556.63,505.75),
precc=c(                                                                                                    ),
marg=c(62777, 45501,  41653, 54583,  74725,  1223310, 1331877,  467442, 1235291,  2149834 ),
precmarg=c(                                                                                                 ),
a=c(1,2,3,4,5,6,22,23,24,25),
b=c(7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,
        48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,
```

```
82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103,104,105,106,107,108,109,110,1
11,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129,130,131,132,133,134,13
5,136,137,138,139,140,141,142,143,144,145,146,147,148,149,150,151,152,153,154,155,156,157,158,159
,160,161,162,163,164,165,166,167,168,169,170,171,172,173,174,175,176,177,178,179,180,181,182,183,
184,185,186,187,188,189,190,191,192,193,194,195,196,197,198,199,200,201,202,203,204,205,206,207,2
08,209,210,211,212,213,214,215,216,217,218,219,220,221,222,223,224,225,226,227,228,229,230,231,23
2,233,234,235,236,237,238,239,240,241,242,243,244,245,246,247,248,249,250,251,252,253,254,255,256
,257,258,259,260,261,262,263,264,265,266,267,268,269,270,271,272,273,274,275,276,277,278,279,280,
281,282,283,284,285,286,287,288,289,290,291,292,293,294,295,296,297,298,299,300,301,302,303,304,3
05,306,307,308,309,310,311,312,313,314,315,316,317,318,319,320,321,322,323,324,325,326,327,328,32
9,330,331,332,333,334,335,336,337,338,339,340,341,342,343,344,345,346,347,348,349,350,351,352,353
,354,355,356,357,358,359,360,361,362,363,364,365,366,367,368,369,370,371,372,373,374,375,376,377,
378,379,380,381,382,383,384),
)

#iNITS
list(
beta = c(
7,-1,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0, 0,0,0,0),

betaa = c(7,-1,0,0,0,0,0,0,0,0)
)
```

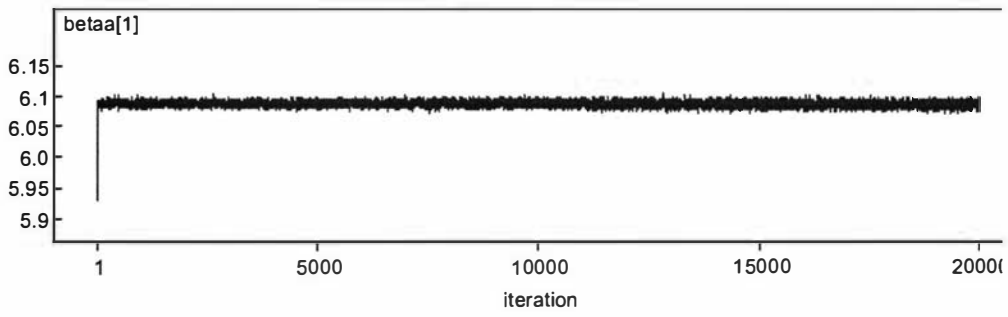
Appendix D

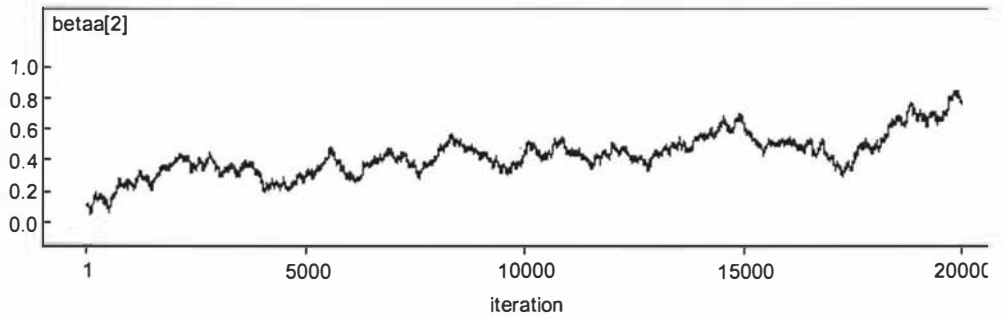
Examples of WinBUGS output.

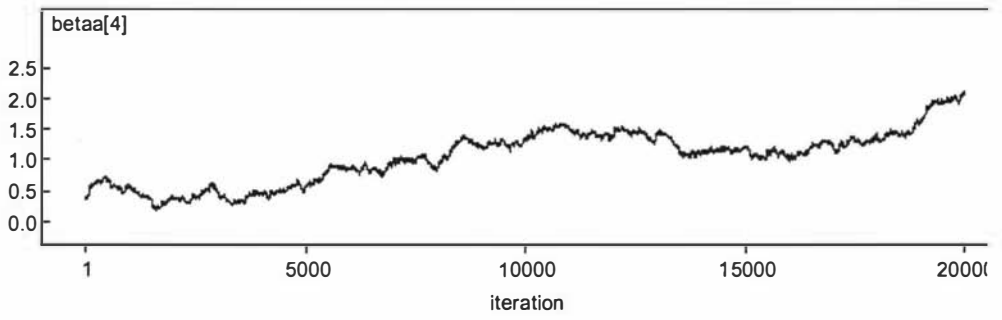
In Chapter Seven, Section 7.3, there is an example of output from WinBUGS which shows good convergence and simulations which are behaving appropriately. This is not always the case. In fitting models to the contingency tables the coefficients are correlated with each other as an adjustment to one can be offset by adjusting another with little change in the final estimates of cell counts.

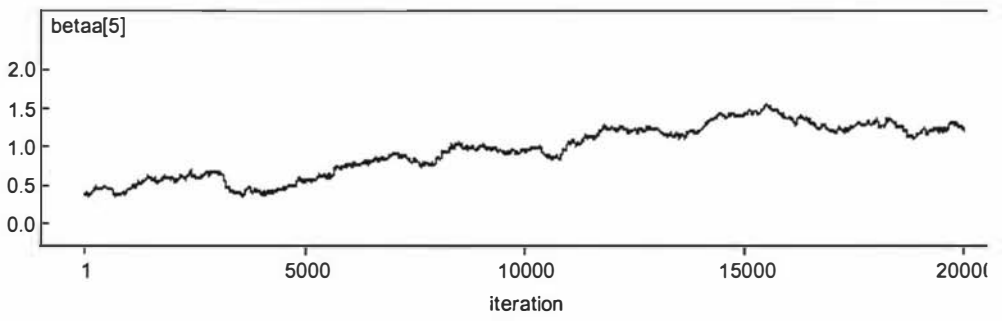
Below is an example of less well behaved chains with a few comments as an introduction to the possible problems which may arise.

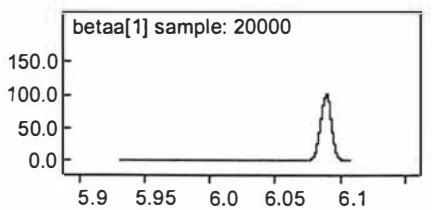
We can see in the traces below that although betaa[1] has converged the other graphs are wandering around and have not. They follow patterns of increasing and decreasing values. The density curves show poor mixing of the chain. They have a number of modes which suggest that the chain is "stuck" in a particular range and then moves on to another. The values are highly correlated to large lags. These are all indicative of poor convergence and problems with the estimates. It can be seen with the estimates themselves that the standard deviations and MC errors are quite large and looking at the traces it is not hard to see why.

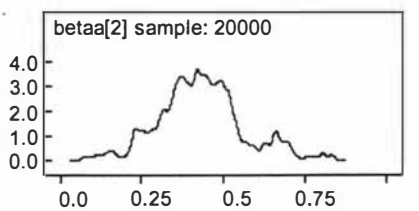


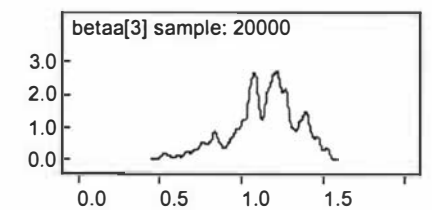

betaa[2]
1.0
0.8
0.6
0.4
0.2
0.0
1
5000
10000
15000
20000
iteration

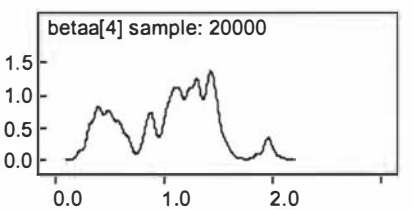

betaa[4]
2.5
2.0
1.5
1.0
0.5
0.0
1
5000
10000
15000
20000
iteration

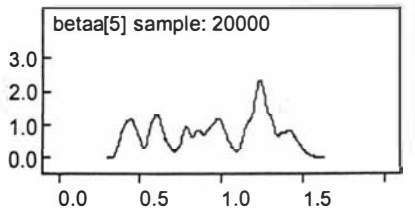

betaa[5]
2.0
1.5
1.0
0.5
0.0
1
5000
10000
15000
20000
iteration

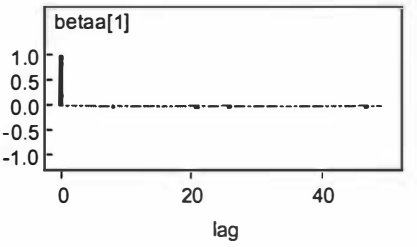

betaa[1] sample: 20000
150.0
100.0
50.0
0.0
5.9
5.95
6.0
6.05
6.1

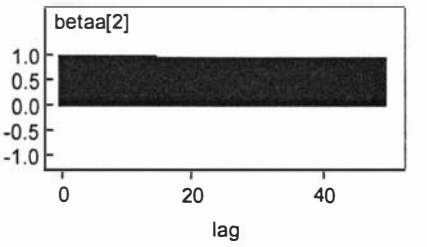

betaa[2] sample: 20000
4.0
3.0
2.0
1.0
0.0
0.0
0.25
0.5
0.75

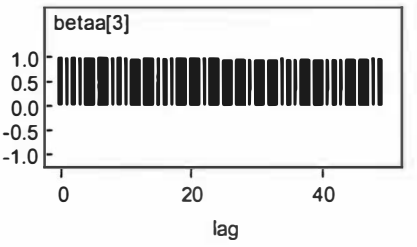

betaa[3] sample: 20000
3.0
2.0
1.0
0.0
0.0
0.5
1.0
1.5

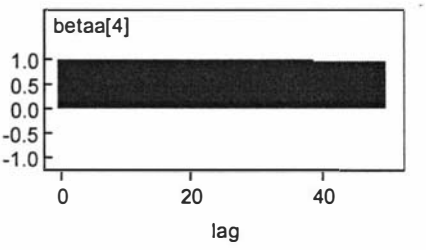

betaa[4] sample: 20000
1.5
1.0
0.5
0.0
0.0
1.0
2.0

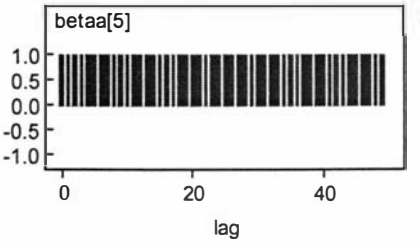












node	mean	sd	MC error	2.5%	median	97.5%	start	sample
betaa[1]	6.089	0.00415	3.281E-5	6.081	6.089	6.097	1	20000
betaa[2]	0.4316	0.1331	0.0111	0.1613	0.4259	0.7119	1	20000
betaa[3]	1.138	0.2023	0.01696	0.6739	1.163	1.474	1	20000
betaa[4]	1.057	0.4238	0.03562	0.3196	1.132	1.954	1	20000
betaa[5]	0.9612	0.3313	0.02787	0.3942	0.9876	1.468	1	20000