A Nested Random Effects Model Analysis of Child Survival in Malawi

A thesis presented to

the University of Waikato

in fulfilment of the requirement for the degree

of

Doctor of Philosophy in Statistics

by

Samuel Osmond Makandi Manda

Department of Statistics

The University of Waikato
Te Whare Wānanga o Waikato

1998

Abstract

This thesis investigates child survival in Malawi using a modified Cox proportional hazards model which includes family and community random effects in addition to the fixed effect of the covariates. The parameters of the model are estimated using the Gibbs sampler, a Bayesian Markov Chain Monte Carlo (MCMC) method. We believe this to be the first time the method has been used for this type of data. The parameters are also estimated using the expectation-maximisation (EM) algorithm, a method that has previously been used to analyse this type of data. Both approaches were implemented in Fortran using the NAG routines.

The child survival data used in this study were collected as part of the 1992 Demographic and Health Survey (DHS) of Malawi. The women respondents were obtained through a two-stage cluster sampling procedure. The respondents were systematically interviewed about biological, social and demographic factors relating to all live births that had occurred during the previous five years.

The results show that the covariates *first birth*, *birth spacing*, *breastfeeding duration*, *maternal age*, *hospital birth* and *father's education* are important determinants of infant and early childhood survival in Malawi. The family and community frailty effect variances are modest in magnitude. The study also shows that child mortality varies more considerably across families than over communities after controlling for the observed covariates. Neglecting frailty biases estimates of the observed covariates slightly downwards, although the subsequent substantive findings are not markedly affected. The strength of the family random effect is grossly overestimated when community random frailty is ignored while that of the community random effect shows stability whether family random frailty is controlled or not.

The Gibbs sampler is shown to be an important alternative to the EM algorithm and other existing methods for estimating parameters in a multilevel hazards model. It allows the full Bayesian inference without the need to evaluate high-dimensional integrals. We obtain a random sample from the complete posterior distribution of all the parameters and hyperparameters whose behaviour can be studied over their range rather than only around the mode. The Gibbs sampler is computational intensive, but this fact has decreasing relevance due to the availability of very powerful computing equipment.

Notes

A number of papers have been produced from the thesis, and their abstracts appear in Appendix B. The titles are:

1. Bolstad, W.M. and Manda, S.O.M. (1998). A Markov Chain Monte Carlo ivestigation of child mortality in Malawi using the proportional hazards model with family and community random effects. *Unpublished*

2. Manda, S.O.M. (1998a). Unobserved family and village effects on infant mortality in Malawi. *GENUS*, Vol. LIV, 1-2 pp. 143-164.

3. Manda, S.O.M. (1998b). Birth interval, breastfeeding and determinants of childhood mortality in Malawi. *Social Science and Medicine*, 48, 3, pp. 301-312.

4. Manda, S.O.M. (1998c). A comparison of methods for analysing a nested frailty model to child survival in Malawi. *Unpublished*

Earlier versions of paper 1 and 4 were presented at the *Bayesian Statistics 6 Conference*, July, 1998, Spain and the *14th Statistical Society of Australia Conference*, July, 1998 Australia respectively.

Acknowledgements

I would like to thank my chief supervisor Dr. William M. Bolstad for all his supervision and continual support in the course of this project. I also received valuable support and constructive comments from Dr. Murray Jorgensen and Dr. A. Dharmalingam, the other members of the supervising panel. Professor Ian Pool offered valuable advice in the early stages of the thesis work.

I also wish to acknowledge the National Statistics Office of Malawi for kindly allowing me to use their data. This project would not have been done without the financial support of the New Zealand Official Development Assistance which I wish to recognize as well.

I am also very thankful to my parents, my brother and sisters for their support and encouragement throughout my life. My friends who encouraged me all the way deserve many thanks.

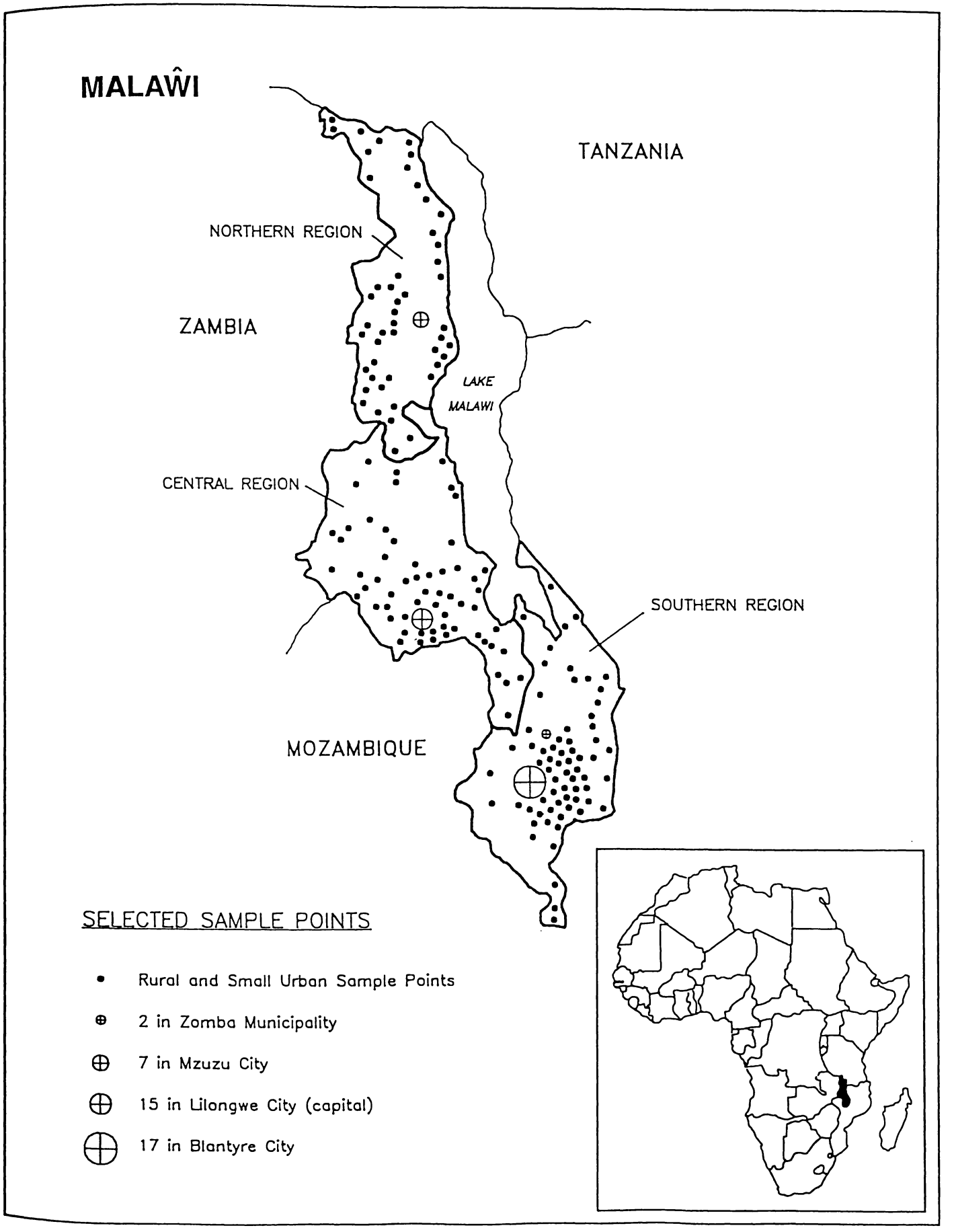
Finally, on a personal note, I dedicate this work to my wife Ellen Ulemu and our daughter Ruth Tilumbenge and son Patson Waliwona for their love and support. They all shared the burden of my limited ability to contribute effectively to our family life during this thesis work.

Contents

List of Tables

List of Figures

Map of Malawi



Chapter 1

Introduction

1.1 Determinants of child mortality

This introductory chapter summarises the literature on determinants of child mortality in the context of poorer and less developed countries. We also point out the shortcomings of many standard analyses for child survival based on the data that can not be assumed independent. The hazard model for survival data that are clustered at two hierarchical levels is defined. A brief description of the data to be used in this thesis is also given. Both frequentist and Bayesian methodology in analysing such models are considered. We then define our research objective and outline the structure of the thesis.

The determinants of childhood mortality have been extensively researched in most parts of the world as evidenced by numerous published articles. Malawi, which has one of the highest infant and child mortality rates in the world, has had very little research. One recent study (Madise, 1993) used data that were not a representative sample of Malawi. The data were purposively collected from only five of the twenty-four districts in Malawi (Srivastava and M'Manga, 1991). As well as this apparent bias, the data were collected in 1988 and as such the results might not relate to current determinants of child mortality. The results presented in this thesis were obtained from highly reliable data collected as part of the Demographic

and Health Survey (DHS) of Malawi conducted in 1992. The World Fertility Survey was not held for Malawi; thus there was no reliable demographic and health data for Malawi until the 1992 DHS.

There are many factors that had been found to determine child mortality (Hill and Pebley, 1989). The relationship between child mortality and many covariates is clear and less problematic. But the influence of the factors *birth interval* and *breastfeeding duration* is confounded by several factors. Thus, in this brief discussion, the focus is on the relationship between child mortality and *birth interval* and *breastfeeding duration.*

Many studies conducted in developing countries have gathered evidence pointing to an inverse relation between *birth intervals* and infant and child mortality rates (Cleland and Sathar, 1984; Forste, 1994; Madise and Diamond, 1995). There are various mechanisms by which *birth intervals* might affect childhood mortality. Rapid succession of births may erode the reproductive and nutritional resources of the mother leading to a higher incidence of premature and weaker births (Pebley and Stupp, 1987). Closely spaced children compete for scarce resources such as food and clothing. An increased transmission of infant and child contagious infections among closely spaced siblings may also occur.

Estimating the magnitude of relationships between child mortality and *birth intervals* is fraught with complications arising from different factors (Cleland and Sathar, 1984). A premature birth reduces both the *preceding birth interval* and its survival chance. If the premature child dies, then in the absence of a statistical control for prematurity, a short *preceding birth interval* might appear to be directly related to death of the child. This bias is often ignored since prematurity is rare; and in any case, prematurity itself may be a result of other risk factors such as short *birth intervals.*

Serious bias of effect of *birth interval* may occur if it is short because parents intentionally want to quickly replace the dead child (so that the result of a short

birth spacing is the outcome of, rather than the cause of, mortality); or the child death has resulted in an early return of fertility due to premature truncation of breastfeeding (Palloni and Millman, 1986). The latter holds in Malawi because the use of *birth control* is not widespread. If mortality risks within a family are correlated, then the child has a higher chance of dying. This would induce an artificially high correlation between infant death and a short *preceding birth interval.* The inclusion, in the analysis, of a variable to control for the survival status of the preceding birth would largely remedy this reverse causality. For a *succeeding birth interval*, the next birth would arrive much sooner than if the child had survived (Rutherford *et al.*, 1989). This reverse relation can be avoided by only considering the *succeeding conception interval* that occurs before the age at which the survival status of the child is determined.

Many studies have also found that effects of *birth intervals* tend to diminish when we control for *breastfeeding duration* (Palloni and Millman, 1986; Retherford *et al.*, 1989). However, the explanatory power of *breastfeeding duration* is largely limited to effects of *succeeding birth interval* through its impact on the return of fertility. Early cessation of breastfeeding may expose the child to greater risks of illness from contaminated water and food in conditions where proper substitutes of food are scarce. In these circumstances, *breastfeeding duration* could explain some effects of *succeeding birth interval* on childhood mortality. In some cases, an early pregnancy may cause premature weaning and not vice versa; the result being that *breastfeeding duration* is positively related to both the *succeeding birth interval* and the child survival.

Both *maternal age* and *birth order* tend to exhibit a U - shaped relation with childhood mortality (Miller *et al.*, 1992; Sastry, 1997). Young mothers have reproductive systems that are not completely mature, and this leads to under-weight and weaker babies, while older mothers have declining maternal resources due to aging. Young mothers are also less likely to have received adequate prenatal care.

High-order births have relatively higher childhood mortality because they are born to older women. First-born children are more likely to be born to young mothers.

In addition to household characteristics such as *electricity* and *toilet facilities*, measures such as *maternal* and *paternal schooling*; and *maternal occupation* are often used to indicate socioeconomic status of a family. Environmental measures such as *urban-rural residence*, *region* and *maternal health services* are proxies for measures of development and customs which otherwise might be lacking in a data set.

1.2 Multilevel approach

Many studies of child mortality are premised on the standard assumption that the survival experience of children are independent (e.g. Manda (1998b) uses standard proportional hazards model analysis of infant and child mortality in Malawi). However, survival data are often clustered into groups such as those arising from a hierarchical structure in large scale multistage demographic sample surveys. For example, the Demographic and Health Survey (DHS) collects child survival data that are clustered at the family, and community levels. Mortality risks of children from the same family or area tend to be more alike than for children chosen at random from the whole population. This might result from the fact that children in the same family may inherit similar genetic factors such as those causing problems in pregnancy. In addition, siblings are likely to share the same physical environment such as poor *toilet facilities* and *drinking water*. Area environmental factors such as child care practices, use of *health facilities*, general standard of hygiene and average social composition all have a direct impact on mortality risks of children. Some such factors (collectively termed *frailty*) may be either unmeasurable or measurable but not available as individual covariates.

Members within a cluster tend to exhibit a positive intra-cluster correlation. This results in the variance being inflated and consequently estimators of variance which do not contain this inflation will be wrong, resulting in confidence intervals which

are too narrow and differences that appear to be more significant than they really are. This dependency must be taken into account to correctly assess the relationship between the survival experience of a child and the explanatory variables. The term that is now widely used to describe statistical tools for analysing hierarchical data is *Multilevel* modelling (Goldstein, 1995). Other terms such as *Hierarchical* modelling or *Random Coefficient* modelling are also employed depending on one's research interest. Most of the theoretical research on random effects modelling has been linked to simple linear mixed models. We describe the simple linear 3-level model in Appendix A.

1.2.1 Related work

Frailty models have been applied to the analysis of event history data in many areas such as unemployment (Heckman and Singer, 1984) and fertility (Larsen and Vaupel, 1993). Only recently, multilevel frailty models for survival analysis have been proposed. However, research on the statistical analysis of survival data for clustered individuals began about two decades ago. Holt and Prentice (1974) and Huster *et al.* (1989) studied survival experiences in twin studies and matched pair experiments respectively. Clayton (1978) proposed a bivariate hazard model that can be interpreted in terms of a proportional hazards model and a gamma distributed random effect. Subsequently, Oakes (1982) and Clayton and Cuzick (1985a) worked on non-parametric estimation of the association parameter for correlated survival times. This stream of work was mirrored by Vaupel *et al.* (1987) who introduced the notion of frailty in demographic models. Heckman and Singer (1984) studied the sensitivity of covariate effects to the choice of frailty distribution and recommended a non-parametric estimator of frailty. A full discussion of heterogeneity in a language familiar to social statisticians is found in Trussell and Rodriguez (1990)

Recently, some work on the statistical analysis of correlated child survival experiences have appeared in the literature. Curtis *et al.* (1993) and Madise and Diamond

(1995) used a logistic model with an additive Gaussian random effect in the analysis of post-neonatal mortality in Brazil and infant mortality in Malawi respectively. Guo and Rodriguez (1992) presented a proportional hazards model with gamma distributed random effect multipliers in the study of child survival in Guatemala. These and other related works have restricted the analysis to a single random effect mainly at the family level. Our contribution focuses on the full Bayesian statistical inference of child survival data hierarchically clustered at family and and community levels. Sastry (1997) used the EM algorithm for estimation in a similar study done on child survival in Northeast Brazil. We use the *Gibbs* sampler, an iterated simulation algorithm, to provide means of making full Bayesian inference. As far as we know, this is the first time Bayesian inference has been used to this type of model and data. The methodology is applied to the analysis of child survival data from Malawi clustered at the family and community levels.

1.2.2 The model and notation

The child survival data analysed in this study are hierarchically clustered at the family and community levels. We provide the general notation of the main model used in this work. Though the model presented is more complicated than a simple linear mixed model (see Appendix A), the basic ideas and structure are the same. Let $X_{ijk}(t)$ denote a possibly time varying design vector for child k in family j in community i and $\beta(t)$ a possibly time-varying covariate effects vector. For simplicity, the time argument is omitted when referring to the design and covariate effect vectors and they are simply denoted by X_{ijk} and β. The baseline hazard and the integrated baseline hazard functions are $\lambda_0(t)$ and $\Lambda_0(t) = \int_0^t \lambda_0(t)dt$. The integrated fixed effect hazard for child k in family j in community i is

$$\Lambda_{ijk}(t) = \int \lambda_0(t) e^{\beta' X_{ijk}} dt$$

assuming a proportional hazards model (Cox, 1972). In this study, the cluster specific random effects operate multiplicatively on the baseline hazard so that they

are interpreted as relative risks. Thus

$$h_{ijk}(t|\beta, b_i, b_{ij}) = b_i b_{ij} \lambda_0(t) e^{\beta' X_{ijk}} \tag{1.1}$$

is the hazard function for child k in family j in community i given fixed effects β, family random effect b_{ij}, and community random effect b_i. The quantity $\exp(\beta' X_{ijk})$ is interpreted as the relative risk associated with covariate X_{ijk}. The associated integrated hazard function for $h_{ijk}(t|\beta, b_i, b_{ij})$ is given by:

$$\begin{aligned} H_{ijk}(t|\beta, b_i, b_{ij}) &= \int_0^t h_{ijk}(t|\beta, b_i, b_{ij}) dt \\ &= b_i b_{ij} \Lambda_{ijk}(t). \end{aligned}$$

The survival function for child k in family j in community i is $S_{ijk}(t|\beta, b_i, b_{ij}) = \exp[-H_{ijk}(t|\beta, b_i, b_{ij})] = \exp[-b_i b_{ij} \Lambda_{ijk}(t)]$ and the time of death density is

$$f_{ijk}(t|\beta, b_i, b_{ij}) = S_{ijk}(t|\beta, b_i, b_{ij}) \times h_{ijk}(t|\beta, b_i, b_{ij}). \tag{1.2}$$

Let t_{ijk} be the time child k in family j in community i leaves the study either by death or censoring (survival past end of study). In the event of death, this child contributes $f_{ijk}(t_{ijk}|\beta, b_i, b_{ij})$ to the likelihood function. Otherwise its contribution is $S_{ijk}(t_{ijk}|\beta, b_i, b_{ij})$. Now let w_{ijk} take on value 1 if child ijk dies during the study or value 0 if censored. Then the contribution to the likelihood of child ijk is

$$\begin{aligned} L_{ijk}(t_{ijk}|\beta, b_i, b_{ij}) &= [f_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{w_{ijk}} \times [S_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{1-w_{ijk}} \\ &= [S_{ijk}(t_{ijk}|\beta, b_i, b_{ij}) \times h_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{w_{ijk}} [\, S_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{1-w_{ijk}} \\ &= [h_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{w_{ijk}} \times S_{ijk}(t_{ijk}|\beta, b_i, b_{ij}) \\ &= \left(b_i b_{ij} \lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right)^{w_{ijk}} \times e^{-H_{ijk}(t_{ijk}|\beta, b_i, b_{ij})} \\ &= [b_i b_{ij}]^{w_{ijk}} \left[\lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right]^{w_{ijk}} \times e^{-b_i b_{ij} \Lambda_{ijk}(t_{ijk})} \end{aligned} \tag{1.3}$$

Therefore, the likelihood of all children in family j in community i is

$$\prod_{k=1}^{K_{ij}} [h_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{w_{ijk}} \times S_{ijk}(t_{ijk}|\beta, b_i, b_{ij})$$

$$= (b_i b_{ij})^{\sum_{k=1}^{K_{ij}} w_{ijk}} \left[\prod_{k=1}^{K_{ij}} \left(\lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right)^{w_{ijk}}\right] \times e^{-\sum_{k=1}^{K_{ij}} H_{ijk}(t_{ijk}|\beta, b_i, b_{ij})}$$

$$= (b_i b_{ij})^{\sum_{k=1}^{K_{ij}} w_{ijk}} \left[\prod_{k=1}^{K_{ij}} \left(\lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right)^{w_{ijk}}\right] \times e^{-b_i b_{ij} \sum_{k=1}^{K_{ij}} \Lambda_{ijk}(t_{ijk})} \quad (1.4)$$

and for all children in community i is

$$\prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} [h_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{w_{ijk}} \times S_{ijk}(t_{ijk}|\beta, b_i, b_{ij})$$

$$= (b_i)^{\sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} w_{ijk}} \left(\prod_{j=1}^{J_i} b_{ij}^{\sum_{k=1}^{k_{ij}} w_{ijk}}\right) \left(\prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left(\lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right)^{w_{ijk}}\right) \times$$

$$e^{-b_i \sum_{j=1}^{J_i} b_{ij} \sum_{k=1}^{K_{ij}} \Lambda_{ijk}(t_{ijk})} \quad (1.5)$$

and finally for all children in the study is

$$\prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} [h_{ijk}(t_{ijk}|\beta, b_i, b_{ij})]^{w_{ijk}} \times S_{ijk}(t_{ijk}|\beta, b_i, b_{ij})$$

$$= \prod_{i=1}^{I} \left[(b_i)^{\sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} w_{ijk}} \prod_{j=1}^{J_i} b_{ij}^{\sum_{k=1}^{k_{ij}} w_{ijk}}\right] \times \left[\prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left(\lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right)^{w_{ijk}}\right] \times$$

$$e^{-\sum_{i=1}^{I} b_i \sum_{j=1}^{J_i} b_{ij} \sum_{k=1}^{K_{ij}} \Lambda_{ijk}(t_{ijk})} \quad (1.6)$$

A less familiar representation of the model based on the theory of rank regression (Clayton and Cuzick, 1985b) transforms the failure times on a scale on which a log-linear exponential failure time model operates. The transformation is in some arbitrary monotone function linking to the integrated baseline hazard function.

However, in this study, we adopt a different approach that not only gives us estimates of the baseline hazard over time, but also correct estimates of regression parameters β. Instead of allowing the baseline hazard to be non-parametric as suggested by Cox (1972), it will be piecewise constant. Thus, suppose that the survival time T has been partitioned into Q intervals $\{A_q = [t_{q-1}, t_q)\ t_0 = 0, t_Q = \infty)\}$. The baseline hazard in each interval is modelled as constant with parameter $\lambda_q = \exp(\vartheta_q)$, so its estimation can be absorbed into the fixed effect vector β. Let y_{ijkq} and e_{ijkq} be the indicator of death and exposure time, respectively, for child k in family j in community i in the q^{th} interval. The survival time for child ijk is then

expressed as a sequence of binary responses y_{ijkq} that indicate if child ijk died in each interval at risk. Under this approach the likelihoods (1.3)-(1.6) are equivalent to a generalised linear model likelihood with a Poisson response y_{ijkq}, with a log link function (Breslow, 1974; Laird and Oliver, 1981; Whitehead, 1980) and offsets $\log e_{ijkq}$ and $\log b_i b_{ij}$. That is, the random effects GLM is

$$\log \mu_{ijkq} = \log e_{ijkq} + \beta' X_{ijkq} + a_i + a_{ij} \tag{1.7}$$

where μ_{ijkq} is the mean of y_{ijkq} for child ijk in interval q and $\log a_i = \log b_i$ and $a_{ij} = \log b_{ij}$.

Before going any further, we need to explain some important considerations. The computational time for fitting this model, even in its standard form, is very considerable. This is so because the number of observations becomes not the number of failure times observed, but the number of children-intervals.

1.3 Random effects generalised linear model

We have seen that the grouped version of the proportional hazards model described above can be expressed in the class of random effects generalised linear models and this formulation will be used in the thesis. Generalised linear models (GLMs), as described in much detail by McCullagh and Nelder (1989), have unified regression methods for various discrete and continuous response variables that can be assumed to be independent. These models have received considerable attention in the applied statistics in the past two decades. The definition of GLM has appeared in many research papers and books, and is now well known. For the benefit of those unfamiliar with these models, a brief description is given here. Let the data set in family ij consist of observed responses y_{ijk} $(k = 1, 2, \ldots, K_{ij})$ and a $K_{ij} \times p$ known matrix X_{ij} of explanatory variables. The response y_{ijk} is assumed to come from a random variable Y independently distributed with mean μ. Following McCullagh and Nelder (1989), the GLMs are then characterised by the following structure:

1. The distribution of an individual response y_{ijk} is assumed to belong to the natural exponential family:

$$f(y_{ijk}) = \exp[\{\theta_{ijk}y_{ijk} - b(\theta_{ijk})\}\frac{\omega_{ijk}}{\phi} + c(y_{ijk}, \phi, \omega_{ijk})]$$

for some known functions $b(.)$ and $c(.)$; for a known positive weight ω_{ijk} that deals with data records which are summaries of raw data; for an unknown scalar parameter θ_{ijk} and for a known scale related parameter ϕ.

2. The explanatory variables in vector x_{ijk} influence y_{ijk} via a linear combination $\eta_{ijk} = \beta' x_{ijk}$ where β is a $p-$dimensional parameter vector and η_{ijk} is called the *linear predictor.*

3. The linear predictor η_{ijk} is related to the mean μ_{ijk} of y_{ijk} by a *link function* g such that $\eta_{ijk} = g(\mu_{ijk})$. If $g(\mu_{ijk}) = \theta_{ijk} = \beta' x_{ijk}$, then the model is a canonical GLM and $g(.)$ is the *canonical link.*

Writing the matrix of explanatory variables as $X_{ij} = (x_{ij1}, \ldots, x_{ijK_{ij}})'$ and the column vector of linear predictors as $\eta_{ij} = (\eta_{ij1}, \ldots, \eta_{ijK_{ij}})'$, the linear model is

$$\eta_{ij} = X_{ij}\beta$$

The mean and variance of y_{ijk} are given by $E(y_{ijk}) = \mu_{ijk} = b'(\theta_{ijk})$ and $V(y_{ijk}) = \phi b''(\theta_{ijk})/\omega_{ijk}$ respectively. The function $V(\mu_{ijk}) = b''(\theta_{ijk}) = b''(b')^{-1}(\mu_{ijk})$ is often termed the *variance function.*

The above family of distributions includes some very familiar models. For the normal distribution, the canonical link is the identity $g(\mu_{ijk}) = \mu_{ijk}$ and we obtain the general linear model. The values of ϕ and ω are σ^2 and 1 respectively. For the Poisson distribution, the canonical link function is $g(\mu_{ijk}) = \log(\mu_{ijk})$, giving rise to the log linear Poisson model. For the Poisson model, ϕ and ω_{ijk} are both equal to 1. For the binomial distribution with mean proportion π_{ijk} the link functions $g(\pi_{ijk}) = \log(\pi_{ijk}/(1-\pi_{ijk}))$, $g(\pi_{ijk}) = \Phi^{-1}(\pi_{ijk})$ and $g(\pi_{ijk}) = \log(-\log(1-\pi_{ijk}))$

result in the logistic, probit and the complimentary log-log models respectively. Here the value of ϕ is 1 and ω_{ijk} equals the sample size.

More recently there has been an increased interest in extending these models to include random effects on the same lines as those extensions to the Gaussian general linear model as described in Bryk and Raudenbush (1992). Such generalisation stems naturally from the need to accommodate more complex error structures such as those arising in multilevel models (Goldstein, 1995). However, inference in random effects GLMs runs into computational problems. This computational burden has limited data analysis in several ways such as using a particular random effects distribution (e.g. the Gaussian).

1.3.1 The frequentist approach

Let the random effects $a = (a_{11}, \ldots, a_{IJ_I}, a_1, \ldots, a_I)'$. The usual approach of extending GLM to include a is by rewriting the regression equation (Longford, 1993; Steel, 1996) for the linear predictor as

$$\eta = X\beta + Za$$

where Z defines the cluster level design matrix with random coefficients in a. This leads to models which specify the conditional distribution of the observed response variable given the random effects and a marginal distribution for the random effects. Let the random effects distribution be $f(a)$ with mean D and variance Σ. The parameters $\Theta = \{D, \Sigma\}$ commonly known as *hyperparameters* must also be estimated from the data. In the rest of this thesis, it is assumed that the random effects a_{ij} and a_i are independent within and between themselves. Then the joint distribution of the observed response vector and the random effect vector is immediate. However, maximum likelihood (ML) based inference must use the marginal distribution of the response vector alone since the random effects are unobservable. Thus we have

$$L(\beta, \Theta) = \prod_{i=1}^{I} \int \left[\prod_{j=1}^{J_i} \int \prod_{k=i}^{K_{ij}} f(y_{ijk}|\beta,\, a_{ij},\, a_i)\, f(a_{ij})\, da_{ij} \right] f(a_i)\, da_i \qquad (1.8)$$

Note that the ML estimates of variance components will be biased downwards since the degrees of freedom lost due to the estimation of the fixed-effects regression parameters are ignored. This is often resolved by the use of restricted or residual maximum likelihood (Longford, 1992). The random effects a are estimated by empirical Bayes estimates, defined as the mean of the distribution

$$f(a|y, \beta = \beta_{ML}, \Theta = \Theta_{ML}) = \frac{g(y|\beta = \beta_{ML}, a) f(a|\Theta = \Theta_{ML})}{\int g(y|\beta = \beta_{ML}, a) f(a|\Theta = \Theta_{ML}) da}$$

where y is the overall observed data and the ML estimates are obtained from (1.8). A defect of the 'plug in' type empirical Bayes estimators is that they ignore the variability in estimating β and Θ, so confidence intervals obtained this way are too short.

Except in the normal linear model with normally distributed random effects, the computation of (1.8) presents substantial problems because the marginal distribution of the response variable is usually not closed. In some instances numerical integration has been used, but for more complicated problems involving more than one dimension these techniques are no longer reliable or feasible. Zhaorang *et al* (1992) and McGilchrist (1994) avoid the need for integration by extending the linear mixed model approach of best linear unbiased prediction (Robinson, 1991) using linearization arguments. Zeger *et al.* (1988) introduced the generalised estimation equations approach whose main focus is with modelling the fixed coefficients rather than exploring the structure of the random component of the model. Breslow and Clayton (1993) construct a Laplace approximation for the marginal quasi-likelihood which is then maximised via linearization methods (see also Goldstein (1995)). Bryk and Raudenbush (1992) describe the use of the EM algorithm in linear mixed models and Steel (1996) describe its use in generalised mixed models. More recently, full Bayesian procedures have become computationally feasible with the development of Markov Chain Monte Carlo (MCMC) methods especially the Gibbs sampler (Zeger and Karim, 1991).

1.3.2 The Bayesian approach

From the Bayesian perspective, the parameters β and Θ are random variables and are treated symmetrically with the observed data y and the unobserved data a. To complete the Bayesian formulation of the random effects GLM, it is only necessary to specify a joint *hyperprior* distribution $g(\beta, \Theta)$ for β and Θ. Thus, the first objective is to derive the joint posterior distribution $g(a, \beta, \Theta|y)$ given by

$$g(a, \beta, \Theta|y) = \frac{g(y|\beta, a)g(a|\Theta)g(\beta, \Theta)}{\int\int\int g(y|\beta, a)g(a|\Theta)g(\beta, \Theta)da d\beta d\Theta}$$

from which, by integration, marginal posteriors $g(\beta, \Theta|y)$ and $g(a|y)$ are found. Analytical evaluation of either $g(\beta, \Theta|y)$ or $g(a|y)$ is typically intractable. In chapter five, we show how this computational difficulty is easily overcome by the Gibbs sampler in a random effects Poisson formulation of the nested frailty survival model.

1.4 The data set

We will be studying child survival data from Malawi, a land locked country in Southern Africa. Malawi is bordered to the north and northeast by Tanzania; to the east, south and southwest by Mozambique; and to the west by Zambia (see map on page x). The country is still in a high birthrate and child mortality stages of development. Estimates from the censuses conducted in 1977 and 1987 show that total fertility rates were 6.9 and 7.1 per woman and infant mortality rates were 165 and 159 per 1000 live births respectively (National Statistics Office (NSO), 1994). The Ministry of Health is making a major effort to reduce these rates through various means such as community mobilisation campaigns.

The National Statistics Office and MACRO International conducted the Demographic and Health Survey (DHS) of Malawi [1] in September, 1992. The survey

[1] A brief descriptive and quality analysis of the data is available from the author. Further details of the survey methodology, sample design, description and quality of the data are contained in the final report of the survey (NSO, 1994).

gathered information to assess the effectiveness of the Mother and Child Health Program in meeting its stated objectives and to show the most effective directions for improving child survival rates. Each of the three regions: Northern, Southern and Central was stratified into urban and rural resulting in six strata overall. A stratified random sample of 225 enumeration areas (EA) was selected, with each stratum having selection probability proportional to number of households it contains. The distribution of the selected EAs is shown in the map of Malawi on page x. A systematic random sample of households was then taken from each selected enumeration area, with the sample size of households in each EA proportional to its number of households. The sampled proportion was greater from the smaller strata because reliable estimates of various demographic and health indicators for the three regions, and urban-rural areas were needed. Thus, the sample is not self weighting at the national level, but it is so within each stratum. We protect against this selection bias by reweighting the responses in any analytical inference (Pfeffermann, 1993).

The selected households formed the DHS sample within which all women aged 15-49 were to be interviewed totalling 4878 respondents. Each respondent provided a complete retrospective birth history for all their live births. The present analysis is limited to births that occurred during the 5 years preceding the survey. This ensures that the results relate to current determinants of childhood mortality. Breastfeeding practices and maternal health services utilisation which are used as covariates in the analyses were not recorded for births outside this time window. Furthermore, current status variables, such as household amenities, would adequately match the experiences of recent births. All multiple [2] births were removed from the sample, as were those without a specified birth month and year. The final sample consisted of the 4838 singleton births that had occurred to 2911 women in 225 communities. Out of this sample, 745 children had died at the time of the survey. For this sample,

[2]Multiple births generally are already at high risk of death; and this might affect modelling of effects of some factors.

estimates of infant (${}_1q_0$), child (${}_4q_1$) and under-five (${}_5q_0$) mortality rates are 118, 115 and 249 per 1000 live births respectively. The survey report gives rates which are based on all births in the period 0-4 years preceding the survey; therefore they are slightly different.

1.5 Objectives

The primary aim of this thesis is the analysis of determinants of child mortality in Malawi. This is achieved by developing and evaluating methods for statistically modelling survival times. The problem of how various risk factors affect child survival is explored using these models. Thus the thesis is both substantive and methodological. With proper data quality evaluation and model specification, the information collected as part of Demographic and Health Surveys is the best available to study the determinants of childhood mortality in Malawi.

It is also recognised that some families might have higher risks of child death than others. It may also be true that some communities have higher rates of child losses than other communities. Therefore, this study also aims to quantify the magnitude and importance of clustering at the family and community levels singly and jointly. This will be achieved by using the nested random frailty effect models. The main focus would then be the full Bayesian inference which has not previously been used on this type of analysis and data.

The estimation of coefficients of the explanatory variables and variance for the family and the community random effects will also be estimated using the expectation-maximisation (EM) algorithm. A subsidiary goal of the thesis is to compare the resulting estimates of the EM algorithm to the full Bayesian estimates.

1.6 Thesis organisation

The thesis is organised into six chapters. This chapter surveys and summarises the literature on determinants of child mortality, introduces the DHS data set, and defines the research objective. In most cases, the analysis of survival time has been reduced to a binary response indicating death or survival in a particular interval. Such an analysis has been common in child mortality studies. Chapter two presents a preliminary analysis of the data using the correlated binary model. It identifies the important determinants of childhood mortality and shows the relative magnitude of the family and the community frailty variances present in the data set. The modified proportional hazards model is first analysed in chapter three using the expectation-maximisation (EM) algorithm. In this chapter, we compare parameters estimates between standard hazards model and the three different frailty models, containing family random effect, community random effect, and both family and community random effects respectively.

As discussed earlier, the major impediment to full Bayesian implementation of random effects GLM has been the difficulty of evaluating the integrals to obtain posterior distributions. Chapter four presents the theory behind the Monte Carlo Markov Chain (MCMC) methods in Bayesian statistics. These provide a relatively straightforward means of making full Bayesian inference in random effects GLMs. Chapter five presents results based on the use of the Gibbs sampler, an MCMC method. Summary and discussion are given in chapter six which also contains conclusions and recommendations for future work.

Chapter 2

Preliminary analysis

2.1 Introduction

In this chapter, we perform preliminary analysis on the Malawi child survival data using the discrete survival model and controlling for the unobserved family and community random effects. This is one of the commonly used methods in survival analysis. The data are not truly discrete; rather the underlying continuous survival time of a child would be made discrete by grouping. The analysis here is preliminary to the main focus of the thesis which is the continuous time survival model. A paper by Manda (1998a) arises from this chapter.

Several authors have investigated the dependency of binary data within a cluster. The beta-binomial model (Griffiths, 1973; Williams, 1975, 1982; Crowder, 1978) is among the earliest. Here covariates can not vary within a cluster, and clustering is at one level. The probability of a positive response is assumed to vary across clusters according to a beta distribution. The marginal distribution of the sum of responses in a cluster is beta-binomial, and its parameters quantify the intra-cluster correlation. The discrete survival model with random effects considered here allows both elementary and cluster level covariates and any number of levels of hierarchy. The association parameter within a cluster is provided by the variance of random effect distribution at that level of clustering.

2.2 Model formulation

Consider the survival experience of child ijk who dies or leaves the study in interval A_{ijkq}. That is, the child experiences a sequence of censoring in $A_1, A_2, \ldots$ and either fails or is finally censored in interval A_{ijkq}. The risk set R_q is the set of children still alive at t_{q-1}, the start of interval A_q. The death indicator y_{ijkq} for child ijk in interval A_q is 1 if the child fails in A_q or 0 otherwise and is defined for all children in R_q. Conditional on a_{ij} and a_i and given k in R_q, we can treat the random variables y_{ijkq} as independent Bernoulli trials across k and q for likelihood estimation and inference (Prentice and Gloeckler, 1978; Aitkin, *et al.*, 1992; Scheike and Jensen, 1997; Xue and Brookmeyer, 1997 and references therein). Let the conditional probability of failure in interval A_q given that child ijk is alive at t_{q-1} be

$$Pr(Y_{ijkq} = 1) = p_{ijkq} = Pr(t_{q-1} \leq t_{ijk} < t_q | t_{ijk} \geq t_{q-1};\ a_{ij}, a_i)$$

It is assumed that the response y_{ijkq} is the expression of an underlying continuous process U, commonly known as latent or tolerance variable (Im and Gianola 1988; Anderson and Aitkin 1985; Pebley, *et al.*, 1996). A child dies if its tolerance exceeds a certain threshold value u. Specifically, if U_{ijkq} represents the underlying latent variable associated with child ijk in interval A_q we have

$$p_{ijkq} = P(U_{ijkq} > u_{ijkq})$$

It is further assumed, without any loss of generality, that U_{ijkq} follows a mixed linear model

$$U_{ijkq} = x'_{ijk}\beta + z'_2 a_{ij} + z'_3 a_i + e_{ijkq}$$

where z_2 is family ij covariate vector with associated random parameter vector a_{ij}; z_3 is community i covariate vector with associated random parameter vector a_i and e_{ijkq} are identically distributed random variables with unimodal density $f(u)$ having variance σ_1^2. The simple case when $z_2 \equiv z_3 \equiv 1$ is only considered, and it is assumed that the random variables a_{ij} and a_i are normal with mean 0 and variances

σ_2^2 and σ_3^2 respectively. Any continuous unimodal distribution can be considered for e_{ijkq}. It is also convenient to assume that the random variables a_{ij}, a_i and e_{ijkq} are independent within and between themselves. See Appendix A for further details in this model formulation.

2.3 Estimation of parameters

The unobserved threshold u_{ijkq} varies over children, and there would be no loss of generality in setting $u_{ijkq} = 0$. This convention is often adopted whatever the distribution of U_{ijkq}. If U_{ijkq} follows a standard logistic distribution then

$$p_{ijkq} = \frac{\exp(x'_{ijk}\beta + a_{ij} + a_i)}{1 + \exp(x'_{ijk}\beta + a_{ij} + a_i)} \tag{2.1}$$

This model for p_{ijkq} gives a *logit* link defined as $\log[p_{ijkq}/(1 - p_{ijkq})]$. The logistic model had been considered by Xue and Brookmeyer (1997) and is similar to the model proposed by Cox (1972) for discrete time data.

The complementary *log-log* link applied to the conditional probabilities gives

$$\log(-\log(1 - p_{ijkq})) = x'_{ijk}\beta + a_{ij} + a_i$$

The complementary *log-log* model can be interpreted mathematically as a grouped version of the usual continuous time Cox hazards regression model (Scheike and Jensen, 1997). This is seen by noting that for the proportional hazards model:

$$p_{ijkq} = 1 - \exp[-e^{\beta_{0q} + x'_{ijk}\beta}]$$

where $\beta_{0q} = \log\{(\int_{t_0}^{t_q} \lambda_0(t) - \int_{t_0}^{t_{q-1}} \lambda_0(t)\}$. Since we will analyse the continuous time proportional hazards model in chapters three and five, there is no need to analyse the discrete complementary *log-log* here, and only the logistic model will be considered. The likelihood function for all responses can be written as

$$L = \prod_{i=1}^{I} \int \left[\prod_{j=1}^{J_i} \int \prod_{k=i}^{K_{ij}} \prod_{q=1}^{ijkq} p_{ijkq}^{y_{ijkq}} (1 - p_{ijkq})^{1 - y_{ijkq}} \, f(a_{ij}) \, da_{ij} \right] f(a_i) \, da_i \tag{2.2}$$

where $f(.)$ are appropriate frailty distributions often chosen on the grounds of identifiability and computational convenience. The maximum likelihood estimates of the frailty distribution parameters and fixed effect parameters are obtained by directly working with the marginal likelihood function (2.2) or by using the EM algorithm (Anderson and Aitkin, 1985; Im and Gianola, 1988) with the unobserved data a_i and a_{ij}. This is not always easy since the computations usually involve complicated high dimensional integrals.

The *Iterative Generalised Least Squares* (IGLS) estimation procedure is illustrated here (see also Goldstein, 1995). Note that the model described for p_{ijkq} implies that the response is nonlinear in the fixed and random parameters. In general for child ijk, the response y_{ijkq} is

$$y_{ijkq} = E(y_{ijkq}) + e_{ijkq}$$

where $E(y_{ijkq})$ is expressed as

$$p_{ijkq} = E(y_{ijkq}) = g(x'_{ijk}\beta + a_{ij} + a_i) \tag{2.3}$$

For the model defined in (2.1), g is the logistic function. We start by linearising g so that the model assumes the form of a standard three-level normal model as described in Appendix A. Suppose g is expanded about z_0 to give

$$g(z_0 + \delta) = g(z_0) + \delta g'(z_0) + \frac{1}{2}\delta^2 g''(z_0) + \cdots \tag{2.4}$$

and if $\delta = \zeta + u$, then

$$\begin{aligned} g(z_0 + \zeta + u) &= g(z_0) + \zeta g'(z_0) + u g'(z_0) + \frac{1}{2}\zeta^2 g''(z_0) \\ &+ \frac{1}{2}u^2 g''(z_0) + \zeta u g''(z_0) + \cdots \end{aligned} \tag{2.5}$$

The idea here is to express the logistic function g as a linear approximation in the fixed effect parameter β and random effects a_i and a_{ij} so that the IGLS procedure can be used. Suppose there exist some reasonable estimates $\beta^{(r)}$, $a_{ij}^{(r)}$ and $a_i^{(r)}$ at some point r. For the next step $(r+1)$, it is possible to use (i) $z_0 = x'_{ijk}\beta^{(r)}$ or (ii) $z_0 =$

$x'_{ijk}\beta^{(r)} + a_{ij}^{(r)} + a_i^{(r)}$, the current estimates of β and $a = (a_{11}, \ldots, a_{IJ_I}, a_1, \ldots, a_I)'$ which are the unobserved variables just like in the EM estimation method. For (i) $z_0 + \zeta = x'_{ijk}\beta^{(r+1)}$ so that $\zeta = x'_{ijk}(\beta^{(r+1)} - \beta^{(r)})$. This gives $E^{(r+1)}(y_{ijkq}) = g(x'_{ijk}\beta^{(r+1)} + a_{ij} + a_i)$ as

$$\begin{aligned} E^{(r+1)}(y_{ijkq}) &\approx g(x'_{ijk}\beta^{(r)}) + (x'_{ijk}(\beta^{(r+1)} - \beta^{(r)})g'(x'_{ijk}\beta^{(r)})) \\ &+ (a_{ij} + a_i)g'(x'_{ijk}\beta^{(r)}) + \frac{1}{2}(a_{ij} + a_i)^2 g''(x'_{ijk}\beta^{(r)}) + \cdots \end{aligned} \quad (2.6)$$

This is essentially a first order expansion for the fixed part about its current estimate and the second order expansion for the random part about zero. The use of approximation (i) is equivalent to the marginal quasi-likelihood (MQL) of Breslow and Clayton (1993). In the case of (ii) $z_0 + \delta = x'_{ijk}\beta^{(r+1)} + a_{ij}^{(r+1)} + a_i^{(r+1)}$ and this gives

$$\begin{aligned} E^{(r+1)}(y_{ijkq}) &\approx g(x'_{ijk}\beta^{(r)} + a_{ij}^{(r)} + a_i^{(r)}) \\ &+ x'_{ijk}(\beta^{(r+1)} - \beta^{(r)})g'(x'_{ijk}\beta^{(r)} + a_{ij}^{(r)} + a_i^{(r)}) \\ &+ ((a_{ij}^{(r+1)} + a_i^{(r+1)}) - (a_{ij}^{(r)}) + a_i^{(r)}))g'(x'_{ijk}\beta^{(r)} + a_{ij}^{(r)} + a_i^{(r)}) \\ &+ \frac{1}{2}((a_{ij}^{(r+1)} + a_i^{(r+1)}) - (a_{ij}^{(r)} + a_i^{(r)}))^2 g''(x'_{ijk}\beta_{(r)}) + a_{ij}^{(r)} + a_i^{(r)}) \\ &+ \cdots \end{aligned} \quad (2.7)$$

and the expansion for the random part is about the current estimate for levels 2 and 3 residuals rather than zero. Breslow and Clayton call this approach a penalised or *predicted* quasi-likelihood (PQL)

In the linear approximations (2.6) and (2.7) it is possible to use first or second-order terms and together with MQL or PQL there are four possible choices of estimation. The first-order approximation is simple and computationally more robust than second-order while PQL is more accurate but computationally less stable than MQL. All these options are available in the program MLn which is used here to estimate all the parameters.

The estimation method described here is only approximate and a recent study (Rodriguez and Goldman, 1995) has shown that the estimates of both fixed effects and random effect variances are biased toward zero whenever the random effects are interestingly larger and the number of level-1 units in any given level of grouping is small.

2.4 Inference

Some modifications are needed to test for a zero random effect variance because the null hypothesis value lies on the boundary of the parameter space. In this situation, for any parameter, Self and Liang (1987) showed that the likelihood ratio test is a mixture of χ^2_0 and χ^2_1 distributions, ie a mixture of a constant and a square of standard normal. Thus, the correct approach is to use an upper tail critical region of a normal distribution to test for the presence of random effects. For instance, the usual ratio z test, for a random effect variance is compared to a critical value of 1.645 at the .05 significant level. The remaining parameters are tested in exactly the same way as in the standard logistic regression.

In the presence of a significant family or community random effect, the interpretation of parameter estimates is slightly different due to the non uniqueness of probability of death for children with a particular set of covariates. The probability depends on both the observed covariates and the unobserved family or community random effects. The odds ratio estimated from the random effects model is therefore specific to a combination of family and community random effects. As noted earlier, the random effect vector has mean 0, hence $x'_{ijk}\beta$ is the mean of $logit(p_{ijkq})$ for child ijk across families and communities. The probability p_{ijkq} of dying for a child can be thought of as an average probability plus that child's family and community random effects. Thus, for two children with the same characteristics, but from different families within the same community, the probability of dying will differ as a result of differences in family random effects. Similarly for two children with the

same characteristics, but from different families and communities, the probability of dying will differ as a result of family and community random effects.

2.5 Application to infant mortality

One approach to the analysis of discrete time survival data is to perform a series of logistic regression models for individuals at risk in each of the intervals. Here only the infancy period is considered: that is, we are interested in the conditional probability of death of a child in the period 0-12 months. Manda (1998b) considers the childhood mortality (0-59 months) using the standard proportional hazards models. We perform logistic regression on infant mortality using additive family and community random effects. To avoid the problem of censoring, only births that were exposed to risk of dying in the entire infant period were analysed.

The effective sample size for the analysis was 3927 with corresponding 499 infant deaths. These 3927 births occurred to 2650 women in 225 communities. Table 2.1 presents the distribution of families and births by number of births in the family. It is seen that 44 percent of families contributed at least 2 births to the sample. The distribution of communities by number of births in the community is displayed in Table 2.2.[1] About 85 percent of communities contributed at least 11 births each. Therefore in the presence of family and community effects on infant mortality, the individual child survival experiences would not be independent.

As a preliminary analysis, the sample was checked for the presence of family and community effects. Tables 2.3 and 2.4 show the distribution of infant deaths per family and community respectively. Only 1.4 percent of the women had at least two infant deaths, but these losses account for a substantial 16 percent of the total number of infant deaths. About 23 per cent of the communities have at

[1]The enumeration areas correspond reasonably to villages and these units provide our measure of a community. It is possible that there was emigration from some communities and this might result in those communities having 10 or fewer births.

Table 2.1: Distribution of families and births by number of births in the family: 1987-92, Malawi DHS

Number of births in a family	% of families	% of births
1	56.2	38.0
2	39.7	53.5
3	3.9	8.0
4	0.2	0.5
Total	100.0	100.0

Table 2.2: Distribution of communities by number of births

Number of births in a community	% of communities
2-10	15.1
11-20	52.9
21-40	32.0
Total	100

least four infant deaths each and these account for 50 percent of all infant deaths. This indicates some presence of clustering of infant deaths in certain families or communities. Consequently a standard logistic model might not be valid in these circumstances.

The results of the standard logistic and random effects logistic models are presented in Table 2.5. The column z-statistic is defined as the ratio of the parameter estimate to its standard error (not shown in the Table). This ratio is standard normal under the assumption that the parameter is zero. Many of the variables described in the previous chapter are not included because they were found not to have had a significant effect either in the univariate logistic models or after controlling for the other variables. Covariates *breastfeeding duration* and *succeeding birth inter-*

Table 2.3: Distribution of infant deaths in the family

Number of deaths in a family	% of families	% of deaths
0	60.0	-
1	15.0	83.4
2	1.3	14.0
3	0.1	1.8
4	0.03	0.2
Total	100.0	100.0

val were considered not appropriate for infant mortality. These two were excluded because nearly all children in Malawi are breastfed during the first year unless they die or are critically ill; and not many births will be followed by another birth within a year. We fit Model I, the standard logistic model that includes all individual and group variables but with no correction for family and community clustering, and Model II, which includes all individual and group variables as well as random effects for family and community. The variance of the family random effects equals 0.38 which is significant ($p < 0.05$), while the community random effect is not.

The fixed effect estimates and their standard errors are generally quite similar in the standard and random effect logistic models. This result is important from a demographic point of view, considering that all major studies on infant mortality risks have ignored dependency of individual child responses. The only noticeable changes occur in the z-statistic for a *previous dead child* and for a short *preceding birth interval*, which reduce from 3.00 to 2.21 and 2.07 to 1.69 respectively. Families that suffer multiple infant losses would have short *preceding birth intervals*, and an infant death is more likely to be followed by another death. In the ordinary logistic model, these two variables act as surrogates for a positive family effect. When family effects are systematically controlled, the importance of these variables is reduced and only affect the birth which immediately follows or precedes.

Table 2.4: Distribution of communities by number of infant deaths

Number of deaths in a community	% of communities	%of deaths
0	18.0	-
1	22.0	10.0
2	20.0	18.0
3	16.0	22.3
4	12.0	22.5
5	6.0	14.0
6	2.0	6.0
7	2.0	5.6
8	1.0	1.6
Total	100	100

The odds ratios of variables not involving interactions are obtained by exponentiating the estimated coefficient. A modification of this approach is used to obtain odds ratio for any variable involved in an interaction, since its effect can not be discerned in isolation. For example, the odds of death for an infant with a *preceding birth interval* of less than 20 are exp(0.27) = 1.31 times as great if the *age of the mother* is less or equal to 35 yaers. However, for mothers aged between 36 to 49 years, the odds increase to exp(0.27 + 0.52) = 2.02 times as great because of the interaction. In a similar manner, the effect of high *birth order* (6 +) for children born to women aged 19 to 35 years is -1.40 + .99 = -0.40 and -1.40 for children born to mothers aged other than 36 to 45 years. The odds of infant death for teenage mothers is constrained to be always one. The effects of *maternal age* of 19 to 35 years on infant deaths is -.21 with a short *preceding birth interval.* Yet if *birth order* is 6 or more the effect associated with *mother's age* 19 to 35 is -0.21 + 0.99 = 0.79.

Probabilities (not shown) of infant deaths by *age of mother* and *birth order* were

also calculated. Controlling for *birth order*, the relationship of infant mortality with *age of mother* exhibits a U-shaped curve. However, when the *age of mother* is controlled, probabilities of infant death were consistently greater for first-births and decreased steadily with increasing *birth order*. This is unexpected, since we would expect the relationship between infant mortality and *birth order* to also be U-shaped. One possible explaination is that if the mother is aged between 36 and 49 years, there is less compression of previous births. Thus, higher-order births might have lower probabilities of dying than other births. On the other hand, older mothers might have learnt from the experience of previous child sicknesses and deaths to be extra careful with the following children- thus increase the survival chances of later births.

The death of a preceding sibling increases the risks of infant mortality. The odds of dying are about 1.35 times more likely than if the preceding sibling had survived. Infant mortality risks are higher for children whose mothers received 7 years or less of education. The odds of dying are .74 in favour of children whose mothers had at least 8 years of education. Despite intense debate on causal mechanism of *mother's education* on infant mortality (Miller *et al.*, 1992; Hobcraft, 1993;), it is generally agreed that *mother's education* influences child health through child care practices and nutrition. Mothers with some education are more likely to seek medical care for their children than mothers without any education. They may also be aware of their child's basic nutritional needs and proper sanitary conditions. *Father's education* also plays a crucial role in child survival. Children whose fathers have attained at least a secondary education are more likely to survive infancy period than other children.

Children whose mothers do not receive *professionally supervised delivery* are 1.26 times as likely to experience infant death as compared to those whose mothers have such deliveries. Evidence from the DHS report indicate that the provision and use of maternal and child health services is associated with reduced risks of infant death.

Mothers who attend antenatal care are adequately protected against tetanus toxoid, which is one of the main causes of infant deaths especially in the neonatal period. In Malawi, it is natural to see a woman having babies without seeing any trained personnel. Such births are exposed to unnecessary early infectious illnesses, and simple complications often result in infant or maternal deaths. Access to health services is very limited because of the distance to the nearest health facility. About 54 per cent of women in a typical rural Malawi village would walk a distance of 5 to 30 kilometres to the nearest health centre (NSO, 1994). Another limitation is the trust most rural women put on traditional medicine. In fact, most rural women believe that early childhood illnesses or deaths result from witchcraft and therefore seek help from a traditional healer (Kavinya, forthcoming).

2.6 Conclusion

The results have shown that biodemographic factors such as the *survival status of the preceding sibling*, the *preceding birth interval* and the *age of the mother* are singly and jointly significant in determining infant mortality. It has also been shown that infant mortality varies inversely with *maternal* or *paternal years of schooling*, and use of *medical services* reduces the incidence of infant mortality.

The results have also shown that the family random variation is significant. This measure of variation among infant deaths across families quantifies the covariance between a pair of siblings. This implies that a child born in a family experiencing previous infant deaths has an elevated risk of infant death compared to a child born in a family which has no previous infant deaths. It is worthwhile to note that familial variations are significant even after controlling for many possible bio-social, biomedical factors. Therefore the estimated family variation can be thought of as a measure of biological, genetic, and parental competence and other households factors not included in the model. Parental competence in organising and caring for their children is different from biological and genetic factors such as inherited

disorders leading to low weight babies or premature births or early infant diseases. These results support the findings for postneonatal mortality in Brazil (Curtis *et al.*, 1993) and for child mortality in Guatemala (Guo, 1993). In the latter study, the family variation was not significant when controls for direct measures of households income and wealth were made. This implies that some of the residual variations in the present study might be due to differences in the family's income and wealth that are not directly measured. Poor families may not be able to purchase drugs to handle minor illnesses and may not feed their children adequately. The end result is that their children are more frail than children of richer families (Das Gupta, 1990).

The analysis further shows that the variance of the community random effect was not significant in determining infant mortality (standard error for community variance equals 0.05). In the infant period, the child is kept within the household, and often with the mother, since traditionally infants are not supposed to be exposed to the wider environment. Thus, familial effects are stronger than community effects.

We have studied child survival based on survival in the infant age. Thus, the analysis has been reduced to the analysis of a binary (0 or 1) dependent variable. This approach wastes information in that only data on individuals observed to be at risk in the entire age interval is employed. Sample proportions are directly possible only when no individuals are censored during the specified period, and this implies all censored observations are thrown away. This leads to biased results, as actual time to death is not part of the analysis. The use of a proportional hazards analysis uses all information including censored observations. This is followed up in the remaining chapters, where child survival in the period 0 to 59 months is analysed.

Table 2.5: Parameter estimates for logistic models without (Model I) and with (Model II) random effects respectively

	Model I		Model II	
Parameter	Coefficient	z-statistic	Coefficient	z-statistic
Constant	-1.69	-5.35	-1.65	-5.24
Previous child died	.42	3.00	.31	2.21
Birth order				
2-5	-.38	-2.05	-.40	-2.16
6+	-1.36	-3.56	-1.40	-3.59
Mother education				
8+ years	-.30	-1.58	-.30	-1.58
Mothers age				
19-35 years	-.21	-1.40	-.21	-1.31
36-49	.60	1.85	.61	1.79
Preceding birth interval				
<= 20 months,none	.31	2.07	.27	1.69
Fathers education				
9+ years	-.24	-1.33	-.24	-1.25
Non skilled delivery	.27	2.30	.27	2.30
Interactions				
Maternal age/interval				
36-49 years/short	.55	1.42	.52	1.31
Other combinations	0.0	-	0.0	-
Maternal age/birth order				
19-35 years/high	.97	2.75	.99	2.70
Other Combinations	0.0	-	0.0	-
Random effect variances				
Family	-	-	.38	1.61
Community	-	-	.00042	0.0

Chapter 3

Analysing the model using the EM algorithm

3.1 Introduction

This chapter presents the first analysis of the multivariate hazard model described in chapter one. The parameters are estimated using the EM algorithm, which is well suited to problems where some of the data are missing. But before doing this, we describe the theory concerning the EM algorithm.

The implementation of standard estimation methods such as the maximum likelihood (ML) is more difficult with incomplete data than when the data set is complete-that is, none of the observations are missing. Yet there are many reasons why particular data sets may be incomplete: some responses may not be observed directly; experiments may fail; values may be reported incorrectly or not at all or the data may be abridged in some way (censored, truncated or grouped). In some applications, such as in mixture models, complete data may be just a conceptual device to formulate the problem for easy estimation. The complete data set x is an augmented form of the observed, or incomplete, set y such that the frequency function for y may be obtained from that of x as a marginal frequency function. That is, if $g(y|\theta)$ and $g_c(x|\theta)$ denote the probability density functions of the observed random

data y and complete random data x respectively, where θ is a vector of unknown parameters then

$$g(y|\theta) = \int g_c(x|\theta)dx$$

where the integral is taken over $\{x : y = f(x)\}$ and the integral is replaced by a sum when the data is discrete.

The treatment of incomplete data may be done in a unified way through the *expectation-maximisation* (EM) algorithm, first presented in its full generality by Dempster *et al.* (1977). The iterative algorithm consists of two conceptually distinct steps at each stage, an expectation, (E) step, and a maximisation, (M) step. These may be kept distinct or merged depending on a particular application. Let $\theta^{(r)}$ denote the estimate of θ obtained at the r^{th} iteration. It is assumed that

$$Q(\theta|\theta^{(r)}) = E\{\log L_c((x|\theta)|y, \theta^{(r)})\}$$

exists for all x and θ, where $L_c(x|\theta) = g_c(x|\theta)$ is the complete data likelihood function. The missing data part of x are represented in the complete likelihood by some functions of their sufficient statistics. Now instead of maximising the observed data (incomplete data) likelihood function $L(y|\theta) = g(y|\theta)$ directly, the EM algorithm proceeds with the updating of $\theta^{(r)}$ as follows:

- E-step: Evaluate $Q(\theta|\theta^{(r)})$ by estimating the functions of sufficient statistics for the missing data by their expected values conditional on the observed data y and the current estimate $\theta^{(r)}$. This may be performed algebraically, if this is feasible, or else numerically. For generalised mixed models, Steele (1996) provides a Laplace's method for analytic approximation and McCulloch (1997) uses a Metropolis algorithm within this step.

- M-step: Select $\theta^{(r+1)}$ as a value of θ that maximises the *pseudo-complete* data log-likelihood $Q(\theta|\theta^{(r)})$ after replacing the functions of the missing data by their conditional expectations. This step is easy to implement in many applications, such as in exponential families, since the computations required are

identical to those of the complete ML. The calculations need not be iterative, although in more complicated practical problems they usually will be.

The basic idea is to find the ML estimator of θ by maximising $\log L_c(x|\theta)$. However due to the incompleteness of the data, we maximise its conditional expectation given the observed (available) data y and the current estimate of θ instead. A distinct feature of the EM algorithm, not shared by many iterative methods, is that each step of the algorithm increases the observed data (incomplete data) likelihood function $L(y|\theta)$ (Dempster *et al.* 1977; Wu, 1983). This implies that, although convergence may be slow, one is assured of finding a local maximum in the long run with a sequence of likelihood values that are bounded from above. However, this does not guarantee convergence to the global maximum. A fuller description and theory of the EM algorithm is found in McLachlan and Krishnan (1996)

3.1.1 Rate of convergence

An attraction of the EM algorithm is its simplicity. It is a gradient method, hence it only requires the first derivatives. Unfortunately, this property implies that the rate of convergence is linear rather than the quadratic rate achieved by the Newton-Raphson procedure. The EM algorithm described above induces the map $\theta^{(r+1)} = M(\theta^{(r)})$. If $\theta^{(r)}$ converges to some point θ^* and $M(\theta)$ is continuous then $\theta^* = M(\theta^*)$. Following Laird *et al.* (1987), a first-order Taylor series expansion of $M(\theta^{(r)})$ about $\theta^{(r-1)}$ gives

$$\theta^{(r+1)} - \theta^{(r)} \approx J^{(r-1)}(\theta^{(r)} - \theta^{(r-1)}) \tag{3.1}$$

where $J^{(r-1)}$ is the Jacobian: a matrix of partial derivatives $\partial M(\theta)/\partial\theta$ evaluated at $\theta^{(r-1)}$. For sufficiently large r, $J^{(r-1)}$ will be close to $J = J^{(\infty)}$. Thus in the neighbourhood of θ^*, the EM algorithm is essentially a linear iteration with rate matrix $J(\theta^*)$ (McLachlan and Krishnan, 1996). The norm of $J(\theta^*)$ will determine the rate of convergence. If J is a square matrix, then convergence of the sequence $\theta^{(r)}$ implies that the absolute values of the eigenvalues of J must not exceed 1.

Equation (3.1) can further be written as:

$$\begin{aligned}\theta^{(r+1)} - \theta^{(r)} &\approx J^{(r-1)})(\theta^{(r)} - \theta^{(r-1)}) \\ &= \left(\prod_{l=1}^{r} J^{(l-1)})\right)(\theta^{(1)} - \theta^{(0)})\end{aligned}$$

Now for large r and $l > 1$

$$\theta^{(r+l+1)} - \theta^{(r+l)} \approx J^{l}(\theta^{(r+1)} - \theta^{(r)}) \tag{3.2}$$

from which the error in iteration $\theta^{(r)}$ can be estimated by:

$$\begin{aligned}\theta^{*} - \theta^{(r)} &= \theta^{(r+1)} - \theta^{(r)} + \theta^{(r+2)} - \theta^{(r+1)} + \theta^{(r+3)} - \theta^{(r+2)} + \cdots \\ &\approx \left(\sum_{l=0}^{\infty} J^{l}\right)(\theta^{(r+1)} - \theta^{(r)}) \\ &= (1 - J)^{-1}(\theta^{(r+1)} - \theta^{(r)}) \qquad (3.3)\end{aligned}$$

since at θ^{*} the eigenvalues of J are all less that 1. Equation (3.3) is a multivariate generalisation of the Aitken acceleration method (Louis, 1982). That is, the speed of the algorithm can be improved by trying

$$\theta^{*(r+1)} = \theta^{(r)} + (I - J)^{-1}(\theta^{(r+1)} - \theta^{(r)})$$

in the next iteration. If inverting $I - J$ is expensive, then a single iteration of the EM algorithm is performed rather than using the Aitken acceleration.

For simple forms of $M(\theta)$, one can directly differentiate it to obtain J. Quite often, in actual practice, the analytic derivatives $\partial M(\theta)/\partial\theta$ are not known explicitly or, even if they are known, they may be difficulty to calculate. Methods discussed in Louis (1982) require further computations for the missing information associated with the conditional distribution of missing variables given the observed data. In most cases J is obtained numerically using the EM code only (Meng and Rubin, 1991). The matrix J serves other purposes such as in the computation of the observed information matrix (see below).

3.1.2 The observed information

One of the criticisms of the EM algorithm is that asymptotic covariance matrix of the parameter vector estimate is not a by-product, as is the case in other methods such as the Newton-Raphson. But this can also be seen as an advantage, since other methods force you to do computations which can be very difficult. Several approaches to estimating the information have been studied and all are based on the standard decomposition of the observed information. Let $S(y|\theta)$ and $S_c(x|\theta)$ be the gradient (score) vector for the observed-data y and complete-data x log likelihood functions respectively, and $I(y|\theta)$ and $I_c(x|\theta)$ be the negatives of the associated second derivatives matrices (information matrices). If $x = (y, x_m)$, where x_m represents the missing part of x, then the conditional density function of x_m (or equivalently the complete data x), given the observed data y, is given by:

$$g(x_m|y;\theta) = g_c(x|\theta)/g(y|\theta) \tag{3.4}$$

This conditional density function provides the missing information matrix defined as:

$$I_{x_m|y} = -\mathrm{E}\{\frac{\partial^2 g(x_m|y,\theta)}{\partial\theta\partial\theta'}|y\}.$$

Now taking logs of (3.4) gives

$$\log L(y|\theta) = \log L_c(x|y;\theta) - \log g(x_m|y,\theta).$$

Taking second derivatives and finding expectations over $f(x_m|y,\theta)$ gives

$$\begin{aligned} -\mathrm{E}\frac{\partial^2 \log L(y|\theta)}{\partial\theta\partial\theta'} &= -\mathrm{E}\frac{\partial^2 \log L_c(x|y,\theta)}{\partial\theta\partial\theta'} - -\mathrm{E}\frac{\partial^2 \log g(x_m|y,\theta)}{\partial\theta\partial\theta'} \\ &= \mathrm{E}I_c(x|y,\theta) - I_{x_m|y}(y|\theta) \end{aligned} \tag{3.5}$$

which can conveniently be written as

$$I_y = I_x - I_{x_m|y}$$

where I_y is the observed information matrix, I_x is the conditional expectation of the complete data information matrix and $I_{x_m|y}$ is the expected information for the conditional distribution of x_m given y. This decomposition of the observed information is an application of the missing information principle: observed information equals complete information less missing information (Louis 1982; Meng and Rubin, 1991). Efron and Hinkley (1978) showed that in most cases the observed information I_y is a more appropriate measure of information than its expectation $\mathrm{E}(I_y)$ over y. The matrix I_x is easily obtained as a by-product of the last E-Step since it equals the negative of the second derivative matrix of the *pseudo-complete* data log-likelihood $Q(\theta|\theta^{(r)})$.

The calculation of $I_{x_m|y}$ can be done directly from its definition. An alternative expression for $I_{x_m|y}$ was established by Louis (1982) who showed that

$$I_{x_m|y} = \mathrm{Cov}\{S_c(x|\theta)|y)\}$$

This result is established by using the definition of $S(y|\theta)$ and straightforward differentiation. Further Louis (1982) and McLachlan and Krishnan (1996) show that

$$S(y|\theta) = \mathrm{E}(S_c(x|\theta)|y))$$

Therefore $\mathrm{Cov}(S_c(x|\theta)|y)$ can be expressed as

$$\mathrm{Cov}(S_c(x|\theta)|y) = \mathrm{E}(S_c(x|\theta)|y)(S_c'(x|\theta)|y) - S(y|\theta)S'(y|\theta)$$

Thus, the observed data (incomplete data) information matrix can be expressed as

$$\begin{aligned} I_y &= I_x - I_{x_m|y} \\ &= I_x - \mathrm{Cov}\{S_c(x|\theta)|y)\} \\ &= I_x - \mathrm{E}(S_c(x|\theta))(S_c'(x|\theta)|y) - S(y|\theta)S'(y|\theta) \end{aligned} \tag{3.6}$$

which can be computed using conditional expectations of only $S_c(x|\theta)$ and $I_c(x|\theta)$. These are the gradient and curvature of the complete data log-likelihood function.

They need only be evaluated at the last iteration of the EM procedure where $\theta = \theta^*$. For the last step

$$I_y = I_x - \mathrm{E}(S_c(x|\theta))(S_c'(x|\theta)|y)$$

since, at θ^*, $S(y|\theta) = 0$.

Alternatively, this problem can be viewed as that of increasing the complete-data covariance-variance matrix to compensate for the missing information. In particular for the exponential family distribution, Dempster *et al* (1977) showed that

$$\begin{aligned} J &= [\mathrm{Var}(s|\theta^*)]^{-1}\mathrm{Var}(s|\theta^*, y) \\ &= I_x^{-1} I_{x_m|y}. \end{aligned}$$

Using the results shown earlier,

$$\begin{aligned} J &= (I_x^{-1}(I_x - I_y)) \\ &= (I - I_x^{-1} I_y)) \end{aligned}$$

It is easily seen that

$$V_y = \frac{V_x}{I - J}$$

where $V_y = I_y^{-1}$ equals $V_x = I_x^{-1}$ inflated by a factor $(I - J)^{-1}$ (Meng and Rubin, 1991). This method does not require extra computation outside the EM algorithm steps

3.2 Incorporating frailty in the Model

In the context of our problem, the family and the community random effects are unobservable. We need to make assumptions on the distribution of these random effects for further progress on the estimation of parameters. The family random effect, b_{ij}, and community random effect, b_i, represent relative risks and, for convenience, each is given a mean of 1. These random effects are modelled as mutually independent and gamma distributed with shape parameters ξ and ψ; inverse scales

ξ and ψ respectively. Thus the densities of b_{ij} and b_i are respectively

$$f(b_{ij}) = \frac{\psi^{\psi}}{\Gamma(\psi)} b_{ij}^{\psi-1} e^{-\psi b_{ij}} \tag{3.7}$$

and

$$f(b_i) = \frac{\xi^{\xi}}{\Gamma(\xi)} b_i^{\xi-1} e^{-\xi b_i}. \tag{3.8}$$

Hence, the relative risk for a family random frailty has mean 1 and variance $1/\psi$, and the relative risk for a community random frailty has mean 1 and variance $1/\xi$.

One of the contentious issues associated with frailty models concerns the form of the frailty distributions. Heckman and Singer (1984), in a study of the factors affecting the duration of unemployment, compared the estimates of covariate parameters for different forms of distribution for frailty. Their results indicated many changes in both sign and magnitude of parameter estimates and argued for the use of nonparametric frailty. Hougaard (1986a, 1986b) pointed out that if a finite mean frailty distribution is assumed, then the marginal hazard is no longer proportional in the covariates, and showed that the assumption of a positive stable distribution for frailty avoids the problem.

Some studies (e.g. Schumacher *et al.*, 1987) have found that, if frailty with finite mean is ignored, the estimates are biased towards zero. That is, allowing for frailty would only mean an increase in magnitude of estimates as opposed to change of signs (Pickles and Crouchley, 1995). This increase would depend on how close the assumed frailty mirrors the true frailty distribution. As well, simulation studies of Pickles and Crouchley (1995) and Sastry (1997) suggest that the choice of a particular parametric frailty distribution is not critical in estimating the regression parameters. Indeed, other studies have found that the hazards model estimates were not markedly different between parametric and nonparametric forms of frailty (Guo and Rodriguez, 1992; Congdon, 1995).

Thus, inference is expected to be fairly similar whether the frailty distribution is assumed to be a gamma or other parametric distribution. We select the gamma over other distributions because it has a flexible shape. It also incorporates easily in

the model computation. As well, the gamma distribution had been used previously in studies on unobserved heterogeneity (Clayton, 1991; Sastry, 1997).

3.2.1 Parameter estimation

For a community i, with response vector t_i and a set of all family random effects $\{b_{ij}\}$, the complete data likelihood (joint distribution) is given by:

$$\begin{aligned} L_i(t_i, b_i, \{b_{ij}\}|\theta) &= \frac{\xi^\xi}{\Gamma(\xi)} b_i^{\xi-1} e^{-\xi b_i} \prod_{j=1}^{J_i} \frac{\psi^\psi}{\Gamma(\psi)} b_{ij}^{\psi-1} e^{-\psi b_{ij}} \times \\ &\quad \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left(b_i b_{ij} \lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right)^{w_{ijk}} e^{-b_i b_{ij} \Lambda_{ijk}(t_{ijk})} \end{aligned} \tag{3.9}$$

where $\theta = (\beta, \xi, \psi)$. If the b_i and b_{ij} were known but variable effects, inference about $\{\beta, \xi, \psi\}$ could be based on the log-likelihood construction of (3.9) namely

$$\begin{aligned} \log L_i(t_i, b_i, \{b_{ij}\}|\theta) &= \sum_j^{J_i} \sum_k^{K_{ij}} w_{ijk} \log(b_i b_{ij} \lambda_0(t_{ijk}) e^{\beta' X_{ijk}}) - b_i b_{ij} \Lambda_{ijk}(t_{ijk}) + \\ &\quad \sum_j^{J_i} \log f(b_{ij}) + \log f(b_i) \end{aligned} \tag{3.10}$$

and this could be maximised over all communities using available standard software. Of course the b_i and b_{ij} are not observed directly. One way to proceed from here would be to integrate out the random effects from (3.9) to obtain the likelihood for the incomplete data which is given by:

$$\begin{aligned} L_i(t_i|\theta) &= \int_0^\infty \int_0^\infty \cdots \int_0^\infty L_i(t_i, b_i, \{b_{ij}\}|\theta) db_{i1} db_{i2} \ldots db_{iJ_i} db_i \\ &= \left[\frac{\xi^\xi}{\Gamma(\xi)} \left(\frac{\psi^\psi}{\Gamma(\psi)}\right)^{J_i} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left(\lambda_0(t_{ijk}) e^{\beta' X_{ijk}}\right)^{w_{ijk}}\right] \times \\ &\quad \int_0^\infty \int_0^\infty \cdots \int_0^\infty b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \times \\ &\quad \prod_{j=1}^{J_i} b_{ij}^{\psi-1+\sum_k w_{ijk}} e^{-b_{ij}(\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk}))} db_{i1} db_{i2} \ldots db_{iJ_i} db_i. \end{aligned}$$

Let

then

$$
\begin{aligned}
L_i(t_i|\theta) &= C_i \int_0^\infty \int_0^\infty \cdots \int_0^\infty b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \times \\
&\quad \prod_{j=1}^{J_i} b_{ij}^{\psi-1+\sum_k w_{ijk}} e^{-b_{ij}(\psi+b_i \sum_k \Lambda_{ijk}(t_{ijk}))} db_{i1} db_{i2} \ldots db_{iJ_i} db_i \\
&= C_i \int_0^\infty b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \times \\
&\quad \left[\prod_{j=1}^{J_i} \int_0^\infty b_{ij}^{\psi-1+\sum_k w_{ijk}} e^{-b_{ij}(\psi+b_i \sum_k \Lambda_{ijk}(t_{ijk}))} db_{ij} \right] db_i \\
&= C_i \int_0^\infty b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \times \\
&\quad \left[\prod_{j=1}^{J_i} \frac{\Gamma(\psi + \sum_k w_{ijk})}{[\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk})]^{\psi+\sum_k w_{ijk}}} \times \right. \\
&\quad \left. \int_0^\infty \frac{[\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk})]^{\psi+\sum_k w_{ijk}}}{\Gamma(\psi + \sum_k w_{ijk})} b_{ij}^{\psi-1+\sum_k w_{ijk}} e^{-b_{ij}(\psi+b_i \sum_k \Lambda_{ijk}(t_{ijk}))} db_{ij} \right] db_i \\
&= C_i [\prod_j^{J_i} \Gamma(\psi + \sum_k w_{ijk})] \int_0^\infty b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \times \\
&\quad \left[\prod_{j=1}^{J_i} \frac{1}{[\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk})]^{\psi+\sum_k w_{ijk}}} \right] db_i \qquad (3.11)
\end{aligned}
$$

where the integral for b_{ij} has disappeared because the integrand is the density of a gamma density with parameters $\Gamma(\psi + \sum_k w_{ijk})$ and $\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk})$. The log-likelihood of (3.11) can be maximised using the Newton-Raphson algorithm. Under this method, the first and second derivatives of (3.11) must be evaluated at each iteration. Even though the information matrix is a by-product, the actual computations are so enormous that they outweigh the method's advantage of rapid convergence.

The fact that the multiplicative frailty model in equations (1.3-1.6) would coincide with a standard hazards model if the random effects were observed makes the fitting of this incomplete data problem ideally suited for the EM algorithm because the available efficient software can be used. For completeness, details of the EM algorithm for our model are given below in 3.2.2 and 3.2.3 (see also Clayton, 1991; Sastry, 1997).

3.2.2 The E-step

From (3.10) it is seen that the complete data log-likelihood is a function of the missing data b_i and b_{ij}. The conditional expectations of b_i, $\log b_i$, b_{ij}, $\log b_{ij}$ and the product $b_i b_{ij}$, given the observed data y and current parameter estimates are required. This is because $L_i(t_i, b_i, \{b_{ij}\}|\theta)$ is a linear function of these quantities. Denote the conditional expectations as $\overline{b_i}$, $\overline{\log b_i}$, $\overline{b_{ij}}$, $\overline{\log b_{ij}}$ and $\overline{b_i b_{ij}}$. Now, consider calculating the conditional expectation of particular community random frailty value b_i. The conditional distribution of b_i is given by:

$$f(b_i|t_i, \theta) = \frac{f(t_i, b_i|\theta)}{L_i(t_i|\theta)}$$

where

$$\begin{aligned} f(t_i, b_i|\theta) &= \int_0^\infty \int_0^\infty \cdots \int_0^\infty L_i(t_i, b_i, \{b_{ij}\}|\theta) db_{i1} db_{i2} \ldots db_{iJ_i} \\ &= C_i[\prod_j^{J_i} \Gamma(\psi + \sum_k w_{ijk})] b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \times \\ &\quad \left[\prod_{j=1}^{J_i} \frac{1}{[\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk})]^{\psi+\sum_k w_{ijk}}}\right]. \end{aligned}$$

Thus the conditional density $f(b_i|t_i, \theta)$ is given by the equation

$$\frac{b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \left[\prod_{j=1}^{J_i} [\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk})]^{-\psi-\sum_k w_{ijk}}\right]}{\int_0^\infty b_i^{\xi-1+\sum_j \sum_k w_{ijk}} e^{-\xi b_i} \left[\prod_{j=1}^{J_i} [\psi + b_i \sum_k \Lambda_{ijk}(t_{ijk})]^{-\psi-\sum_k w_{ijk}}\right] db_i}. \tag{3.12}$$

The conditional expectation of b_i, given the data $\{t_i, w_i\}$ and the current parameter estimates is

$$E(b_i|t_i, \theta) = \int_0^\infty b_i f(b_i|t_i, \theta)$$

which can be approximated numerically. The density (3.12) is also used to calculate the conditional expectation of $\log b_i$. Similar arguments apply for the conditional means of b_{ij} and $\log b_{ij}$. For the product of b_i and a particular b_{ih}, we need to integrate out the remaining family random effects b_{ij}, $j \neq h$ from $L_i(t_i, b_i, \{b_{ij}\}|\theta)$ to obtain the joint marginal distribution of b_i, b_{ih}, t_i and w_i.

3.2.3 The M-step

This step proceeds by substituting the conditional expectations from the E-step into the complete log-likelihood function (3.10) to obtain $L_i(t_i, \overline{\{b_i\}}, (\overline{\{b_{ij}\}}|\theta)$ which is then maximised with respect to θ. It is clear from (3.10) that the maximisation process separates itself into three distinct parts. Two of these involve ξ and ψ and are one-dimensional. These are easy to solve. For instance, for the Newton-Raphson algorithm, the maximum likelihood estimates for ψ would require the following derivatives:

$$\frac{\partial Q(\beta,\xi,\psi)}{\partial \psi} = \sum_i^I \sum_j^{J_i} \left(\overline{\log b_{ij}} - \overline{b_{ij}}\right) + \sum_i^I \sum_j^{J_i} \left(1 + \log(\psi) - \frac{d \log \Gamma(\psi)}{d\psi}\right)$$

$$\frac{\partial^2 Q(\beta,\xi,\psi)}{\partial \psi^2} = \sum_i^I \sum_j^{J_i} \left(\frac{1}{\psi} - \frac{d^2 \log \Gamma(\psi)}{d\psi^2}\right)$$

with similar derivatives for ξ.

The M-step for β is essentially an ordinary proportional hazards regression in which the baseline hazard is multiplied by $\overline{b_i b_{ij}}$. Since the baseline hazard is piecewise exponential, this entails augmenting the explanatory variables in the linear part by $\log(\overline{b_i b_{ij}})$: that is treating $\log(\overline{b_i b_{ij}})$ as an offset. This is very helpful in practice because it means we can use standard Poisson regression routines which are readily available.

We also see that in estimating parameters ξ and ψ, a lot of information is missing. On the other hand, there is ample information for parameters in β. Thus, convergence of the EM for ξ and ψ would be far slower than for the parameter vector β (Dempster *et al*, 1977). This discrepancy in convergence rates may be corrected by running several steps of an inner iteration for ξ and ψ for each main EM step. An inner iteration procedure was not run nor were the acceleration routines discussed in (3.1.1). We obtained reasonable convergence without them, thus avoiding the additional computational complexity.

3.3 Application to the data

The overall data has already been described in chapter one. The analyses in this chapter and chapter 5 are restricted to the mortality experience of 4838 live singleton births in the preceding 5 years, of which 745 had died at the time of the survey. The 4838 births occurred in 2911 families of 225 communities. Births that belong to families with two or more births account for about 73% of all births. About 3% of all families contribute at least two deaths and these deaths account for 22% of the total number of deaths. The number of births per community ranges from 3 to 50. The average number of births and deaths per community is 21.3 and 3.3 respectively. It is clear from the data that births and deaths tend to cluster in certain families and communities

Preliminary analysis showed that the *age of the child*, the *sex of the child, birth order*, *birth spacing*, *whether the preceding child died*, *breastfeeding duration*, *maternal age*, *father's education* and *birth in a hospital* are the fixed effect predictors that should be included in the model. These predictor variables are summarised in Table 3.1. The EM algorithm was implemented in FORTRAN 77 using the NAG subroutines (particularly G02GCF,D01AMF,D01BBF and D01FBF) on a single SUN SPARC station 2. The EM code was also used to estimate the rate matrix and the observed covariance matrix.

We started by estimating the parameters in a standard hazards model which does not allow for clustering of observations at either level. A χ^2 goodness of fit test showed that the standard model gave a good fit to the data. The logarithms of the piecewise constant baseline hazards were estimated as fixed effects. Exponentiating them gives the baseline hazards which are shown in Figure 3.1[1].

The baseline rate shows a general decline with increasing age of the child. In months 1 to 5, the child faces about a fifth of the risk faced in the first month of life. There is a slight increase in mortality for months 5 to 12; after which it drops steadily

[1]Other definitions of age intervals were used but produced virtually identical results.

Table 3.1: Descriptive statistics of covariates used in the analysis

Variable	Percent		Variable	Percent
Child's sex			*Father's education*	
Female	48.7		9+ years	16.3
Male	51.3		*Birth in hospital*	62.3
Previous child dead	13.2		*Succeeding conception interval*[a]	
Preceding birth interval			$<$ 12 mths	06.5
First Births	19.5		*Breastfeeding duration*[b]	
$<$ 18 mths	07.8		$\leq$ 5 mths	03.0
18 - 24 mths	12.3			
	median	mean	std.dev.	
Birth order	3	4	2.6	
Mother's age	26.3	27.5	7.46	

[a]Based on 3428 cases
[b]Based on 4132 cases

for the rest of the study period. The increase in mortality for months 5 to 12 may be due to mothers' loss of memory or the tendency to rounding off ages of death to nearest 6 months. But a larger part of the increase may result from exogenic factors such as infectious diseases, sanitation and nutrition. These factors are causes of later infant mortality, and are supposed to be susceptible to prevention and treatment. But, in Malawi, due to the poor living conditions, the exogenic factors are beyond the control of most households. However, in most developed countries, these factors are better controlled resulting in a continual decrease in mortality from birth.

The results of the other predictor variables in the standard hazards model are shown in Table 3.2. Most of the covariates are statistically significant, and all parameter estimates of the effects are in the expected direction. Being female shows a slightly reduced risk, and the *death of the previous child* indicates a substantial and

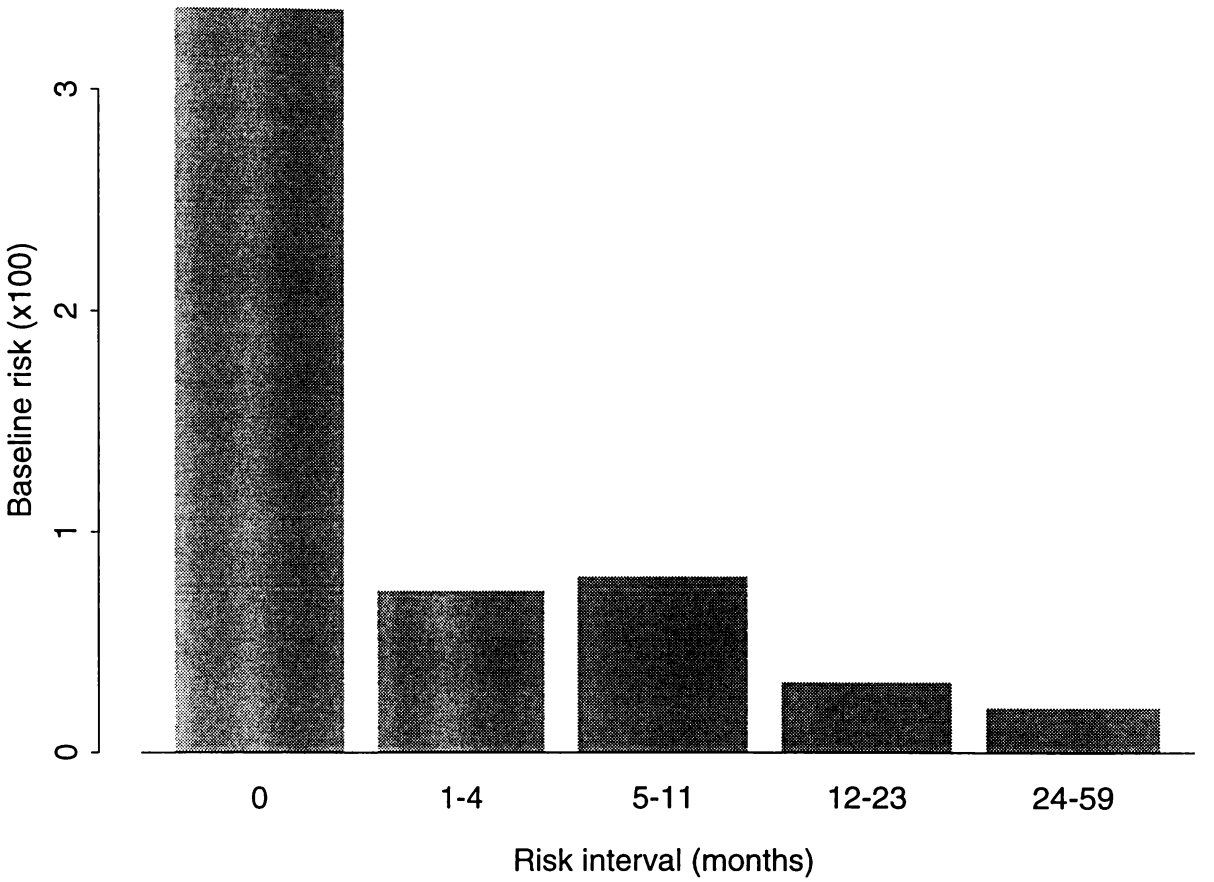


Figure 3.1: *Estimates of the baseline hazards*

significant additional risk. First-birth and short *preceding birth interval* are highly associated with increased risks of childhood mortality. The risk also decreases with increasing *level of education of the father*. A *hospital birth* reduces the risk. Both short *succeeding birth interval* and short *breastfeeding duration* greatly increases the risk of death. The estimated effects of the *mother's age* are shown in Figure 3.2. Very young and old *maternal age* are related to increased risk. However, the estimated risks of *maternal age* are strongly correlated with other factors such as *birth order* and birth intervals. For instance, most old mothers are giving birth to children in the high *birth order*, and these have a decreased risk which partially overcomes

the increased maternal risk. Therefore, the effect of *maternal age* should not be considered in isolation of these other factors.

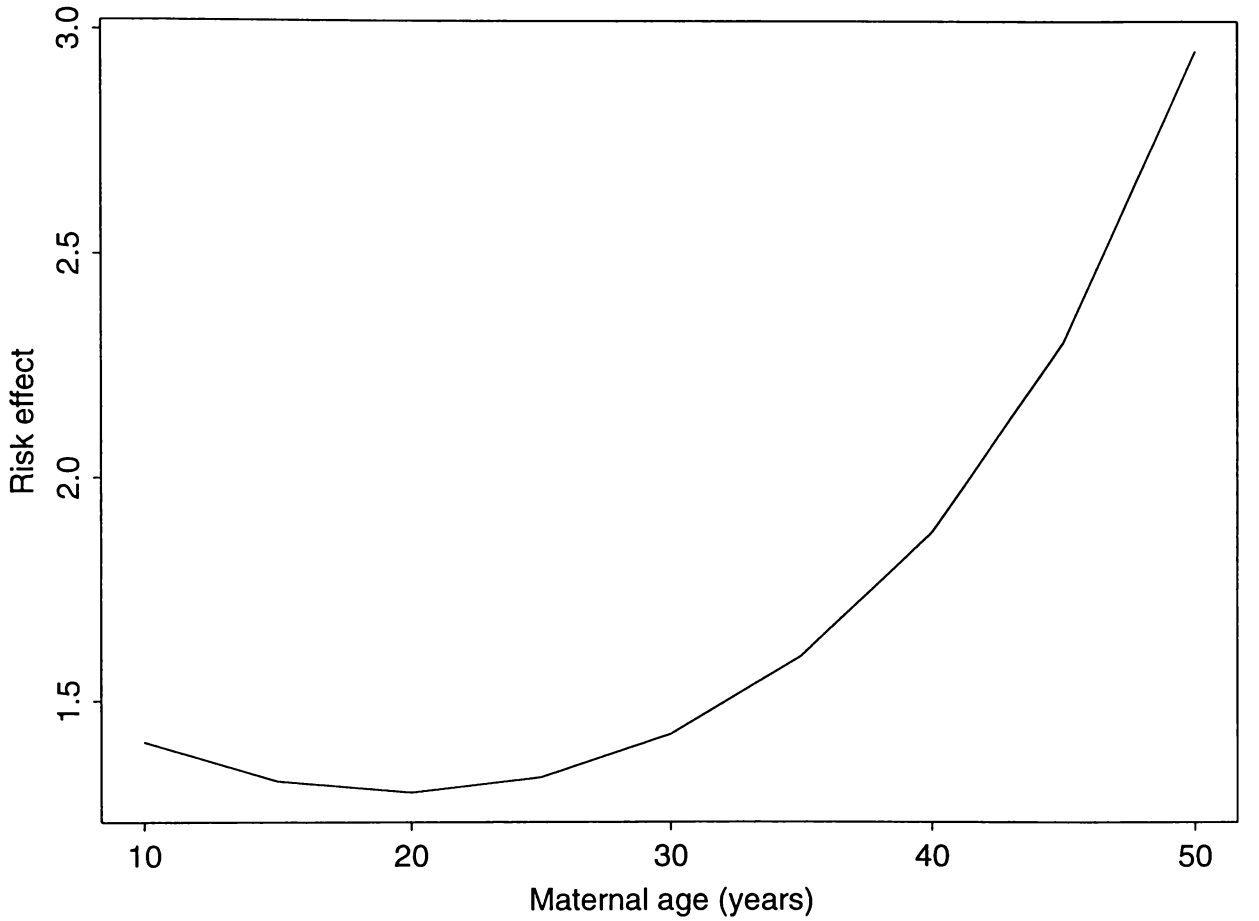


Figure 3.2: *Estimate of risk effect of maternal age*

Then we proceeded to fit three different multilevel hazards models. The results are presented in Table 3.3. Models I and II control for a single random effect at the family and community levels respectively, and Model III allows for family random effects nested within community clustering random effects. Both Models I and II significantly improve the fit to data over the standard hazards model (Table 3.2) and Model III offer significant improvements over both Models I and II.

Based on Model I, the family frailty model, the variance of the family random

Table 3.2: Results for the standard hazards model

Parameter	Estimate	Standard Error
Female child		
(0-no, 1-yes)	-0.086	0.073
Previous child		
(0-alive, 1-dead)	0.197	0.107
Birth order	-0.070	0.027
Preceding birth interval		
First Births	0.557	0.124
$<$ 18 mths	0.476	0.125
18 - 24 mths	0.374	0.110
Mother's age		
age	0.012	0.011
$(age-26)^2$	0.001	0.001
Father's education		
(9+ years)		
(0-no, 1-yes)	-0.449	0.146
Birth in hospital		
(0-no,1-yes)	-0.280	0.074
Succeeding conception		
$<$ *12 mths*		
(0-no,1-yes)	1.416	0.198
Breastfeeding duration		
$<$ *5 mths*		
(0-no,1-yes)	0.853	0.261

effect is estimated at 0.772 and is greater than 0 at very high significance level. This means that the unobserved family effects have a large influence on mortality risk of children. This value implies a substantial association between the survival times of a pair of siblings after controlling for observed fixed effects covariates as, based on Oakes (1982), it corresponds to a rank correlation of 0.278. An even more gloomy picture emerges when considering the conditional hazards ratios. Following Clayton (1978), the conditional hazard for a child at any time given that a sibling died at age t_1 months is 0.772 higher than if the sibling had survived to t_1 months; and the excess risks doubles if it is known that two siblings had died.

The estimates of the constant baseline hazards are very similar to those found in the standard hazards model. The effect of *father's education*, *hospital birth* and *succeeding conception interval* are slightly magnified while that of the *breastfeeding duration* is reduced. The most notable change occurs in the estimate of the effect of a *previous child death*. Its risk changes from a significant 21 percent in excess in the standard model to a 6 percent reduction in the family frailty model. In the standard hazards model, the variable indicating whether the preceding child died acts as proxy for unobserved family effects. Its positive effect in the ordinary model confirms this. Therefore, one would expect its positive role to diminish once the family-specific random effect is included in the model. In the family random effect model, its coefficient changes to an insignificant negative effect. Even though the negative effect is not significant, it may suggest that a child death actually lowers the risk for an index child either through reduced competition for family resources, or because death is such a traumatic event that it may induce changes in parental behaviour. A less likely scenario is that the death of a preceding sibling removes the source of infection, if the principal reason for the adverse effects of short *birth spacing* is the increased spread of infection among sibling of similar age.

Model II indicates that the unobserved community effects are important determinants of child mortality in Malawi. The variance of the community random effect

Table 3.3: Child survival data from Malawi: EM parameter estimates

	Model I		Model II		Model III	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
Female child						
(0-no, 1-yes)	-0.084	0.078	-0.093	0.073	-0.091	0.078
Previous child						
(0-alive, 1-dead)	-0.059	0.139	0.178	0.108	0.084	0.138
Birth order	-0.078	0.030	-0.069	0.026	-0.072	0.029
Preceding birth interval						
First Births	0.556	0.137	0.568	0.126	0.566	0.137
< 18 mths	0.437	0.135	0.474	0.126	0.460	0.125
18 - 24 mths	0.379	0.118	0.389	0.111	0.390	0.110
> 24 mths[a]						
Mother's age						
age	0.011	0.011	0.011	0.010	0.011	0.011
$(age-26)^2$	0.001	0.007	0.001	0.001	0.001	0.001
Father's education						
(9+ years)						
(0-no, 1-yes)	-0.481	0.158	-0.449	0.148	-0.462	0.158
Birth in hospital						
(0-no,1-yes)	-0.314	0.083	-0.295	0.077	-0.307	0.083
Succeeding conception						
(<12 mths:0-no,1-yes)	1.477	0.207	1.424	0.199	1.450	0.207
Breastfeeding duration						
(<5mths: 0-no,1-yes)	0.806	0.274	0.819	0.263	0.808	0.274
Random effect variances						
Family	0.772	0.286			0.247	0.106
Community			0.075	0.0.030	0.066	0.032

[a]Omitted category

is estimated at 0.076 and clearly significantly greater than 0. This value implies a very modest association between lifetimes of children in the same community as it corresponds to a rank correlation of 0.037. The changes in the estimates of the parameter effects are similar to Model I; only that in Model II they are smaller in magnitude. The effect of the *preceding child death* is to raise the relative risks by 18 percent- very close to the excess risk found in the standard model. The apparent stability of the effect of the *preceding child death* in the community frailty model provides another indication that it acts as proxy to the family frailty.

The results for the main model (Model III) show that the family random effect variance is now 0.247, still significantly greater than 0, but only 32 percent of its value from Model I. The community random effect variance has changed little from Model II. It is 0.066, again significantly greater than 0, unlike in the infant mortality (Chapter two). These values show that family random effects are considerably more variable than community random effects. The distribution of the family and community random effects are shown in Figure 3.3.

The fixed effect estimates of Model III are similar to those found from Models I and II. The most notable change from the standard model is that the effect of *preceding child death* becomes insignificant for the models including family random effect (I and III). The other fixed effects are roughly similar in magnitude to those of the standard models, although the standard errors are generally magnified in frailty models, indicating that the coefficients are now estimated with less precision. This is consistent with the theory that positively correlated responses contain less information than are independent responses. However, several covariates slightly increase their significance.

3.4 Conclusion

The analysis shows that even after controlling for the observed covariates, the survival experiences of children in Malawi vary considerably across families and to a

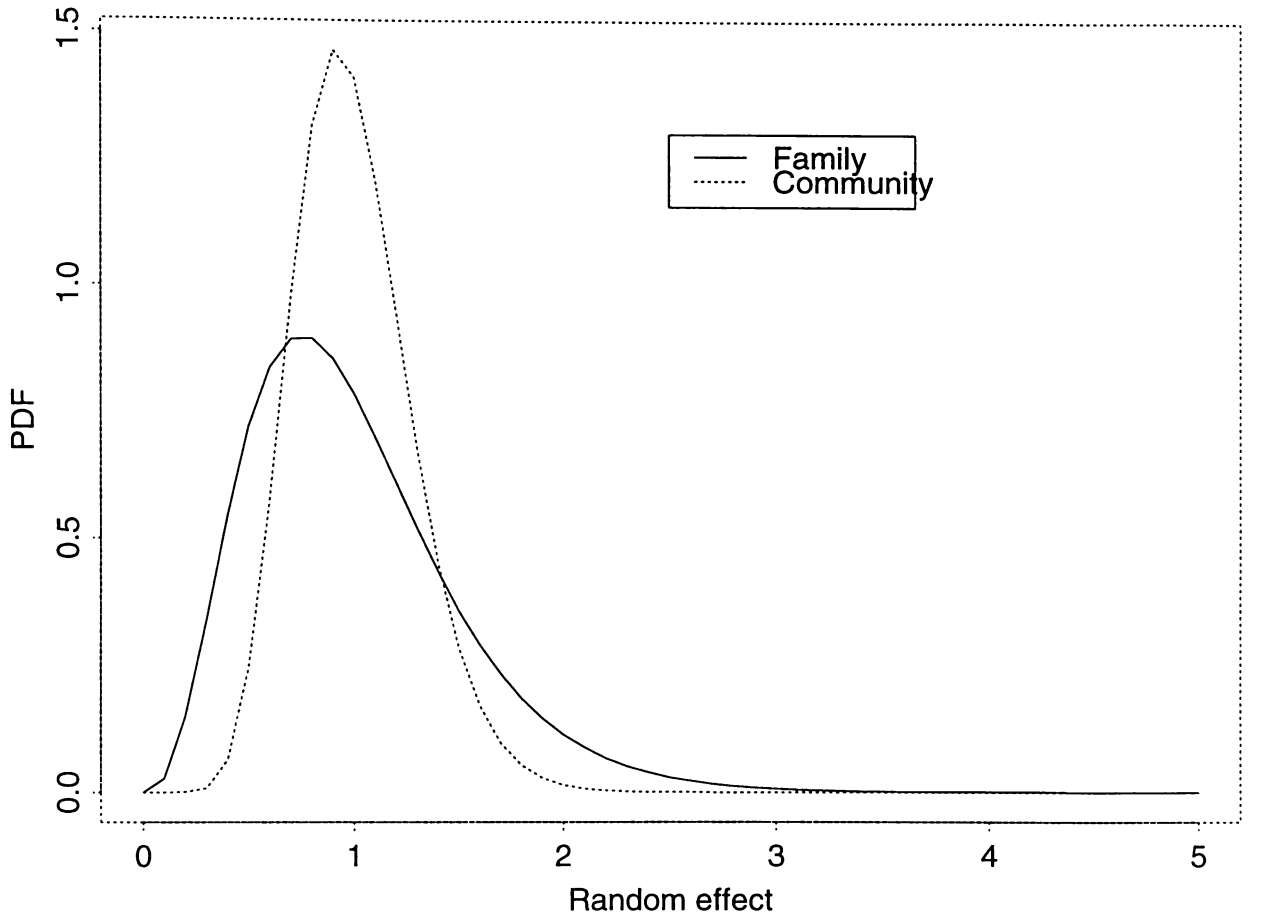


Figure 3.3: *Estimate of family and community random effect distributions*

lesser extent across communities. It also shows, that neglecting frailty biases the estimated regression effects of observed covariates slightly downwards. However, this does not markedly affect the subsequent substantive findings. This is reassuring, considering that most studies of childhood mortality have not allowed for heterogeneity of family or community random effects. Similar conclusions may not hold in other contexts where survival models are used- such as employment duration and migration.

The strength of the family random effect is grossly overstated when community frailty is ignored. This bias comes in because the family random effect also covers

the effects of some factors common to both children in the same family and to all children in the same community. On the other hand, there is very little bias in the estimate of the community random effect variance in the absence of family random effect. The near absence of upward bias may indicate that the community random effect includes family effects for those families living in that community.

The result, that there are only very modest differences in the estimates of the observed fixed effects between the standard and frailty models is worth commenting on. There are two broad factors that can explain this result. Firstly, it may be that the distributions assumed for the random effects are inappropriate. In this regard, the results can be sensitive only if the proportion surviving the period is low and the variance of the random effect is large, of magnitude 2 or above (Sastry, 1997). In this study, we are dealing with one of the highest under-five mortality rates in the world. However, our results and others (Madise, 1993) have found random effect variances much smaller than 2, the level at which estimates may become unstable due to the assumed random effect distribution. Secondly, the survival rate over the period under consideration can also affect the estimates of the parameters. This can be understood by noting that individuals from high risk clusters will fail earlier than those from low risk clusters. As time goes by, the observations from low risk clusters will dominate the sample and the apparent risk, in the absence of the clustering effects, will decline over time (Guo and Rodriguez, 1992). For the data analysed here, the selection may be rapid due to both highly significant values of random effects variance, particularly at the family level. But the selection process itself covers a period of time, and the amount of selection will depend on the length of the study period. The survival study period from 0 to 59 months, that we are using in this study, is relatively small, so large selection biases are unlikely to occur.

Chapter 4

Markov chain simulation in Bayesian statistics

4.1 Introduction

In the previous chapter, the EM algorithm was used to obtain the modes of the model parameter distribution. We encountered enormous problems finding the exact analytical solution and finally resorted to an approximate solution. Given the complexity of the problem, the approximations may not be adequate. In chapter five, the posterior distribution of the model parameters will be found using a Bayesian approach. The posterior, which summarises our knowledge of θ given the observations y, is used to perform inferences on θ. Instead of determining the posterior analytically or numerically, we use Markov Chain Monte Carlo (MCMC) methods which allow us to obtain a sample from the posterior. This chapter provides the theoretical background for these sample based methods for exploring and summarising posterior distributions. Some implementation issues associated with MCMC methods are also discussed.

In making inferences about the parameter vector θ, after observing data y, all relevant sample information is contained in the likelihood function $f(y|\theta)$. However, in the Bayesian approach, in addition to specifying a model for the observable y given

θ, we suppose that θ is a random quantity with a prior distribution $g(\theta)$. By Bayes theorem the posterior distribution $g(\theta|y)$ of θ is given by

$$g(\theta|y) = \frac{f(y|\theta)g(\theta)}{\int f(y|\theta)g(\theta)d\theta}. \tag{4.1}$$

Often the posterior is presented as

$$g(\theta|y) \propto f(y|\theta)g(\theta). \tag{4.2}$$

That is, the posterior distribution of θ is proportional to the product of the prior distribution and the likelihood function. The incomplete form (4.2) of the posterior density provides the shape of $g(\theta|y)$, and consequently the modes and relative probabilities at any two locations. The determination of the exact posterior density involves the evaluation of a complex and often high-dimensional integral

$$\int f(y|\theta)g(\theta)d\theta.$$

A closed form of the posterior density can be found only in a few cases, such as where the density for y is a member of the exponential family and the prior density is from the conjugate family. Otherwise the posterior density can sometimes be approximated analytically by a normal density or by more complicated asymptotic techniques, such as Laplace methods (Tierney and Kadane, 1986; Carlin and Louis, 1996). Numerical integration of the required integral is often difficult and can be very inaccurate where there is a very high dimensional parameter space. Various versions of standard Monte Carlo importance sampling are another approach.

Recently, most applied Bayesians have turned to Monte Carlo sample based methods for exploring the posterior density. These methods provide more complete information and are comparatively easy to compute for very high-dimensional models. Bolstad (1997) points out that the biggest advantage of these methods is that an applied statistician can use more realistic models without the need of complicated analytical or numerical solutions. Further Smith and Robert (1993) show that exploratory data analysis (EDA) techniques are useful for exploring the posterior

distribution from a sample of points. Marginal distributions of some of the components of θ can be found by a kernel density estimate from the appropriate sample. In some cases, better estimates of marginal distributions can be found using known forms of posterior marginal distributions. Bolstad (1997) and Brooks (1998) provide a comprehensive review of some of the common MCMC methods with associated implementation issues.

The Monte Carlo methods can be split into two classes. In one class, we have methods that generate a sample directly from a posterior density having a known shape. These are noniterative methods that include *Sampling Importance Resampling* and *Acceptance and Rejection* sampling and are the subject of section 4.2. The other class includes indirect Monte Carlo methods. Here the sample is not drawn from the posterior distribution directly. Rather these methods produce a Markov chain, which has, as its *long run* density, the posterior distribution. The *Substitution* sampler, the *Metropolis-Hastings* algorithm, and the *Gibbs* sampler are methods that produce these chains. The Markov chain runs a long time until it approaches the limiting distribution. Any value taken after this *burn-in* time approximates a random draw from the posterior density. Markov Chain Monte Carlo methods, together with techniques for assessing thier convergence, are presented in sections 4.3 and 4.4

4.2 Noniterative Monte Carlo methods

Unlike Monte Carlo Markov Chains methods, noniterative procedures generate a non-Markov sample whose successive observations are statistically independent unless correlation is introduced as a variance reduction device. There is no notion of the algorithm converging, but the sample size need to be sufficiently large. There are several noniterative Monte Carlo methods and here we shall discuss only the two: *Sampling Importance Resampling* (SIR) and *Acceptance Rejection* (AR).

4.2.1 Sampling Importance Resampling (SIR)

This method was first presented by Rubin(1987) and further discussed by Smith and Gelfand (1992). Suppose, we want to obtain a sample of n' values from the posterior distribution $g(\theta|y)$ from which it is difficult to simulate directly. However, a sample $\{\theta_1, \theta_2, \ldots, \theta_{N'}\}$ is available from some approximate density $g_0(\theta)$. The following quantities can be computed for each value in the sample:

$$q_j = \frac{f(y|\theta_j)g(\theta_j)}{g_0(\theta_j)} \quad \& \quad w_j = \frac{q_j}{\sum_{j=1}^{N'} q_j}$$

The quantity w_j is called the *importance weight* for value j in the sample. A second sample $\{\theta_1^*, \theta_2^*, \ldots, \theta_{n'}^*\}$ is then taken with replacement from $\{\theta_1, \theta_2, \ldots, \theta_{N'}\}$ using weights w_j. The second sample can be shown to be approximately distributed as the posterior density $g(\theta|y)$ (Smith and Gelfand, 1992). This approximation improves as N' becomes large. This is a weighted bootstrap since some points are resampled more often than others due to the unequal weighting. However, unlike the usual bootstrap, the parameters are sampled rather than the data. Thus, it is sometimes called the Bayesian bootstrap.

4.2.2 Acceptance and Rejection sampling (AR)

This is a classical simulation technique. The objective, as before, is to generate a sample of values from $g(\theta|y)$. Let $g_0(\theta)$ be an approximate easily sampled starting distribution. Further, suppose that there exists a known positive constant M such that $f(y|\theta)g(\theta) \leq Mg_0(\theta)$. The density $g_0(\theta)$ is called the *envelope density*. The rejection method proceeds as outlined below:

1. Draw θ_j from $g_0(\theta)$

2. Draw U from a uniform (0,1) distribution

3. If $U \leq f(y|\theta_j)g(\theta_j)/Mg_0(\theta_j)$ then accept θ_j, otherwise reject θ_j

4. Return to step (1) and repeat until the desired sample $\{\theta_1, \theta_2, \ldots, \theta_{n'}\}$ is obtained. Members of this sample will then be random variables from $g(\theta|y)$.

If $f(y|\theta)g(\theta) = Mg_0(\theta)$, then $g_0(\theta)$ is proportional to $g(\theta|y)$ and an *i.i.d* sample is obtained from $g(\theta|y)$. The problem with this algorithm is in finding an M that works well. Let p be the probability of getting a candidate θ_j accepted. Then

$$\begin{aligned} p &= Pr[U \leq f(y|\theta_j)g(\theta_j)/Mg_0(\theta_j)] \\ &= \int Pr[U \leq f(y|\theta_j)g(\theta_j)/Mg_0(\theta_j)]g_0(\theta_j)d\theta_j \\ &= \int [f(y|\theta_j)g(\theta_j)/Mg_0(\theta_j)]g_0(\theta_j)d\theta_j \\ &= \frac{1}{M}\int f(y|\theta_j)g(\theta_j)d\theta_j \\ &= \frac{c}{M} \end{aligned}$$

where c is the normalising constant for the posterior $g(\theta|y)$. Thus, the number of iterations required to accept a single θ_j is a geometric random variable with mean $p^{-1} = M/c$. This suggests that M should be chosen as small as possible. On the other hand, we must ensure that M is large enough for $Mg_0(\theta)$ to dominate $f(y|\theta)g(\theta)$ without M being so large that it leads to an inefficient algorithm with many rejections. Even then, without extensive analysis of the tails of posterior and approximating densities, we can not be certain that $Mg_0(\theta)$ dominates $f(y|\theta)g(\theta)$.

The efficiency of the algorithm can be improved by choosing $g_0(\theta)$ that is similar in shape to the posterior, but with heavier tails, such as the student's t distribution with low degrees of freedom. Gilks and Wild (1992) propose, for low dimensional problems, what is known as *adaptive rejection sampling* to remedy the problem of finding a tight envelope function $g_0(\theta)$. If $g(\theta|y)$ was available for selection as the $g(\theta)$ density, then the minimal acceptable value for M would be 1 obtaining an acceptance probability of 1.

In both SIR and AR, if the prior is proper, it can play the role of $g_0(\theta)$. In this case, the calculations involve only the likelihood function. However, the prior is usually not a very efficient starting distribution. For many problems, especially

when there are a large number of parameters, it may be quite inefficient to conduct direct simulation.

4.3 Markov Chains Monte Carlo methods

The following notation will be used. It is assumed that the parameter vector is divided into b sub-blocks $(\theta_0, \theta_1, \ldots, \theta_b)$. In many cases, it may be natural to work with a complete breakdown of θ into all its d scalar components ($b = d$). In some cases the block θ_k may contain sub-vectors or matrices of parameters. The set of all other parameters not in block θ_k will be denoted by θ_{-k}. A sequence of θ values will be denoted by $(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(r)}, \ldots)$.

Now suppose, we wish to sample from the posterior distribution $g(\theta|y)$ which is known up to a multiplicative constant. The usual approach to Markov chain theory on a continuous state space is to start with a *transition kernel* P defined by

$$P(\theta^{(r)}, A) = P\{\theta^{(r+1)} \in A|\theta^{(r)}\}$$

for $\theta^{(r)} \in \Omega$ (the parameter space) and A a *measurable set.* The problem is to find conditions under which iterations of the transition kernel $P(\theta, A)$ converge to the posterior density $g(\theta|y)$. If $g(\theta|y)$ is the long run or *invariant* distribution of a transition kernel $P(\theta, A)$ then it satisfies the *steady state equation*

$$\int_A g(\theta|y)d\theta = \int g(\theta|y)P(\theta, A)d\theta \tag{4.3}$$

In the present context, the invariant density $g(\theta|y)$ is known but the transition kernel is unknown. The MCMC methods find a transition kernel $P(\theta, A)$ for a Markov chain which has the posterior density as its invariant distribution. Hence, a draw from the chain after a long *burn-in* time, approximates a draw from the desired posterior density. For this to hold, the constructed Markov chain must satisfy *irreducibility*, *aperiodity* and *positive recurrence* properties. The process is started from an arbitrary $\theta^{(0)}$ and is iterated long enough- say c times. The determination

of such a c is discussed in section (4.4). The simulated observations $\{\theta^{(r)} : r = c+1, \ldots, N\}$ constitute a dependent sample from $g(\theta|y)$. For any function $h(\theta)$ of θ, we can use the sample to estimate any characteristic of $h(\theta)$, say $\mathrm{E}[h(\theta)]$. Thus, an estimate of $\mathrm{E}[h(\theta)]$ is given as

$$\bar{h} = \frac{1}{N-c} \sum_{r=c+1}^{N} h(\theta^{(r)})$$

which is an *ergodic* average. Other parameters for $h(\theta)$ are also estimated by their sample equivalents.

We now briefly describe an important result in Markov chain theory which is the basis for most of the MCMC methods. Suppose that $q(\theta, \theta')$ is a *candidate generating* density: that is, it generates a candidate θ' given θ. If $q(\theta, \theta')$ satisfies the *reversibility* condition

$$g(\theta|y)q(\theta, \theta') = g(\theta'|y)q(\theta', \theta) \tag{4.4}$$

for all θ, θ' in the parameter space, then $g(\theta|y)$ is the long run distribution of transition kernel given by:

$$P(\theta, A) = \int_A q(\theta, \theta')d\theta' + s(\theta)\delta_A(\theta)$$

where $s(\theta) = 1 - \int q(\theta, \theta')d\theta'$ is the probability that the chain remains at θ; and $\delta_A(\theta) = 1$ if $\theta \in A$ and 0 otherwise. To see this, evaluate the right hand side of (4.3) to give:

$$\int g(\theta|y)P(\theta, A)d\theta = \int \int_A q(\theta, \theta')d\theta' g(\theta|y)d\theta + \int g(\theta|y)s(\theta)\delta_A(\theta)d\theta$$

Since $\delta_\theta(A) = 1$ if $\theta \in A$ and 0 otherwise, the second part of the right hand side of the above equation only needs to be integrated over A to give:

$$\begin{aligned}
\int g(\theta|y)P(\theta, A)d\theta &= \int_A \int g(\theta|y)q(\theta, \theta')d\theta d\theta' + \int_A g(\theta|y)s(\theta)d\theta \\
&= \int_A \int q(\theta', \theta)g(\theta'|y)d\theta d\theta' + \int_A g(\theta|y)s(\theta)d\theta \\
&= \int_A g(\theta'|y)(1 - s(\theta'))d\theta' + \int_A g(\theta|y)s(\theta')d\theta \\
&= \int_A g(\theta|y)d\theta
\end{aligned} \tag{4.5}$$

From the preceding discussion, it is apparent that $q(\theta, \theta')$ must satisfy the reversibility condition in order to construct a Markov chain whose long run distribution is the known posterior density $g(\theta|y)$. In the following subsections, we describe how the three MCMC methods: the Metropolis-Hastings algorithm; the Substitution sampler; and the Gibbs sampler each find a transition kernel whose r^{th} iteration converges to the specified posterior density $g(\theta|y)$ for a large r.

4.3.1 The Metropolis-Hastings (M-H) algorithm

The Metropolis-Hastings algorithm is a MCMC method that has been widely used in the mathematical physics and image restoration fields for several years. It is now being adopted with increasing frequency by the general statistical community. The simplest form was developed by Metropolis *et al.* (1953). Hastings (1970) provided an important generalisation. We now provide a description of the algorithm.

In most cases the candidate generating density $q(\theta, \theta')$ does not satisfy the reversibility condition (4.4). For some θ and θ'

$$g(\theta|y)q(\theta, \theta') > g(\theta'|y)q(\theta', \theta) \tag{4.6}$$

implying that the probability of a move from θ to θ' is higher than that of θ' to θ. This can be corrected by introducing a probability $\alpha(\theta, \theta')$ that the move from θ to θ' is made. Now the reversibility condition becomes

$$g(\theta|y)q(\theta, \theta')\alpha(\theta, \theta') = g(\theta'|y)q(\theta', \theta)\alpha(\theta', \theta)$$

If the move is not made, θ is retained as a value from $g(\theta|y)$. The probability $\alpha(\theta', \theta)$ is set as large as possible in order to always move in the less probable direction. The reversibility condition further reduces to

$$g(\theta|y)q(\theta, \theta')\alpha(\theta, \theta') = g(\theta'|y)q(\theta', \theta)$$

from which the probability $\alpha(\theta, \theta')$ can be found. Thus, in order for $q(\theta, \theta')\alpha(\theta, \theta')$ to satisfy the reversibility condition, the probability of a move from θ to θ' must be

$$\alpha(\theta, \theta') = \min\left\{\frac{g(\theta'|y)q(\theta', \theta)}{g(\theta|y)q(\theta, \theta')}, 1\right\}, \text{ whenever } g(\theta|y)q(\theta, \theta') >, 0 \tag{4.7}$$

and, by construction, $g(\theta|y)$ is the long run distribution of the Markov chain having transition kernel given by:

$$P(\theta, A) = \int_A q(\theta, \theta')\alpha(\theta, \theta')d\theta' + s(\theta)\delta_A(\theta)$$

It is important to note that $g(\theta|y)$ only enters $\alpha(\theta, \theta')$ through the ratio $g(\theta'|y)/g(\theta|y)$. This is quite crucial, since it implies that $g(\theta|y)$ need only be known up to a multiplicative constant for the M-H implementation.

Hence, the M-H algorithm initiated with $\theta^{(0)}$ can be summarised as follows:

- Repeat for $r = 1, 2, \ldots, N$
- Draw θ' from $q(\theta^{(r-1)}, \theta')$ and U from a uniform (0,1) distribution
- Calculate $\alpha(\theta^{(r-1)}, \theta')$
- If $U < \alpha(\theta^{(r-1)}, \theta')$ then set $\theta^{(r)} = \theta'$, else set $\theta^{(r)} = \theta^{(r-1)}$
- Return to the value $\theta^{(r)}$.

The chain converges to the invariant distribution $g(\theta|y)$, after running long enough, so that the effect of an arbitrary starting value is small and negligible. Any value after this long run approximates a random variable from $g(\theta|y)$. If the candidate generating density is the true posterior $g(\theta|y)$, then $\alpha(\theta, \theta')$ is always 1 so θ' is always accepted.

A successful implementation of the M-H algorithm depends on a suitable choice of the candidate generating density $q(\theta, \theta')$. In most cases, this density is chosen from a family of distributions that require the specification of parameters such as the location and spread. One might choose $q(\theta, \theta') = g_1(\theta' - \theta)$ where g_1 is some density function. The candidate θ' is, thus, generated from the process $\theta' = \theta + \theta_*$ where θ_* is the noise and is distributed as g_1. This is referred to as a *random walk chain* (Tierney, 1994). If g_1 is symmetric about the origin, then the algorithm reduces to its original form as developed by Metropolis *et al.* (1953). The multivariate normal

and the multivariate-t densities are the possible choices for g_1. Alternatively, as suggested by Hastings (1970), we might instead use an independent chain, where $q(\theta, \theta') = g_2(\theta')$. This generates a candidate independent of the current value resulting in an independent chain. Note that the acceptance probability $q(\theta, \theta')$ now becomes

$$\alpha(\theta, \theta') = \frac{g(\theta'|y)g_2(\theta)}{g(\theta|y)g_2(\theta')}$$

The function $g(\theta|y)/g_2(\theta)$ is the importance weight that would be used in SIR scheme with the starting distribution g_2. A candidate θ' with low weight would generally be rejected. As before, g_2 should be a good match for $g(\theta|y)$, but perhaps with heavier tails. This would result in an acceptance rate between 0.25 and 0.5 as suggested by Robert *et al* (1994). Tierney (1994) and Chib and Greenberg (1995) provide guidance on the proper choice of candidate generating densities.

4.3.1.1 M-H Acceptance-Rejection chains

Recall that in the A-R algorithm, we require $Mg_0(\theta)$ to dominate the posterior density $g(\theta|y)$. A suitable sized M may be hard to find in some applications. Tierney (1994) provides a remedy for this by using the A-R scheme to drive an independent M-H chain. Define $C = \{\theta : g(\theta|y) < Mg_0(\theta)\}$ and suppose that the value θ' has come through the A-R step. Since θ and θ' can each be in C or in C^c there are four possible cases: $\theta \in C$ and $\theta' \in C$; $\theta \in C^c$ and $\theta' \in C$; $\theta \in C$ and $\theta' \in C^c$; and $\theta \in C^c$ and $\theta' \in C^c$. By using the M-H algorithm, the probability of a move from θ to θ' is given by:

$$\alpha(\theta, \theta') = \begin{cases} 1 & \text{if } \theta \in C \\ Mg_0(\theta)/g(\theta|y) & \text{if } \theta \in C^c \text{ and } \theta' \in C \\ \min\{g(\theta'|y)g_0(\theta)/g(\theta|y)g_0(\theta'), 1\} & \text{if } \theta \in C^c \text{ and } \theta' \in C^c \end{cases}$$

The general proof of this result is given by Chib and Greenberg (1995). Notice that for $\theta \in C$, the probability of move to θ' is 1 regardless of where θ' lies.

4.3.1.2 Metropolis-Hastings Blockwise algorithm

Hastings (1970) discussed the possibility of applying the M-H algorithm in turn to sub-blocks θ_k rather than simultaneously to all elements of θ. Let $P_k(\theta_k, A_k|\theta_{-k})$ be the conditional transition kernel for the M-H algorithm applied to sub-blocks θ_k holding all other sub-blocks fixed. The important result is that the product of these conditional kernels

$$P(\theta, A) = \prod_{k=1}^{b} P_k(\theta_k, A_k|\theta_{-k})$$

has the posterior density as its long run distribution (Hastings, 1970). This *product of kernels* principle allows successive draws from each of the kernels, instead of having to run each of the kernels to convergence for every value of the conditioning blocks. Further, it is often much easier to find several conditional kernels that converge to their respective conditional posterior densities, than to find one kernel that converges to the joint posterior density. This result gives rise to several important special cases, of which the Gibbs sampler is the most popular. The Gibbs sampler, always samples from the correct full posterior distribution, thus the probability of acceptance is 1. The Gibbs sampler is the subject of subsection 4.3.3.

4.3.1.3 Sampling from the M-H chain

The convergence of a Markov chain to $g(\theta|y)$ can be exploited in various ways to obtain a sample from $g(\theta|y)$. Gelfand and Smith (1990) suggest generating n' independent chains each of length c, with starting values sampled from an overdispersed distribution, and then using the final value $\theta^{(c)}$ from each chain. If c is large enough, it yields an approximate *i.i.d.* sample from $g(\theta|y)$. Apart from ensuring the independence of observations in the sample, comparisons of several seemingly converged chains might reveal genuine differences, if the chains have not yet approached convergence. However, a large number of simulated values are wasted since only n' of the cn' are used and ergodic computations are not possible.

Geyer (1992) suggests generating only one M-H chain, but to continue sampling

for an additional n' iterations after convergence at c. Although the resulting sample will not be independent, it will still be the case that the empirical distribution of $\theta^{(r)}$ converges to $g(\theta|y)$ for large c and n'. Further, one long run has the advantage of reducing the dependency on initial value. The dependency of values in one long run may be reduced by extracting every m^{th} observation in the chain. For m, relatively smaller than c, this would yield an approximate *i.i.d.* sample from $g(\theta|y)$. The danger of one long run is that the chain may stay in a small subset of the parameter space for a long time giving rise to an unrepresentative sample. For large computationally expensive problems a less wasteful approach is to use all realisations of $\theta^{(r)}$ for $r \geq c$. Clearly, the selection of a sampling plan is a major implementation issue in MCMC methods. Further implementation issues are discussed in section 4.4.

4.3.2 Substitution sampling

Tanner and Wong (1987) developed an algorithm known as *data augmentation* since the observed data y are augmented with the missing data z. It determines the posterior density $g(\theta|y)$ using ideas similar to those of the EM algorithm. Basically, there is a system of two integral equations of the form:

$$\begin{aligned} g(\theta|y) &= \int g(\theta|z,y)g(z|y)dz \\ g(z|y) &= \int g(z|\theta',y)g(\theta'|y)d\theta' \end{aligned} \tag{4.8}$$

where $g(\theta|z,y)$ and $g(z|\theta',y)$ are known and $g(\theta|y)$ is to be determined. The system (4.8) can be solved by substituting the second equation into the first, and then changing the order of integration to obtain

$$\begin{aligned} g(\theta|y) &= \int g(\theta|z,y) \int g(z|\theta',y)g(\theta'|y)d\theta'dz \\ &= \int h(\theta,\theta')g(\theta'|y)d\theta' \end{aligned} \tag{4.9}$$

where $h(\theta,\theta') = \int g(\theta|z,y)g(z|\theta',y)dz$. Hence, $g(\theta|y)$ is a fixed point of the equation (4.9) since if the true density $g(\theta|y)$ is inserted on the right hand side, $g(\theta|y)$ is

obtained on the left. However, the true $g(\theta|y)$ may be analytically difficult to obtain. We may only have $g_0(\theta)$ as some appropriate starting estimate of $g(\theta|y)$. From this

$$g_1(\theta) = \int h(\theta, \theta^{(0)}) g_{(0)}(\theta^{(0)}) d\theta^{(0)}$$

is calculated. Tanner and Wong (1987) showed that the iterative process

$$g_r(\theta) = \int h(\theta, \theta^{(r-1)}) g_{(r-1)}(\theta^{(r-1)}) d\theta^{(r-1)}$$

converges to the posterior density $g(\theta|y)$ at exponential rate. The same results apply for z as well. Notice that, the iterative process of obtaining the true $g(\theta|y)$ involves integration, which in general will not be possible. Instead, a sampling based method is used for generating a sequence of random variables from each of the successive distributions. This is achieved in the following manner: firstly, draw $\theta^{(0)}$ from $g_{(0)}(\theta)$ and then draw $z^{(1)}$ from $g(z|\theta^{(0)}, y)$. This cycle is completed by drawing $\theta^{(1)}$ from $g(\theta|z^{(1)}, y)$. It is easily shown (e.g. Carlin and Louis, 1996, pp. 161) that the marginal distributions of $z^{(1)}$ and $\theta^{(1)}$ are $g_1(z)$ and $g_1(\theta)$ respectively. This process is repeated for a long time. By virtue of convergence of $g_r(\theta)$ and $g_r(z)$ to their respective true densities, we have that $\theta^{(r)} \longrightarrow \theta \sim g(\theta|z, y)$ and $z^{(r)} \longrightarrow z \sim g(z|\theta, y)$. This algorithm successively substitutes values for θ and z in turn at each cycle- hence the name Substitute sampler.

Methods for generating a sample from the posterior distribution, using the Markov chain output, have already been described. Once we have a sample, the estimate of the posterior marginal density $g(\theta|y)$ can be estimated directly by a kernel density estimate. Alternatively, as suggested by Gelfand and Smith (1990), the known form of the marginal density $g(\theta|z, y)$ cab be used to obtain a better estimate given by

$$\hat{g}(\theta|y) = \frac{1}{n'} \sum_{j=1}^{n'} g(\theta|z_j^{(c)}, y)$$

This is a Monte Carlo estimate of $g(\theta|y) = \int g(\theta|z, y) g(z|y) dz$.

4.3.3 Gibbs sampling

In general, when the parameter space dimension is d, a univariate substitution algorithm would require sampling from a chain of $d(d-1)$ distributions, which is likely to be impractical for large d. Fortunately, in many situations a much simpler alternative is provided by an algorithm, which has come to be known as the Gibbs sampler. The Gibbs sampler is an adaptation of the M-H algorithm and is discussed in detail by Geman and Geman (1984) in the context of spatial statistics. Gelfand and Smith (1990) generated new interest in the Gibbs sampler by revealing its potential in a wide variety of conventional statistical problems. They also developed the connection between the Gibbs sampler and Substitution sampling.

Suppose that all the full conditional posterior distributions

$$g_k(\theta_k|\theta_1,\ldots,\theta_{k-1},\theta_{k+1},\ldots,\theta_b,y)$$

are available for sampling. The joint posterior distribution $g(\theta|y)$ is fully determined by its conditionals (Besag, 1974; Gelman and Speed, 1993). The Gibbs sampler is a method for generating a random variable from the joint posterior density $g(\theta|y)$ in the following manner. Given arbitrary starting values $\{\theta_1^{(0)},\theta_2^{(0)},\ldots,\theta_b^{(0)}\}$, the algorithm proceeds as follows:

- Draw $\theta_1^{(1)}$ from $g_1(\theta_1|\theta_2^{(0)},\ldots,\theta_b^{(0)},y)$
- Draw $\theta_2^{(1)}$ from $g_2(\theta_2|\theta_1^{(1)},\theta_3^{(0)},\ldots,\theta_b^{(0)},y)$ then continue in the same fashion until the first cycle is completed by drawing $\theta_b^{(1)}$ from $g_b(\theta_b|\theta_1^{(1)},\theta_2^{(1)},\ldots,\theta_{b-1}^{(1)},y)$

This completes a transition from $\theta^{(0)}$ to $\theta^{(1)}$. A continuation of this process produces a sequence $(\theta^{(1)},\theta^{(2)},\ldots,\theta^{(r)},\ldots)$ which is a realisation of a Markov chain, with *transition* probability of moving from $\theta^{(r-1)}$ to $\theta^{(r)}$ given by

$$P(\theta^{(r-1)},\theta^{(r)} = \prod_{k=1}^{b} g_k(\theta_k|\theta_1^{(r-1)},\ldots,\theta_{k-1}^{(r-1)},\theta_{k+1}^{(r)}\ldots,\theta_b^{(r)},y)$$

After c such cycles we obtain $(\theta_1^{(c)},\theta_2^{(c)},\ldots,\theta_b^{(c)})$. Geman and Geman (1984) showed that under mild conditions the joint distribution of $(\theta_1^{(c)},\theta_2^{(c)},\ldots,\theta_b^{(c)})$ converges to

$g(\theta|y)$ at an exponential rate. Thus, to obtain samples from the joint posterior distribution $g(\theta|y)$, it only needs the ability to sample from the b corresponding full conditional distributions. Posterior marginal density estimates will be available through a kernel density estimate. Gelfand and Smith (1990) suggest, for n' independent generated chains, a better estimate of marginal density $g_k(\theta_k|y)$ given by

$$\hat{g_k}(\theta_k|y) = \frac{1}{n'}\sum_{j=1}^{n'} g_k(\theta_k|\theta_{j1}^{(c)}, \theta_{j2}^{(c)}, \ldots, \theta_{j(k-1)}^{(c)}, \theta_{j(k+1)}^{(c)}, \ldots, \theta_{jb}^{(c)}, y).$$

This is a less variable estimate, since the known shape of the full posterior conditional density of θ_k is used.

Gelfand *et al.* (1990) illustrate the use of the Gibbs sampler in various normal data models. Smith and Robert (1993) give examples showing how to implement the Gibbs sampler on several models, such as constrained parameter models, hierarchical models, generalised linear models and time series models. Gilks *et al.* (1993) review applications of the Gibbs sampler in medicine involving longitudinal, spatial, covariate and survival models. Zeger and Karim (1991) provide an application to generalised linear mixed model. Further applications are provided by Brooks (1998).

4.3.3.1 Relationship to M-H and Substitution algorithms

Define the M-H blockwise transition kernel as

$$P_k(\theta_k, A_k|\theta_{-k}) = \int_{A_k} g_k(\theta_k|\theta_{-k})$$

Then the Gibbs sampler comes as a special case of the blockwise M-H algorithm. In this case, we are sampling from the full conditional distribution for each block. Hence, the acceptance probability $\alpha(\theta, \theta') = 1$ for all θ and θ'.

In substitution sampling, let the missing data $z = \theta_{-k}$, then substitution sampling is equivalent to the Gibbs sampler with a different visitation order (Gelfand and Smith, 1990). One important consideration, in both algorithms, is the choice of the level at which the components for conditionals are chosen. This choice depends on the correlation structure of $g(\theta|y)$. If highly correlated scalar components are

treated individually, there could be a very slow convergence of the chain as a result of very little movement at each conditional random variate generation step. When correlated scalars are blocked together, this problem is avoided, but at the expense of having to draw from a multivariate conditional distribution.

4.3.3.2 Hierarchical models

It is recommended that the first step in the analysis of any hierarchical model is the construction of a directed graph. This graph shows all parameters including hyperparameters and the observed data as nodes in a directed graph. Arrows run into nodes from their directed influences (*parents*). There may be constant nodes which are fixed by the design of the study. These are represented by rectangles. On the other hand, stochastic nodes for parameters that are given a distribution, are denoted as circles in the graph. These stochastic nodes may be observed, in which case they are data (fixed), and thus denoted by a rectangle

A directed graph represents conditional independence assumptions on the nodes. Based on these assumptions, the joint distribution $f(\theta, y)$ of all quantities (parameters and data) can easily be expressed as the product of the conditional distributions:

$$f(\theta, y) = \prod_{w \in \{\theta, y\}} g(w|\text{parents of } w) \tag{4.10}$$

Hence a full specification of the model only needs the provision of the conditional distribution of each node given its *parents*. In Gibbs sampling, we are interested in the conditional distributions $g_k(\theta_k|\theta_{-k}, y)$. Using (4.10), these conditional distributions have the form

$$\begin{aligned} g_k(\theta_k|\theta_{-k}, y) &\propto f(\theta_k, \theta_{-k}, y) \\ &\propto \text{terms in } f(\theta, y) \text{ containing } k \\ &= g(\theta_k|\text{parents of } \theta_k) \prod_{w \in \{\theta_k\}} g(w|\text{parents of } w) \end{aligned} \tag{4.11}$$

Thus, the full conditional distribution for θ_k has a prior component $g_k(\theta_k|$ parents of $\theta_k)$

and a likelihood component coming from each *child* of θ_k. Further, it is seen that for a hierarchical model, the full conditional posterior distribution for any node depends only on the values of its *parents, children* and *co-parents* of *children* of that node.

So far, we have assumed that the prior distribution for each parameter sub-block is chosen to be conjugate with the corresponding likelihood term that precedes it in the hierarchy. This enables all of the full conditional distributions necessary for implementing the Gibbs sampler to be reduced analytically to closed form distributions. If these closed conditionals are members of familiar distributions, then sampling from them is easily done by highly efficient sampling routines. Bolstad (1997) exploits this property in a hierarchical normal regression model, where all of the Gibbs sampling conditionals were found in closed forms. But in many applications we may want to use a prior that is not conjugate or the conjugate prior does not exist such as in nonlinear growth curves and generalised linear mixed models.

Fortunately, the non-conjugacy problem for Gibbs sampling in hierarchical models is nothing but the usual computational problem of sampling from an unstandardised density. This follows from (4.11), where we see that, provided the likelihoods and priors are available in closed forms, any full conditional will be available, up to a constant of proportionality. This suggests, using direct Monte Carlo methods that do not require evaluation of the normalising constant (see section 4.2). Unfortunately, using these direct methods within a Gibbs sampler may be rather tedious to program and prohibitively expensive to implement. When a substantial number of unstandardised full conditionals is encountered, the problem may be solved more easily by using another Markov chain Monte Carlo method such as the Metropolis-Hastings.

Lastly, care must be taken when using improper priors, since proper posteriors will not always result. Recent work by Hobart and Casella (1996) show that, if improper priors are used for variance components in a hierarchical linear mixed model, the joint posterior may be improper. The danger is that the full conditional

posterior distributions can easily be found by conjugacy analysis. Gibbs Markov chains constructed from such improper posteriors will be either *null recurrent* or *transient*. Hence, the convergence properties associated with chains having proper posterior will not hold. Yet results from these null Gibbs chains may still look satisfactory.

4.4 Implementing issues

The theoretical properties of a Markov chain constructed to have a specified distribution as its long run distribution are well founded from the theory of general state space Markov chains (Tierney, 1994; Roberts, 1996). For many distributions, convergence to the limiting distribution is fairly rapid. However, there are several practical implementation issues that need to be considered before using such a chain. One such important issue is the choice of a sampling plan which has already been discussed in the context of the Metropolis-Hastings algorithm.

The determination of *burn-in* length is another implementation issue. Basically, this involves the determination of the convergence of the iterates $\theta^{(r)}$ to a random variable distributed as $g(\theta|y)$. Visual inspection of plots of the output $\{\theta^{(r)} : r = 1, \ldots, N\}$ is the most obvious and commonly used method for determining *burn-in* length. Tanner and Wong (1987) propose the use of quantile plots to monitor performance and convergence of a MCMC method. Regardless of the quality of starting values, a chain should be run long enough, so that the effects of its starting position are negligible. More formal diagnostic tools for convergence and the determination of c have been proposed. These use a variety of complicated theoretical methods and approximations, but all are based on the Monte Carlo output in some way. Raftery and Lewis (1996) and Gelman (1996) describe two of these methods using a single chain and multiple chains, respectively. Others, notably Geyer (1992), object to the idea of determining the length of *burn-in*, arguing that it is likely to be a very small fraction of N which is usually sufficiently large for

adequate precision in the estimates.

Another consideration is to determine the total sample size N or run length for adequate accuracy in the estimates. For an *i.i.d.* sample from a posterior density $g(\theta|y)$, N can easily be estimated by using a reasonable estimate of the standard error of $h(\theta)$. For dependent samples, as is the case in the iterates $\{\theta^{(r)}\}$, the estimation of the standard error of $h(\theta)$ is complicated. In such a situation, observations are generally positively correlated and a larger sample size will be required. If the series $\{h(\theta_{(r)})\}$ can be approximated by a first-order *autoregressive* process, then the asymptotic standard deviation of the sample mean $\bar{h}$ is

$$\frac{\sigma}{\sqrt{N}}\sqrt{\frac{1+\rho}{1-\rho}}$$

where σ is the posterior standard deviation of $h(\theta)$ and ρ is the autocorrelation of the series $h(\theta_{(r)})$. A rough guess for ρ can thus be used to adjust the sample size for dependency in the series. Informally, one can run several chains in parallel with different starting values, and compare the estimates $\bar{h}$. If they do not agree adequately, then N must be increased.

It is also important to monitor the performance of the samplers to ensure that they are not exhibiting any unusual behaviour. Monitoring sample paths of estimates is also useful for this purpose, as is monitoring autocorrelations of the parameters. Time series methods may also be useful for determining whether a chain exhibits any unusual features. In a Metropolis-Hastings chain, it is also important to keep track of the number of rejected candidates. This number, in the case of an independent chain, can be related to the total variation of distances between the posterior density $g(\theta|y)$ and the candidate generating density $g_0(\theta)$.

Further, as with other optimising statistical methods, such as maximum likelihood, inappropriate modelling, such as poor parameterisation, can cause convergence failures. The assumed model may contain parameters that are not identified, or may not fit the data. Sometimes, due to programming errors, the invariant distribution of the simulated process may not be the same as the specified target

posterior distribution $g(\theta|y)$, or the simulated observations may not converge to any proper distribution. Poor starting values may result in slow convergence of the chain, which can remain in a region heavily influenced by the starting distribution for many iterations.

Finally, an important consideration in selecting a Markov chain method is the cost of implementing and using the method. The cost includes coding the method, generating the chain, and storing and processing the results. The importance of these varies from application to application, and different chains may be optimal for different applications.

4.5 Conclusion

Markov chain Monte Carlo methods have increased the use of Bayesian statistics. In particular, they offer straightforward analysis of samples from the posterior instead of exact analysis of the posterior which would require subtle and sophisticated numerical and analytic approximation techniques. Their applications in statistical problems is enormous. Simple uses include exploratory data analysis on the posterior density using simulated output.

The ease of implementation can vary from one application to another. As a result, no single Markov chain method will dominate all others in all applications. It is important to be able to select a method with suitable characteristics. Despite being discovered first, the use of Metropolis-Hastings algorithm has not grown as quickly as the use of the Gibbs sampler. This may be due to the relative ease of implementing the Gibbs sampler. In many applications, a hybrid algorithm incorporating steps of Metropolis-Hastings and the Gibbs samplers may offer the best method.

Chapter 5

A full Bayesian analysis of the model

5.1 Introduction

In chapter three, we showed how the model parameters can be estimated using the EM algorithm. The model framework presented there is already hierarchical, and it is fully specified from a frequentist viewpoint. In order to cast the problem as a full Bayesian hierarchical model, it is necessary only to specify priors for the fixed effect parameter β and for the hyperparameters ξ and ψ. The prior distribution of β will be assumed multivariate normal with mean vector d_0 and diagonal covariance matrix $D_0 = v_0 I$, where v_0 is a very large number. We let the prior mean $d_0 = 0$, since the fixed effects represent logarithms of relative risks due to the baseline hazards and included factors and therefore are not located too far from 0. Thus, the prior will be nearly flat over the region under consideration. The specification of diffuse priors for precision parameters ξ and ψ is more difficult. The standard noninformative prior for a generic precision component τ is $f(\tau) \propto 1/\tau$, or equivalently $f(\sigma^2) \propto 1/\sigma^2$ for a generic variance component σ^2. In our model, the improper prior $f(\sigma^2) \propto 1/\sigma^2$ leads to a posterior distribution that is also improper.

Gamma priors are often specified for precision components due to their conjugate

family status. It seems reasonable to specify a prior density that is finite for large values of τ and also decreases monotonically in τ. Such a prior density will not preclude very small variances and also will have adequate coverage over the range that is reasonable. Thus, the hyperparameters ξ and ψ are each independently drawn from a Gamma (κ_0, ν_0) prior where, $\kappa_0 = 1$ and $\nu_0 = 1$, which is equivalent to variance prior having mode just below 1. This leads to a posterior marginal density of frailty variance with the mode depending on the observed data. Furthermore, the prior median of each random effect variance is fixed at 1.4; a plausible value based on Manda (1998a) and other studies that show very modest variances of family or community random frailty.

The Bayesian hierarchical model for survival data, studied here, is closely related to that used in Clayton (1991), Gustafson (1997). However, we differ on many aspects, such as in the treatment of the baseline hazards. Clayton (1991) uses the cumulative baseline hazard as a parameter arising from an *independent increment* gamma process prior. Gustafson (1997) uses the Cox partial likelihood resulting from integrating out the cumulative baseline hazard function with respect to a gamma process prior. We use a piecewise constant baseline hazard whose logarithm is a parameter arising from a multivariate normal distribution prior. One obvious advantage of this approach is that inference on the baseline hazards can be performed. We use a different shape of the priors for precision components as from those of the previous authors, and our computations involve an unusually high dimensional parameter space. Moreover, we are not aware of any previous implementation of a Bayesian model to child mortality data hierarchically clustered at two levels.

5.2 Joint distribution of the model

The full Bayesian model described here is essentially a random effects Poisson regression:

$$\begin{aligned}
y_{ijkq} &\sim \text{Poisson}(\mu_{ijkq}) \\
\log \mu_{ijkq} &= \log e_{ijk} + \beta' X_{ijkq} + a_i + a_{ij} \\
\beta &\sim MVN(d_0, D_0) \\
b_i = \exp(a_i) &\sim Ga(\xi, \xi) \\
b_{ij} = \exp(a_{ij}) &\sim Ga(\psi, \psi) \\
\xi &\sim Ga(\kappa_0, \nu_0) \\
\psi &\sim Ga(\kappa_0, \nu_0)
\end{aligned}$$

where $MVN(d, s)$ generically denotes a multivariate normal distribution with mean vector d and covariance matrix s; and $Ga(d, s)$ generically denotes a gamma distribution with mean d/s and variance d/s^2. Here, we assume independence between $\{y_{ijk}\}$ given all model parameters; between $\{b_i\}$ given the hyperparameter ξ; between $\{b_{ij}\}$ given the hyperparameter ψ; between $\{b_i\}$ and $\{b_{ij}\}$ and between β, ξ and ψ. The appropriate directed graph is shown in Figure 5.1.

Assuming independence among all the priors, the joint distribution of all the parameters, hyperparameters and the data is given by:

$$\begin{aligned}
& f(\text{data}, \beta, b, \xi, \psi) \\
= \; & f(\beta) f(\xi) f(\psi) \times \\
& \left\{ \prod_{i=1}^{I} f(b_i|\xi) \prod_{j=1}^{J_i} f(b_{ij}|\psi) \prod_{k=1}^{K_{ij}} L_{ijk}(t_{ijk}|\beta, b_i, b_{ij}) \right\}. \qquad (5.1)
\end{aligned}$$

where $b = \{\{b_i\}, \{b_{ij}\}\}$. After inserting appropriate terms, the joint density simplifies to

$$\begin{aligned}
& (2\pi)^{-p/2} |D_0|^{-1/2} \exp\{-\frac{1}{2}(\beta - d_0)' D_0^{-1} (\beta - d_0)\} \times \\
& \frac{\nu_0^{\kappa_0}}{\Gamma(\kappa_0)} \xi^{\kappa_0 - 1} e^{-\nu_0 \xi} \times \frac{\nu_0^{\kappa_0}}{\Gamma(\kappa_0)} \psi^{\kappa_0 - 1} e^{-\nu_0 \psi} \times
\end{aligned}$$

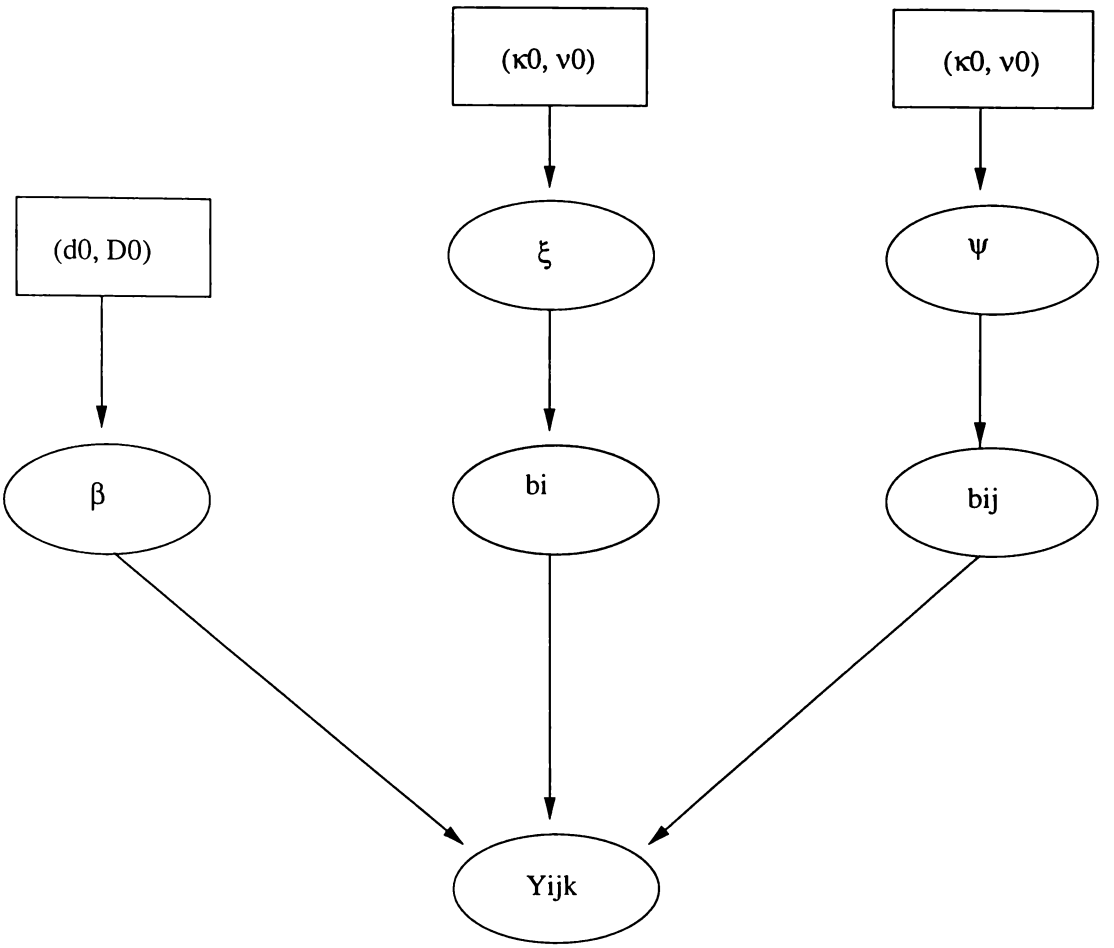


Figure 5.1: *A directed graph model for the child mortality data.*

$$\left\{\prod_{i=1}^{I}\frac{\xi^{\xi}}{\Gamma(\xi)}\left(\frac{\psi^{\psi}}{\Gamma(\psi)}\right)^{J_i}\left(\prod_{j=1}^{J_i}\prod_{k=1}^{K_{ij}}[\lambda_0(t_{ijk})e^{\beta' X_{ijk}}]^{w_{ijk}}\right) b_i^{\sum_{j,k} w_{ijk}+\xi-1} e^{-\xi b_i} \times \right.$$
$$\left.\prod_{j=1}^{J_i} b_{ij}^{\sum_k w_{ijk}+\psi-1} \exp\left(-b_i b_{ij}\sum_k \Lambda_{ijk}(t_{ijk}) - \psi b_{ij}\right)\right\} \qquad (5.2)$$

where p is the number of elements in β ($p = 17$ in our example). For Bayesian statistical inference, we require the joint posterior density of all the parameters and the hyperparameters given the observed data. Unfortunately, these can not be obtained analytically; nor is it practical to numerically obtain them because of the dimension of the parameter space. However, the conditional distributions are relatively simple to find. Thus, the Gibbs sampler can be used to obtain a sample from the full posterior distribution. The model parameters will be estimated from

this sample.

5.3 The Gibbs conditional distributions

We now specify the conditional distributions from which simulated values are to be drawn.

- The community random effect: $f(b_i|\text{data}, \beta, \{b_{ij}\}, \xi, \psi)$

$$\begin{aligned} &\propto b_i^{\sum_{j,k} w_{ijk}+\xi-1} e^{-\xi b_i} e^{-b_i\left(\sum_j b_{ij} \sum_k \Lambda_{ijk}(t_{ijk})\right)} \\ &= b_i^{\sum_{j,k} w_{ijk}+\xi-1} e^{-\left(\xi+\sum_j b_{ij} \sum_k \Lambda_{ijk}(t_{ijk})\right) b_i} \end{aligned}$$

  which is the *kernel* of a gamma distribution with shape $\sum_{j,k} w_{ijk}+\xi$ and inverse scale $\sum_j b_{ij} \sum_k \Lambda_{ijk}(t_{ijk}) + \xi$. This node will be sampled directly.

- The family random effect: $f(b_{ij}|\text{data}, \beta, \{b_i\}, \xi, \psi)$

$$\propto b_{ij}^{\sum_k w_{ijk}+\psi-1} e^{-\left(b_i \sum_k \Lambda_{ijk}(t_{ijk})+\psi\right) b_{ij}}$$

  which also is just the kernel of a gamma distribution with shape $\sum_k w_{ijk} + \psi$ and inverse scale $b_i \sum_k \Lambda_{ijk}(t_{ijk}) + \psi$.

- The community random effect inverse variance: $f(\xi|\text{data}, \beta, \{b_i\}, \{b_{ij}\}, \psi)$ is

$$\propto \left[\frac{\xi^\xi}{\Gamma(\xi)}\right]^I \left[\prod_{i=1}^I b_i\right]^{\xi-1} \exp\left[-\xi(\nu_0 + \sum b_i)\right] \xi^{\kappa_0-1}$$

  This full conditional does not simplify, and therefore methods are required for sampling from an arbitrary complex full conditional distribution. Fortunately, this conditional is *log-concave* in ξ and the adaptive rejection sampling method can be used.

- The family random effect inverse variance: $f(\psi|\text{data}, \beta, \{b_i\}, \{b_{ij}\}, \xi)$ is proportional to

$$\prod_{i=1}^I \left\{ \left[\frac{\psi^\psi}{\Gamma(\psi)}\right]^{J_i} \left[\prod_{j=1}^{J_i} b_{ij}\right]^{\psi-1} \exp\left[-\psi(\nu_0 + \sum_j b_{ij})\right] \right\} \psi^{\kappa_0-1}$$

  This also will be sampled by adaptive rejection sampling.

- The baseline and fixed effects: $f(\beta|\text{data}, \{b_i\}, \{b_{ij}\}, \xi, \psi)$. This conditional distribution is proportional to

$$|D_0|^{-1/2} \exp\{-\frac{1}{2}(\beta - d_0)' D_0^{-1}(\beta - d_0)\} \times$$
$$\prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left[(\lambda_0(t_{ijk}) e^{\beta' X_{ijk}})^{w_{ijk}} e^{-b_i b_{ij} \Lambda_{ijk}(t_{ijk})}\right]$$

which does not simplify further. Thus, methods of sampling from a non-closed full conditional distribution are required. A Taylor series expansion of $\log f(\beta|\text{data}, b, \xi, \psi)$ about the posterior mode $\tilde{\beta}$ gives

$$\begin{aligned}
& \log f(\beta|\text{data}, b, \xi, \psi) \\
= \; & \log f(\tilde{\beta}|\text{data}, b, \xi, \psi) + (\beta - \tilde{\beta}) \frac{\partial}{\partial \beta} \log f(\beta|\text{data}, b, \xi, \psi)\Big|_{\beta=\tilde{\beta}} \\
& + (\beta - \tilde{\beta})' \frac{\partial^2}{\partial \beta^2} \log f(\beta|\text{data}, b, \xi, \psi)\Big|_{\beta=\tilde{\beta}} (\beta - \tilde{\beta}) \\
= \; & \log f(\tilde{\beta}|\text{data}, b, \xi, \psi) - -\frac{1}{2}(\beta - \tilde{\beta})' \frac{\partial^2}{\partial \beta^2} \log f(\beta|\text{data}, b, \xi, \psi)\Big|_{\beta=\tilde{\beta}} (\beta - \tilde{\beta})
\end{aligned} \tag{5.3}$$

where the linear term in the expansion is zero since the log posterior density has zero derivative at its mode. As a function of β, the first term is just a constant and the last term is proportional to the logarithm of a normal density. This process yields the approximation

$$f(\beta|\text{data}, b, \xi, \psi) \approx N(\tilde{\beta}, [I(\tilde{\beta})]^{-1})$$

where $I(\tilde{\beta})$ is the observed information

$$I(\tilde{\beta}) = -\frac{\partial^2}{\partial \beta^2} \log f(\beta|\text{data}, b, \xi, \psi)$$

Thus, in larger samples $f(\beta|\text{data}, b, \xi, \psi)$ may be approximated by a normal distribution having mean equal to the posterior mode and covariance matrix equal to minus the inverse of the second derivative matrix of the log posterior evaluated at the mode $\tilde{\beta}$. This can be simplified further if a flat prior is adopted for β. This in effect replaces the posterior mode $\tilde{\beta}$ by the ML estimate $\hat{\beta}$ and $I(\hat{\beta})$ is the usual

observed Fisher information matrix. That is, to sample from $f(\beta|\text{data}, b, \xi, \psi)$, we find $\hat{\beta}$ and $[I(\hat{\beta})]^{-1}$ by performing a Poisson regression of y_{ijkq} on x_{ijkq} using the simulated values $\log b_i$ and $\log b_{ij}$ as offsets and then generate a random variable β from multivariate Gaussian distribution $N(\hat{\beta}, [I(\hat{\beta})]^{-1})$. We inserted a Metropolis step using the normal approximation as the candidate generating density, ensuring that our accepted values come from the exact distribution. The acceptance rate was almost 44%, which is very high considering that the dimension of β for our problem is 17. This indicates that the normal approximation used to generate the candidates nearly matches the conditional density.

5.4 Application to the data

As in chapter three, we fitted the Gibbs sampler to the three different frailty models. However, only results from the model containing both the family and community random effects are given here.

The resultant model required the estimation of a total of 3155 parameters: 17 fixed effects, 2911 family-specific random effects, 225 community-specific random effects and the inverse scales for both family and community frailty distributions. The Gibbs sampler method was implemented in Fortran 77 with NAG subroutines on a single SUN SPARC station 2. 5 parallel Gibbs sampler chains were run from independent starting positions for $2n = 2000$ iterations. We updated b_i, b_{ij}, ξ and ψ, 5 times for each simulated β. Sampling them equally is inefficient since it leads to successive simulated values of b_i, b_{ij}, ξ and ψ that are highly autocorrelated while those of β that are nearly independent.

Practical monitoring of the sampling process for each of the 3155 parameters was not feasible. We monitored all the fixed effects parameters; some family and community random effects; and the inverse scales ξ and ψ from all the 5 chains. The median and 95^{th} percentile of Gelman and Rubin's (1992) scale reduction factor (GR) for each monitored variable were calculated. GR compares the *between chain* variation

to the *within chain* variation and should be close to 1 if the Gibbs sampler is close to the target distribution. All the parameters had GR values very close to 1. Figure 5.2 shows traces of sampled values for some monitored components. For each, all five chains showed convergence to the same node. This is encouraging, considering that the variance components are typically more problematic parameters in Gibbs or other MCMC sampling. Furthermore, autocorrelation of most chains tended to dissipate by third to tenth lag. Thus, we took 1000 iterations as satisfactory *burn in* time.

We simulated a further 2000 values from each chain and took every 20th value after the *burn in* time. This resulted in 750 nearly independent values from the posterior distribution. Boxplots of the baseline hazards are shown in Figure 5.3. Boxplots of the remaining estimates of fixed effects are presented in Figures 5.4 and 5.5 on the log scale, where no risk is represented by 0. The medians of all these fixed effect boxplots match the respective modes obtained from the EM algorithm. Therefore the results are not discussed here.

However, the variances of the family and community random effects are larger than the EM estimates. The posterior median and means of the family and community random effect variances are 0.761, 0.789 and 0.166, 0.170 respectively. 95% credible intervals for the variances are (0.340, 1.291) and (0.010, 0.261) respectively. In the EM algorithm, the modes for the variances were estimated at 0.247 and 0.066, respectively, for the family and community random effects.

Figure 5.6 displays posterior marginal distributions for the random effects variances and Figure 5.7 shows two bivariate distributions to illustrate that higher dimensional marginal posterior distributions can easily be estimated using Gibbs sampling.

5.5 Conclusion

The Gibbs sampler offers the full Bayesian inference of the proportional hazards model with two nested levels of clustering without the evaluation of high-dimensional integrals. The approach yields a sample from a complete posterior distribution of all the parameters and hyperparameters. Thus, their behaviour can be studied over their range rather than just around the mode. We have successfully used the Gibbs sampler to study the determinants of child mortality in Malawi using child survival data hierarchically clustered at the family and the community levels. The methodology could easily be extended to non-gamma random effects and other statistical problems. Some specific examples include logistic-Gaussian (Zeger and Karim, 1991) and regression models with ARMA errors (Chib and Greenberg, 1994). It is obvious that the Gibbs sampler is computationally intensive. However, with the present very powerful computing equipment, this is of decreasing concern.

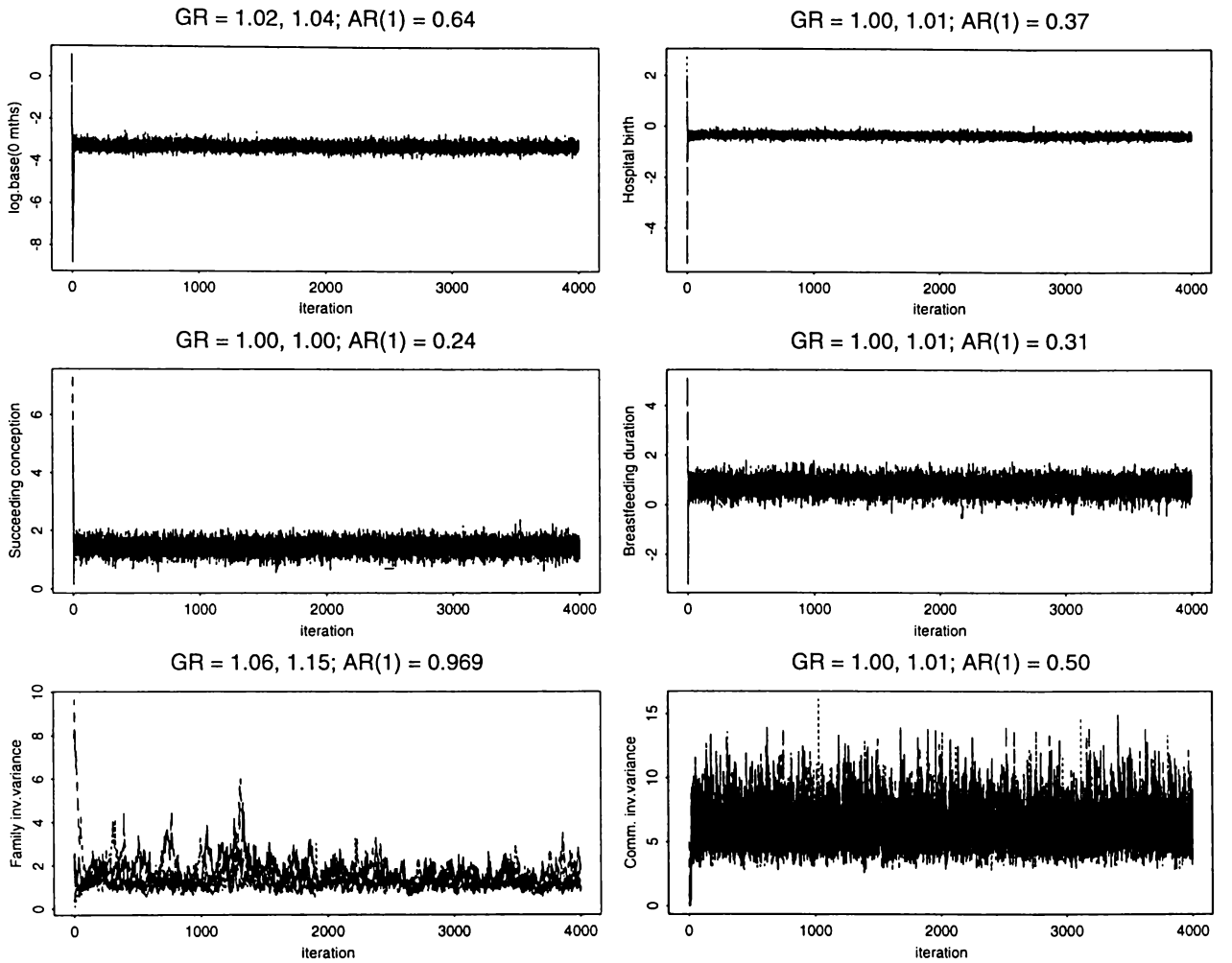


Figure 5.2: *Convergence monitoring plots for a selection of fixed effects and the two variance components. In each case, all five independent chains are plotted. Also included are the median and* 97.5^{th} *percentile of GR statistic for the first 2000 iterations; and the first-order autocorrelation AR(1), estimated from the first chain.*

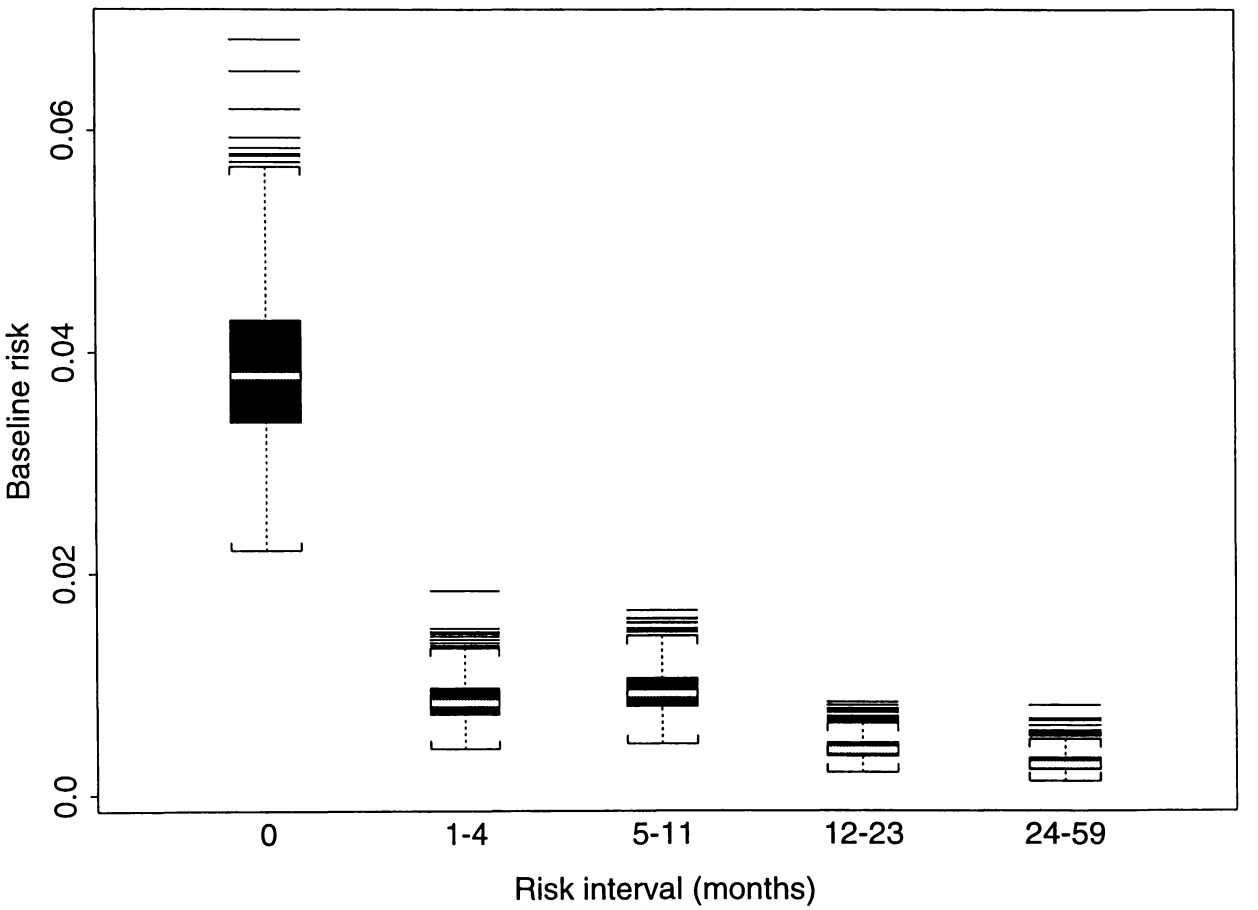


Figure 5.3: *Boxplots of the baselines hazards.*

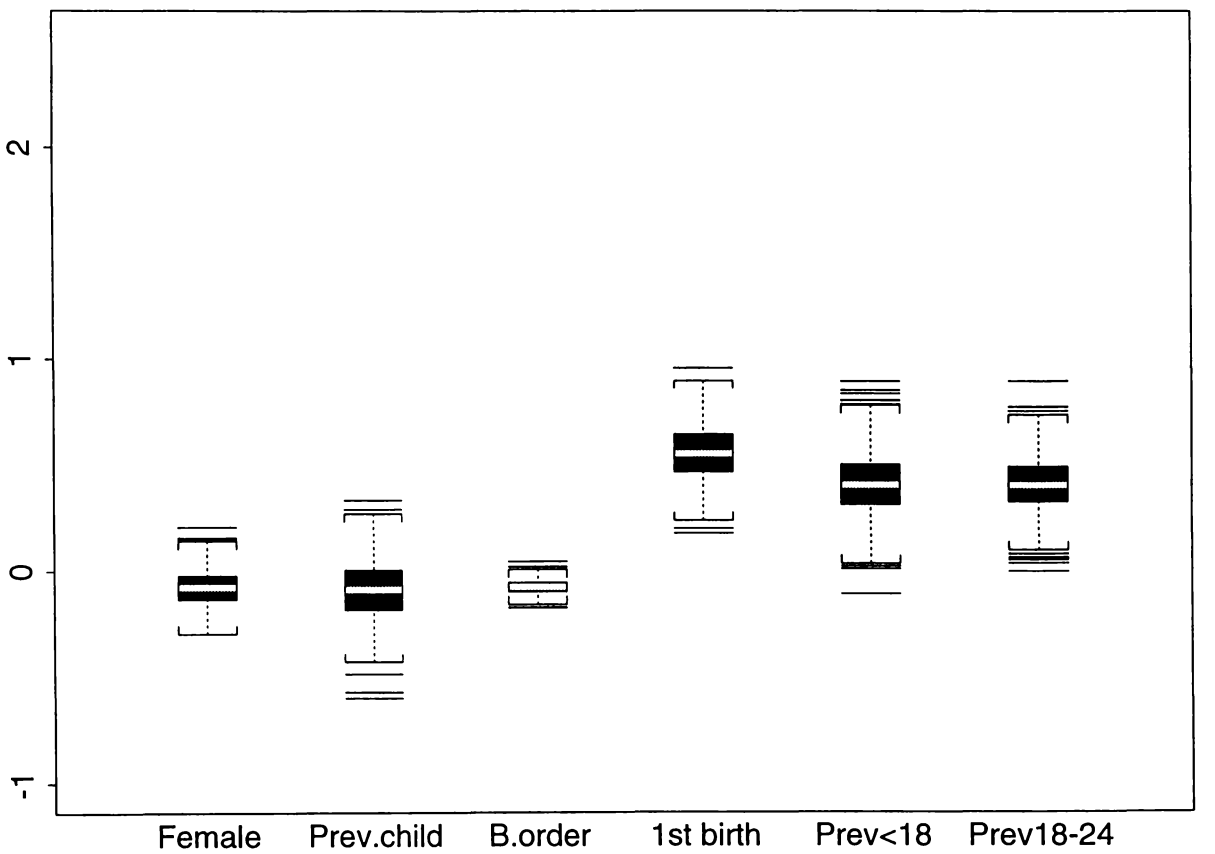


Figure 5.4: *Boxplots of logarithms of 6 fixed effect hazard rates.*

Figure 5.5: *Boxplots of logarithms of the remaining fixed effect hazard rates.*

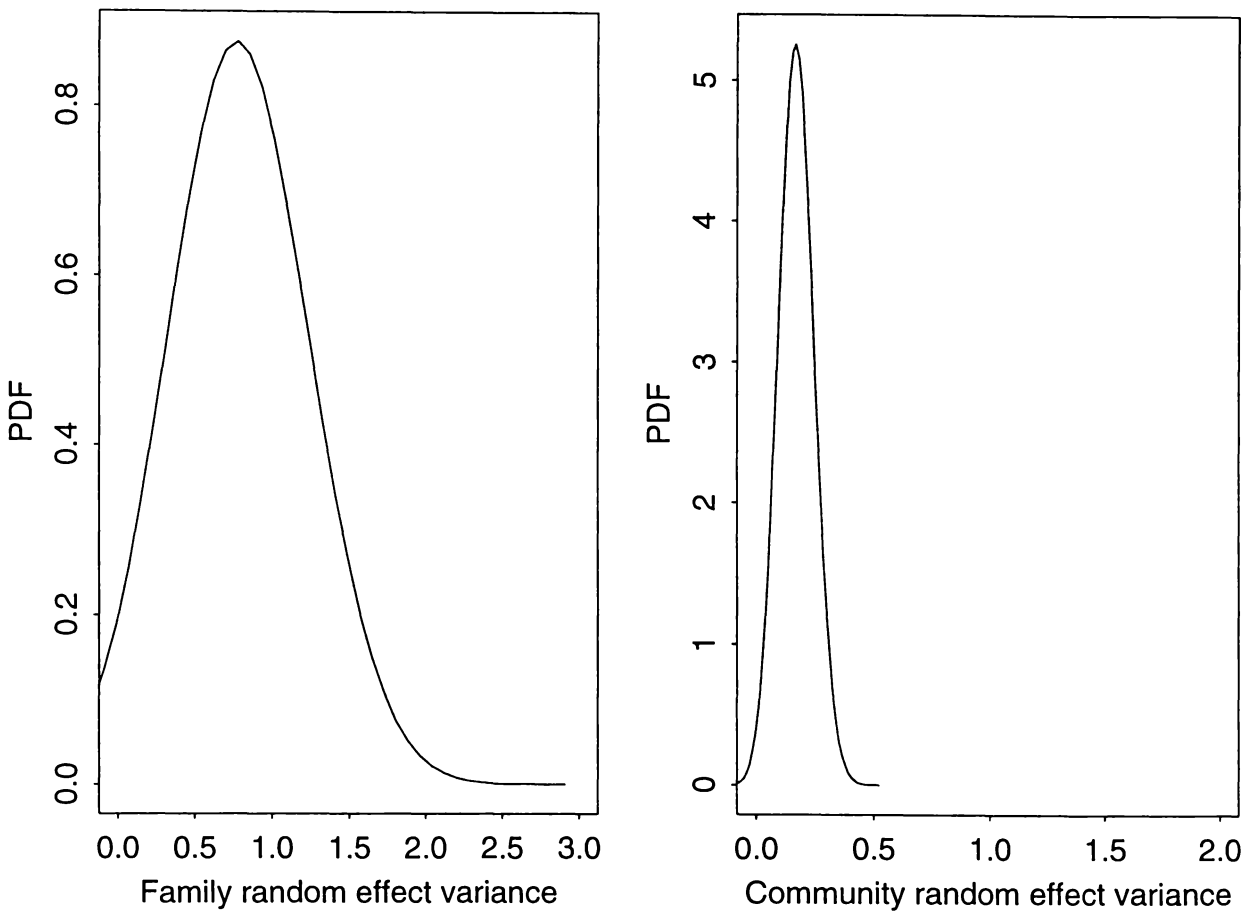


Figure 5.6: *The marginal posterior distribution of random effect variances.*

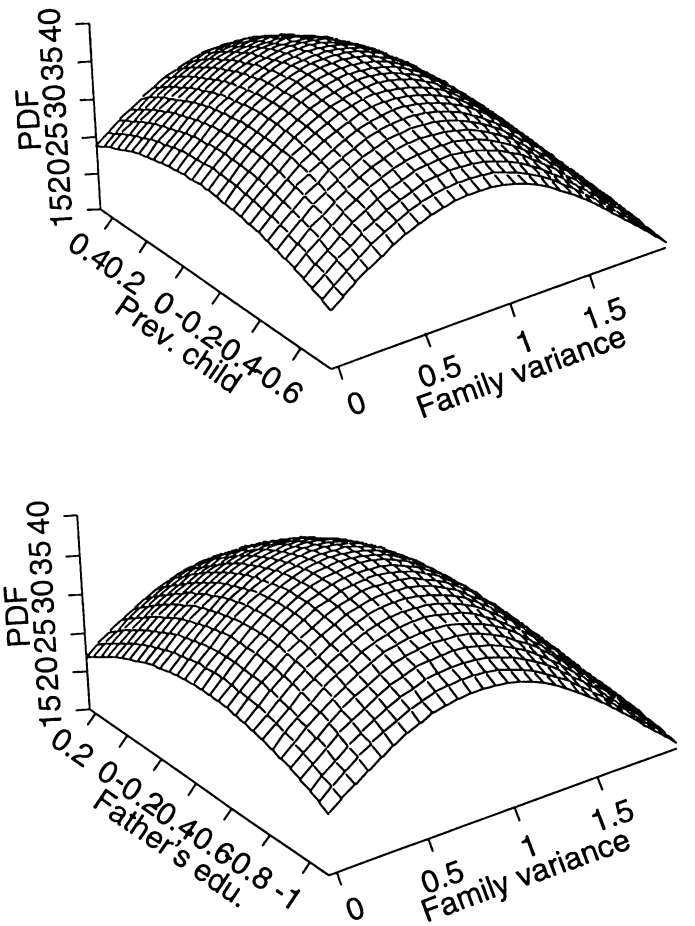


Figure 5.7: *Two bivariate distributions.*

Chapter 6

Conclusions

6.1 Thesis outline

This thesis has introduced the use of Bayesian MCMC methodology on the proportional hazards model with family and community random frailty effects to investigate the determinants of child mortality in Malawi. The resulting estimates have been compared to those obtained using the expectation-maximisation (EM) algorithm. To a lesser extent, the thesis has also covered substantive issues relating to child survival in Malawi. The child survival data used in the study were obtained from the women respondents in the 1992 Malawi Demographic and Health Survey (DHS).

The literature on determinants of child mortality was reviewed and a brief description of the DHS data was presented in chapter one. Also in the same chapter, the model was formulated and the thesis objectives were defined. In chapter two, the data was reduced into a binary response indicating death or survival during infant period, and analysed mortality using standard and random effect logistic models, the latter controlling for the random effect of the family and the community. Some substantive issue arising from the data, were also discussed in chapter two. In chapter three, the nested frailty model for survival data was analysed using the expectation-maximisation (EM) algorithm. A brief background on the theory

and problems in implementing the EM algorithm was also presented. The theory behind the Monte Carlo Markov Chain (MCMC) methods in Bayesian statistics is presented in chapter four. In chapter five, all the parameters of the model were estimated by implementing the full Bayesian inference via the Gibbs sampler.

6.2 Main conclusions

6.2.1 Substantive

The standard hazards model results suggested the following: The *death of the previous child* indicates substantial and significant additional risk. Being female shows a slightly reduced risk, and *first-birth* and short *preceding birth interval* are highly associated with increased risks of child mortality. The risk also decreases with an increasing *level of education of the father.* A *hospital birth* reduces the risk. Both short *succeeding birth interval* and short *breastfeeding duration* greatly increases the risk of death. Very young and old *maternal age* are related to increased risk.

These finding have obvious policy implications. Malawi has a high infant mortality rate, and these findings support the long-held assertion that the use of contraception to increase *birth spacing* could lead to substantial reduction in infant mortality. Short *birth intervals* are particularly harmful to mothers' health and thus increase the risk of death of a child, and in some cases maternal death. National figures show that while 94.6 percent of all married women knew some form of modern medical method of contraception, only 7 per cent used any method (NSO, 1994). It is quite likely that many women are aware of the benefits of contraception, but because of the social, economic and cultural factors they do not use (Zulu, 1998). It is also possible that some women are not fully aware of the advantages of contraception. In such situations *family planning educational* programs could contribute towards increasing the proportion of women adopting contraception, resulting in longer *birth intervals.* At a more general approach, policies and programs are needed that are aimed at

weakening and eventually removing the social, economic and cultural barriers to the use of contraception. One such program would be the training of traditional midwives and healers and then absorbing them into the public health system.

6.2.2 Methodological

The study has shown that neglecting frailty biases the estimated regression effects of observed covariates slightly downwards. However, this does not markedly affect the subsequent substantive findings, except for the effect of the *previous child's death.* This is reassuring, considering that most studies of childhood mortality have not allowed for heterogeneity of family or community random effects. We have found that in models involving family effect, the risk of a *preceding sibling death* changes from excessive in the standard model to a reduction in the frailty models. In the standard hazards model, the variable indicating whether the preceding child died has often been used to act as proxy for unobserved family effects. Its positive effect in the ordinary model may indeed confirm this.

We have found that the estimates of fixed effect parameters from the Gibbs sampler are similar to those of the EM algorithm. The only differences occur in the estimates of variance components. The random effects variances are biased towards zero in the EM algorithm. This is a feature of ML estimates of variance components as the degrees of freedom lost due to the estimation of the fixed-effects regression parameters are ignored. We found the strength of the family random effect grossly overstated when the community random effect is not included in the model. Both estimation methods clearly show this. On the other hand, there is very little change in the estimate of the community random effect variance when the family random effects are omitted.

The Gibbs sampler approach that has been used here to analyse a nested frailty model for survival data offers a better alternative to the EM algorithm and other existing methods. Unlike these other methods, the Gibbs sampler approach allows

the full inference of the model without the need to evaluate high-dimensional integrals. The sampling-based approach yields complete posterior distribution of all the parameters and hyperparameters whose behaviour can be studied over their range rather than only around the mode. Furthermore, the Gibbs approach is more believable because it incorporates the increase in uncertainty due to using the estimated rather than the actual random effects and the variability in estimating the fixed effects and hyperparameters.

6.3 Future work

The child survival data analysed in this thesis is based on a five year window. While there are adequate sample sizes for family and community units, the average number of children per family is only 1.7, giving us very limited information on family-specific random effect. Future studies might consider using a longer window, say 10 years, to increase the average level-1 units per family. On the other hand, most covariates included in the analyses tend to have similar values for children in the same family, but not necessarily so among children residing in the same community. Therefore, it would be important to include community-level covariates in future studies to see how they relate to the community random effect variance.

In some contexts the estimated cluster-specific random effects are useful in their own right. For instance, in animal breeding they are used to predict the genetic values of sires (Im and Gianola, 1988). For the DHS data, they are estimates of the unobserved risk factor associated with a particular community and family. We have not considered them, but only their distributions. In future, these estimates may help to identify high risk communities or families. Once their locations are known, the risk values could be plotted on a map with varying degree of intensity. The resulting plot might suggest causes of high mortality which might have been unrecorded. Features found in those plots could help policy makers to identify other high risk families or communities before they experience multiple child mortality.

We have relied on empirical evidence that estimates of the fixed effects parameters would be fairly insensitive to forms of distribution assumed for frailty. We have selected the gamma over other distributions because of its flexible shape and the relative ease with which it incorporates in the model computation. As well, the gamma distribution had been used previously in many studies on unobserved heterogeneity. It would be worthwhile to consider other forms, such as the log-normal and log-logistic for the random effects distribution. Unfortunately, we lose some of the analytical tractability in implementing the Gibbs sampling procedure. In this scenario, one often encounters a substantial number of unstandardised full conditionals. These concerns can be addressed by using other Markov chain Monte Carlo methods such as the Metropolis-Hastings.

Finally, another area worth investigating in future is the coverage properties particularly for variance components between the full Bayesian and maximum likelihood approaches. This, of course, would require extensive simulation with known variance components. It is expected that a Bayesian approach would provide better accurate coverage properties because it takes into account the increased uncertainity in the estimated random effects.

Appendix A

The simple linear 3-level model

A.1 The model and notation

The type of layout adopted in this thesis is a three-way nested layout with children nested within families which are nested within communities. A given data point y_{ijk} represents the response of child k in family j in community i. There are I communities, where community i contains J_i families and each of the J_i families contains K_{ij} children. In our example children are level-1 units, families are level-2 units and communities are referred to as level-3 units. For a simple 3-level model, the level-1 unit has the following model equation:

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + a_i + a_{ij} + e_{ijk} \qquad \text{(A.1)}$$

where x_{ijk} is the explanatory variable for child ijk, a_i and b_{ij} are community and family random effects that are independently normally distributed with means 0 and variances σ_c^2 and σ_f^2 respectively; and e_{ijk} is the level-1 residual having a normal distribution with mean 0 and variance σ_0^2. Model (A.1) is often termed *variance components* model because the variance of the response about the fixed part is $\sigma_c^2 + \sigma_f^2 + \sigma_0^2$, the sum of a level 1, a level 2 and a level 3 variance.

In the above formulation, responses in the same family are subjected to the same family random effect a_{ij} and the same community random effect a_i; similarly

responses in the same community are subjected to the same community random effect a_i. The covariance between a pair of responses in the same family is measured by $\sigma_c^2 + \sigma_f^2$ and for responses within the same community by σ_3^2 since the residuals are assumed to be independent. The correlations between two children in the same community or the same family are

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_f^2 + \sigma_0^2}$$

and

$$\rho_f = \frac{\sigma_c^2 + \sigma_f^2}{\sigma_c^2 + \sigma_f^2 + \sigma_0^2}$$

which are referred to as the *intra-community correlation*, which measures the proportion of variance between communities, and the *intra-family correlation* measuring variance between families.

Suppose we focus on the covariance matrix resulting from the model. For responses of three children in the same family, the covariance structure obtained is

$$\begin{bmatrix} \sigma_c^2 + \sigma_f^2 + \sigma_0^2 & \sigma_c^2 + \sigma_f^2 & \sigma_c^2 + \sigma_f^2 \\ \sigma_c^2 + \sigma_f^2 & \sigma_c^2 + \sigma_f^2 + \sigma_0^2 & \sigma_c^2 + \sigma_f^2 \\ \sigma_c^2 + \sigma_f^2 & \sigma_c^2 + \sigma_f^2 & \sigma_c^2 + \sigma_f^2 + \sigma_0^2 \end{bmatrix}$$

using the above expression. For two families in the same community, each with two children, the resulting covariance matrix is

$$\begin{bmatrix} \sigma_c^2 + \sigma_f^2 + \sigma_0^2 & \sigma_c^2 + \sigma_f^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 + \sigma_f^2 & \sigma_c^2 + \sigma_f^2 + \sigma_0^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_f^2 + \sigma_0^2 & \sigma_c^2 + \sigma_f^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_f^2 & \sigma_c^2 + \sigma_f^2 + \sigma_0^2 \end{bmatrix}$$

reflecting the fact that the covariance between children in different families in the same community is σ_c^2. The independent assumption between and among all residual components has resulted in the covariance matrix to have *block diagonal* structure. Clearly, the covariance matrix can be constructed easily for any number of levels 2 and 3 units.

A.2 Estimation

If the values of the variance components are known, then the application of the usual *Generalised Least Squares* (GLS) estimation procedure is immediate to obtain

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'V^{-1}X)^{-1}X^{-1}V^{-1}Y \tag{A.2}$$

where X is the design matrix of explanatory variables, Y is the vector of responses, and V is the block diagonal covariance matrix of the responses. However, the variances are not known, and must be estimated from the data. Several procedures such as Fisher Scoring, Newton-Raphson and the EM algorithm iteratively compute maximum likelihood estimates. These manifest themselves in several software packages. Kreft *et al.* (1994) offer a comprehensive review of these programs, and it happens that in most cases the differences are negligible.

We now provide an overview of the *Iterative Generalised Least Squares* (IGLS) method which forms the basis for estimation in chapter 2. Suppose the variance components are all zero, then an estimate of the fixed-effect vector β can be obtained from an *Ordinary Least Squares* (OLS) fit. From these reasonable estimates we can form the residual $\tilde{r}_{ijk} = y_{ijk} - \hat{\beta}_0 - \hat{\beta}_1 x_{ijk}$. The cross products matrix $\{\tilde{r}_{ijk}\}\{\tilde{r}_{ijk}\}'$ has expected value equal to the matrix V. The two matrices can be rearranged as vectors by stacking columns on top of one another to get:

$$\begin{pmatrix} \tilde{r}^2_{111} \\ \tilde{r}_{112}\tilde{r}_{111} \\ \\ \tilde{r}^2_{IJK} \end{pmatrix} = \begin{pmatrix} \sigma^2_c + \sigma^2_f + \sigma^2_0 \\ \sigma^2_c + \sigma^2_f \\ \\ \sigma^2_c + \sigma^2_f + \sigma^2_0 \end{pmatrix} + R \tag{A.3}$$

where R is a residual vector. The above equation as a regression is

$$\begin{pmatrix} \tilde{r}^2_{111} \\ \tilde{r}_{112}\tilde{r}_{111} \\ \\ \tilde{r}^2_{IJK} \end{pmatrix} = \sigma^2_c \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma^2_f \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma^2_0 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix} + R. \tag{A.4}$$

We use the estimated covariance matrix of the vector $\tilde{r}^2_{111}, \tilde{r}_{112}\tilde{r}_{111}, \cdots, \tilde{r}^2_{IJK}$ to use the GLS to obtain estimates of σ^2_c, σ^2_f and σ^2_0. We then return to (A.2) to obtain a new estimate of β. This procedure is alternated between the random and fixed effect parameter estimates until convergence (Goldstein, 1995).

Appendix B

Abstract of papers from the thesis

A Markov Chain Monte Carlo Investigation of Child Mortality in Malawi using the Proportional Hazards Model with Family and Community Random Effects

William M. Bolstad and Samuel O. Manda

Abstract

The Malawi Demographic and Health Survey conducted in 1992 collected the retrospective birth histories for a national sample of 4878 women aged between 15 and 49 years. The sample was randomly selected by a two-stage sampling design. The data consist of biological, demographic, and social variables collected for each birth. This paper models infant and early childhood survival using family and community random effects multipliers on the fixed effect proportional hazards model, which allows the dependence between observations in the same family and community into model. A Markov chain Monte Carlo sample from the posterior distribution of the parameters given the data is found. The standard errors of the fixed effect estimates are correct, unlike those found in the standard model which are underestimated because of the ignored correlation structure.

Keywords: Random effects, Gibbs sampler, Metropolis algorithm.

Unobserved Family and Village Effects on Infant Mortality in Malawi

Samuel O.M. Manda

Abstract

A three level variance components model is used to investigate the importance and magnitude of family and community random effects on infant mortality in Malawi. The results show that only the family random effect is significant in determining infant mortality, even in the presence of controls for a number observed individual and family characteristics. The results also show that biodemographic and t o a lesser extent sociodemographic factors are important determinants of infant mortality. However, the adverse effects of a prior death and a short preceding interval are magnified in the absence for controls for unobserved family random effect.

Birth Intervals, Breastfeeding and Determinants of Childhood Mortality in Malawi

Samuel O.M. Manda

Abstract

Childhood mortality in Malawi is analysed by employing proportional hazards models. The analysis uses highly reliable data collected from the 1992 Demographic and Health Survey (DHS) of Malawi. The results show that the substantial birth interval and maternal age effects are largely limited to the infant period. The influence of social and economic variables on the mortality risk and on the relationship between biodemographic variables and mortality risk is much enhanced with increasing age of the child. It has also been found that consideration of breastfeeding status of the child does not significantly alter interpretation of effects of preceding birth interval length on mortality risk, but does partially diminish the succeeding birth interval effect. The results are discussed and then summarised in the context of policy implications for Malawi. The paper addresses a very important issue in Malawi and it adds valuable insights to the base of knowledge in childhood mortality in sub-Sahara Africa.

Keywords: Childhood mortality; Malawi; Proportional hazards model.

A Comparison of Methods for Analysing a Nested Frailty Model to Child Survival in Malawi

Samuel O.M. Manda

Abstract

Many demographic and health studies collect child survival times that are clustered at the family and the community levels. We assume that each cluster has a specific unobservable random frailty effect which induces a positive dependence between the survival times in that cluster. We model the survival time using the family and community random frailty effect multipliers on the proportional hazards model. We then examine the estimation of fixed effect parameters as well as the association parameter at both levels using Markov chain Monte Carlo and the expectation-maximisation (EM) methods. We compare the two methods using child survival data collected from women respondents in the 1992 Demographic and Health Survey of Malawi. The women to be interviewed were selected by a three-stage cluster sampling design. We found that the two methods led to very similar estimates of the fixed effect parameters. However the random effect variances estimated by the methods were different. Specifically, the EM estimates are smaller than those of the Gibbs sampler. Both estimation methods clearly show that the survival experiences vary considerably across families and to a lesser extent across communities.

Key words: Bayesian inference; EM algorithm; Frailty; Gibbs sampling; Hierarchical survival analysis; Proportional hazards model.

Bibliography

[1] Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1992), *Statistical modelling in GLIM*, Oxford University Press. Oxford.

[2] Albert, J. (1993), "Teaching Bayesian statistics using sampling methods and Minitab", *The American Statistician*, 47, 3, pp. 181-191.

[3] Anderson, D.A., and Aitkin, M. (1985), "Variance component models with binary response: Interviewer variability" *Journal of the Royal Statistical Society*, Series B, 47, 2, pp. 203-210.

[4] Bedrick, E.J., Christensen, R., and Johnson, W. (1996), "A new perspective on priors for generalised linear models", *Journal of the American Statistical Association*, 91, pp. 1450-1460.

[5] Beseg, J. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)", *Journal of the Royal Statistical Society*, Series B, 36, pp. 192-236.

[6] Bolstad, W.M. (1997), "Monte Carlo method in Bayesian statistics", *New Zealand Statistician*, 32, 1, pp. 2-17.

[7] Bolstad, W.M., and Manda, S.O.M. (1998), "A Markov Chain Monte Carlo investigation of child mortality in Malawi using the Cox proportional hazards model with family and community random effects", *Unpublished*

[8] Boney, G.E. (1987), "Logistic regression fro dependent binary regression" *Biometrics*, 34, 4, pp. 951-973.

[9] Breslow, N. (1974), "Covariance analysis of censored survival data", *Biometrics*, 30, pp. 89-99.

[10] Breslow, N.E. and Clayton, D.G. (1993), "Approximate inference in generalised linear mixed models", *Journal of the American Statistical Association*, 88, 421, pp. 9-25

[11] Brooks, S.P. (1998), "Markov chain Monte Carlo methods and its application", *The Statistician*, 47, 1, pp. 69-100.

[12] Bryk, A.S., and Raudenbush, S.W. (1992), *Hierarchical linear models*, SAGA Publications. Newbury Park.

[13] Carlin, B.P. (1996), "Hierarchical longitudinal modelling", In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 303-319. Chapman and Hall. London.

[14] Carlin, B.P., and Louis, A.T. (1996), *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall. London.

[15] Casella, G., and George, E.I. (1992), "Explaining the Gibbs sampler", *The American Statistician*, 46, 3, pp. 167-174.

[16] Chib, S., and Greenberg, E. (1994), "Bayes inference in regression models with ARMA(p.q) errors", *Journal of Econometrics*, 64, pp. 183-206.

[17] Clayton, D.G. (1978), "A model for association in bivariate life tables and its application in epidemiological studies in familial tendency in chronic disease incidence", *Biometrika*, 61, 1, pp. 141-151.

[18] Clayton, D.G., and Cuzick, J. (1985a), "Multivariate generalisations of the proportional hazards model", *Journal of the Royal Statistical Society*, Series A, 148, pp. 82-117.

[19] Clayton, D.G., and Cuzick, J. (1985b), "The EM algorithm for Cox's regression model using GLIM", *Journal of the Royal Statistical Society*, Series C, 34, pp. 148-156.

[20] Clayton, D.G. (1991), "A Monte Carlo method for Bayesian inference in frailty models", *Biometrics*, 47, pp. 467-485.

[21] Cleland, J.G., and Sathar, Z.A (1984), "The effects of birth spacing in childhood mortality in Pakistan", *Population Studies*, 34, pp. 401-416.

[22] Congdon, P. (1994), "Analysing mortality in London: Life tables with frailty", *The Statistician*, 43, 2, pp. 277-308.

[23] Cox, D.R (1972), "Regression models and Life Tables (with discussion)", *Journal of the Royal Statistical Society*, Series B, 34, pp. 187-220.

[24] Curtis, S.L., Diamond, I., and McDonald, J.W (1993), "Birth interval and family effects on postneonatal mortality in Brazil", *Demography*, 30, 1, pp. 33-43.

[25] Crowder, M.J. (1978), "Beta-Binomial ANOVA for proportions", *Applied Statistics*, 27, 1, pp. 34-37.

[26] Das Gupta, M. (1990), "Death clustering, mothers' education and the determinants of child mortality in rural Punjab, India", *Population Studies*, 44, pp. 489-505.

[27] Dellarportas, P., and Smith, A.F.M. (1993), "Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling", *Journal of Applied Statistics*, 42, pp. 443-459.

[28] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society*, B, 39, pp. 1-38.

[29] Efron, B., and Hinkley, D.V. (1978), "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information", *Biometrika*, 65, 3, pp. 457-487

[30] Forste, R. (1994), "The effects of breastfeeding and child mortality in Bolivia", *Population Studies*, 48, 3, pp. 397- 511.

[31] Frankenberg, E. (1995), "The effects of access to health care on infant mortality in Indonesia", *Health Transition Review*, 5, 2, pp. 143-163.

[32] Gelfand, A.E. and Smith, A.F.M. (1990). "Sampling-based approaches to calculating marginal densities", *Journal of the American Statistical Association*, 85, pp.398-409.

[33] Gelfand A.F., Hills, S.E., Racino-Poon, A., Smith A.F.M (1990). "Illustration of Bayesian inference in normal data models using Gibbs sampling", *Journal of the American Statistical Association*, 85, pp.972-985.

[34] Gelman, A. (1996), "Inference and monitoring convergence", In *Markov Chain Monte Carlo in practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 131-143. Chapman and Hall. London.

[35] Gelman, A., and Rubin, D.B. (1992), "Inference from iterative simulation using multiple sequences (with comment)", *Statistical Science*, 7, pp. 457-511.

[36] Geman, S., and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp. 721-740.

[37] Geyer, C.J. (1992), "Practical Markov Chain Monte Carlo", *Statistical Science*, 7, 4, pp. 473-483

[38] Gilks, W.R., and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling", *Journal of Applied Statistics*, 41, pp. 337-348

[39] Gilks, W.R., Spiegelhalter, D.J, Best, N.G., McNeil, A.J., Sharples, L.D., and Kirby, A.J. (1993), "Modelling complexity: application of Gibbs sampling in medicine", *Journal of the Royal Statistical Society*, B, 55, pp. 39-102.

[40] Goldstein, H. (1995), *Multilevel statistical models*, Arnold. London.

[41] Griffiths, D.A (1973), "Maximum likelihood estimation for the Beta-Binomial distribution and an application to household distribution of the total number of cases of a disease", *Biometrics*, 29, pp. 637-648

[42] Guo, G., and Rodriguez, G. (1992), "Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala", *Journal of the American Statistical Association*, 87, 420, pp. 969-976.

[43] Guo, G. (1993). "Use of sibling data to estimate family mortality effects in Guatemala", *Demography*, 30, pp. 15-32.

[44] Gustafson, P. (1997), "Large hierarchical Bayesian analysis of multivariate survival data", *Biometrics*, 55, pp. 230-242.

[45] Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, 57, pp. 97-109.

[46] Heckman, J., and Singer, B. (1984), "A method for minimising the impact of distributional assumption in econometric models for duration data", *Econometrica*, 52, pp. 271-320.

[47] Hill, K., and Pebley, A.R. (1989), "Child mortality in the developing world", Population and Development Review, 15, pp. 657-687.

[48] Hobcraft, J. (1993), "Women's education, child welfare and child survival: A review of evidence", *Health Transition Review*, 3, pp. 159-175.

[49] Hobert, J.P., and Casella, G. (1996), "The effect of improper priors on Gibbs sampling in hierarchical linear mixed models", *Journal of the American Statistical Association*, 91, 436, pp. 1461-1473.

[50] Holt, J.D., and Prentice, R.L (1974), "Survival analysis in twin studies and matched pair experiments", *Biometrika*, 61, pp. 17-30.

[51] Holt, D., and Scott, A.J. (1981), "Regression analysis using survey data", *The Statistician*, 30, 3, pp. 169-178.

[52] Hougaard, P. (1986a), "Survival models for heterogeneous populations derived from stable distributions", *Biometrika*, 73, 2, pp. 387-396.

[53] Hougaard, P. (1986b), "A class of multivariate failure time distributions", *Biometrika*, 73, 2, pp. 671-678.

[54] Kavinya, A.M. (in progress), *Social-cultural determinants of women's reproductive-decision making in Malawi*, Ph.D thesis, Population Studies Centre, University of Waikato.

[55] Kreft, I.G., de Leeuw, J., and van der Leeden, R. (1994), "Comparing five different statistical packages for hierarchical linear regression: BMDP-5V, GENMOD, HLM, ML3, and VARCL", *American Statistician*, 48, pp. 324-335.

[56] Huster, W.J., Brookmeyer, R., and Self, S.A. (1989), "Modelling paired survival data with covariates", *Biometrics*, 45, pp. 145-156.

[57] Im, S., and Gianola, D. (1988), "Mixed models for binomial data with an application to lamb mortality", *Applied Statistics*, 37, 2, pp. 196-204.

[58] Kupper, L.L., and Haseman, J.K (1978), "The use of correlated Binomial model for the analysis of certain toxicological experiments", *Biometrics* 34, 1, pp. 69-76.

[59] Laird, N., and Oliver, D. (1981), "Covariance analysis of censored survival data using log-linear analysis techniques", *Journal of the American Statistical Association*, 76, 374, pp. 231-240.

[60] Laird, N., Lance, N., and Stram, D. (1987), "Maximum likelihood computations with repeated measures: An application of the EM algorithm", *Journal of the American Statistical Association*, 82, 397, pp. 97-105.

[61] Lantz, P., Partin, M., and Palloni, A (1992), "Using retrospective surveys for estimating the effects of breastfeeding and childspacing on infant and child mortality", *Population Studies*, 46, pp. 121-139.

[62] Larsen, U., and Vaupel, J.W. (1993), "Hutterite fecundability by age and parity: Strategies for frailty modelling of event histories", *Demography*, 30, pp. 81-102.

[63] Longford, N.T. (1993), *Random coefficient models*, Oxford. Oxford.

[64] Louis, T.A. (1982), "Finding the observed information matrix when using the EM algorithm", *Journal of the Royal Statistical Society*, Series B, 44, pp. 226-233

[65] Madise, N.J. (1993), *Birth spacing in Malawi and its impact on under five mortality*, Ph.D thesis, University of Southampton.

[66] Madise, N.J., and Diamond, I. (1995), "Determinants of infant mortality in Malawi: An analysis to control for death clustering within families", *Journal of Biosocial Science*, 27, pp. 95-106.

[67] National Statistical Office of Malawi (1994), *Malawi Demographic and Health Survey 1992*, Zomba, Malawi.

[68] Manda, S.O.M. (1998a), "Unobserved family and community effects on infant mortality in Malawi", *GENUS*, Vol. LIV, 1-2, pp. 143-164.

[69] Manda, S.O.M. (1998b), "Birth interval, breastfeeding and determinants of childhood mortality in Malawi", *Social Science and Medicine*, 48, 3, 301-312.

[70] Manda, S.O.M. (1998c), "A comparison of methods for analysing a nested frailty model to child survival in Malawi", *Unpublished*

[71] McCullagh, P., and Nelder, J.A. (1989), *Generalised linear models*, Chapman and Hall, London.

[72] McGilchrist, C.A. (1994), "Estimation in generalised mixed models", *Journal of the Royal Statistical Society*, Series B, 56, pp. 61-69.

[73] McLachlan, G.J., and Krishnan, T. (1996), *The EM algorithm and extensions*, John Wiley, New York.

[74] Meilijson, I. (1989), "A fast improvement on the EM algorithm on its own terms", *Journal of the Royal Statistical Society*, Series B, 51, pp. 127-138

[75] Meng, X.L., and Rubin, D.B. (1991), "Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm", *Journal of the American Statistical Association*, 86, 416, pp. 899-909

[76] Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of state calculations by fast computing machines", *Journal of Chemical Physics*, 21, pp. 1087-1092.

[77] Miller, J.E., Trussell, J., Pebley, A.R., and Vaughan, B (1992), "Birth spacing and child mortality in Bangladesh and the Philippines", *Demography*, 29, 2, pp. 305-318.

[78] Oakes, D. (1982), "A model for association in bivariate survival data", *Journal of the Royal Statistical Society*, Series B, 44, 3, pp. 414-422.

[79] Palloni, A., and Millman, S. (1986), "Effects of inter-births intervals and breastfeeding on infant and early childhood mortality", *Population Studies*, 40, pp. 215-236.

[80] Pebley, A.R., and Stupp, P.W. (1987), "Reproductive patterns and child mortality in Guatemala", *Demography*, 24, pp. 43-60.

[81] Pfeffermann, D. (1993), "The role of sampling weights when modelling survey data", *International Statistical Review*, 61, 2, pp. 317-337.

[82] Pickles, A., and Crouchley, R. (1995), "A comparison of frailty models for multivariate survival data", *Statistics in Medicine*, 14, pp. 1447-1461.

[83] Prentice, R.L. (1986), "Binary regression using an extended Beta Binomial distribution with discussion of correlation induced by covariates measurement errors", *Journal of American Statistical Association*, 81, 394, pp. 321-327.

[84] Prentice, R.L., and Gloeckler, L.A. (1978), "Regression analysis of grouped survival data with application to breast cancer data", *Biometrics*, 34, pp. 57-67.

[85] Retherford, R.D., Choe, M.K., Thapa, S., Gubhaju, B.B. (1989), "To what extend does breastfeeding explain birth-interval effects on early childhood mortality", *Demography*, 26, pp. 439-450.

[86] Roberts, G.O. (1996), "Markov chain concepts related to sampling algorithms", In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 45-57. Chapman and Hall. London.

[87] Robinson, G.K. (1991), "That BLUP is a good thing: Estimation of random effects", *Statistical Science*, 6, 1, pp. 15-51.

[88] Rodriguez, G., and Goldman, N. (1995), "An assessment of estimation procedures for multilevel models with binary responses", *Journal of the Royal Statistical Society*, Series A, 159, 1, pp. 73-89.

[89] Rubin, D.B. (1987), Comment of "The calculation of posterior distributions by data augmentation" by M.A. Tanner and W.H. Wong, *Journal of the American Statistical Association*, 82, pp. 543-546.

[90] Sastry, N. (1997), "A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil", *Journal of the American Statistical Association*, 92, 438, pp. 426-435.

[91] Scheike, T.H., and Jensen, T.K. (1997), "A discrete survival model with random effects: An application to time to pregnancy", *Biometrics*, 53, pp. 318-329.

[92] Schumacher, M., Olscheuski, M., and Schmoor, C (1987), "The impact of heterogeneity of comparisons of survival times", *Statistics in Medicine*, 6, pp. 773-784.

[93] Self, S.G., and Liang, K. (1987), "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions", *Journal of the American Statistical Association*, 82, 398, pp 605-610.

[94] Smith, A.F.M., and Roberts, G.O. (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", *Journal of the Royal Statistical Society*, B, 55, 1, pp. 3-23.

[95] Smith, A.F.M., and Gelfand (1992), "Bayesian statistics without tears: a Sampling-Resampling perspective", *The American Statistician*, 46, 2, pp. 84-88.

[96] Steel, B.M. (1996), "A modified EM algorithm for estimation in generalised mixed models", *Biometrics*, 52, pp. 1295-1310.

[97] Srivastava, M.L., and M'Manga, W.R. (1991), *Traditional and modern methods of child spacing in Malawi: Knowledge, attitude and practice*, University of Malawi. Zomba.

[98] Tanner, M.A., and Wong, W.H. (1987), "The calculation of posterior distributions by data augmentation (with comment)", *Journal of the American Statistical Association*, 82, pp. 528-550.

[99] Tierney, L. (1994), "Markov chain for exploring posterior distributions", *The Annals of Statistics*, 22, pp. 1701-1762.

[100] Tierney, L., and Kadane, J.B. (1986), "Accurate approximations for posterior moments and marginal densities", *Journal of the American Statistical Association*, 81, pp. 82-86.

[101] Trussell, J., and Rodriguez, G. (1990), "Heterogeneity in Demographic Research", In *Convergent Issues in Genetics and Demography*, (eds. J. Adams, D. Lam, A. Hermalin, and P. Smouse), pp. 111-132, Oxford University Press. London.

[102] Vaupel, J.W., Manton, K.G., and Stallard, E. (1987), "The impact of heterogeneity in individual frailty on the dynamics of mortality", *Demography*, 16, pp. 439-454

[103] Whitehead, J. (1980), "Fitting Cox's regression model to survival data using GLIM", *Applied Statistics*, 29, 3, pp. 268-275.

[104] Williams, D.A. (1975), "The Analysis of binary responses from toxicological experiment involving reproduction and teratogenicity", *Biometrics*, 34, 4, pp. 949-952

[105] Williams, D.A. (1982), "Extra Binomial variation in logistic models", *Journal of Applied Statistics*, 31, 2, pp. 144-148

[106] Wu, C.F.J. (1983), "On the convergence properties of the EM algorithm", *The Annals of Statistics*, 11, 1, pp. 95-103.

[107] Xue, X., and Brookmeyer, R. (1997), "Regression analysis of discrete time survival data under heterogeneity", *Statistics in Medicine*, 16, pp. 1983-1993.

[108] Zeger, S. L., and Karim, M.Z. (1991), "Generalised linear models with random effects; A Gibbs sampling approach", *Journal of the American Statistical Association*, 86, 413, pp. 79-86.

[109] Zeger, S.L., Liang, K.-Y., and Albert, R.S. (1988), "Models for longitudinal data: A generalised estimating equation approach", *Biometrics*, 44, pp. 1049-1060.

[110] Zhaorang, J., McGilchrist, C.A., and Jorgensen, M.A. (1992), "Mixed models discrete regression", *Biom. Journal*, 6, pp. 691-700.

[111] Zulu, E. (1998), *The role of men and women in decision-making about reproductive issues in Malawi*, Working Paper, 2, African Population Policy Research Centre, Nairobi.