The Bayesian Analysis of Data Arising from Complex Sampling Designs

Michelle Ross

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Jonathan C. Wakefield, Chair

Norman Breslow

Kenneth Rice

Program Authorized to Offer Degree:
School of Public Health and Community Medicine - Biostatistics

University of Washington

Abstract

The Bayesian Analysis of Data Arising from Complex Sampling Designs

Michelle Ross

Chair of the Supervisory Committee:
Professor Jonathan C. Wakefield
Biostatistics and Statistics

The majority of this thesis concerns the development of Bayesian methods for two-phase studies. A two-phase study is a study design in which limited information (including outcome) is known for a large "first-phase" study population, and more extensive information is known for a "second-phase" subset of these individuals. The two-phase study design we are considering is one in which a simple random or case-control sample is taken from a population at phase I, and is cross-classified with respect to outcome and confounder variables. At phase II, individuals are sampled within the cells of the cross-classified data, with additional data collected on exposure variables. Clearly such a design requires specialized methods of analysis to acknowledge the non-random (outcome-dependent) sampling scheme. The benefit of the two-phase design is that large efficiency gains are possible by judicial choice of the phase I confounder variables and the phase II sample sizes. A number of likelihood-based methods have been developed for the analysis of two-phase data, but we describe a Bayesian approach, which has previously been unavailable. The benefits of a Bayesian approach include relaxation of the reliance on asymptotic inference, and the potential to model data with complex dependencies, for example through the introduction of random effects. The proposed approach uses a log-linear model for the disease-exposure-confounder relationship, and specifies a multivariate normal prior distribution on a reduced set of main effect and interaction terms in the log-linear model. We extend the methodology to include random effects terms in the log-linear model to perform different kinds of smoothing. In particular,

we are interested in the use of two-phase studies in a spatial epidemiological context where one may wish to account for confounding by location by the introduction of spatial random effects. We assign independent normal priors on the non-spatial random effects, and an intrinsic conditional autoregressive (ICAR) prior on the collection of spatial random effects. Random effects can also be included in the log-linear model to smooth the cell probabilities in large contingency tables, particularly in the case of sparse data. The Bayesian two-phase approach is illustrated using data collected on Wilms tumour in children, and data on infant mortality in North Carolina.

In the last part of the thesis, we consider small area estimation in the context of the developing world. There is a distinct lack of accurate, timely, full-coverage civil registration data in the developing world, and as such, vital statistics cannot be obtained from these countries. This data is needed to formulate good public health programs, develop regional, national, and global policies and implement and evaluate public health actions. We describe an integrated data collection and statistical analysis framework for improved mortality monitoring in areas without comprehensive vital records systems. In particular, we propose the use of statistically informed sampling to increase the efficiency of sampling and to ensure that sufficient data is collected on rare populations. To do so, we use existing information from democratic surveillance system sites to construct a mortality model based on village-level characteristics. On the basis of this model, we subsequently predict the number of deaths of interest in each village in the study region, and sample proportionately in each village. The sampled deaths are then modeled as a function of known demographic factors and village-level characteristics, and we use spatial smoothing to tune the model to each village and exploit similarities of risk in neighbouring villages. The method is illustrated using a simulated data set based on a real democratic surveillance system site in Agincourt, South Africa.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

There are several individuals who have helped me to complete this dissertation and I wish to express my sincerest thanks. Firstly, I would like to thank my advisor Jon Wakefield. Over the past five years, Jon has been my advisor, my mentor and my friend. He taught me how to approach and conduct methodological research, and has been a true source of support and inspiration for me. I look forward to continued collaboration.

I would also like to thank my committee members, Ken Rice, Ross Prentice and Norm Breslow, for their insightful thoughts and discussion regarding this research. I would especially like to thank Ken Rice and Norm Breslow for agreeing to be part of my reading committee, and providing excellent comments and feedback.

I owe a debt of gratitude to my undergraduate advisor and Masters thesis supervisor, Russell Steele, UW Department of Statistics alumnus, for I never would have considered a career in Biostatistics without him. My classmates have also provided me tremendous support, both academic and not, during this five year journey. In particular, I would like to thank Erin Gabriel, Brett Hanscom, Cici Chen and Albert Kim for their advice, encouragement and support. I would also like to thank my parents, Brian and Shirley Ross, who always believed in me and constantly supported me in so many ways. I would especially like to thank my mom for encouraging me to continue to work hard, especially during the past year. She has been a guiding light and inspiration for me during some truly difficult times.

Finally, I would like to thank my beloved husband, Matthew Bryan, whose constant love, support and encouragement made it possible for me to always keep going. I could not have done this without him.

DEDICATION

For my dad,

Brian Ross

1953-2011.

Chapter 1

INTRODUCTION

The majority of this thesis concerns the development of Bayesian methods for two-phase studies. We begin by discussing two-phase studies within the context of other, perhaps more familiar, epidemiological studies, followed by a brief description of the small area estimation problem that we also consider.

1.1 An Overview of Case-Control Studies

In a 'cohort study', a group of disease-free individuals is selected and exposure levels are determined. This group is then followed over time to determine which individuals develop outcomes of interest (for example, binary disease states). This type of study is well suited to study rare exposures and multiple outcomes. However, cohort studies can be costly and can take a long time to complete. In a 'case-control study', on the other hand, a group of cases and a group of controls are selected, and their exposure status is measured. This type of study is useful for studying rare outcomes (such as a rare disease) and multiple exposures, particularly for diseases that take a long time to develop. However, it can be difficult to select an appropriate group of controls, as the controls should be selected from the same underlying population that gives rise to the cases, if selection bias is to be avoided (Breslow and Day, 1980, Section 1.5).

A 'matched case-control study' involves matching cases with a given number of controls based on a set of covariates (for example, age, gender, race). Frequency matching may reduce confounding, but the primary potential benefit of matching is a gain in efficiency. To this same end, multiple controls can be sampled from different strata so that the control group has roughly the same strata distributions as the cases, as in a 'stratified case-control design'. In each of these scenarios, the resulting control group is not a random sample from the whole population, but a random sample from the subpopulations formed by the

stratification or matching variables (Breslow and Day, 1980, Section 2.8).

As described by Ernster (1994), in a nested case-control study, a defined population is selected, and we identify cases that have already occurred (as in a 'retrospective case-control study') or as they occur (as in a 'prospective case-control study'). As in a cohort study, in nested case-control studies we follow the cohort as it moves through time. For each case, a given number of controls is selected from among those in the population who are non-cases by the time of outcome occurrence in the case. Controls should be time-matched to cases in the sense that they be matched on age, date of entry into the cohort, length of time in the cohort, or a combination of these measures. This matching allows for control of the confounding effects of time in the analysis (Wacholder et al., 1992). Further, controls may eventually become cases and an individual from the cohort may be selected as a control for more than one case. An advantage of the nested case-control design is that it is often less expensive and less time-consuming than cohort studies.

A case-cohort study resembles a nested case-control study in that a defined population is identified, however, in a case-cohort study, all cases in the cohort are selected, along with a random sample of the entire cohort (termed the 'subcohort'). Further, in this study design, cases are not matched to individuals in the comparison group. The case-cohort design has the advantage that the same subcohort can be used to study multiple outcomes. As in the nested case-control study, the case-cohort design can be less expensive and less time-consuming than other study designs.

1.2 Two-Phase Studies

It is often the case that outcome data, along with some exposure and/or confounder information is available on an entire study population of interest. Due to high costs, missing data or feasibility, additional information is only available on a small subset of individuals. A study design in which limited information (including outcome) is known for a large "first-phase" study population, and more extensive information is known for a "second-phase" subset of these individuals is referred to as a 'two-phase study'. Such a study may be conducted by design in order to increase efficiency. A two-phase study may be thought of as a generalization of stratified case-control studies nested within a cohort.

In a two-phase study, at phase I, a sample is taken from a population either via simple random or case-control sampling. In the design we will consider, these phase I data are then cross-classified with respect to the binary outcome variable, y, and to strata of *confounder* variables that we label z. The outcome variable is disease status. At phase II, individuals are randomly sampled within the cells of the cross-classified data, with additional data collected on *exposure* variables x. We assume that x is discrete, taking one of m_x possible values. We emphasize that although we refer to z and x as confounder and exposure variables, respectively, z need not be a confounder and x need not be an exposure variable. For example, in the Wilms tumour example that we will describe in Section 2.8.1, the phase I variable z is a mismeasured surrogate for the exposure variable of interest. In general, the z variable can include any variable available for all subjects at phase I, and the x variable represents the additional data collected on the phase II individuals.

Due to the sampling scheme of the phase II data, the phase II sample is a non-representative sample of the entire first-phase study population. Analyzing the second-phase data using conventional methods therefore leads to biased estimates of the parameters of interest, so specialized methods of analysis are required. The benefit of the two-phase design is that large efficiency gains are possible by judicial choice of the phase I confounder variables and the phase II sample sizes.

A number of likelihood-based methods have been developed for the analysis of two-phase data, but we describe a Bayesian approach that has previously been unavailable. A Bayesian approach offers the possibility of improved small sample properties, since the asymptotic arguments of likelihood-based approaches are not required, and it also offers the possibility of modeling complex dependencies in the data through the introduction of random effects. For example, Wakefield and Haneuse (2008) describe the use of two-phase studies in a spatial epidemiological context, in which one may wish to reduce confounding by location by using spatial random effects.

1.3 Small Area Estimation

In general, small area estimation tackles the problem of providing reliable estimates of one or more variables of interest in a set of small geographic areas. The term "small area"

generally refers to any area for which direct estimates of the variables of interest cannot be produced with adequate precision (Rao, 2003, Section 1.1). Here, we consider small area estimation in the context of the developing world.

With less than a third of the world's population covered by accurate data on births and deaths, the need to commit resources to the registration of births and deaths, and to certify the causes of death is evident (Horton, 2007; Setel et al., 2007). Adult, child, and maternal mortality data are critical to formulating good public health programs, developing regional, national, and global policies, and implementing and evaluating public health actions. Setel et al. (2007) maintain that "barely a third of countries outside North America and Europe have the capacity to obtain usable mortality statistics", leaving most of the world's poor as "unseen, uncountable, and hence uncounted". In 2007, *The Lancet* published a series of articles titled "Who Counts?" (AbouZahr et al., 2007; Boerma and Stansfield, 2007; Hill et al., 2007; Horton, 2007; Mahapatra et al., 2007; Setel et al., 2007), which details this "scandal of invisibility" resulting from the lack of accurate, timely, full-coverage civil registration that affects much of the developing world.

Currently, the most common data collection methods include sample surveys and surveillance systems. Sample surveys aim to provide information about a finite population by observing a representative subset of the population. Sample surveys often collect a vast amount of information on a large group of individuals, possibly belonging to different subpopulations of interest, but they generally do not provide any longitudinal information at the individual level since the same individuals are often not revisited (Särndal et al., 1992, Section 1.1-1.2). On the other hand, surveillance systems provide detailed longitudinal data on individuals or households describing various demographic events (such as births, deaths, marriages and divorces) on a non-representative group of individuals (Setel et al., 2007). The monitoring system we propose combines the benefits of both of these data collection methods to provide useful indicators for large populations over prolonged periods of time, so that we can monitor change and determine reasons for the observed changes, for example changes in mortality due to the introduction of an intervention.

1.4 Outline of Thesis

Chapter 2 begins by providing basic notation and describing and critiquing approaches to the analysis of two-phase data. We go on to describe Bayesian outcome-dependent sampling methods, Bayesian methods for contingency tables and Bayesian computational techniques. An overview of spatial modeling and disease mapping methods is also provided. We then discuss the data collection methods currently in place to monitor health information in the developing world, in particular sample surveys and surveillance systems, and conclude the chapter with a discussion of two motivating examples.

Chapter 3 considers the Bayesian analysis of two-phase data for the case of independent outcomes. We start by describing in detail methods for the simplest case of assessing the association between a binary outcome variable and a discrete exposure variable. We extend these methods to the case of assessing the association between a binary outcome variable and a discrete exposure in the presence of discrete confounder variables.

The methods described in Chapter 3 are extended in Chapter 4 to include random effects to handle correlated outcomes, for example. In this case, the confounder variables can represent geographical location and the models incorporate spatial and non-spatial random effects.

Chapter 5 describes the integrated data collection and statistical analysis framework that we propose for improved mortality monitoring in regions without comprehensive vital records systems.

Finally, Chapter 6 provides an overview of the research and a discussion of the issues that remain to be investigated.

Chapter 2

NOTATION AND STATISTICAL BACKGROUND

In this chapter, we provide background information and a review of previous literature for the topics considered in the thesis. In Section 2.1, we provide basic notation and in Section 2.2 we outline proposed approaches to the analysis of two-phase data. Section 2.3 reviews Bayesian approaches to outcome-dependent sampling. This is followed by a discussion of Bayesian methods for contingency tables in Section 2.4 and Bayesian computation methods in Section 2.5. An overview of spatial modeling and disease mapping methods is provided in Section 2.6 and sample survey methodology is discussed in Section 2.7. Motivating and literature examples are discussed in Section 2.8. The chapter concludes with a brief review of the small area estimation problem in Section 2.9.

2.1 Notation

Let N_{yxz} denote the number of individuals in the population with outcome y, exposure level x, and confounder level z. We begin by describing the phase I study in which we observe the response-confounder margin $\boldsymbol{N}^{\mathrm{y\cdot z}} = \{N_{y\cdot z}, y = 0, 1, z = 1, \ldots, m_z\}$, where we use the dot notation to indicate summation over that index. Table 2.1 provides a representation of the cross-classification of the phase I data.

Table 2.1: Representation of phase I data from a two-phase sampling scheme.

	$Z=1$	$Z=2$	$\cdots$	$Z=m_z$	
$Y=0$	$N_{0\cdot 1}$	$N_{0\cdot 2}$	$\cdots$	$N_{0\cdot m_z}$	$N_{0\cdot\cdot}$
$Y=1$	$N_{1\cdot 1}$	$N_{1\cdot 2}$	$\cdots$	$N_{1\cdot m_z}$	$N_{1\cdot\cdot}$
	$N_{\cdot\cdot 1}$	$N_{\cdot\cdot 2}$	$\cdots$	$N_{\cdot\cdot m_z}$	$N_{\cdots}$

At phase II, sample sizes $n_{y\cdot z}$ are decided upon and individuals are drawn at random from the $N_{y\cdot z}$ in each of the $2m_z$ outcome by stratum combinations and the exposure information is measured for each individual. Note that the $n_{y\cdot z}$ are fixed by design and must be smaller than the corresponding (random) $N_{y\cdot z}$ values. Let n_{yxz} denote the number of individuals in the phase II sample with outcome y, exposure level x, and confounder level z. At phase II, we observe the phase II sample sizes $\boldsymbol{n}^{\mathrm{y\cdot z}} = \{n_{y\cdot z}, y = 0, 1, z = 1, \ldots, m_z\}$, and the phase II outcomes $\boldsymbol{n}^{\mathrm{yxz}} = \{n_{yxz}, y = 0, 1, x = 1, \ldots, m_x, z = 1, \ldots, m_z\}$. Variables super-scripted with x, y, z will represent vectors throughout. Table 2.2 provides a representation of the cross-classification of the phase II data.

Table 2.2: Representation of phase II data from a two-phase sampling scheme.

	$Z=1$		$Z=2$		$\cdots$	$Z=m_z$		
	$Y=0$	$Y=1$	$Y=0$	$Y=1$		$Y=0$	$Y=1$	
$X=1$	n_{011}	n_{111}	n_{012}	n_{112}	$\cdots$	n_{01m_z}	n_{11m_z}	
$X=2$	n_{021}	n_{121}	n_{022}	n_{122}	$\cdots$	n_{02m_z}	n_{12m_z}	
$\vdots$								
$X=m_x$	n_{0m_x1}	n_{1m_x1}	n_{0m_x2}	n_{1m_x2}	$\cdots$	$n_{0m_xm_z}$	$n_{1m_xm_z}$	
	$n_{0\cdot1}$	$n_{1\cdot1}$	$n_{0\cdot2}$	$n_{1\cdot2}$	$\cdots$	$n_{0\cdot m_z}$	$n_{1\cdot m_z}$	$n_{\cdots}$

The aim is to estimate the confounder-adjusted association between disease and exposure. We assume a linear logistic disease model of the form

$$\log\left(\frac{p_{xz}}{1-p_{xz}}\right) = \boldsymbol{W}\boldsymbol{\beta}, \tag{2.1}$$

where $p_{xz} = \mathrm{pr}(Y = 1|X = x, Z = z, \boldsymbol{\beta})$ is the probability of disease in exposure group x and confounder stratum z. Depending on the context, the design matrix $\boldsymbol{W}$ may contain terms for both main effects and interactions. Let p denote the dimension of $\boldsymbol{\beta}$. To re-iterate, the goal is to efficiently estimate the parameter vector $\boldsymbol{\beta}$.

2.2 Two-Phase Studies

Two-phase studies were introduced by Neyman (1938), who described estimation methods for what he termed "double sampling". The development of two-phase studies within the context of epidemiology (where the binary outcome represents disease status) began with the papers of Walker (1982) and White (1982). White proposes a two-stage approach in studies of the relationship between a rare disease and a rare exposure in an effort to gain efficiency of odds ratio estimation over the standard case-control scheme. For binary Y and X, and discrete Z, $Z = 1, \ldots, m_z$, let

$$\begin{aligned} \psi_z &= \frac{\mathrm{pr}(Y{=}1|X{=}1,Z{=}z)\mathrm{pr}(Y{=}0|X{=}0,Z{=}z)}{\mathrm{pr}(Y{=}0|X{=}1,Z{=}z)\mathrm{pr}(Y{=}1|X{=}0,Z{=}z)} \\ &= \frac{\mathrm{pr}(Z{=}z|Y{=}1,X{=}1)\mathrm{pr}(Z{=}z|Y{=}0,X{=}0)}{\mathrm{pr}(Z{=}z|Y{=}0,X{=}1)\mathrm{pr}(Z{=}z|Y{=}1,X{=}0)} \times \frac{\mathrm{pr}(Y{=}1,X{=}1)\mathrm{pr}(Y{=}0,X{=}0)}{\mathrm{pr}(Y{=}0,X{=}1)\mathrm{pr}(Y{=}1,X{=}0)}, \end{aligned}$$

denote the odds ratio relating disease and exposure for confounder level z. White shows that a consistent estimate of ψ_z is

$$\widehat{\psi}_z = \frac{(n_{11z}/n_{11\cdot})(n_{00z}/n_{00\cdot})N_{11\cdot}N_{00\cdot}}{(n_{01z}/n_{01\cdot})(n_{10z}/n_{10\cdot})N_{01\cdot}N_{10\cdot}}.$$

Since the paper of White (1982), the methodology has become more advanced with weighted likelihood (Flanders and Greenland, 1991; Reilly and Pepe, 1995; Whittemore, 1997), pseudo-likelihood (Breslow and Cain, 1988; Scott and Wild, 1991; Schill, Jockel, Drescher, and Timm, 1993) and non-parametric maximum likelihood (Breslow and Holubkov, 1997a,b; Scott and Wild, 1997) approaches being suggested.

The phase I data are available through one of two sampling mechanisms: either case-control or simple random sampling. In the case-control sampling at phase I scheme, $N_{1\cdot\cdot}$ cases and $N_{0\cdot\cdot}$ controls are sampled independently, and the individuals are then classified with respect to the strata of confounder variables. As detailed in Holubkov (1995), the complete likelihood for the data from a two-stage sampling scheme has two components: one from the phase I sample, the other from the phase II sample. At phase I, the random quantities are $\boldsymbol{N}^{\mathrm{y\cdot z}}$, the response-confounder margin, whereas at phase II the random quantities are $\boldsymbol{n}^{\mathrm{y\cdot z}}$, the phase II sample sizes, and $\boldsymbol{n}^{\mathrm{yxz}}$, the phase II observations. For a given disease status, $\boldsymbol{N}^{\mathrm{y\cdot z}}$ is Multinomial$(N_{y\cdot\cdot}, \{\mathrm{pr}(Z = z|Y = y)\})$. Hence, the phase I

component of the likelihood is proportional to

$$\prod_{y=0}^{1}\prod_{z=1}^{m_z}\mathrm{pr}(Z=z|Y=y)^{N_{y\cdot z}}.$$

Similarly, the phase II component of the likelihood is given by

$$\prod_{y=0}^{1}\prod_{z=1}^{m_z}\prod_{x=1}^{m_x}\mathrm{pr}(X=x|Z=z,Y=y)^{n_{yxz}}.$$

Assuming the sampling of the phase II individuals does not depend on the parameters in the model or on the unobserved $\boldsymbol{N}^{\mathrm{yxz}}$, the complete likelihood is given by

$$\begin{aligned}\prod_{y=0}^{1}\mathrm{pr}(\boldsymbol{N}^{\mathrm{y\cdot z}},\boldsymbol{n}^{\mathrm{yxz}}|\boldsymbol{N}^{\mathrm{y\cdot\cdot}}) \quad\propto\quad & \left(\prod_{y=0}^{1}\prod_{z=1}^{m_z}\mathrm{pr}(Z=z|Y=y)^{N_{y\cdot z}}\right)\\ & \times\left(\prod_{y=0}^{1}\prod_{z=1}^{m_z}\prod_{x=1}^{m_x}\mathrm{pr}(X=x|Z=z,Y=y)^{n_{yxz}}\right). \qquad (2.2)\end{aligned}$$

To determine the form of the likelihood under the logistic model, the likelihood is parameterized in a particular way. Consider a sampling scheme where individuals are sampled from the population, with cases being sampled with probability $N_{1\cdot\cdot}/N_{\cdots}$ and controls with probability $N_{0\cdot\cdot}/N_{\cdots}$. Let pr^1 denote probabilities under this first-stage sampling scheme. Then, we define $\mathrm{pr}^1(Z=z,Y=y)$ as the joint disease-confounder distribution conditional on a subject being sampled at phase I according to the sampling scheme just described:

$$\begin{aligned}\mathrm{pr}^1(Z=z,Y=y) &= \frac{N_{y\cdot\cdot}}{N_{\cdots}}\mathrm{pr}(Z=z|Y=y),\\ \Rightarrow Q_z &= \sum_{y=0}^{1}\frac{N_{y\cdot\cdot}}{N_{\cdots}}\mathrm{pr}(Z=z|Y=y),\end{aligned}$$

where Q_z is the marginal distribution of Z conditional on having been sampled at phase I, and where pr denotes probabilities under sampling from the population. Note that $\mathrm{pr}^1(Z=z|Y=y)=\mathrm{pr}(Z=z|Y=1)$. Using Bayes theorem, we find

$$\begin{aligned}\mathrm{pr}(Z=z|Y=y) &= \mathrm{pr}^1(Y=y|Z=z)\mathrm{pr}^1(Z=z)\frac{N_{\cdots}}{N_{y\cdot\cdot}}\\ &\equiv p_y(z)Q_z\frac{N_{\cdots}}{N_{y\cdot\cdot}}. \qquad (2.3)\end{aligned}$$

To formulate pr^1 in terms of the logistic model, we use (2.3) to obtain

$$\log\left(\frac{\mathrm{pr}^1(Y=1|Z=z)}{\mathrm{pr}^1(Y=0|Z=z)}\right) = \log\left(\frac{N_{1\cdot\cdot}}{N_{0\cdot\cdot}}\right)+\log\left(\frac{\mathrm{pr}(Z=z|Y=1)}{\mathrm{pr}(Z=z|Y=0)}\right)$$

so that

$$\mathrm{logit}\left(\mathrm{pr}^1(Y=1|Z=z)\right) = \log\frac{N_{1\cdot\cdot}}{N_{0\cdot\cdot}}+\delta_z,$$

where $\delta_z = \log(\mathrm{pr}(Z=z|Y=1)/\mathrm{pr}(Z=z|Y=0))$. That is, δ_z represents the log-odds ratio of disease comparing levels of Z in the population.

Analogously, for phase II we define the stratum-specific marginal distribution for X assuming a sampling scheme where cases and controls are sampled within stratum z with probabilities $n_{1\cdot z}/n_{\cdot\cdot z}$ and $n_{0\cdot z}/n_{\cdot\cdot z}$, respectively. Letting pr^2 denote probabilities under this second-stage sampling scheme, this is given as

$$\mathrm{pr}^2(X=x|Z=z) = \sum_{y=0}^{1}\frac{n_{y\cdot z}}{n_{\cdot\cdot z}}\mathrm{pr}(X=x|Y=y,Z=z),$$

where $\mathrm{pr}^2(X=x|Z=z)$ is denoted by $q_z(x)$ and $\mathrm{pr}^2(X=x|Y=y,Z=z) = \mathrm{pr}(X=x|Y=y,Z=z)$. Using Bayes theorem, we have

$$\begin{aligned}\mathrm{pr}(X=x|Y=y,Z=z) &= \mathrm{pr}^2(Y=y|X=x,Z=z)\mathrm{pr}^2(X=x|Z=z)\frac{n_{\cdot\cdot z}}{n_{y\cdot z}}\\ &\equiv p_{yx}(z)q_z(x)\frac{n_{\cdot\cdot z}}{n_{y\cdot z}}. \qquad (2.4)\end{aligned}$$

To formulate pr^2 in terms of the logistic model, we use (2.4) to obtain

$$\begin{aligned}\log\left(\frac{\mathrm{pr}^2(Y=1|X=x,Z=z)}{\mathrm{pr}^2(Y=0|X=x,Z=z)}\right) &= \log\left(\frac{n_{1\cdot z}}{n_{0\cdot z}}\right)+\log\left(\frac{\mathrm{pr}(X=x|Y=1,Z=z)}{\mathrm{pr}(X=x|Y=0,Z=z)}\right)\\ &= \log\left(\frac{n_{1\cdot z}}{n_{0\cdot z}}\right)+\log\left(\frac{\mathrm{pr}(Y=1|X=x,Z=z)}{\mathrm{pr}(Y=0|X=x,Z=z)}\right)\\ &\quad -\log\left(\frac{\mathrm{pr}(Y=1|Z=z)}{\mathrm{pr}(Y=0|Z=z)}\right)\end{aligned}$$

so that

$$\mathrm{logit}\left(\mathrm{pr}^2(Y=y|X=x,Z=z)\right) = \frac{n_{1\cdot z}}{n_{0\cdot z}}+\gamma-\delta_z+\boldsymbol{x}^T\boldsymbol{\beta},$$

where $\gamma = \beta_0 - \log(\mathrm{pr}(Y=1)/\mathrm{pr}(Y=0))$.

Substituting (2.3) and (2.4) into (2.2), the complete likelihood is proportional to

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\delta}, \gamma, \mathbf{Q}_z, \mathbf{q}_z(\cdot)) &= \prod_{y=0}^{1} P(\boldsymbol{N}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{yxz}} | \boldsymbol{N}^{\mathrm{y\cdot\cdot}}) \\ &\propto \left(\prod_{y=0}^{1} \prod_{z=1}^{m_z} (p_y(z) Q_z)^{N_{y\cdot z}} \right) \times \left(\prod_{y=0}^{1} \prod_{z=1}^{m_z} \prod_{x=1}^{m_x} (p_{yx}(z) q_z(x_{yxz}))^{n_{yxz}} \right) \\ &= L_1(\boldsymbol{\delta}, \mathbf{Q}_z) \times L_2(\boldsymbol{\delta}, \gamma, \boldsymbol{\beta}, \mathbf{q}_z(\cdot)), \end{aligned} \tag{2.5}$$

where $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_{m_z}\}$. The parameters $\boldsymbol{\delta}$, γ, $\boldsymbol{\beta}$, $\{Q_z\}$, and $\{q_z(\mathbf{x})\}$ are restricted by the fact that (2.3) and (2.4) are probability densities for $y = 0, 1$ and $z = 1, \ldots, m_z$. These constraints are referred to as the phase I and phase II constraints and can be written as

$$\begin{aligned} \frac{N_{y\cdot\cdot}}{N_{\cdot\cdot\cdot}} &= \sum_{z=1}^{m_z} p_y(z) Q_z, \qquad y = 0, 1 \\ \frac{n_{y\cdot z}}{n_{\cdot\cdot z}} &= \sum_{x=1}^{m_x} p_{yz}(\mathbf{x}) q_z(\mathbf{x}), \quad y = 0, 1, z = 1, \ldots, m_z. \end{aligned}$$

Breslow and Holubkov (1997a) solve the constrained maximum likelihood estimation problem using a system of Lagrangian equations. Before we discuss the Lagrangian equations, we present two simpler approaches.

The simpler of the two approaches is the weighted likelihood approach proposed by Flanders and Greenland (1991) that has its origins in sampling theory. In this approach, the logistic regression model (2.1) is fit to the phase II data using the inverse sampling fractions as prior weights, where the sampling fractions are given by $n_{y\cdot z}/N_{y\cdot z}$, thereby adjusting for the outcome-dependent sampling scheme. This is equivalent to solving a weighted score equation. The resulting estimate has been referred to as the Horvitz-Thompson estimate (Breslow and Chatterjee, 1999). Note that this method ignores the phase I data except to provide a weighting scheme for the analysis of the phase II data, and both sets of constraints are ignored. A more general estimating equation technique was introduced by Reilly and Pepe (1995) but results in the same estimating equation as Flanders and Greenland (1991) in the two-phase setting.

The second approach is the pseudo-likelihood approach in which the nuisance parameters Q_z and $q_x(z)$, $x = 1, \ldots, m_x, z = 1, \ldots, m_z$, of the two-phase likelihood are ignored, and the

resulting simplified likelihood is maximized. This likelihood is referred to as the pseudo-likelihood (Besag, 1975):

$$\begin{aligned} L^*(\boldsymbol{\delta}, \gamma, \boldsymbol{\beta}) &= \left(\prod_{y=0}^{1} \prod_{z=1}^{m_z} p_y(z)^{N_{y\cdot z}} \right) \times \left(\prod_{y=0}^{1} \prod_{z=1}^{m_z} \prod_{x=1}^{m_x} p_{yx}(z)^{n_{yxz}} \right) \\ &= L_1^*(\boldsymbol{\delta}) L_2^*(\boldsymbol{\delta}, \gamma, \boldsymbol{\beta}). \end{aligned}$$

Breslow and Cain (1988) extend the results of White (1982) to incorporate continuous exposure and confounder variables and use conditional maximum likelihood under a logistic disease model to obtain parameter estimates. They maximize the phase I component of the pseudo-likelihood, $L_1^*(\boldsymbol{\delta})$, with respect to $\boldsymbol{\delta}$. The resulting estimates, $\hat{\boldsymbol{\delta}}$, are plugged in to the phase II component of the pseudo-likelihood, $L_2^*(\boldsymbol{\delta}, \gamma, \boldsymbol{\beta})$, which is then maximized to find estimates for $(\gamma, \boldsymbol{\beta})$. Using the results of Hsieh, Manski, and McFadden (1985), they derive the asymptotic properties of their estimates. Schill et al. (1993) use the same model as Breslow and Cain (1988), but obtain a slightly different estimator. Their estimate is obtained by maximizing the pseudo-likelihood jointly with respect to all parameters. In general, the estimates obtained from either method are not maximum likelihood, as the phase II constraints are not satisfied. However, if stratum-specific intercepts are included in the underlying disease model, then both methods yield the same results, and in fact, these estimates are maximum likelihood since the phase II constraints are satisfied in this case.

The non-parametric maximum likelihood approach offers a solution to the complete two-phase likelihood (2.5) under the phase I and phase II constraints. Following the notation of Breslow and Holubkov (1997a), let $\theta = (\boldsymbol{\beta}, \boldsymbol{\delta}, Q, q)$ and let $\mathbf{h}^T = (h_{00}, h_{01}, h_{10}, \ldots, h_{m_z 0}, h_{m_z 1})$ denote the vector of constraints, where

$$\begin{aligned} h_{0y} &= N_{y\cdot\cdot} - N_{\cdot\cdot\cdot} \sum_{z=1}^{m_z} p_y(z) Q_z = 0 \qquad y = 0, 1 \\ h_{zy} &= n_{y\cdot z} - n_{\cdot\cdot z} \sum_{x=1}^{m_x} p_{yz}(x) q_z(x) = 0 \quad y = 0, 1; z = 1, \ldots, m_z. \end{aligned}$$

Let H denote the corresponding matrix of partial derivatives $\frac{\partial h^T}{\partial \theta}$, and define $\lambda^T = (\Lambda_{00}, \Lambda_{01}, \lambda_{10}, \ldots, \lambda_{m_z 1})$ to be a vector of Lagrange multipliers, where Λ represents

the phase I constraints. The system of equations

$$\begin{aligned} \frac{\partial \log(L(\boldsymbol{\beta}, \boldsymbol{\delta}, \gamma, \mathbf{Q}_z, \mathbf{q}_z(\cdot)))}{\partial \theta} + H\lambda &= 0 \\ \mathbf{h} &= 0 \end{aligned}$$

involves the $p + m_z(m_x + 4) + 3$ original variables $\boldsymbol{\beta}, \boldsymbol{\delta}, \gamma, \mathbf{Q}_z, \mathbf{q}_z(\cdot), \Lambda, \lambda$, and is solved for θ and λ. Equivalently, a more complex set of equations involving only $p + m_z + 1$ variables may be solved. These concentrated Lagrangian equations are given by

$$\begin{aligned} U_1(\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}) &= \sum_{z=1}^{m_z} \sum_{x=1}^{m_x} \left(n_{1xz} - \frac{(n_{1\cdot z} - T_z) n_{0\cdot z} n_{\cdot xz} p_{1z}(x)}{n_{0\cdot z} n_{1\cdot z} - T_z(n_{0\cdot z} - n_{\cdot\cdot z} p_{0z}(x))} \right) x_k = 0, \\ U_{2z}(\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}) &= T_z - \sum_{x=1}^{m_x} \left(n_{yxz} - \frac{(n_{1\cdot z} - T_z) n_{0\cdot z} n_{\cdot xz} p_{1z}(x)}{n_{0\cdot z} n_{1\cdot z} - T_z(n_{0\cdot z} - n_{\cdot\cdot z} p_{0z}(x))} \right) = 0, \quad z = 1, \cdots, m_z, \end{aligned}$$

where $T_z = N_{1\cdot z} - N_{\cdot\cdot z} p_1(z)$, which can be further simplified into a single estimating equation:

$$U(\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}) = \sum_{z=1}^{m_z} T_z \tilde{x}_{z0} + \sum_{z=1}^{m_z} \sum_{k=1}^{m_x} \left(n_{1kz} - \frac{(n_{1\cdot z} - T_z) n_{0\cdot z} n_{\cdot kz} p_{1z}(k)}{n_{0\cdot z} n_{1\cdot z} - T_z(n_{0\cdot z} - n_{\cdot\cdot z} p_{0z}(k))} \right) \tilde{x}_{zk} = 0,$$

where $\tilde{x}_{zk}^T = (x_k^T, e_z^T)$ for e_z the m_z-dimensional vector with 1 in the zth location and 0 elsewhere, and $\tilde{x}_{z0}^T = (0, e_z^T)$. There are several iterative algorithms that have been proposed to obtain the corresponding maximum likelihood estimates. One straightforward algorithm involves setting

$$\delta_z = \log \left(\frac{n_{1\cdot z} N_{0\cdot z}}{n_{0\cdot z} N_{1\cdot z}} \right) \tag{2.6}$$

as the starting value for δ_z, and using these as stratum-specific offsets to fit a logistic regression to the phase II data. This corresponds precisely to the Breslow and Cain pseudo-likelihood estimator. Hence, by choosing (2.6) as the starting values for $\boldsymbol{\delta}$, we are choosing the Breslow and Cain pseudo-likelihood estimator as the starting value for $\boldsymbol{\beta}$. Let F_z denote the sum of the fitted values in the zth stratum from the resulting logistic fit. Then, δ_z is recalculated using

$$\delta_z = \log \left(\frac{F_z}{n_{\cdot\cdot z} - F_z} \right) + \log \left(\frac{N_{0\cdot z} + n_{1\cdot z} - F_z}{N_{1\cdot z} - n_{1\cdot z} + F_z} \right).$$

The logistic regression of the phase II data is re-fit using these new values of δ_z as offsets to obtain the updated $\boldsymbol{\beta}$ vector. This process is iterated until sufficient convergence of the estimated $\boldsymbol{\beta}$ has been achieved (Breslow and Chatterjee, 1999).

In the case of simple random sampling at phase I, $N_{\cdot\cdot z}$ individuals are sampled independently from each stratum z and are classified with respect to outcome. Under this sampling scheme, for a given stratum, $\boldsymbol{N}^{\mathrm{y\cdot z}}$ is Multinomial$(N_{\cdot\cdot z}, \{\mathrm{pr}(Y = y|Z = z)\})$. Hence, the complete likelihood is given by

$$\begin{aligned}\prod_{y=0}^{1} \mathrm{pr}(\boldsymbol{N}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{yxz}}|\boldsymbol{N}^{\cdot\cdot\mathrm{z}}) \quad &\propto \quad \left(\prod_{z=1}^{m_z}\prod_{y=0}^{1} \mathrm{pr}(Y = y|Z = z)^{N_{y\cdot z}}\right) \\ &\times \left(\prod_{y=0}^{1}\prod_{z=1}^{m_z}\prod_{x=1}^{m_x} \mathrm{pr}(X = x|Z = z, Y = y)^{n_{yxz}}\right), \qquad (2.7)\end{aligned}$$

as shown in Scott and Wild (1991).

To determine the form of the likelihood under the logistic model, reparametrization is only required at phase II. Now, letting $\delta_z^\star = \log(\mathrm{pr}(Y = 1|Z = z)/\mathrm{pr}(Y = 0|Z = z))$ and $\mathrm{logit}(p_y^\star(z)) = \delta_z^\star$, the complete likelihood is proportional to

$$\begin{aligned}L(\boldsymbol{\beta}, \boldsymbol{\delta}, \gamma, \mathbf{Q}_z, \mathbf{q}_z(\cdot)) \quad &= \quad \left(\prod_{y=0}^{1}\prod_{z=1}^{m_z} \left(p_y^\star(z)\right)^{N_{y\cdot z}}\right) \times \left(\prod_{y=0}^{1}\prod_{z=1}^{m_z}\prod_{x=1}^{m_x} (p_{yx}(z)q_z(x))^{n_{yxz}}\right) \\ &= \quad L_1(\boldsymbol{\delta}^\star) \times L_2(\boldsymbol{\delta}^\star, \gamma, \boldsymbol{\beta}, \mathbf{q}_z(\cdot)), \qquad (2.8)\end{aligned}$$

where $p_{yx}(z)$ and $q_z(x)$ are defined as before. Despite the redefined δ_z and $p_y(z)$, the preceding methodology still applies. Estimates $\widehat{\delta}_z^\star$ and $\widehat{\boldsymbol{\beta}}^\star$ are related to the parameter estimates $\widehat{\delta}_z$ and $\widehat{\boldsymbol{\beta}}$ assuming case-control sampling via

$$\begin{aligned}\widehat{\delta}_z^\star \quad &= \quad \widehat{\delta}_z - \log\frac{N_{1\cdot\cdot}}{N_{0\cdot\cdot}} + \log\frac{\mathrm{pr}(Y = 1)}{\mathrm{pr}(Y = 0)} \\ \widehat{\beta}_0^\star \quad &= \quad \widehat{\beta}_0 - \log\frac{N_{1\cdot\cdot}}{N_{0\cdot\cdot}} + \log\frac{\mathrm{pr}(Y = 1)}{\mathrm{pr}(Y = 0)},\end{aligned}$$

and $\widehat{\beta}_i = \widehat{\beta}_i^\star$, $i = 1, \ldots, p-1$ (Breslow and Chatterjee, 1999). Scott and Wild (1991) considered simple random sampling at phase I and showed analogous results to those of Schill et al. (1993) for the pseudo-likelihood approach, while the results of Scott and Wild (1997) in the case of simple random sampling at phase I are equivalent to those of Breslow and Holubkov (1997a,b) for the non-parametric maximum likelihood approach .

2.3 Bayesian Outcome-Dependent Methods

The difficulty with deriving the likelihood for case-control studies is that the random variables are the exposure variables, $\boldsymbol{X}$, rather than the outcome variable, Y. In many applications, the vector of exposure variables will be high-dimensional and may consist of continuous, discrete or a mixture of continuous and discrete variables. As such, specifying the joint distribution of exposures given the outcome is difficult (Ghosh, Song, Forster, Mitra, and Mukherjee, 2012). However, this is not such a problem in frequentist analyses since Prentice and Pyke (1979) proved that the same asymptotic inference is obtained whether the retrospective ($p(\boldsymbol{X}|Y)$) or prospective ($p(Y|\boldsymbol{X})$) model is fit. In particular, they showed that the likelihoods arising from the prospective and retrospective models are identical for the odds ratios, provided that the underlying distribution of the covariates is unrestricted (Staicu, 2010; Prentice and Pyke, 1979). The intercept parameter is affected by the reparameterization from the retrospective to the prospective model, however. The intercept is aliased with a term involving the disease prevalences in the population, which are generally not known. Hence, applying the prospective model to case-control data generally yields incorrect estimation (and inference) with respect to the intercept term.

The Bayesian analyses of outcome-dependent data that have previously been proposed tend to be rather complicated since they involve the use of the retrospective likelihood. The retrospective likelihood typically involves many nuisance parameters and involves fitting the models using Markov chain Monte Carlo methods, which requires problem-specific algorithms and computer code (Seaman and Richardson, 2004). Zelen and Parker (1986), Nurminen and Mutanen (1987), Marshall (1988), and Ashby, Hutton, and McGee (1993) examined the single binary exposure case. Müller and Roeder (1997) and Müller, Parmigiani, Schildkraut, and Tardella (1999) consider case-control studies with errors in variables, allowing for any number of continuous exposure variables. Seaman and Richardson (2001) generalize the single binary exposure variable method to deal with any number of categorical or discretized continuous exposure variables. Further, the authors adapted the methods of Müller and Roeder (1997) to the categorical variables scenario, and showed that it is equivalent to the approach of Zelen and Parker (1986).

Seaman and Richardson (2004) showed that the Bayesian analysis of case-control studies may be performed using the prospective likelihood, as in classical frequentist analysis, by proving the equivalence of the two posterior distributions assuming specific prior distributions. Let Y_{ij} denote the number of individuals with disease status $i, i = 0, 1$ and exposure $X = z_j, j = 1, \ldots, J$. Interest lies in estimating the log odds ratio of disease, where this vector of unknown parameters is denoted by δ. Suppose the Y_{ij} are independently distributed as $Y_{ij} \sim \text{Poi}(\lambda_{ij})$, where

$$\log \lambda_{ij} = i \log \alpha + \log \beta_j + i\boldsymbol{\delta}^T z_j,$$

for α the baseline log odds of disease and $\beta_j / \sum_{k=1}^{J} \beta_k$ the probability of exposure z_j. Under independent improper priors for α and β_j, $p(\alpha) \propto \alpha^{-1}$, $p(\beta_j) \propto \beta_j^{a_j - 1}$, an independent proper prior for $\boldsymbol{\delta}$, $p(\boldsymbol{\delta})$, and suitable regularity conditions, the posterior density of $(\omega, \boldsymbol{\delta})$, where $\omega = \log \alpha$, is given by

$$p(\omega, \boldsymbol{\delta}|y) \propto p(\boldsymbol{\delta}) \prod_{j=1}^{J} \frac{(\exp(\omega + \boldsymbol{\delta}^T z_j))^{y_{1j}}}{(1 + \exp(\omega + \boldsymbol{\delta}^T z_j))^{y_{\cdot j} + a_j}}, \tag{2.9}$$

and the posterior density of $(\boldsymbol{\theta}, \boldsymbol{\delta})$, where $\theta_j = \beta_j / \sum_{k=1}^{J} \beta_k$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\delta}|y) \propto p(\boldsymbol{\delta}) \prod_{j=1}^{J} \boldsymbol{\theta}^{a_j - 1} \prod_{i=0}^{1} \left(\frac{\prod_{j=1}^{J} (\theta_j \exp(i\boldsymbol{\delta}^T z_j))^{y_{ij}}}{(\sum_{j=1}^{J} \theta_j \exp(i\boldsymbol{\delta}^T z_j))^{y_{i\cdot}}} \right). \tag{2.10}$$

They go on to prove that the marginal posterior densities of $\boldsymbol{\delta}$ obtained from joint densities (2.9) and (2.10) are the same. In the context of case-control studies, fitting the prospective model with a uniform prior for ω corresponds to (2.9), while (2.10) corresponds to fitting the retrospective model with the prior $\theta \sim \text{Dir}(a_1, \ldots, a_J)$. Thus, the posterior marginal distribution for $\boldsymbol{\delta}$ is the same whether the prospective or retrospective model is fit. The advantage is that the prospective model can more easily be fitted, and involves fewer nuisance parameters than the retrospective model. These methods can also be used for continuous exposure variables by pretending that only exposure values actually observed may be observed (Gustafson et al., 2002; Seaman and Richardson, 2004).

Ghosh, Zhang, and Mukherjee (2006) extend the work of Seaman and Richardson (2004) to stratified case-control studies, where some of the exposure variables could be missing

completely at random. In particular, they prove that a Bayesian analysis of stratified case-control data that uses the prospective likelihood and assumes a uniform prior for the log odds that an individual with baseline exposure is diseased, is exactly equivalent to an analysis that uses the retrospective likelihood and assumes a uniform prior distribution for the exposure probabilities in the control group (Ghosh et al., 2006). Staicu (2010) extends the result of posterior equivalence of retrospective and prospective models established in Seaman and Richardson (2004) for a particular class of priors, to a more general class of priors. In particular, a prior distribution is assumed for the retrospective model parameters, and an induced prior on the prospective model parameters can be determined so that identical marginal posterior densities for the log odds ratio parameter can be obtained by fitting a prospective model as by fitting the retrospective model. Byrne and Dawid (2012) also extend the posterior equivalence of retrospective and prospective models established in Seaman and Richardson (2004) for a particular class of priors, to a more general class of priors and arrive at the same general class of priors as Staicu (2010). Byrne and Dawid (2012) also describe the extension to stratified case-control designs. Ghosh et al. (2012) extend the results of Ghosh et al. (2006) and Seaman and Richardson (2004) to the case of multiple, possibly ordered, disease states that accommodates general link functions (such as the probit link).

2.4 Bayesian Methods for Contingency Tables

Consider data in the form of a contingency table representing the cross-classification of N individuals, where the cell counts in the table follow a multinomial distribution and assume the cell probabilities are modeled according to a hierarchical log-linear model. We use the term hierarchical model to denote a model such that whenever higher order effects are included, all lower order effects are also included in the model. The Bayesian analysis of contingency tables using log-linear models has a long history beginning with early papers of Good (1956) and Lindley (1964), who describe the selection of prior distributions for the log-linear parameters and cell probabilities of multinomial data, respectively. Lindley (1964) used a Dirichlet prior distribution for multinomial probabilities and showed that the posterior distributions of differences of log probabilities (such as the log odds ratio) are approximately joint normal (Theorem 1 of Lindley (1964)). Good (1965) described an

alternative (hierarchical) approach to specifying the parameters of a Dirichlet prior distribution. In particular, the approach treats the Dirichlet parameters as unknown, and specifies a second-stage prior distribution for them. However, this hierarchical approach does not produce a conjugate Dirichlet form for the posterior distribution (Agresti, 2010, Section 11.2.6). An important subclass of the class of hierarchical log-linear models is the class of discrete graphical models (Massam, Liu, and Dobra, 2009). For graphical models, the cell probabilities can be expressed in terms of marginal and conditional probabilities. Independent Dirichlet prior distributions can be assigned to these probabilities which induce independent Dirichlet posterior distributions (Agresti and Hitchcock, 2005). Dawid and Lauritzen (1993) describe a 'hyper Dirichlet' distribution that is conjugate for multinomial sampling. Under this prior, the cliques of a decomposable graphical model have marginal Dirichlet distributions (Madigan et al., 1995). The hyper Dirichlet has been used in several studies. For example, Madigan, York, and Allard (1995) use the hyper Dirichet prior distribution to perform a Bayesian graphical model analysis of data on spina bifida in the United States from 1970–1973. The use of the hyper Dirichlet prior is limited to decomposable models, however. The decomposable models have corresponding graphs which are *chordal*, meaning the graph contains no cycles of length ≥ 4 without a chord (Madigan et al., 1995). Also in their 1995 paper, Madigan et al. used Dirichlet priors to perform a Bayesian graphical model analysis on double sampling data of Down's syndrome in newborns from Norway, where the available data consist of two binary test results and maternal age. In particular, they perform graphical model comparison and construct posterior distributions for parameters of interest by averaging over relevant models. The full details of the analysis are provided in York et al. (1995). Massam et al. (2009) propose a flexible family of conjugate priors (termed the 'Diaconis-Ylvisaker' conjugate priors) for the class of hierarchical log-linear models, which includes the class of discrete graphical models. The Diaconis-Ylvisaker priors they describe are a generalization of the hyper Dirichlet prior to nondecomposable graphical and hierarchical log-linear models.

Although the Dirichlet prior distribution for the cell probabilities of multinomial data is a computationally convenient choice, it does not allow sufficient structure to be imposed on the probabilities (Knuiman and Speed, 1988). An alternative approach is to specify multi-

variate normal prior distributions for the log-linear parameters. Good (1956) used a normal prior distribution for the interaction parameters of a two-way table to obtain smoothed estimates of the probabilities for cells that had small observed frequencies. Similarly, Leonard (1975) assigned independent normal prior distributions for the individual main effect and interaction terms to smooth estimates in sparse two-way tables, and used a uniform hyperprior for the means and inverse chi-squared hyperpriors for the variance parameters. More recently, Knuiman and Speed (1988) proposed assigning structured multivariate normal prior distributions on collections of the log-linear parameters in order to incorporate constraints on main effects and interaction parameters of the log-linear model directly into the prior distribution. This approach was later refined by Dellaportas and Forster (1999). King and Brooks (2001) specify a general multivariate normal prior distribution on the log-linear parameters to induce a multivariate log-normal prior on the corresponding expected cell counts of a contingency table.

Care should be taken when assigning priors to the log-linear parameters since the form for the induced prior of the cell probabilities is not of convenient form and hence it is not straightforward to determine the implications for the probabilities. This is especially true in the case of so-called uninformative prior distributions (for example, normal prior distributions with zero mean and large variances). Assigning uninformative priors for the log-linear parameters can induce informative priors on the cell probabilities, where most of the prior mass may be placed on the boundary values of 0 and 1. Evans and Jang (2011) provide examples to illustrate this phenomenon in logistic regression models. We provide an illustrated example in Figure 2.1. Uninformative priors can be especially problematic when the sample size is small, and hence should not be used in sparse data situations (Galindo-Garre et al., 2004). A common approach in contingency tables with zero counts is to add small counts to each cell of the table and analyse the adjusted data. This procedure is essentially equivalent to prior information being incorporated into the analysis, which leads to a Bayesian approach. Good was among the first to consider such an approach and describes specifying a hierarchical prior on the cell probabilities in Good (1967). The approach assigns a Dirichlet prior to the cell probabilities and a log Cauchy hyperprior distribution is specified for the Dirichlet parameter. Gelman, Jakulin, Pittau, and Su (2008)

also highlight the advantages of a Bayesian approach in sparse data situations and describe a proper, yet vague, prior distribution based on the Student-t distribution. Another prior that can be used in the analysis of contingency table data with zero counts is a normal distribution with moderate variances (rather than using a normal distribution with large variance), for example, as is done by Weiss et al. (1999) in the analysis of death penalty charging data from 1990 to 1994 for Los Angeles County.

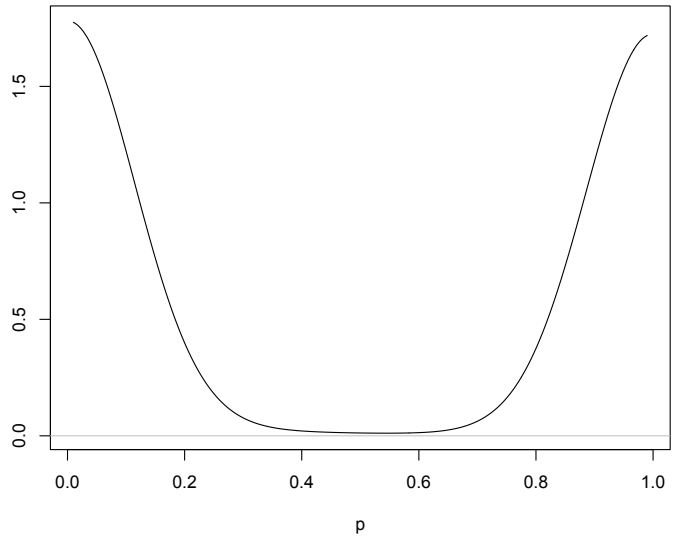


Figure 2.1: Induced prior distribution on p for the logistic regression model $\text{logit}(p) = \beta$, where $\beta \sim N(0, 100)$.

2.5 Bayes Computation

There are several methods available for evaluating the integrals that arise in Bayesian analyses. Typically, we are interested in reporting summaries of the posterior distribution such as the posterior mean or median, or the $100 \times q\%$ quantile, $0 < q < 1$. For a posterior distribution $p(\theta|\mathbf{y})$, the posterior mean is given by

$$\text{E}(\theta|\mathbf{y}) = \int_{\theta} \theta p(\theta|\mathbf{y}) d\theta,$$

while the $100 \times q\%$ quantile, denoted by $\theta(q)$, is the solution to

$$\int_{-\infty}^{\theta(q)} p(\theta|\mathbf{y})d\theta,$$

for θ univariate. We follow the presentation in Wakefield (2012).

2.5.1 Conjugacy

In conjugate situations, the prior and the likelihood are chosen such that the posterior distribution belongs to the same family as the prior distribution. Analytical evaluation of the posterior is more straightforward in this case as the posterior is available in closed-form. Likelihoods belonging to an exponential family have natural conjugate prior distributions (Gelman et al., 2004). Let $\mathbf{T}(\mathbf{Y})$ denote a sufficient statistic of fixed dimension for a particular likelihood $p(\cdot|\boldsymbol{\theta})$. Then, we have

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{t}) \propto p(\mathbf{t}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Consider an exponential family of the form

$$p(y_i|\boldsymbol{\theta}) = f(y_i)g(\theta)\exp(\boldsymbol{\lambda}(\boldsymbol{\theta})^T\mathbf{u}(y_i)),$$

where $\boldsymbol{\lambda}(\boldsymbol{\theta})$ and $\mathbf{u}(y_i)$ generally have the same dimension as $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}(\boldsymbol{\theta})$ is the natural parameter. Define the conjugate prior density as

$$p(\boldsymbol{\theta}) = c(\eta, \boldsymbol{\nu}) \times g(\boldsymbol{\theta})^n \exp(\boldsymbol{\lambda}(\boldsymbol{\theta})^T\boldsymbol{\nu}),$$

for *a priori* specified η and $\boldsymbol{\nu}$. Then, the resulting posterior distribution is given by

$$p(\boldsymbol{\theta}|\mathbf{y}) = c(\eta + n, \boldsymbol{\nu} + \mathbf{t}) \times g(\boldsymbol{\theta})^{\eta+n} \exp(\boldsymbol{\lambda}(\boldsymbol{\theta})^T(\boldsymbol{\nu} + \mathbf{t}(\mathbf{y}))),$$

where $\mathbf{t}(\mathbf{y}) = \sum_{i=1}^n \mathbf{u}(y_i)$, demonstrating conjugacy. Note that η can be viewed as a prior sample size yielding a sufficient statistic $\boldsymbol{\nu}$.

2.5.2 Analytical Approximations

Analytical approximations can be used to evaluate integrals of interest, and these methods include the use of Laplace approximations. Suppose that

$$I = \int_{-\infty}^{\infty} \exp(nh(\theta))d\theta$$

is the integral of interest for a scalar θ. Using a Taylor series expansion about the mode of $h(\cdot)$, denoted by $\widetilde{\theta}$, we can rewrite I as

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \exp\left(n \sum_{k=0}^{\infty} \frac{(\theta - \widetilde{\theta})^k}{k!} h^{(k)}(\widetilde{\theta}) \right) d\theta \\ &\approx \exp(nh(\widetilde{\theta})) \int_{-\infty}^{\infty} \exp\left(\frac{nh^{(2)}}{2}(\widetilde{\theta})(\theta - \widetilde{\theta})^2 \right) d\theta, \end{aligned}$$

ignoring the higher order terms in the Taylor series, using $h^{(1)}(\widetilde{\theta}) = 0$ and where $h^{(k)}(\widetilde{\theta})$ denotes the k-th derivative of h evaluated at $\widetilde{\theta}$. Letting $\widetilde{\nu} = -1/h^{(2)}(\widetilde{\theta})$, we have

$$\widehat{I} = \exp(nh(\widetilde{\theta}))(2\pi\widetilde{\nu}/2)^{1/2}$$

as an estimate for I, which is known as the Laplace approximation. For multivariate $\boldsymbol{\theta}$ of dimension p, the Laplace approximation is given by

$$\widehat{I} = \exp(nh(\widetilde{\boldsymbol{\theta}}))(2\pi/n)^{p/2}|\widetilde{\boldsymbol{\nu}}|^{1/2},$$

where $\widetilde{\boldsymbol{\theta}}$ is the mode of h and $\widetilde{\boldsymbol{\nu}}$ is the $p \times p$ matrix with (i,j)-th element

$$\left. \frac{\partial^2 h}{\partial\theta_i \partial\theta_j} \right|_{\widetilde{\boldsymbol{\theta}}}.$$

In a Bayesian context, these methods can be used to approximate the posterior expectation of a positive function of interest. In practice, the accuracy of the approximation is unknown since it is difficult to perform error assessment. Another difficulty of this method is the evaluation of the derivatives of h.

2.5.3 Quadrature

Numerical integration rules exist for evaluating integrals of the form

$$I = \int f(t)dt$$

using the weighted sum

$$\widehat{I} = \sum_{i=1}^{m} w_i f(t_i),$$

where the points t_i and the weights w_i define the integration rule. Gauss rules are optimal rules in the sense that

$$\sum_{i=1}^{m} w_i p(t_i) = \int w(t)p(t)dt$$

for a polynomial, $p(t)$, of degree $2m-1$. Gauss-Hermite rules correspond to the weight function $w(t) = e^{-t^2}$ and are accurate for evaluating integrals of the form

$$I = \int f(t)e^{-t^2}dt$$

provided $f(t)$ can be well approximated by a polynomial of degree $2m-1$. For a two-dimensional parameter $\boldsymbol{\theta}$, the integral

$$I = \int f(\boldsymbol{\theta})d\boldsymbol{\theta} = \int\int f(\theta_1,\theta_2)d\theta_2 d\theta_1 = \int f^*(\theta_1)d\theta_1,$$

for $f^*(\theta_1) = \int f(\theta_1,\theta_2)d\theta_2$, can be approximated by

$$\widehat{I} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_i u_j f(\theta_{1i},\theta_{2j})$$

since

$$\widehat{I} = \sum_{i=1}^{m_1} w_i \widehat{f}^*(\theta_{1i}),$$

where $\widehat{f}^*(\theta_{1i}) = \sum_{j=1}^{m_2} u_j f(\theta_{1i},\theta_{2j})$. This is known as the Cartesian rule. In general, the points of the rule need to be located and scaled appropriately, where μ denotes the location parameter and σ denotes the scale parameter, however these are unknown in practice. An adaptive Gauss-Hermite rule is one in which μ and σ are estimated followed by estimation of I. These rules can provide accurate integration, however they suffer from the curse of dimensionality as m^p points are required for p parameters, assuming m points for each parameter.

2.5.4 Integrated Nested Laplace Approximations

The integrated nested Laplace approximations (INLA) computational approach combines Laplace approximations and numerical integration (Rue, Martino, and Chopin, 2009). We

consider a model where $\boldsymbol{\theta}_1$ denotes parameters assigned $N_G(\mathbf{0}, \Sigma)$ priors and $\boldsymbol{\theta}_2$ denotes the remainder of the parameters, where the dimension of $\boldsymbol{\theta}_1$ is G, the dimension of $\boldsymbol{\theta}_2$ is V and Σ depends on elements in $\boldsymbol{\theta}_2$. The posterior distribution is given by

$$\begin{aligned}\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) &\propto \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_2)\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &\propto \pi(\boldsymbol{\theta}_2)\,|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{\theta}_1^T\Sigma\boldsymbol{\theta}_2 + \sum_{i=1}^{n}\log p(y_i|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right), \qquad (2.11)\end{aligned}$$

where interest lies in the posterior marginal distributions $\pi(\theta_{1g}|\mathbf{y})$, $g = 1, \ldots, G$, and $\pi(\theta_{2v}|\mathbf{y})$, $v = 1, \ldots, V$. Numerical integration techniques are applied to $\boldsymbol{\theta}_2$, while analytical approximations are used for $\boldsymbol{\theta}_1$ (applied to the exponent of (2.11), conditioning on particular values of $\boldsymbol{\theta}_2$). We can evaluate

$$\pi(\theta_{1g}|\mathbf{y}) = \int \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}) \times \pi(\boldsymbol{\theta}_2|\mathbf{y})d\boldsymbol{\theta}_2$$

using the approximation

$$\begin{aligned}\widetilde{\pi}(\theta_{1g}|\mathbf{y}) &= \int \widetilde{\pi}(\theta_{1g}|\boldsymbol{\theta}_2, \mathbf{y}) \times \widetilde{\pi}(\boldsymbol{\theta}_2|\mathbf{y})d\boldsymbol{\theta}_2 \\ &\approx \sum_{k=1}^{K} \widetilde{\pi}(\theta_{1g}|\boldsymbol{\theta}_2^{(k)}, \mathbf{y}) \times \widetilde{\pi}(\boldsymbol{\theta}_2^{(k)}|\mathbf{y}) \times \Delta_k,\end{aligned}$$

where Laplace approximations are applied to evaluate $\widetilde{\pi}(\theta_{1g}|\boldsymbol{\theta}_2, \mathbf{y})$. The mode of $\widetilde{\pi}(\boldsymbol{\theta}_2|\mathbf{y})$ is obtained and the Hessian is approximated in order to produce the grid of points $\{\boldsymbol{\theta}_2^{(k)}, k = 1, \ldots, K\}$, with associated weights Δ_k, used for the numerical integration.

The INLA approach provides highly accurate results with a low computational cost providing results in seconds or minutes (Rue et al., 2009). The approach gives easy access to posterior marginal distributions, Bayes factors, various prediction measures and DIC. One disadvantage of this approach is that its computational cost is exponential with respect to the number of hyperparameters. This may not be such a problem in practice, however, as the number of hyperparameters is usually small (Rue et al., 2009).

We now address several sampling-based approaches for evaluating integrals.

2.5.5 Importance Sampling Monte Carlo

Suppose we want to evaluate the integral

$$I = \int f(\theta) d\theta.$$

We can trivially rewrite I to achieve an approximately constant function,

$$I = \int \frac{f(\theta)}{g(\theta)} g(\theta) d\theta = E_g \left(\frac{f(\theta)}{g(\theta)} \right),$$

which we can estimate using

$$\widehat{I}_m = \frac{1}{m} \sum_{t=1}^{m} \frac{f(\theta^{(t)})}{g(\theta^{(t)})},$$

where $\theta^{(t)} \sim_{iid} g(\cdot)$, and

$$\sqrt{m}(\widehat{I}_m - I) \rightarrow_d N(0, \text{var}(f/g)),$$

for $\text{var}(f/g) = E((f/g)^2) - I^2$, which can be estimated by $\widehat{\text{var}}(f/g) = \frac{1}{m} \sum_{t=1}^{m} \left(\frac{f(\theta^{(t)})}{g(\theta^{(t)})} \right)^2 - \widehat{I}_m^2$. This approach provides both an estimate of I and a measure of uncertainty, which allows us to construct confidence intervals. The choice of $g(\cdot)$ should be such that it closely mimics f, so that the Monte Carlo estimator has low variance.

Finding a suitable function $g(\cdot)$ is critical to efficient use of importance sampling. In particular, when the support of θ is infinite, g must dominate in the tails, otherwise the variance will be infinite (resulting in an estimate which will not be useful in practice). In addition, $g(\cdot)$ should be computationally inexpensive to sample from.

2.5.6 Direct Sampling using the Rejection Algorithm

Suppose we want to sample from

$$f(x) = \frac{f^*(x)}{\int f^*(x) dx}$$

and we have a proposal distribution $g(\cdot)$ such that

$$M = \sup_x \frac{f^*(x)}{g(x)} < \infty.$$

The rejection algorithm

1. Generate $U \sim \mathrm{U}(0,1)$ and, independently, $X \sim g(\cdot)$.

2. Accept X if

$$U < \frac{f^*(X)}{Mg(X)},$$

otherwise return to 1,

produces accepted points with distribution $f(x)$. They have acceptance probability

$$p_a = \frac{\int f^*(x)dx}{M}.$$

This algorithm is easily implemented to sample from the posterior distribution, however a proper prior distribution is required. The efficiency of the algorithm depends on the correspondence between the likelihood and the prior distribution. As the likelihood becomes increasingly concentrated, prior points are less likely to be accepted and so the algorithm will become less efficient.

2.5.7 Markov Chain Monte Carlo

The goal of Markov Chain Monte Carlo (MCMC) is to create a Markov chain over the parameter space, where the stationary distribution corresponds to the posterior distribution of interest. For a 2-dimensional parameter vector (θ_1, θ_2), suppose we are interested in the following posterior distribution:

$$p(\theta_1, \theta_2|\mathbf{y}) \propto \ell(\theta_1, \theta_2) \times \pi(\theta_1, \theta_2).$$

The Gibbs sampler proceeds as follows. Since direct sampling from the posterior may be difficult, the conditional distributions $p(\theta_1|\theta_2, \mathbf{y})$ and $p(\theta_2|\theta_1, \mathbf{y})$ are used for sampling. For $(\theta_1^{(t)}, \theta_2^{(t)})$ the current point at iteration t, Gibbs sampling proceeds by alternately generating from the distributions

$$\begin{aligned} \theta_1^{(t+1)} &\sim p(\theta_1|\theta_2^{(t)}, \mathbf{y}) \\ \theta_2^{(t+1)} &\sim p(\theta_2|\theta_1^{(t+1)}, \mathbf{y}), \end{aligned}$$

which produces the sample $(\theta_1^{(0)}, \theta_2^{(0)}), (\theta_1^{(1)}, \theta_2^{(1)}), \ldots, (\theta_1^{(t)}, \theta_2^{(t)}), \ldots$. While this algorithm is not sampling directly from the posterior, the limiting stationary distribution is the posterior distribution. The generated points can therefore be used to calculate quantities of interest. This method is easily applied to a more general $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. In this case, the Gibbs sampling algorithm is given by

$$\begin{aligned}
\boldsymbol{\theta}_1^{(t+1)} &\sim p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}, \mathbf{y}) \\
\boldsymbol{\theta}_2^{(t+1)} &\sim p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}, \mathbf{y}) \\
&\vdots \\
\boldsymbol{\theta}_k^{(t+1)} &\sim p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(t+1)}, \mathbf{y}),
\end{aligned}$$

where $\boldsymbol{\theta}_j, j = 1, \ldots, k$, may be univariate or multivariate.

A more general approach is the Metropolis-Hastings algorithm (Tierney, 1994). Let $\boldsymbol{\theta}^{(t)}$ denote the current point at iteration t. To generate the new point, $\boldsymbol{\theta}^{(t+1)}$, the Metropolis-Hastings algorithm proceeds as follows:

- Sample a proposal value, $\boldsymbol{\theta}^*$ from a proposal distribution, $g(\cdot|\boldsymbol{\theta}^{(t)})$.

- Calculate the acceptance probability,

$$r = \frac{p(\boldsymbol{\theta}^*|\mathbf{y})}{p(\boldsymbol{\theta}^{(t)}|\mathbf{y})} \times \frac{g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}.$$

- Set

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{(t)} & \text{otherwise.} \end{cases}$$

If the proposal distribution is symmetric in the sense that $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}) = g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)$, then the acceptance probability, r, simplifies to

$$r = \frac{p(\boldsymbol{\theta}^*|\mathbf{y})}{p(\boldsymbol{\theta}^{(t)}|\mathbf{y})}.$$

An initial number of iterations may be removed as "burn-in", and inference can be based on the remaining samples, say $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$. Expectations of functions $g(\boldsymbol{\theta})$, $\mu = E(g(\boldsymbol{\theta}))$,

with respect to the posterior distribution can be approximated using $\widehat{\mu}_m = \frac{1}{m}\sum_{t=1}^{m} g(\boldsymbol{\theta}^{(t)})$, and an associated 95% credible interval can be obtained by using the 2.5% and 97.5% quantiles of $g(\boldsymbol{\theta}^{(t)})$. It is important to ensure that the chain has converged to its limiting stationary distribution and that m is sufficiently large (depends on the dependence in the chain) in order to obtain reliable Monte Carlo estimates.

Auxiliary data methods for computing posteriors were introduced in an early paper by Tanner and Wong (1987). Using auxiliary data z, the posterior distribution for a parameter of interest β can be written as

$$\begin{aligned} p(\beta|y) &= \int p(\beta, z|y)dz \\ &= \int p(\beta|z, y)p(z|y)dz. \end{aligned}$$

The idea is to introduce augmented 'data' such that it is more straightforward to sample from $p(\beta|z, y)$ than from $p(\beta|y)$. Ideally, one would be able to sample from $p(\beta|z, y)$ directly. A Gibbs sampler can easily be extended to include an extra full conditional for z (Damien, Wakefield, and Walker, 1999). Auxiliary data methods can also easily be incorporated into an MCMC algorithm by including a Gibbs or Metropolis-Hastings step for sampling the auxiliary variables z to the Gibbs or Metropolis-Hastings step for sampling the parameters β (Besag and Green, 1993).

We can instead view the introduction of auxiliary variables as a missing data problem. We present one approach based on count data. Suppose we wish to sample from within an $I \times J$ contingency table, conditioning on known row and column totals. It is possible to sample directly from the full conditional distribution, which requires enumerating all possible tables with the given row and column totals. However, this can be computationally expensive as the summation will generally be over a vast space. Instead, we can use the MCMC approach proposed by Diaconis and Sturmfels (1998). For an $I \times J$ table, we randomly select a pair of rows and a pair of columns, and the table is modified in the four intersecting cells. Following the notation of Wakefield, Haneuse, Dobra, and Teeple (2011), let $n = (n_{11}, n_{12}, n_{21}, n_{22})$ denote the entries in the resulting 2×2 table with known row totals, $(n_{i\cdot})_{1\leq i\leq 2}$, and column totals, $(n_{\cdot j})_{1\leq j\leq 2}$, and let $\mathcal{S}$ denote the set of tables with the

same row and column totals,

$$\mathcal{S} = \{n' = (n'_{ij})_{1 \leq i,j \leq 2} : n'_{1\cdot} = n_{1\cdot}, n'_{2\cdot} = n_{2\cdot}, n'_{\cdot 1} = n_{\cdot 1}, n'_{\cdot 2} = n_{\cdot 2}\}.$$

Consider all possible pairwise differences of tables in $\mathcal{S}$, denoted by $\mathcal{M} = \{n' - n'' : n', n'' \in \mathcal{S}\}$, where the elements of $\mathcal{M}$ are called moves. We want to define moves so that all tables in $\mathcal{S}$ can be visited. Any two tables, n' and n'', in $\mathcal{S}$ can be connected through the move $n'' - n' \in \mathcal{M}$. However, not all moves in $\mathcal{M}$ are required to connect any two tables in $\mathcal{S}$. By removing moves from $\mathcal{M}$ such that the remaining moves still connect $\mathcal{S}$, we obtain a Markov basis. A Markov basis for $\mathcal{S}$ allows construction of a Markov chain to move around $\mathcal{S}$.

Diaconis and Sturmfels (1998) prove that a Markov basis associated with an $I \times J$ contingency table is the simplest Markov basis, that is, one containing two entries equal to one, two entries equal to minus one and the remaining entries equal to zero. For a 2×2 table, let $g = (1, -1, -1, 1)$. Then, $\mathcal{M} = \{g, -g\}$ is a Markov basis for $\mathcal{S}$.

Let n' denote the current table. A Markov chain proceeds by drawing a move g' from the uniform distribution on $\mathcal{M}$ and the proposed table is $n^* = n' + g'$. If n^* contains negative entries, we stay at the current table, n'. If not, then $n^* \in \mathcal{S}$ and we accept n^* with probability

$$\min\left(1, \frac{\text{pr}(n^*)}{\text{pr}(n')}\right).$$

This method provides reliable results in a reasonable amount of time. The Markov basis method analysis takes a fraction of the time to run compared to the full enumeration method (as found in Wakefield et al. (2011) as well as for the examples considered in this thesis). The difference in speed of the algorithms is especially apparent for large row and column totals, since the full enumeration method requires summation over a vast space.

2.6 Bayesian Spatial Data Methods

Suppose we wish to study the risk of health outcomes in a particular study region consisting of n non-overlapping areas. Let p_i denote the risk in area i, $i = 1, \ldots, n$. For Y_i the number

of cases in area i, we can use the model

$$Y_i|p_i \sim \text{Bin}(N_i, p_i),$$

where N_i is the number of individuals at risk in area i. The maximum likelihood estimates for the area-specific risks, $\widehat{p}_i = Y_i/N_i$, can be highly unstable in the case of sparse data (Elliott et al., 2000, Section 7.2). Hence, global and/or local smoothing of the risks can be used to achieve more reliable estimates using hierarchical models. The smoothing can be carried out by incorporating spatial and non-spatial random effects in models that specify how we believe the p_i vary across the study region. However, we begin by describing models that do not use spatial information.

2.6.1 Non-spatial Models

We begin by describing the Beta-Binomial model, which is given by

$$\begin{aligned} Y_i|p_i &\sim_{\text{ind}} \text{Bin}\,(N_i, p_i) \\ p_i|\alpha,\beta &\sim_{\text{iid}} \text{Beta}(\alpha,\beta), \qquad i = 1,\ldots,n. \end{aligned}$$

The marginal distribution of $Y_i|\alpha,\beta$ is Beta-Binomial with mean and variance given by

$$\begin{aligned} \text{E}[Y_i|\alpha,\beta] &= \frac{N_i\alpha}{\alpha+\beta} \\ &= N_i p \\ \text{var}(Y_i|\alpha,\beta) &= \frac{N_i\alpha\beta(\alpha+\beta+N_i)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= N_i p(1-p)\left(\frac{\alpha+\beta+N_i}{\alpha+\beta+1}\right), \end{aligned}$$

ie. overdispersion of the data relative to a pure Binomial distribution (Williams, 1975). To fit the model, we may use a fully Bayesian approach, where prior distributions must be specified on all parameters. Alternatively, an empirical Bayes approach can be used where estimates $\widehat{\alpha}$ and $\widehat{\beta}$ are first obtained (through maximum likelihood estimation, for example) and inference is then based on the posterior mean resulting from the plug-in prior $p_i|\widehat{\alpha},\widehat{\beta}$. The posterior mean is then given by a weighted combination of the prior mean, $\widehat{\alpha}/(\widehat{\alpha}+\widehat{\beta})$,

and the sample mean, Y_i/N_i:

$$\begin{aligned}\widehat{p}_i &= \frac{Y_i + \widehat{\alpha}}{N_i + \widehat{\alpha} + \widehat{\beta}} \\ &= w_i \times \frac{Y_i}{N_i} + (1 - w_i) \times \left(\frac{\widehat{\alpha}}{\widehat{\alpha} + \widehat{\beta}}\right),\end{aligned}$$

where $w_i = \frac{N_i}{N_i+\alpha+\beta}$ (Wakefield, 2007). Unfortunately, this approach cannot be easily extended to incorporate spatial dependence.

We may also specify a Binomial logistic-normal model, which is given by

$$\begin{aligned} Y_i|p_i &\sim_{\text{ind}} \text{Bin}\,(N_i, p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{x}_i^T\boldsymbol{\beta} + V_i \\ V_i|\sigma_v^2 &\sim_{\text{iid}} N(0, \sigma_v^2), \end{aligned} \tag{2.12}$$

where $\boldsymbol{x}_i$ is a vector of area-level covariates, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)$ is a vector of fixed effects, and $V_i, i = 1, \ldots, n$, are area-specific random effects that capture the residual log-odds ratios in area i and do not display a spatial pattern. One benefit of using a Binomial logistic-normal model is that (2.12) naturally extends to allow for the addition of spatially structured random effects (Elliott et al., 2000, Section 7.2;Wakefield, 2007). Empirical Bayes is not a convenient approach for this model (Wakefield, 2007). Hence, to fit this model, we use a fully Bayesian approach and specify prior distributions on all parameters.

2.6.2 Prior Choices for Non-Spatial Models

We need to specify prior distributions for $\boldsymbol{\beta}$ and σ_v^2. An improper prior, $\pi(\boldsymbol{\beta}) \propto 1$, is often used, however this can result in an improper posterior distribution. An 'uninformative prior', which corresponds to placing a normal distribution with mean zero and large variance to each $\beta_j, j = 0, \ldots, J$, can also be used. More informative prior distributions will be beneficial in situations where there are many covariates, or when there is high dependence among the elements of x (Wakefield, 2007). In such cases, we can assign lognormal prior distributions to $\exp(\beta_j)$. The lognormal distribution is a convenient choice as two quantiles of the distribution can be specified and we can use the following relationships to solve for

μ and σ, the two parameters of the lognormal distribution:

$$\begin{aligned} \mu &= \log(\theta_1)\left(\frac{z_{q_2}}{z_{q_2}-z_{q_1}}\right) - \log(\theta_2)\left(\frac{z_{q_1}}{z_{q_2}-z_{q_1}}\right) \\ \sigma &= \frac{\log(\theta_1)-\log(\theta_2)}{z_{q_1}-z_{q_2}} \end{aligned} \tag{2.13}$$

where θ_1 and θ_2 are the q_1 and q_2 quantiles of the lognormal distribution, respectively, and z_q is the qth quantile of a standard normal distribution (Wakefield, 2007).

We focus on assigning priors to the precision variable $\tau_v = \sigma_v^{-2}$ for conjugacy reasons. A Gamma(a, b) distribution is a convenient choice as it yields the marginal distribution for V_i in closed form. (An equivalent prior for σ_v^2 is the inverse Gamma distribution.). In particular, with this choice of prior, the marginal distribution of V_i is $t_{2a}(0, b/a)$, the Student's t-distribution with $2a$ degrees of freedom, location zero, and scale b/a. As outlined in Wakefield (2007), a and b can be determined by specifying the degrees of freedom, d, and the range, $\exp(\pm R)$, within which the residual relative risks lie with probability q and then using the relationship

$$\pm t^d_{q/2}\sqrt{b/a} = \pm R,$$

where t^d_q is the qth quantile of a Student's t-distribution with d degrees of freedom. From this, we obtain

$$\begin{aligned} a &= d/2 \\ b &= \frac{R^2 d}{2(t^d_{q/2})^2}. \end{aligned} \tag{2.14}$$

As noted in Wakefield (2007), the prior distribution that is chosen must allow all reasonable levels of variability in the residual relative risks, and particularly, must not exclude small values. The prior distributions $\tau_v \sim$ Gamma$(1, 0.0260)$ and $\tau_v \sim$ Gamma$(0.5, 0.0005)$ will often be suitable in a disease mapping context, though priors should be considered on a case-by-case basis. The Gamma$(1, 0.0260)$ distribution, for example, arises by assuming *a priori* the residual log-relative risks follow a Student's t-distribution with 2 degrees of freedom, where 95% of these risks fall in the interval (0.5, 2.0). Using this prior, we obtain a 0.01 quantile of 0.075 for σ_v, while we obtain a 0.01 quantile of 0.012 for σ_v using a Gamma$(0.5, 0.0005)$ prior distribution on τ_v. The Gamma$(0.001, 0.001)$ prior distribution

that had previously been suggested, however, results in a 0.01 quantile of 6.42 for σ_v and, hence, should be avoided (Kelsall and Wakefield, 1999).

2.6.3 Spatial Models

Consider the model

$$\begin{aligned} Y_i|p_i &\sim_{\text{ind}} \text{Bin}\,(N_i, p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{x}_i^T\boldsymbol{\beta} + g(\boldsymbol{S}_i, \boldsymbol{\gamma}) + U_i + V_i \\ V_i|\sigma_v^2 &\sim_{\text{iid}} N(0, \sigma_v^2), \end{aligned} \tag{2.15}$$

where $\boldsymbol{x}_i$ is a vector of area-level covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ is a vector of fixed effects, $\boldsymbol{S}_i = (S_{i1}, S_{i2})$ denotes spatial location (centroid of area i), $g(\boldsymbol{S}_i, \boldsymbol{\gamma})$ is a regression model that may be included to capture large-scale spatial trend, V_i denote non-spatial random effects and U_i denote spatially structured area-specific random effects. The goal is to model $\boldsymbol{U} = (U_1, \ldots, U_n)$ allowing for dependence between U_i and U_j for $i \neq j$. Since we expect areas that are close together to have more similar residual odds, or log odds, of disease than those that are not close, we want our model to exploit this information to obtain more reliable relative risk estimates in each area. One approach would be to specify the joint distribution of $\boldsymbol{U}$. Alternatively, we may specify the univariate conditional distributions $U_i|U_j, j \neq i, i = 1, \ldots, n$ (Elliott et al., 2000, Section 7.2;Wakefield, 2007).

Let $\boldsymbol{U} \sim N_n(\boldsymbol{0}_n, \sigma_u^2\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite symmetric matrix. Let $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. Then,

$$U_i|U_j = u_j, j \neq i \sim N\left(\sum_{j=1}^{n} W_{ij}u_j, \sigma_u^2 D_{ii}\right), \tag{2.16}$$

where $W_{ii} = 0$, $W_{ij} = -Q_{ij}/Q_{ii}$, $(i \neq j)$ and $D_{ii} = Q_{ii}^{-1}$ (Besag and Kooperberg, 1995; Elliott et al., 2000, Section 7.2). Since $\boldsymbol{Q}$ is symmetric, we have that

$$W_{ij}D_{jj} = W_{ji}D_{ii}.$$

Hence, in the specification of the joint distribution of $\boldsymbol{U}$, we either need to directly specify the elements of the covariance matrix $\boldsymbol{\Sigma}$, or we need to specify the weights W_{ij} and D_{ij} in

(2.16) in the conditional approach. Let $\boldsymbol{D}$ be an $n \times n$ diagonal matrix with elements D_{ii} and $\boldsymbol{W}$ a matrix of weights with elements $W_{ij}, i, j = 1, \ldots, n$. Then, we can relate the joint and conditional approaches using $\boldsymbol{Q} = \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{W})$ (Elliott et al., 2000, Section 7.2).

Conditional Models

A common model to assign to the spatial random effects is the intrinsic conditional autoregressive (ICAR) distribution considered by Besag, York, and Mollié (1991). This model specifies a distribution for U_i conditional on knowing the spatial effects for its neighbouring areas, U_j:

$$U_i | U_j, j \in \delta_i \sim N\left(\overline{U}_i, \frac{\omega_u^2}{m_i}\right), \tag{2.17}$$

where δ_i is the set of neighbours of area i, m_i is the number of neighbours of area i, $\overline{U}_i$ is the mean of the spatial effects of the neighbours, and ω_u^2 is the conditional variance which determines the amount of spatial variation. In this model, we have $D_{ii} = m_i^{-1}$ and $W_{ij} = m_i^{-1}$ for neighbouring areas and $W_{ij} = 0$ otherwise. Under this specification, $\boldsymbol{Q}$ is singular as each row sums to 0, meaning $\boldsymbol{Q}$ has rank $n - 1$. This model is a limiting form of (2.16).

The joint specification for $\boldsymbol{U}$ corresponding to (2.17) is given by:

$$\begin{aligned} p(\boldsymbol{U}|\tau_u) &\propto \tau_u^{(n-1)/2} \exp\left(-\frac{\tau_u}{2}\boldsymbol{U}^T\boldsymbol{K}\boldsymbol{U}\right) \\ &= \tau_u^{(n-1)/2} \exp\left(-\frac{\tau_u}{2}\sum_{i\sim j}(U_i - U_j)^2\right), \end{aligned} \tag{2.18}$$

where $\tau_u = \omega_u^{-2}$, $i \sim j$ denotes all pairs of neighbouring areas i and j, and $\boldsymbol{K}$ is an $n \times n$ matrix representing the neighbourhood structure of the areas in which the (i, j)th off-diagonal element is -1 if areas i and j are neighbours and 0 otherwise, and the ith diagonal element is m_i (Besag et al., 1991). This prior distribution is improper as the rank of $\boldsymbol{Q}$ is $n-1$. In addition, the "density" only contains differences of U_i and so if the same constant is added to each U_i, the exact same density is obtained. Hence, the overall level of the model is unspecified (Knorr-Held and Rue, 2002). Since the U_i are not centered, if an intercept term is included in (2.15), we must impose an additional constraint on (2.18) for identifiability.

We can do so by placing an improper flat prior on the intercept and imposing a sum-to-zero constraint on $\boldsymbol{U}$ by centering the U_i about their mean (Besag and Kooperberg, 1995).

With this conditional model, a rule for determining neighbours must be specified and there are many possible definitions that could be used. One possibility is defining neighbours based on contiguity, meaning that we define two areas as neighbours if they share a common border. This definition has been used by several authors such as Besag et al. (1991), Clayton and Kaldor (1987), Richardson et al. (1995), Bernardinelli et al. (1997), and Waller et al. (1996). This is a reasonable definition to use if all the areas are of similar size and arranged in a regular pattern, however it is less appealing otherwise (Elliott et al., 2000, Section 7.2;Wakefield, 2007). Another possibility is to allow the neighbourhood structure to depend on the distance between area centroids. Using this definition, some others have suggested that the distance within which areas are considered neighbours can be determined using an exploratory analysis (Cressie and Chan, 1989).

Rather than defining $\boldsymbol{K}$ using a neighbourhood scheme, we can use distance-based weights with weights decreasing with increasing distances between area centroids (Elliott et al., 2000, Section 7.2). For example, Best et al. (1999) use distance-based weights defined as $w_{ij} = w_{ji} = \exp(-d_{ij}/\delta)$, where d_{ij} is the distance between the geographic centroids of areas i and j, and δ is a constant chosen to achieve particular weights for areas a given distance apart. In their example, Best et al. (1999) use $\delta = 33$ to give a relative weight of 1% ($w_{ij} = 0.01$) for area centroids that lie 152 km apart, which corresponds to the mean distance between area centroids for their study region.

An alternative choice for modeling the conditional distribution of U_i is the Laplace distribution:

$$U_i|U_j, j \in \delta_i \propto \tau_u \exp\left(-\tau_u \sum_{j\in\delta_i} |U_i - U_j|\right),$$

where δ_i is the set of neighbors of area i (Besag et al., 1991; Best et al., 1999). This distribution is centered at the median of the spatial effects of the neighbours of area i, rather than the mean as in (2.17). As discussed in Besag et al. (1991), this distribution may be a more appropriate choice when we expect discontinuities in disease rates between areas

(perhaps due to differing environmental policies, land-use boundaries, or physical features such as lakes or mountains (Best et al., 1999)).

Joint Models

We use a multivariate Normal distribution to model the joint distribution of $\boldsymbol{U}$:

$$\boldsymbol{U} \sim N(\mathbf{0}_n, \sigma_u^2 \boldsymbol{\Sigma}), \tag{2.19}$$

where $\mathbf{0}_n$ denotes a vector of zeros of length n and $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite correlation matrix with off-diagonal elements, Σ_{ij}, describing the correlation between U_i and U_j.

There are several structured forms for $\boldsymbol{\Sigma}$ that may be assumed. One common choice is to assume that the dependence is a function of the distance between area centroids. Letting d_{ij} represent the distance between area centroids i and j, $\Sigma_{ij} = f(d_{ij}, \phi)$, where $\phi > 0$ determines the extent of the correlation (Elliott et al., 2000, Section 7.2). One possible choice for $f(d_{ij}, \boldsymbol{\phi})$ is

$$f(d_{ij}, \phi) = \exp(-\phi d_{ij}).$$

Note that this model assumes the correlation is the same in all spatial directions. A more general choice for $f(d_{ij}, \phi)$ is the powered exponential family:

$$f(d_{ij}, \phi) = \exp\left(-\left(\phi d_{ij}\right)^{\kappa}\right),$$

where $\kappa \in (0, 2]$. There are some theoretical difficulties associated with using this family, such as a nearly singular covariance matrix when $\kappa = 2$, and abrupt changes in continuity properties (see comments by Handcock, Kent, Stein and Webster to the discussion of Diggle et al. (1998)). As suggested in the discussion of Diggle et al. (1998) and Handock and Stein (1993), the Matérn class may be a more preferable choice of family since a broader range of continuity behavior can be obtained (Matérn, 1986). Using this family, we have

$$f(d_{ij}, \phi) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\phi' d_{ij}\right)^{\kappa} B\left(\phi' d_{ij}\right),$$

where $\phi' = 2\phi\sqrt{\kappa}$ and $B = (\cdot)$ is the modified Bessel function of order κ.

By including a non-spatial random effect along with the ICAR spatial random effect in (2.15), we can determine the strength of the spatial dependence by examining the relative

contributions of U_i and V_i to the posterior distribution (Besag et al., 1991). If the majority of the variability was non-spatial and we failed to include a non-spatial random effect, we may erroneously conclude that spatial dependence was present (Leroux et al., 1999; Wakefield, 2007). Hence, it is strongly recommended that a non-spatial random effects term be included along with the spatial random effect in (2.15) when modeling the U_i using an ICAR model. It is not necessary to include a separate non-spatial random effect term when using the joint model (2.19), however, since the strength of spatial dependence is captured by ϕ, where $\phi \to 0$ implies no spatial dependence.

2.6.4 Prior Choices for Spatial Models

It is important to note that ω_u^2 is a *conditional* variance whose magnitude determines the amount of spatial variation (Wakefield, 2007). The variance parameters σ_v^2 and ω_u^2 are not directly comparable as they are on different scales. The amount of spatial variation can be estimated using the empirical variance, $\text{var}(U_i)$, however. Because of the conditional interpretation of ω_u^2, care must be taken when a specifying prior distribution.

We may specify a Gamma(a, b) prior distribution for $\tau_u = \omega_u^{-2}$, as described in Section 2.6.2 (Elliott et al., 2000, Section 7.2). However, different candidate prior distributions should be examined using simulations to evaluate whether the realizations conform to our prior expectations. That is, realizations from the candidate prior distributions must exhibit the required amount of smoothing (Fong, Rue, and Wakefield, 2010). Since (2.18) is not a proper density, we cannot directly simulate from this prior. However, as described in Fong et al. (2010), samples can easily be generated using Algorithm 3.1 of Rue and Held (2005). This algorithm is reproduced as Algorithm 1 below using our notation, where δ_j represents the j^{th} eigenvalue and $\boldsymbol{e}_j$ represents the j^{th} eigenvector. The marginal variances are available as the diagonal elements of the matrix $\sum_j \delta_j^{-1} \boldsymbol{e}_j \boldsymbol{e}_j^T$.

In the joint model, a Gamma(a, b) prior distribution for σ_u^{-2} can be used (Elliott et al., 2000, Section 7.2). In the case where Σ is the identity matrix in (2.19), we may prefer to specify priors for the total precision, $\tau_T = (\sigma_u^2 + \sigma_v^2)^{-1}$, and the proportion of the total residual variation that is spatial, $p = \sigma_u^2/(\sigma_u^2 + \sigma_v^2)$ (Wakefield, 2007). In this case, the

Algorithm 1 Sampling from an improper GMRF with mean zero.

for $j = 2$ to m_z **do**

$w_j \sim N(0, \delta_j^{-1})$

end for

return $\boldsymbol{U} = w_{k+1}\boldsymbol{e}_{k+1} + w_{k+2}\boldsymbol{e}_{k+2} + \cdots + w_{m_z}\boldsymbol{e}_{m_z}$

residual relative risk, e^{U+V}, is lognormal with parameters 0 and $\sigma_u^2+\sigma_v^2$. Hence, by specifying priors to τ_T and p, the amount of total residual variability can be controlled (Wakefield, 2007). A Gamma(a, b) prior distribution can be specified for τ_T, while a Beta(c, d) can be specified on p. Then, (τ_T, p) can be transformed to (σ_v^2, σ_u^2) using

$$\begin{aligned} \sigma_v^2 &= (1-p)\tau_T^{-1} = (1-p)(\sigma_v^2 + \sigma_u^2) \\ \sigma_u^2 &= p\tau_T^{-1} = p(\sigma_v^2 + \sigma_u^2). \end{aligned}$$

In the joint model, we also need to specify a prior for ϕ. As described in Wakefield (2007), a prior distribution can be specified for the distance at which the correlations fall to a half:

$$d_{1/2} = -\log 2/\phi.$$

We assign a lognormal prior distribution to $d_{1/2}$ using the method described by (2.13).

2.7 Sample Survey Methodology

In survey samples, a population is specified whose data values are unknown but regarded as fixed. The observed sample is random, however, since it depends on the random selection of individuals from this fixed population (Lumley, 2010, Section 1.1). The analysis of survey samples is generally design-based, where the goal is to estimate features of the fixed population. As described in Lumley (2010), the procedure for taking samples from a population (the sampling method) must have the following properties:

- Every individual in the population must have a non-zero probability of ending up in the sample (denoted π_i for individual i);

- The probability π_i must be known for every individual who does end up in the sample;
- Every pair of individuals in the sample must have a non-zero probability of both ending up in the sample (denoted π_{ij} for the pair of individuals (i, j));
- The probability π_{ij} must be known for every pair that does end up in the sample,

where the first two properties are needed to get valid population estimates and the last two are needed to estimate the accuracy of the estimates. It is helpful to think of an individual sampled with probability π_i representing $1/\pi_i$ individuals in the population, where $1/\pi_i$ is the sampling weight.

For some variable of interest X, let X_i denote the measurement of X on individual i. Then, we denote the weighted observation $\check{X}_i$ by

$$\check{X}_i = \frac{1}{\pi_i} X_i.$$

Then, for a sample of size n, the Horvitz-Thompson estimator, $\hat{T}_X$, for the population total, T_X, of X is

$$\hat{T}_X = \sum_{i=1}^{n} \frac{1}{\pi_i} X_i = \sum_{i=1}^{n} \check{X}_i, \tag{2.20}$$

with variance estimate

$$\widehat{\text{var}}\left[\hat{T}_X\right] = \sum_{i,j} \left(\frac{X_i X_j}{\pi_{ij}} - \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} \right)$$

(Horvitz and Thompson, 1952; Lumley, 2010, Section 1.1).

2.7.1 Simple Random Sampling

For a simple random sample without replacement of size n from a population of size N, each possible subset of n individuals is equally likely to be sampled. Hence, all the sampling weights are equal to N/n (Korn and Graubard, 1999, Section 2.2; Lumley, 2010, Section 2.1). From (2.20), we have that the Horvtiz-Thompson estimator of the population total is

$$\hat{T}_X = \sum_{i=1}^{n} \check{X}_i = \frac{N}{n} \sum_{i=1}^{n} X_i,$$

with variance given by

$$\text{var}\left(\hat{T}_X\right) \quad = \quad \left(1 - \frac{n}{N}\right) \times N^2 \times \frac{S^2}{n},$$

where $1 - n/N$ is the finite population correction factor and S^2 is the variance of X, which can be estimated using the sample variance. For samples in which n is much smaller than N, the finite population correction factor can often be ignored (Korn and Graubard, 1999, Section 2.2; Lumley, 2010, Section 2.1).

The population mean can also be estimated by dividing the estimated total by the population size N:

$$\widehat{\mu}_X = \frac{1}{N} \sum_{i=1}^{n} \check{X}_i = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

which is simply the sample average, with estimated variance

$$\widehat{\text{var}}\left(\widehat{\mu}_X\right) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n},$$

where s^2 is the sample variance of X (Korn and Graubard, 1999, Section 2.2; Lumley, 2010, Section 2.1).

2.7.2 Stratified Sampling

Compared to simple random sampling, the precision of estimates can be increased using stratified sampling, where the population is divided into disjoint strata and simple random samples are drawn independently from each strata. Hence, a prespecified number of observations from each stratum end up in the sample, resulting in a less variable sample, and thus more precise estimates (Lumley, 2010, Section 2.2). To employ stratified sampling, the stratum membership must be known for every individual in the population.

Suppose the population is divided into L strata, where $N_l, l = 1, \ldots, L$, is the known population size of stratum l, and a sample of size n_l is taken from the lth stratum. Then, the Horvitz-Thompson estimator of the population total is

$$\hat{T}_X = \sum_{l=1}^{L} \frac{N_l}{n_l} \sum_{i=1}^{n_l} X_i,$$

which is simply the sum of the estimated totals in each stratum. Its variance is given by

$$\text{var}\left(\hat{T}_X\right) = \sum_{l=1}^{L} N_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{S_l^2}{n_l},$$

where S_l^2 is the variance of X in stratum l and can be estimated using the sample variance of X in stratum l (Lumley, 2010, Section 2.2).

The population mean is estimated by dividing the estimated population total by the population size, N:

$$\widehat{\mu}_X = \frac{1}{N} \sum_{l=1}^{L} \frac{N_l}{n_l} \sum_{i=1}^{n_l} X_i, \tag{2.21}$$

with estimated variance

$$\widehat{\text{var}}\left(\widehat{\mu}_X\right) = \frac{1}{N^2} \sum_{l=1}^{L} N_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{s_l^2}{n_l},$$

where s_l^2 is the sample variance in stratum l (Korn and Graubard, 1999, Section 2.2; Lumley, 2010, Section 2.2).

Note that the stratified mean given in (2.21) is an example of a weighted mean

$$\widehat{\mu}_X = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}, \tag{2.22}$$

where $w_i = 1/\pi_i$ is the sample weight associated with the ith individual (Korn and Graubard, 1999, Secion 2.2). A general principal of sampling theory is that one obtains unbiased or approximately unbiased estimators of parameters by using weighted estimators, with weights the inverse of the inclusion probabilities, π_i (Korn and Graubard, 1999, Section 2.2).

2.7.3 Probability-Proportional-to-Size Sampling

In probability-proportion-to-size sampling, some information is known for each individual in the population, and this information is taken to be a continuous "size" variable, Z. The inclusion probabilities are then taken to be proportional to Z:

$$\pi_i = \frac{nZ_i}{\sum_{j=1}^{N} Z_j}.$$

Under this sampling scheme, there are two estimates for the population mean. One is given by the weighted mean in (2.22) with weights $w_i = 1/\pi_i$, and the other is the Horvitz-Thompson estimator given by $\frac{1}{N}\sum_{i=1}^{n} X_i/\pi_i$, which is not equal to $\frac{1}{\sum_{i=1}^{n} 1/\pi_i}\sum_{i=1}^{n} X_i/\pi_i$.

To calculate the variance of the Horvitz-Thompson estimator, both the inclusion probabilities, π_i, and the joint inclusion probabilities, π_{ij}, need to be known. Then, the variance is

$$\text{var}\left(\widehat{\mu}_X\right) = \frac{1}{N^2}\sum_{i=1}^{N} X_i^2 \frac{1-\pi_i}{\pi_i} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} X_i X_j \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j},$$

which can be estimated using

$$\text{var}\left(\widehat{\mu}_X\right) = \frac{1}{N^2}\sum_{i=1}^{n} X_i^2 \frac{1-\pi_i}{\pi_i^2} + 2\sum_{i=1}^{n}\sum_{j>i}^{n} X_i X_j \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}.$$

Gains in efficiency can be obtained if the size variable is correlated with the variable of interest (Korn and Graubard, 1999, Section 2.2). As described in Cochran (1977, Section 9A.4,p.255), using previously obtained values of X_i as the measures of size may be a good strategy.

2.7.4 Multistage Sampling

In multistage sampling, groups of individuals, called clusters, are sampled at the first stage and are called "Primary Sampling Units" (PSUs). Then, at the second and subsequent stages, subsamples are drawn from within the PSUs. It is important to note that every individual in the population must be in only one PSU, and the subsampling probabilities for any given PSU must not depend on which other PSUs were sampled (Lumley, 2010, Section 3.1). In this case, the inclusion probabilities are the products of the PSU-level inclusion probabilities from the first stage of sampling times the conditional inclusion probabilities from the later stages of sampling (Korn and Graubard, 1999, Section 2.3).

Simple Random Sampling of PSUs

Suppose the population consists of K PSUs of sizes $N_1, \ldots, N_K$, and a simple random sample of k PSUs is taken. In the second stage of sampling, a simple random sample of $n_i = \tau N_i$ is drawn from the ith sampled PSU, where the proportion sampled, τ, is the

same regardless of its size. The inclusion probability is the same for all individuals in the population, $\pi_i = \tau k/K$ (Korn and Graubard, 1999, Section 2.3). Then, assuming the finite population correction factors are negligible, the population mean can be estimated using a weighted estimator:

$$\widehat{\mu}_X = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}X_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{k} n_i},$$

and a variance estimator is

$$\widehat{\text{var}}\,(\widehat{\mu}_X) = \frac{1}{\bar{n}^2 k(k-1)}\sum_{i=1}^{k} n_i^2(\bar{X}_i - \widehat{\mu}_X)^2,$$

where $\bar{n} = \sum_{i=1}^{k} n_i/k$ and $\bar{X}_i$ is the mean of the n_i sampled observations from the ith PSU.

Stratified Probability-Proportional-to-Size Sampling of PSUs

Suppose that the population is divided into L strata containing K_l PSUs in the lth stratum. From the lth stratum, k_l PSUs are sampled with probability-proportional-to-size sampling for some PSU-level size variable with inclusion probabilities denoted by $\pi_{l1}, \pi_{l2}, \ldots, \pi_{lk_l}$. The additional sampling stages within PSUs can be quite complicated, involving multiple additional stages and further stratification, for example (Korn and Graubard, 1999, Section 2.3). Let n_{li} denote the number of individuals sampled from the ith sampled PSU from stratum l and w_{lij} denote the sample weights for $j = 1, \ldots, n_{li}$ after the sampling is complete. Then, the estimator for the population mean is the weighted estimator

$$\widehat{\mu}_X = \frac{\sum_{l=1}^{L}\sum_{i=1}^{k_l}\sum_{j=1}^{n_{li}} w_{lij}X_{lij}}{\sum_{l=1}^{L}\sum_{i=1}^{k_l}\sum_{j=1}^{n_{li}} w_{lij}},$$

with estimated variance given by

$$\widehat{\text{var}}\,(\widehat{\mu}_X) = \frac{\sum_{l=1}^{L}\frac{k_l}{k_l-1}\sum_{i=1}^{k_l}\left(W_{li}(\bar{X}_{li} - \widehat{\mu}_X) - \frac{1}{k_l}\sum_{s=1}^{k_l} W_{ls}(\bar{X}_{ls} - \widehat{\mu}_X)\right)^2}{\left(\sum_{l=1}^{L}\sum_{i=1}^{k_l} W_{li}\right)^2},$$

where $\bar{X}_{li}$ is the weighted mean of the observations in the ith sampled PSU in the lth stratum, and W_{li} is the sum of the sample weights of these sampled observations (Korn and Graubard, 1999, Section 2.3).

Although this cluster sampling decreases precision for a specified sample size, it can increase sample size and precision for a specified cost (Lumley, 2010, Section 3.1).

2.8 Motivating Data

We describe two data sets that will be used to illustrate the proposed methods.

2.8.1 US National Wilms Tumour Study

Wilms tumour is an embryonal cancer of the kidney primarily affecting children. The National Wilms Tumour Study (NWTS) began in 1969 and five clinical trials have been subsequently performed. We consider data from the third and fourth clinical trials, following Breslow and Chatterjee (1999). There are two Wilms tumour histologies, favourable (FH) and unfavourable (UH), so named due to their respective prognoses. The definitive histologic diagnosis (termed central histology) for each child's tumour is performed by the individual pathologist who initially defined the FH and UH subtypes. During the study, a histologic diagnosis is also made by the pathologist on duty at the time of treatment (termed institutional histology). As in Breslow and Chatterjee (1999), we view institutional histology as a surrogate measure of clinical histology. The population data consist of a total of 4,088 children diagnosed with Wilms tumour. The full data is provided in Table 2.3.

These data are ideal for methods evaluation since analysis of these complete data provide a benchmark with which analyses based on subsets of the data may be compared. As in Breslow and Chatterjee (1999), we take the phase I data as the bottom line of Table 2.3, so that the confounder is the binary (favourable/unfavourable) institutional histologic (IH) diagnosis. This variable is available on all children. At phase II, we assume that additional data (cancer stage and central histology) are collected on 316 non-cases and 415 cases with a favourable IH, and 255 non-cases and 156 cases with an unfavourable IH. The phase II data is provided in Table 2.4 and reveal that we have only sampled approximately 10% of the largest favourable IH non-case group. (In fact, stage was also known for all patients and could have been used to further stratify the phase I data to obtain more efficient estimates (Breslow and Chatterjee, 1999).) The aim is to estimate the association between relapse and the two covariates cancer stage and central histology. We return to these data in Section 3.5.1.

Table 2.3: Number of Wilms tumour non-cases and cases by institutional histology (IH), central histology (CH) and stage of disease: full data.

		Favourable IH		Unfavourable IH	
		Non-cases	Cases	Non-cases	Cases
	Stage I	1363	91	15	1
Favourable	Stage II	816	117	11	2
CH	Stage III	693	100	18	3
	Stage IV	307	60	23	3
	Stage I	37	9	64	16
Unfavorable	Stage II	25	14	51	33
CH	Stage III	14	15	60	57
	Stage IV	7	9	13	41
Total		3262	415	255	156

Table 2.4: Number of Wilms tumour non-cases and cases by institutional histology (IH), central histology (CH) and stage of disease: phase II data. The two-phase design we analyze takes all of the unfavourable IH non-cases and cases and all of the favourable IH cases, but takes a sample of 316 of the available 3,262 favourable IH non-cases.

		Favourable IH		Unfavourable IH	
		Non-cases	Cases	Non-cases	Cases
	Stage I	115	91	15	1
Favourable	Stage II	86	117	11	2
CH	Stage III	79	100	18	3
	Stage IV	27	60	23	3
	Stage I	2	9	64	16
Unfavorable	Stage II	3	14	51	33
CH	Stage III	3	15	60	57
	Stage IV	1	9	13	41
Total		316	415	255	156

2.8.2 North Carolina Infant Mortality Data

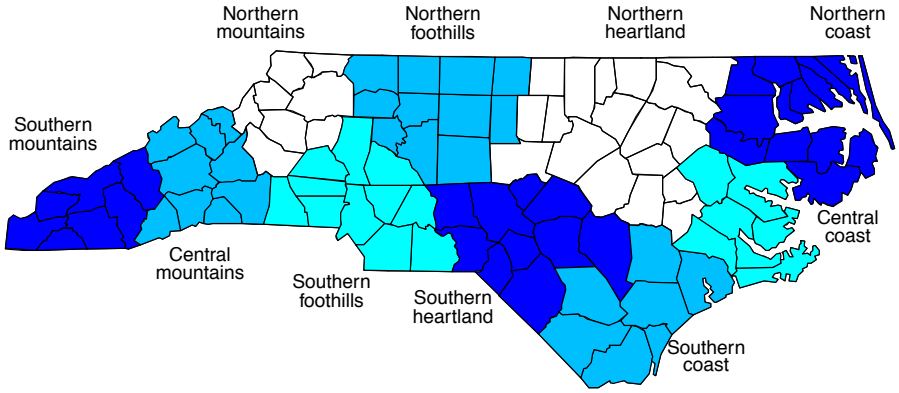


Figure 2.2: Grouping the 100 counties of North Carolina into 10 regions.

Vital statistics are collected annually for all North Carolina residents by the North Carolina State Center for Health Services (SCHS). Infant mortality data from all 100 counties in North Carolina for the years 2000-2004 was considered, where 699,035 infants were born and 5,854 died over these 5 years. An infant's race, gender and low birth weight status was also collected, where low birth weight is defined to be a birth weight of less than 2,500 g. These data are provided in Table 2.5, collapsed across counties. The goal is to estimate the association between infant mortality and birth weight, controlling for gender and race. In particular, we are interested in the potential effect modification by race of the infant mortality-birth weight association.

The spatial component makes this data set ideal for illustrating the Bayesian random effects models (Wakefield and Haneuse, 2008).

Due to the relatively small number of cases in some counties, the counties of North Carolina were grouped into 10 regions based on contiguous counties. These regions are shown in Figure 2.2. The complete data cross-classified by race, gender, birth weight and region is given in Table 2.6. We assume that, at phase I, we have observed the infant mortality by region counts so that the confounder is region. At phase II, we assume that

additional data are collected on an infant's race, gender and birth weight on 200 non-cases and 50 cases from each region. The phase II data is provided in Table 2.7. We return to these data in Section 4.4.2.

Table 2.5: Numbers of births, deaths and probability of infant mortality (×100) by race, gender and low birth weight status for North Carolina, 2000-2004: full data

	Whites			Non-Whites		
Birth weight and gender	No. of deaths	No. of births	Prob. of infant mortality (×100)	No. of deaths	No. of births	Prob. of infant mortality (×100)
Normal birth weight						
Female	425	226,103	0.19	229	82,713	0.28
Male	609	240,581	0.25	302	87,313	0.35
Low birth weight						
Female	892	19,047	4.7	995	13,748	7.2
Male	1,175	17,360	6.8	1,227	12,170	10.1

2.9 Small Area Estimation

In small area estimation, we wish to estimate characteristics of interest, such as counts or mean number of individuals with a particular outcome of interest, for small areas where only small samples are available. Often, data arising from sample surveys are used for estimation in such cases. For a binary outcome and covariates $\boldsymbol{x}_i$ for an area i, $i = 1, \ldots, I$, suppose we want to estimate the population total in area i, T_i. Let N_i be the total number of individuals, y_i be the number of sampled outcomes and n_i be the number of sampled individuals in area i. If we assume the logistic regression model

$$\text{logit}(p(\boldsymbol{x}_i)) = \boldsymbol{x}_i\boldsymbol{\beta},$$

then we can estimate T_i using

$$\widehat{T}_i = y_i + (N_i - n_i) \times \widehat{p}(\boldsymbol{x}_i), \tag{2.23}$$

where $\text{logit}(\widehat{p}(\boldsymbol{x}_i)) = \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}$ for $\widehat{\boldsymbol{\beta}}$ the maximum likelihood estimate for $\boldsymbol{\beta}$ (Royall, 1970). Standard model-based approaches are subject to potentially large biases, however (Rao, 2003, Section 2.1). Alternatively, a design-based approach could be used. In this case, the sampling weights w_i need to be known, and the methods described in Section 2.7 can be applied to obtain $\widehat{p}(\boldsymbol{x}_i)$. In fact, the design-based estimate for T_i would still be given by (2.23) (Rao, 2003, Section 2.5).

However, since only small samples are generally available, the two estimators just described lead to unacceptably large standard errors (Rao, 2003, Section 4.1). Random effects terms may be included in the logistic regression model to borrow strength across areas to obtain more precise estimates. Fay and Herriot (1979) proposed such a model using independent and identically distributed random effects. In the case of a binary outcome, this model can be expressed as

$$\text{logit}(p(\boldsymbol{x}_i)) = \boldsymbol{x}_i\boldsymbol{\beta} + V_i, \tag{2.24}$$

where $\text{E}(V_i) = 0$ and $\text{var}(V_i) = \sigma_v^2$. Normality of the V_i is often assumed. The random effects can be estimated using linear mixed effects models (Singh, Shukla, and Kundu, 2005). The Fay-Herriot model can be extended to allow for spatial correlation between areas. In this approach, a spatial random effect term, U_i, is added to (2.24) and either the joint or conditional models described in Section 2.6.3 can easily be assumed as prior distributions (Best et al., 2005). For example, Singh et al. (2005) illustrate various random effects (both spatial and non-spatial) models using data from the National Sample Survey Organisation (NSSO) of the Ministry of Statistics and Programme Implementation (Government of India).

Sampling weights are rarely included in these types of smoothing models, which are crucial when data arise from complex surveys. Congdon and Lloyd (2010) offer one approach. The authors incorporate spatial random effects into a weighted likelihood to derive small area prevalence estimates for diabetes. Chen, Lumley, and Wakefield (2011) develop a method to incorporate sampling weights for binary data, and consider empirical Bayes beta-binomial models, and also normal hierarchical models in which spatial random effects can be included.

Table 2.6: Number of live and dead infants by North Carolina region, race, sex and low birth weight status: full data.

		Normal Birth Weight				Low Birth Weight			
		Female		Male		Female		Male	
		White	Non-white	White	Non-white	White	Non-white	White	Non-white
Southern Mountains	Alive	4427	473	4788	514	356	32	320	29
	Dead	9	1	15	2	9	2	35	4
Central Mountains	Alive	14504	1198	15492	1316	1291	168	1229	167
	Dead	35	4	50	7	59	19	96	12
Northern Mountains	Alive	10476	795	11122	847	992	110	853	100
	Dead	35	2	40	2	49	4	54	7
Southern Foothills	Alive	54853	17325	58687	18520	4445	2694	3898	2260
	Dead	105	56	153	73	216	182	272	241
Northern Foothills	Alive	40334	11923	42487	12560	3370	1872	3060	1682
	Dead	77	28	114	35	187	157	216	202
Southern Heartland	Alive	48617	23924	51734	25315	3519	3697	3113	3198
	Dead	83	70	107	86	154	273	207	357
Southern Coast	Alive	22207	14178	23726	14899	1783	2135	1570	1847
	Dead	31	46	57	56	87	193	128	213
Northern Heartland	Alive	17189	5577	18158	5727	1403	845	1252	664
	Dead	32	6	43	22	72	50	89	71
Central Coast	Alive	8987	4112	9520	4341	682	744	573	584
	Dead	9	7	22	13	36	71	50	74
Northern Coast	Alive	4084	2979	4258	2972	314	456	317	412
	Dead	9	9	8	6	23	44	28	46

Table 2.7: Number of live and dead infants by North Carolina region, race, sex and low birth weight status: phase II data.

		Normal Birth Weight				Low Birth Weight			
		Female		Male		Female		Male	
		White	Non-white	White	Non-white	White	Non-white	White	Non-white
Southern Mountains	Alive	92	3	79	6	9	2	9	0
	Dead	6	0	9	1	7	2	23	2
Central Mountains	Alive	84	5	89	9	7	3	3	0
	Dead	6	1	10	0	7	3	20	3
Northern Mountains	Alive	86	6	87	3	11	0	7	0
	Dead	12	0	14	1	9	0	14	0
Southern Foothills	Alive	64	30	68	28	3	0	4	3
	Dead	1	5	8	2	9	6	6	13
Northern Foothills	Alive	67	25	69	16	9	5	6	3
	Dead	4	2	7	3	9	8	9	8
Southern Heartland	Alive	53	30	71	28	5	5	5	3
	Dead	4	2	7	4	3	17	2	11
Southern Coast	Alive	49	38	65	35	4	2	3	4
	Dead	1	3	4	2	5	13	8	14
Northern Heartland	Alive	70	18	87	15	2	5	2	1
	Dead	4	2	4	2	8	5	11	14
Central Coast	Alive	73	26	54	26	11	4	2	4
	Dead	1	1	3	4	10	9	11	11
Northern Coast	Alive	51	43	52	36	5	5	6	2
	Dead	4	2	3	2	8	13	6	12

Chapter 3

BAYESIAN ANALYSIS OF TWO-PHASE DATA FOR INDEPENDENT DATA

In this chapter, we consider the Bayesian analysis of two-phase data in the case of independent data with simple random or case-control sampling of the first phase data. In Section 3.1, we describe in detail methods for the simplest case of assessing the association between a binary outcome variable and a single discrete exposure variable. In Section 3.2, we describe the extension of these methods to the case of assessing the association between a binary outcome variable and a discrete exposure in the presence of discrete confounder variables. In Section 3.4, the methods are compared with existing approaches using a simulation study and the sparse data situation is illustrated using simulated data. We conclude the chapter by applying the methods to several different data sets including the Wilms tumour data set described in Chapter 2.

3.1 Discrete Exposure Variables

We begin with the case where there is no confounder variable z, and hence the z index is dropped to simplify notation. This example does not represent a situation of practical interest, but it does provide an easy introduction to the methods that appear in the next section.

Let y denote disease status, with $y = 1$ denoting a case, and $y = 0$ denoting a non-case, and let x denote exposure status, with $x = 1, \ldots, m_x$. Let N_{yx} denote the (unobserved) number of individuals with disease status y, and exposure status x. At phase I, we observe the response margin $\boldsymbol{N}^{\mathrm{y}\cdot} = \{N_{y\cdot}, y = 0, 1\}$ obtained via simple random or case-control sampling. At phase II, $n_{y\cdot}$ individuals are randomly sampled from $N_{y\cdot}$, $y = 0, 1$, and their exposure status is measured. At phase II, we observe the phase II sample sizes $\boldsymbol{n}^{\mathrm{y}\cdot} = \{n_{y\cdot}, y = 0, 1\}$ and the phase II outcomes $\boldsymbol{n}^{\mathrm{yx}} = \{n_{yx}, y = 0, 1, x = 1, \ldots, m_x\}$. Table 3.1 displays the two-phase study design.

Table 3.1: Two-phase study design for a single discrete exposure: the observed data are $N_{y\cdot}$ at phase I, and $n_{y\cdot}$ and n_{yx} at phase II; the N_{yx} entries are unobserved in a two-phase study, denoted by $[\cdot]$.

Phase I Data

	$Y=0$	$Y=1$
$X=1$	$[N_{01}]$	$[N_{11}]$
$X=2$	$[N_{02}]$	$[N_{12}]$
$\vdots$		
$X=m_x$	$[N_{0m_x}]$	$[N_{1m_x}]$
	$N_{0\cdot}$	$N_{1\cdot}$

Phase II Data

	$Y=0$	$Y=1$
$X=1$	n_{01}	n_{11}
$X=2$	n_{02}	n_{12}
$\vdots$		
$X=m_x$	n_{0m_x}	n_{1m_x}
	$n_{0\cdot}$	$n_{1\cdot}$

3.1.1 The Likelihood

We specify a joint log-linear model for the binary random variable y and the univariate discrete random variable x to cover both phase I sampling mechanisms. Consider $\mu_{yx} = E[N_{yx} \mid \boldsymbol{\lambda}]$, the mean of the disease-exposure count in cell (y, x) of the $2 \times m_x$ contingency table, $y = 0, 1; x = 1, ..., m_x$. We specify a saturated model given by:

$$\log(\mu_{yx}) = \mu + \lambda_y^Y + \lambda_x^X + \lambda_{yx}^{YX} \tag{3.1}$$

with identifiability gained through the sum-to-zero constraints

$$\sum_y \lambda_y^Y = \sum_x \lambda_x^X = 0,$$
$$\sum_y \lambda_{yx}^{YX} = \sum_x \lambda_{yx}^{YX} = 0.$$

On the way to deriving the likelihood, we first consider the distribution of the phase I data. Under simple random sampling at phase I, the joint probabilities are denoted $r_{yx} = \text{pr}(Y = y, X = x \mid \boldsymbol{\lambda})$, while for case-control sampling we denote $r_{x|y} = \text{pr}(X = x \mid Y = y, \boldsymbol{\lambda})$. Under simple random sampling, the intercept μ cancels when calculating the

probabilities

$$\begin{aligned} r_{yx} &= \frac{\mu_{yx}}{\sum_u \sum_v \mu_{uv}} \\ &= \frac{\exp(\mu + \lambda_y^Y + \lambda_x^X + \lambda_{yx}^{YX})}{\sum_u \sum_v \exp(\mu + \lambda_u^Y + \lambda_v^X + \lambda_{uv}^{YX})} \\ &= \frac{\exp(\lambda_y^Y + \lambda_x^X + \lambda_{yx}^{YX})}{\sum_u \sum_v \exp(\lambda_u^Y + \lambda_v^X + \lambda_{uv}^{YX})}. \end{aligned} \tag{3.2}$$

Under case-control sampling, both the intercept μ and main effect term for Y, λ_y^Y, cancel when calculating the probabilities

$$\begin{aligned} r_{x|y} &= \frac{\mu_{yx}}{\sum_v \mu_{yv}} \\ &= \frac{\exp(\mu + \lambda_y^Y + \lambda_x^X + \lambda_{yx}^{YX})}{\sum_v \exp(\mu + \lambda_y^Y + \lambda_v^X + \lambda_{yv}^{YX})} \\ &= \frac{\exp(\lambda_x^X + \lambda_{yx}^{YX})}{\sum_v \exp(\lambda_v^X + \lambda_{yv}^{YX})}. \end{aligned} \tag{3.3}$$

Under simple random sampling at phase I, we condition on the grand total $N_{\cdot\cdot}$ since it is fixed by design and the likelihood is given by

$$\begin{aligned} \mathrm{pr}(\boldsymbol{N}^{\mathrm{y}\cdot}, \boldsymbol{n}^{\mathrm{y}\cdot}, \boldsymbol{n}^{\mathrm{yx}} | N_{\cdot\cdot}, \boldsymbol{\lambda}) &\propto \prod_{y=0}^{1} \mathrm{pr}(Y=y)^{N_{y\cdot}} \times \prod_{y=0}^{1} \prod_{x=1}^{m_x} \mathrm{pr}(X=x|Y=y)^{n_{yx}} \\ &= \mathrm{pr}(\boldsymbol{N}^{\mathrm{y}\cdot} | N_{\cdot\cdot}, \boldsymbol{\lambda}) \times \mathrm{pr}(\boldsymbol{n}^{\mathrm{yx}} | \boldsymbol{n}^{\mathrm{y}\cdot}, \boldsymbol{\lambda}). \end{aligned} \tag{3.4}$$

The first term in (3.4) is the likelihood corresponding to

$$\boldsymbol{N}^{\mathrm{y}\cdot} | N_{\cdot\cdot}, \boldsymbol{\lambda} \sim \mathrm{Bin}\left(N_{\cdot\cdot}, \mathrm{pr}(Y=1)\right),$$

and the second term of (3.4) is the likelihood corresponding to

$$\boldsymbol{n}^{\mathrm{yx}} | \boldsymbol{n}^{\mathrm{y}\cdot}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Multinomial}\left(n_{y\cdot}, \{\mathrm{pr}(X=x|Y=y)\}_{x=1,\ldots,m_x}\right), \quad y=0,1,$$

where $\mathrm{pr}(X=x|Y=y) = \frac{r_{yx}}{\sum_{v=1}^{m_x} r_{yv}}$. In the case where the exposure variable x is binary, the second term of (3.4) is given by independent binomial distributions:

$$\boldsymbol{n}^{\mathrm{yx}} | \boldsymbol{n}^{\mathrm{y}\cdot}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Bin}\left(n_{y\cdot}, \mathrm{pr}(X=1|Y=y)\right), \quad y=0,1.$$

Similarly, under case-control sampling at phase I, we condition on $N_{0\cdot}$ and $N_{1\cdot}$, which are fixed by design, and the likelihood is

$$\begin{aligned}
\mathrm{pr}(\boldsymbol{N}^{\mathrm{y}\cdot}, \boldsymbol{n}^{\mathrm{y}\cdot}, \boldsymbol{n}^{\mathrm{yx}} | N_{0\cdot}, N_{1\cdot}, \boldsymbol{\lambda}) \quad &\propto \quad \prod_{y=0}^{1} \prod_{x=1}^{m_x} \mathrm{pr}(X = x | Y = y)^{n_{yx}} \\
&\sim_{\mathrm{ind}} \quad \mathrm{Multinomial}\left(n_{y\cdot}, \{\mathrm{pr}(X = x | Y = y)\}_{x=1,\ldots,m_x}\right), y = 0, 1, \\
&= \quad \mathrm{pr}(\boldsymbol{n}^{\mathrm{yx}} | \boldsymbol{n}^{\mathrm{y}\cdot}, \boldsymbol{\lambda}). \qquad (3.5)
\end{aligned}$$

As in the simple random sampling at phase I situation, a special case occurs when the exposure x is a binary variable. In this case, (3.5) is given by independent binomial distributions:

$$\boldsymbol{n}^{\mathrm{yx}} | \boldsymbol{n}^{\mathrm{y}\cdot}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Bin}\left(n_{y\cdot}, \mathrm{pr}(X = 1 | Y = y)\right), \quad y = 0, 1.$$

In many situations, we are interested in the coefficients from the linear logistic disease model

$$\begin{aligned}
\log\left[\frac{\mathrm{pr}(Y = 1 \mid X = x)}{\mathrm{pr}(Y = 0 \mid X = x)}\right] &= \beta_0 + \beta_2 \mathbf{1}[x = 2] + \cdots + \beta_{m_x} \mathbf{1}[x = m_x] \qquad (3.6) \\
&= \boldsymbol{W}\boldsymbol{\beta},
\end{aligned}$$

where $\boldsymbol{W}$ is the $m_x \times m_x$ design matrix. This disease model can be equivalently expressed using the log-linear model in (3.1), if we rewrite the left hand side of (3.6) as

$$\begin{aligned}
\log\left[\frac{\mathrm{pr}(Y = 1 \mid X = x)}{\mathrm{pr}(Y = 0 \mid X = x)}\right] &= (\lambda_1^Y + \lambda_x^X + \lambda_{1x}^{YX}) - (\lambda_0^Y + \lambda_x^X + \lambda_{0x}^{YX}) \\
&= (-\lambda_0^Y - \lambda_{0x}^{YX}) - (\lambda_0^Y + \lambda_{0x}^{YX}) \\
&= -2\lambda_0^Y - 2\lambda_{0x}^{YX}.
\end{aligned}$$

In terms of model (3.6), the intercept, β_0, is the log odds of disease given $x = 1$, so that all the indicator terms $\mathbf{1}[x = 2], \ldots, \mathbf{1}[x = m_x]$ are zero. Hence,

$$\beta_0 = -2(\lambda_0^Y + \lambda_{01}^{YX}).$$

For the xth main effect term, β_x is the log odds ratio comparing individuals with exposure level x to individuals with exposure level $x = 1$:

$$\begin{aligned}
\beta_x &= \log\left[\frac{\mathrm{pr}(Y = 1 \mid X = x)}{\mathrm{pr}(Y = 0 \mid X = x)}\right] - \log\left[\frac{\mathrm{pr}(Y = 1 \mid X = 1)}{\mathrm{pr}(Y = 0 \mid X = 1)}\right] \\
&= -2(\lambda_0^Y + \lambda_{0x}^{YX}) + 2(\lambda_0^Y + \lambda_{01}^{YX}) \\
&= -2(\lambda_{0x}^{YX} - \lambda_{01}^{YX}),
\end{aligned}$$

for $x = 2, \ldots, m_x$.

With respect to (3.1), let $\boldsymbol{\lambda}^Y = \lambda_0^Y$, $\boldsymbol{\lambda}^X = (\lambda_1^X, ..., \lambda_{m_x-1}^X)$,and $\boldsymbol{\lambda}^{YX} = (\lambda_{01}^{YX}, ..., \lambda_{0,m_x-1}^{YX})$. Then, we have $\lambda_{0m_x}^{YX} = -\sum_{i=1}^{m_x-1} \lambda_{0i}^{YX}$ using the sum-to-zero constraints on $\boldsymbol{\lambda}$. Hence, we see that $\boldsymbol{\beta}$ only depends on the subset of $\boldsymbol{\lambda}$, $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX})$ and we describe the relationship via

$$\begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_{m_x} \end{pmatrix} = \begin{pmatrix} -2 & -2 & 0 & 0 & \cdots & 0 \\ 0 & 2 & -2 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 2 & 0 & \cdots & 0 & -2 \\ 0 & 4 & 2 & 2 & \cdots & 2 \end{pmatrix} \begin{pmatrix} \lambda_0^Y \\ \lambda_{01}^{YX} \\ \lambda_{02}^{YX} \\ \vdots \\ \vdots \\ \lambda_{0,m_x-1}^{YX} \end{pmatrix}, \tag{3.7}$$

i.e. $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\Lambda}^Y$, where $\boldsymbol{C}$ is an $m_x \times m_x$ invertible matrix. Due to the sum-to-zero constraints, for β_{m_x} we have

$$\begin{aligned} \beta_{m_x} &= -2(\lambda_{0m_x}^{YX} - \lambda_{01}^{YX}) \\ &= -2\left(-\sum_{i=1}^{m_x-1} \lambda_{0i}^{YX} - \lambda_{01}^{YX}\right) \\ &= 2\sum_{i=2}^{m_x-1} \lambda_{0i}^{YX} + 4\lambda_{01}^{YX}. \end{aligned}$$

We perform computation for $\boldsymbol{\lambda}$ and then transform to $\boldsymbol{\beta}$ using $\boldsymbol{C}$. In the phase I case-control situation, we can no longer estimate the intercept β_0 since λ_y^Y cancels in the calculation of $r_{x|y}$ due to the conditioning on disease status. In the frequentist development, an intercept is present but is not interpretable unless information is available on the sampling frequencies of the cases and controls. Thus, in the phase I case-control situation, $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^{YX})$ and the first row and column of $\boldsymbol{C}$ are removed.

3.1.2 Prior Specification

We specify prior distributions on the reduced sets of parameters $\boldsymbol{\lambda}^Y$, $\boldsymbol{\lambda}^X$,and $\boldsymbol{\lambda}^{YX}$. Under a log-linear specification, a convenient choice of prior for $\boldsymbol{\lambda}$ is the Dirichlet, but this choice

is restrictive since there are insufficient free parameters. Specifically, there is a single parameter only to control the precision for all elements. We suppose there is simple random sampling at phase I. The prior we adopt assumes

$$\pi(\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^{YX}) = \pi(\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX})\pi(\boldsymbol{\lambda}^X).$$

Since $\boldsymbol{\beta}$ depends on $\boldsymbol{\Lambda}^Y$, these parameters have special status as parameters of interest. The prior distribution we place on $\boldsymbol{\Lambda}^Y$ induces a prior on $\boldsymbol{\beta}$, and hence should be carefully chosen. We focus on assigning priors to $\boldsymbol{\beta}$. We assume a multivariate normal prior distribution for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\mu}$ is a vector of length m_x and $\boldsymbol{\Sigma}$ is a square matrix of order m_x, with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ chosen based on the context. The induced multivariate normal prior for $\boldsymbol{\Lambda}^Y$ has mean $\boldsymbol{C}^{-1}\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$.

For $\boldsymbol{\lambda}^X$, we adapt the specification proposed by Knuiman and Speed (1988), and refined by Dellaportas and Forster (1999). This specification allows us to incorporate the sum-to-zero constraints we have imposed on $\boldsymbol{\lambda}^X$ directly into its prior distribution by using a structured multivariate normal prior rather than using univariate normal priors on the elements of $\boldsymbol{\lambda}^X$. We let $\boldsymbol{\lambda}^X \sim N(\mathbf{0}_{m_x-1}, \alpha_a^2 \boldsymbol{V}_a)$ where $\mathbf{0}_{m_x-1}$ is a vector of zeros of length $m_x - 1$, and

$$\boldsymbol{V}_a = \left(\boldsymbol{I}_{m_x-1} - \frac{1}{m_x}\boldsymbol{J}_{m_x-1}\right),$$

where $\boldsymbol{I}_{m_x-1}$ is the $(m_x - 1) \times (m_x - 1)$ identity matrix and $\boldsymbol{J}_{m_x-1}$ is the $(m_x - 1) \times (m_x - 1)$ matrix containing all ones. Dellaportas and Forster (1999) discuss the choice of α_a^2, with an invariance argument suggesting $\alpha_a^2 = k \times m_x$. We adopt a value of $k = 1$, corresponding closely to a unit information prior (Kass and Wasserman, 1995); see Section 3.4 of Dellaportas and Forster (1999). Since we often will not have any prior information on $\boldsymbol{\lambda}^X$, using a data driven prior is pragmatic.

We assign the same prior distributions in the phase I case-control situation, however we need only specify priors for $\boldsymbol{\lambda}^X$ and $\boldsymbol{\lambda}^{YX}$ as $\boldsymbol{\lambda}^Y$ cannot be estimated in this setting since we condition on disease status.

3.1.3 Computation

The likelihoods given by (3.4) and (3.5) are complicated functions of $\boldsymbol{\lambda}$ with no closed form expressions. We therefore use a Metropolis-Hastings algorithm to compute the posterior for $\boldsymbol{\lambda}$, sampling the complete $\boldsymbol{\lambda}$ vector via a single Metropolis-Hastings step using a multivariate random walk based on a normal proposal.

Sampling the complete $\boldsymbol{\lambda}$ vector via a multivariate normal proposal requires a covariance matrix that closely mimics the dependence within the posterior distribution. Again, we suppose we have simple random sampling at phase I and a saturated log-linear model. (Case-control sampling follows in a straightforward fashion). The phase II data $\boldsymbol{n}^{\mathrm{yx}}$ provide information on the $y \times x$ table, but these data are not a representative sample due to the phase I sampling. However, we can correct for the bias by creating expected counts $N^{\star}_{yx} = E[N_{yx}] = N_{\cdot\cdot} \times \widehat{\mathrm{pr}}(Y = y, X = x)$, with

$$\begin{aligned} \widehat{\mathrm{pr}}(Y = y, X = x) &= \widehat{\mathrm{pr}}(X = x \mid Y = y) \times \widehat{\mathrm{pr}}(Y = y) \\ &= \frac{n_{yx}}{n_{y\cdot}} \times \frac{N_{y\cdot}}{N_{\cdot\cdot}} \end{aligned}$$

to give $N^{\star}_{yx} = n_{yx}N_{y\cdot}/n_{y\cdot}$. A saturated log-linear model is fitted to the $N^{\star}_{yx}$ data using maximum likelihood estimation, and the variance-covariance matrix of the proposal for $\boldsymbol{\lambda}$ is taken as a constant, c, times the asymptotic variance-covariance matrix of the maximum likelihood estimate. The constant is chosen to achieve acceptance rates of approximately 25–30% (Roberts et al., 1997). This results in small step sizes for $\boldsymbol{\lambda}$.

3.2 Discrete Exposure and Confounder Variables

We extend the case of a discrete exposure variable to include discrete confounder variables, z, that take on finitely many values. As we will see in Chapter 4 when we consider dependent data, z may represent geographic area.

3.2.1 The Likelihood

As described in Section 3.1, we consider both simple random and case-control sampling at phase I and specify a joint log-linear model for the binary random variable y and the

univariate discrete random variables x and z to cover both sampling mechanisms. We emphasize that in the simple random sampling at phase I scheme, we assume that $N_{...}$ individuals are sampled and then cross-classified with respect to the outcome variable y and to the strata of confounder variables, z. In the case-control sampling at phase I scheme, we assume $N_{1..}$ cases and $N_{0..}$ controls are sampled independently, and the individuals are then classified with respect to the strata of confounder variables, z.

Let $\mu_{yxz} = \mathrm{E}[N_{yxz} \mid \boldsymbol{\lambda}]$ denote the mean of the disease-exposure-confounder count in cell (y, x, z) of the $2 \times m_x \times m_z$ contingency table, $y = 0, 1; x = 1, ..., m_x; z = 1, ..., m_z$. We begin by specifying a saturated model, however this is not always desirable, a point to which we return in Section 3.6. The saturated log-linear model is:

$$\log(\mu_{yxz}) = \mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ} \tag{3.8}$$

with identifiability gained through the sum-to-zero constraints

$$\begin{aligned}
\sum_y \lambda_y^Y = \sum_x \lambda_x^X = \sum_z \lambda_z^Z = 0,\\
\sum_y \lambda_{yx}^{YX} = \sum_x \lambda_{yx}^{YX} = 0,\\
\sum_y \lambda_{yz}^{YZ} = \sum_z \lambda_{yz}^{YZ} = 0,\\
\sum_x \lambda_{xz}^{XZ} = \sum_z \lambda_{xz}^{XZ} = 0,\\
\sum_y \lambda_{yxz}^{YXZ} = \sum_x \lambda_{yxz}^{YXZ} = \sum_z \lambda_{yxz}^{YXZ} = 0.
\end{aligned}$$

We first consider the distribution of the phase I data. Under simple random sampling at phase I, the joint probabilities are given by $r_{yxz} = \mathrm{pr}(Y = y, X = x, Z = z \mid \boldsymbol{\lambda})$, where

$$\begin{aligned}
r_{yxz} &= \frac{\mu_{yxz}}{\sum_u \sum_v \sum_w \mu_{uvw}} \\
&= \frac{\exp(\mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ})}{\sum_u \sum_v \sum_w \exp(\mu + \lambda_u^Y + \lambda_v^X + \lambda_w^Z + \lambda_{uv}^{YX} + \lambda_{uw}^{YZ} + \lambda_{vw}^{XZ} + \lambda_{uvw}^{YXZ})} \\
&= \frac{\exp(\lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ})}{\sum_u \sum_v \sum_w \exp(\lambda_u^Y + \lambda_v^X + \lambda_w^Z + \lambda_{uv}^{YX} + \lambda_{uw}^{YZ} + \lambda_{vw}^{XZ} + \lambda_{uvw}^{YXZ})}.
\end{aligned} \tag{3.9}$$

The intercept μ cancels when the r_{yxz} probabilities are calculated. Under case-control

sampling, we have $r_{xz|y} = \text{pr}(X = x, Z = z \mid Y = y, \boldsymbol{\lambda})$, where

$$\begin{aligned} r_{xz|y} &= \frac{\mu_{yxz}}{\sum_v \sum_w \mu_{yvw}} \\ &= \frac{\exp(\mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ})}{\sum_v \sum_w \exp(\mu + \lambda_y^Y + \lambda_v^X + \lambda_w^Z + \lambda_{yv}^{YX} + \lambda_{yw}^{YZ} + \lambda_{vw}^{XZ} + \lambda_{yvw}^{YXZ})} \\ &= \frac{\exp(\lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ})}{\sum_v \sum_w \exp(\lambda_v^X + \lambda_w^Z + \lambda_{yv}^{YX} + \lambda_{yw}^{YZ} + \lambda_{vw}^{XZ} + \lambda_{yvw}^{YXZ})}. \end{aligned}$$

Hence, under case-control sampling, the intercept μ and main effect term for Y, λ_y^Y, cancel when the $r_{xz|y}$ probabilities are calculated.

The observed data consist of three vectors of counts: the response-confounder margin observed at phase I, $\boldsymbol{N}^{\text{y·z}} = \{N_{y\cdot z}, y = 0, 1; z = 1, ..., m_z\}$, the phase II sample sizes $\boldsymbol{n}^{\text{y·z}} = \{n_{y\cdot z}, y = 0, 1; z = 1, ..., m_z\}$ and the phase II outcomes $\boldsymbol{n}^{\text{yxz}} = \{n_{yxz}, y = 0, 1; x = 1, ..., m_x; z = 1, ..., m_z\}$. Under simple random sampling at phase I, we condition on the grand total $N_{\cdots}$ only, which is fixed by design. Using analogous arguments from Proposition 1 of Holubkov (1995), the derivation of the likelihood proceeds as follows. We begin by introducing $\boldsymbol{N}^{\text{yxz}}$ as auxiliary variables to facilitate the procedure:

$$\text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{\cdots}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{N}^{\text{yxz}} \in \mathcal{S}_{yxz}} \text{pr}(\boldsymbol{N}^{\text{yxz}}, \boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{\cdots}, \boldsymbol{\lambda}), \quad (3.10)$$

where $\mathcal{S}_{yxz}$ represents all possible configurations of the internal cells, $\boldsymbol{N}^{\text{yxz}}$, such that the column totals, $\boldsymbol{N}^{\text{y·z}}$, and the phase II outcomes, $\boldsymbol{n}^{\text{yxz}}$, are respected:

$$\mathcal{S}_{yxz} = \left\{ N_{yxz} : n_{yxz} \leq N_{yxz} \leq N_{y\cdot z} - \sum_{v=x+1}^{m_x} N_{yvz} - \sum_{v'=1}^{x-1} n_{yv'x} \right\}_{y=0,1; 2 \leq x \leq m_x - 1; 1 \leq z \leq m_z} \quad (3.11)$$

and $\mathcal{S}_{ym_xz} = \{N_{yxz} : n_{yxz} \leq N_{yxz} \leq N_{y\cdot z} - \sum_{v=1}^{m_x - 1} n_{yvz}\}$. The right-hand side of (3.10) can be decomposed into 3 parts:

$$\begin{aligned} \text{pr}(\boldsymbol{N}^{\text{yxz}}, \boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{\cdots}, \boldsymbol{\lambda}) &= \text{pr}(\boldsymbol{N}^{\text{y·z}} | N_{\cdots}, \boldsymbol{\lambda}) \times \text{pr}(\boldsymbol{N}^{\text{yxz}} | \boldsymbol{N}^{\text{y·z}}, N_{\cdots}, \boldsymbol{\lambda}) \\ &\quad \times \text{pr}(\boldsymbol{n}^{\text{y·z}} | \boldsymbol{N}^{\text{yxz}}, \boldsymbol{N}^{\text{y·z}}, N_{\cdots}, \boldsymbol{\lambda}) \\ &\quad \times \text{pr}(\boldsymbol{n}^{\text{yxz}} | \boldsymbol{N}^{\text{yxz}}, \boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, N_{\cdots}, \boldsymbol{\lambda}) \\ &= \text{pr}(\boldsymbol{N}^{\text{yxz}} \mid N_{\cdots}, \boldsymbol{\lambda}) \times \text{pr}(\boldsymbol{n}^{\text{y·z}} | \boldsymbol{N}^{\text{yxz}}, \boldsymbol{N}^{\text{y·z}}, \boldsymbol{\lambda}) \\ &\quad \times \text{pr}(\boldsymbol{n}^{\text{yxz}} | \boldsymbol{N}^{\text{yxz}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{\lambda}). \end{aligned} \quad (3.12)$$

The first term on the right hand side of (3.12) is the likelihood corresponding to

$$\boldsymbol{N}^{\text{yxz}}|\boldsymbol{\lambda} \sim \text{Multinomial}_{2\times m_x \times m_z}(N_{\cdots}, \boldsymbol{r}^{yxz}),$$

where $\boldsymbol{r}^{yxz}$ is the vector of probabilities with entries

$$r_{yxz} = \text{pr}(Y = y, X = x, Z = z)$$

defined by the log-linear model (3.9).

The second term of (3.12) is determined by the investigator, though $n_{y\cdot z} \leq N_{y\cdot z}$ and the latter are random, so that technically the $n_{y\cdot z}$ are random variables also (Schill et al., 1993). However, since this term does not depend on $\boldsymbol{\lambda}$, it need not be considered for inference on $\boldsymbol{\lambda}$.

The third term in (3.12) is the likelihood corresponding to independent multivariate hypergeometric distributions, and does not depend on $\boldsymbol{\lambda}$:

$$\boldsymbol{n}^{\text{yxz}}|\boldsymbol{N}^{\text{yxz}}, \boldsymbol{n}^{\text{y·z}} \sim_{\text{ind}} \left(\frac{\prod_{x=1}^{m_x} \begin{pmatrix} N_{yxz} \\ n_{yxz} \end{pmatrix}}{\begin{pmatrix} N_{y\cdot z} \\ n_{y\cdot z} \end{pmatrix}} \right), y = 0, 1; 1 \leq z \leq m_z.$$

Hence, we can rewrite the likelihood as

$$\begin{aligned}
\text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}}|N_{\cdots}, \boldsymbol{\lambda}) &\propto \sum_{\boldsymbol{N}^{\text{yxz}} \in \mathcal{S}_{yxz}} \text{pr}(\boldsymbol{N}^{\text{yxz}} \mid N_{\cdots}, \boldsymbol{\lambda}) \times \text{pr}(\boldsymbol{n}^{\text{yxz}}|\boldsymbol{N}^{\text{yxz}}, \boldsymbol{n}^{\text{y·z}}) \\
&\propto \sum_{\boldsymbol{N}^{\text{yxz}} \in \mathcal{S}_{yxz}} \left(N_{\cdots}! \prod_{y=0}^{1} \prod_{x=1}^{m_x} \prod_{z=1}^{m_z} \frac{r_{yxz}^{N_{yxz}}}{N_{yxz}!} \right) \\
&\quad \times \prod_{y=0}^{1} \prod_{z=1}^{m_z} \frac{\frac{N_{y1z}!}{n_{y1z}!(N_{y1z}-n_{y1z})!} \cdots \frac{N_{ym_xz}!}{n_{ym_xz}!(N_{ym_xz}-n_{ym_xz})!}}{\frac{N_{y\cdot z}!}{n_{y\cdot z}!(N_{y\cdot z}-n_{y\cdot z})!}} \\
&\propto \sum_{\boldsymbol{N}^{\text{yxz}} \in \mathcal{S}_{yxz}} \left(\prod_{y=0}^{1} \prod_{z=1}^{m_z} (N_{y\cdot z} - n_{y\cdot z})! \prod_{x=1}^{m_x} \frac{\left(\frac{r_{yxz}}{r_{y\cdot z}}\right)^{N_{yxz}-n_{yxz}}}{(N_{yxz} - n_{yxz})!} \right) \\
&\quad \times \left(\prod_{y=0}^{1} \prod_{z=1}^{m_z} \left(\prod_{x=1}^{m_x} r_{yxz}^{n_{yxz}} \right) r_{y\cdot z}^{N_{y\cdot z}-n_{y\cdot z}} \right), \qquad (3.13)
\end{aligned}$$

where $r_{y\cdot z} = \sum_{x=1}^{m_x} r_{yxz}$. We can interchange the summation and multiplication in the first term of (3.13) since the bounds on N_{yxz} in the summation are separate for each (y, z) combination. Therefore, we have

$$
\begin{aligned}
(3.13) \quad = \quad & \left(\prod_{y=0}^{1} \prod_{z=1}^{m_z} \left(\prod_{x=1}^{m_x} \left(\frac{r_{yxz}}{r_{y\cdot z}} \right)^{n_{yxz}} \right) r_{y\cdot z}^{N_{y\cdot z}} \right) \\
& \times \left(\prod_{u=0}^{1} \prod_{w=1}^{m_z} \left(\sum_{\boldsymbol{N}^{uxw} \in \mathcal{S}_{yxz}} (N_{u\cdot w} - n_{u\cdot w})! \prod_{x=1}^{m_x} \frac{\left(\frac{r_{uxw}}{r_{u\cdot w}} \right)^{N_{uxw} - n_{uxw}}}{(N_{uxw} - n_{uxw})!} \right) \right). \quad (3.14)
\end{aligned}
$$

In the second term on the right-hand side of (3.14), each summation is the sum of all possible realizations of a Multinomial$(N_{u\cdot w} - n_{u\cdot w}, \frac{r_{uxw}}{r_{u\cdot w}})$ distribution. Hence, the entire term is equal to 1. Therefore, the likelihood is given by

$$
\begin{aligned}
\mathrm{pr}(\boldsymbol{N}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{yxz}} | N_{\cdots}, \boldsymbol{\lambda}) \quad \propto \quad & \prod_{y=0}^{1} \prod_{z=1}^{m_z} \mathrm{pr}(Y = y, Z = z)^{N_{y\cdot z}} \\
& \times \prod_{y=0}^{1} \prod_{z=1}^{m_z} \prod_{x=1}^{m_x} \mathrm{pr}(X = x | Z = z, Y = y)^{n_{yxz}} \\
= \quad & \mathrm{pr}(\boldsymbol{N}^{\mathrm{y\cdot z}} | N_{\cdots}, \boldsymbol{\lambda}) \times \mathrm{pr}(\boldsymbol{n}^{\mathrm{yxz}} | \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{\lambda}). \quad (3.15)
\end{aligned}
$$

The first term in (3.15) is the likelihood corresponding to is

$$
\boldsymbol{N}^{\mathrm{y\cdot z}} | N_{\cdots}, \boldsymbol{\lambda} \sim \mathrm{Multinomial}_{2 \times m_z} \left(N_{\cdots}, \{ \mathrm{pr}(Y = y, Z = z) \}_{y=0,1; 1 \le z \le m_z} \right),
$$

where $\mathrm{pr}(Y = y, Z = z) = \sum_{x=1}^{m_x} r_{yxz}$. The second term of (3.15) is the likelihood corresponding to

$$
\boldsymbol{n}^{\mathrm{yxz}} | \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Multinomial}_{m_x} \left(n_{y\cdot z}, \{ \mathrm{pr}(X = x | Z = z, Y = y) \}_{1 \le x \le m_x} \right), \quad y = 0, 1; 1 \le z \le m_z,
$$

where $\mathrm{pr}(X = x | Z = z, Y = y) = \frac{r_{yxz}}{\sum_{v=1}^{m_x} r_{yvz}}$. In the case where the exposure variable x is binary, the second term of (3.15) is given by independent binomial distributions:

$$
\boldsymbol{n}^{\mathrm{yxz}} | \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Bin} \left(n_{y\cdot z}, \mathrm{pr}(X = 1 | Z = z, Y = y) \right), y = 0, 1; 1 \le z \le m_z.
$$

Under case-control sampling at phase I, we condition on $N_{0\cdot\cdot}$ and $N_{1\cdot\cdot}$, which are fixed by design, and the likelihood is

$$\begin{aligned}\mathrm{pr}(\boldsymbol{N}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{yxz}} | N_{0\cdot\cdot}, N_{1\cdot\cdot}, \boldsymbol{\lambda}) &\propto \prod_{y=0}^{1}\prod_{z=1}^{m_z} \mathrm{pr}(Z=z|Y=y)^{N_{y\cdot z}} \\ &\quad \times \prod_{y=0}^{1}\prod_{z=1}^{m_z}\prod_{x=1}^{m_x} \mathrm{pr}(X=x|Z=z,Y=y)^{n_{yxz}} \\ &= \mathrm{pr}(\boldsymbol{N}^{\mathrm{y\cdot z}} | N_{1\cdot\cdot}, N_{0\cdot\cdot}, \boldsymbol{\lambda}) \times \mathrm{pr}(\boldsymbol{n}^{\mathrm{yxz}} | \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{\lambda}), \quad (3.16)\end{aligned}$$

using Proposition 1 of Holubkov (1995). The first term of (3.16) is the likelihood corresponding to

$$\boldsymbol{N}^{\mathrm{y\cdot z}} | N_{1\cdot\cdot}, N_{0\cdot\cdot}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Multinomial}_{m_z}\left(N_{y\cdot\cdot}, \{\mathrm{pr}(Z=z|Y=y)\}_{1\le z\le m_z}\right), y=0,1,$$

where $\mathrm{pr}(Z=z|Y=y) = \sum_{x=1}^{m_x} r_{xz|y}$. The second term of (3.16) is the likelihood corresponding to

$$\boldsymbol{n}^{\mathrm{yxz}} | \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Multinomial}_{m_x}\left(n_{y\cdot z}, \{\mathrm{pr}(X=x|Z=z,Y=y)\}_{1\le x\le m_x}\right), \\ y=0,1; 1\le z\le m_z,$$

where $\mathrm{pr}(X=x|Z=z,Y=y) = \frac{r_{xz|y}}{\sum_{v=1}^{m_x} r_{vz|y}}$. When the confounder variable z is binary, the first term of (3.16) is given by independent binomial distributions:

$$\boldsymbol{N}^{\mathrm{y\cdot z}} | N_{1\cdot\cdot}, N_{0\cdot\cdot}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Bin}\left(N_{y\cdot\cdot}, \mathrm{pr}(Z=1|Y=y)\right), y=0,1,$$

however the second term remains the same. Additionally, as in the simple random sampling at phase I situation, if the exposure variable x is binary, the second term of (3.16) is the likelihood corresponding to

$$\boldsymbol{n}^{\mathrm{yxz}} | \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{\lambda} \sim_{\mathrm{ind}} \mathrm{Bin}\left(n_{y\cdot z}, \mathrm{pr}(X=1|Z=z,Y=y)\right), y=0,1; 1\le z\le m_z.$$

We would like to use the coefficients of the log-linear model to make inference about odds ratios, which can be done by relating the $\boldsymbol{\lambda}$ parameters from the log-linear model in (3.8) to the coefficients from fitting the linear logistic disease model

$$\log\left[\frac{\mathrm{pr}(Y=1 \mid X=x, Z=z)}{\mathrm{pr}(Y=0 \mid X=x, Z=z)}\right] = \boldsymbol{W}\boldsymbol{\beta}, \quad (3.17)$$

as described in Section 3.1.1. Depending on the context, the design matrix $\boldsymbol{W}$ may contain terms for both main effects and interactions. Note that if interactions are not included then λ_{yxz}^{YXZ} is zero in the log-linear model. As in (3.7), there is a direct correspondence between $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$:

$$\begin{aligned}
\beta_0 &= -2(\lambda_0^Y + \lambda_{01}^{YX} + \lambda_{01}^{YZ} + \lambda_{011}^{YXZ}) \\
\beta_x &= -2(\lambda_{0x}^{YX} - \lambda_{01}^{YX}) - 2(\lambda_{0x1}^{YXZ} - \lambda_{011}^{YXZ}) \\
\beta_z &= -2(\lambda_{0z}^{YZ} - \lambda_{01}^{YZ}) - 2(\lambda_{01z}^{YXZ} - \lambda_{011}^{YXZ}) \\
\beta_{xz} &= -2(\lambda_{0xz}^{YXZ} - \lambda_{0x1}^{YXZ} - \lambda_{01z}^{YXZ} + \lambda_{011}^{YXZ}),
\end{aligned}$$

for $x = 2, \ldots, m_x$ and $z = 2, \ldots, m_z$. With respect to (3.8), let

$$\begin{aligned}
\boldsymbol{\lambda}^Y &= \lambda_0^Y, \\
\boldsymbol{\lambda}^X &= (\lambda_1^X, ..., \lambda_{m_x-1}^X), \\
\boldsymbol{\lambda}^Z &= (\lambda_1^Z, ..., \lambda_{m_z-1}^Z), \\
\boldsymbol{\lambda}^{YX} &= (\lambda_{01}^{YX}, ..., \lambda_{0,m_x-1}^{YX}), \\
\boldsymbol{\lambda}^{YZ} &= (\lambda_{01}^{YZ}, ..., \lambda_{0,m_z-1}^{YZ}), \\
\boldsymbol{\lambda}^{XZ} &= (\lambda_{11}^{XZ}, ..., \lambda_{m_x-1,m_z-1}^{XZ}) \quad \text{and} \\
\boldsymbol{\lambda}^{YXZ} &= (\lambda_{011}^{YXZ}, ..., \lambda_{0,m_x-1,m_z-1}^{YXZ}).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\lambda_{0m_x}^{YX} &= -\sum_{i=1}^{m_x-1} \lambda_{0i}^{YX}, \\
\lambda_{0m_z}^{YZ} &= -\sum_{j=1}^{m_z-1} \lambda_{0j}^{YZ}, \\
\lambda_{0m_x1}^{YXZ} &= -\sum_{i=1}^{m_x-1} \lambda_{0i1}^{YXZ}, \\
\lambda_{01m_z}^{YXZ} &= -\sum_{j=1}^{m_z-1} \lambda_{01j}^{YXZ}, \quad \text{and} \\
\lambda_{0m_xm_z}^{YXZ} &= -\sum_{i=1}^{m_x-1} \sum_{j=1}^{m_z-1} \lambda_{0ij}^{YXZ}
\end{aligned}$$

using the sum-to-zero constraints on $\boldsymbol{\lambda}$. Analogous to the case of no confounder variable z in Section 3.1 , $\boldsymbol{\beta}$ depends only on the subset of $\boldsymbol{\lambda}$, $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{YXZ})$ and we describe the relationship via $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\Lambda}^Y$, where $\boldsymbol{C}$ denotes the invertible $(m_x-1)(m_z-1)+(m_x-1)+(m_z-1)+1$ square transformation matrix. For example, let $m_x = 3$ and $m_z = 3$. Then, we have $\boldsymbol{\Lambda}^Y = (\lambda_0^Y, \lambda_{01}^{YX}, \lambda_{02}^{YX}, \lambda_{01}^{YZ}, \lambda_{02}^{YZ}, \lambda_{011}^{YXZ}, \lambda_{021}^{YXZ}, \lambda_{012}^{YXZ}, \lambda_{022}^{YXZ})$ and

$$\boldsymbol{C} = \begin{pmatrix} -2 & -2 & 0 & -2 & 0 & -2 & 0 & 0 & 0 \\ 0 & 2 & -2 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 4 & 2 & 0 & 0 & 4 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 2 & 0 & -2 & 0 \\ 0 & 0 & 0 & 4 & 2 & 4 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2 & 2 & 2 & -2 \\ 0 & 0 & 0 & 0 & 0 & -4 & -2 & 4 & 2 \\ 0 & 0 & 0 & 0 & 0 & -4 & 4 & -2 & 2 \\ 0 & 0 & 0 & 0 & 0 & -4 & 0 & 0 & 2 \end{pmatrix}.$$

We perform computation for $\boldsymbol{\lambda}$ and then transform to $\boldsymbol{\beta}$ using $\boldsymbol{C}$. In the phase I case-control situation, we can no longer estimate the intercept β_0 since λ_y^Y cancels in the calculation of $r_{xz|y}$ due to the conditioning on disease status. In this case, $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{YXZ})$ and the first row and column of $\boldsymbol{C}$ are removed.

3.2.2 Prior Specification

We suppose there is simple random sampling at phase I. We specify prior distributions on the reduced sets of parameters $\boldsymbol{\lambda}^Y$, $\boldsymbol{\lambda}^X$, $\boldsymbol{\lambda}^Z$, $\boldsymbol{\lambda}^{YX}$, $\boldsymbol{\lambda}^{YZ}$, $\boldsymbol{\lambda}^{XZ}$ and $\boldsymbol{\lambda}^{YXZ}$. The prior we adopt assumes

$$\pi(\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{XZ}, \boldsymbol{\lambda}^{YXZ}) = \pi(\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{YXZ})\pi(\boldsymbol{\lambda}^X)\pi(\boldsymbol{\lambda}^Z)\pi(\boldsymbol{\lambda}^{XZ}).$$

Since $\boldsymbol{\beta}$ depends on $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{YXZ})$, these parameters have special status as parameters of interest. The prior distribution we place on $\boldsymbol{\Lambda}^Y$ induces a prior on $\boldsymbol{\beta}$, and hence should be carefully chosen. We assume a multivariate normal prior distribution for $\boldsymbol{\beta}$

with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\mu}$ is a vector of length $(m_x-1)(m_z-1)+(m_x-1)+(m_z-1)+1$ and $\boldsymbol{\Sigma}$ is a square matrix of order $(m_x-1)(m_z-1)+(m_x-1)+(m_z-1)+1$, with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ chosen based on the context. (Specific examples are given later in this chapter). The induced multivariate normal prior for $\boldsymbol{\Lambda}^Y$ has mean $\boldsymbol{C}^{-1}\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$. We may assign $\boldsymbol{\mu}=\mathbf{0}$ and $\boldsymbol{\Sigma}$ a diagonal matrix with large variances if uninformative priors are desired. In other situations, for example when we have sparse data, more informative choices will be desirable.

We propose a slightly different approach for placing priors on $\boldsymbol{\lambda}^X$, $\boldsymbol{\lambda}^Z$ and $\boldsymbol{\lambda}^{XZ}$, which will usually be nuisance parameters. For these parameters, we use the specification proposed by Knuiman and Speed (1988), and refined by Dellaportas and Forster (1999), in which independent multivariate normal priors are placed on the collections of main effects and interactions. For $a \in \{X, Z, XZ\}$, let $\boldsymbol{\lambda}^a \sim N(\mathbf{0}_{d_a}, \alpha_a^2 \boldsymbol{V}_a)$ where $d_a = |\boldsymbol{\lambda}_a|$, $\mathbf{0}_{d_a}$ is a vector of zeros of length d_a, and

$$\boldsymbol{V}_a = \frac{1}{m_x m_z} \prod_{\gamma \in a} |I_\gamma| \otimes \left(\boldsymbol{I}_{|I_\gamma|-1} - \frac{1}{|I_\gamma|} \boldsymbol{J}_{|I_\gamma|-1} \right)$$

where I_γ is the set of levels of factor γ, $\boldsymbol{I}_{|I_\gamma|}$ is the $|I_\gamma| \times |I_\gamma|$ identity matrix and $\boldsymbol{J}_{|I_\gamma|}$ is the $|I_\gamma| \times |I_\gamma|$ matrix containing all ones. As in the case with no confounding variables, we use $\alpha_a^2 = m_x m_z$.

In the phase I case-control situation, we specify prior distributions for the parameters $\boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{XZ}, \boldsymbol{\lambda}^{YXZ}$ only. We assign the same prior distributions for these parameters as in the phase I simple random sampling situation described above. We need not specify a prior distribution for $\boldsymbol{\lambda}^Y$ as this parameter is eliminated in this setting since we condition on disease status.

3.2.3 Computation

As in the case where no confounder variables are present, the likelihoods given by (3.15) and (3.16) are complicated functions of $\boldsymbol{\lambda}$ with no closed form expressions. As detailed in Section 3.2.4, we have experimented with an auxiliary variable scheme in which the full data N_{yxz} are introduced into this model. This leads to simplified conditional forms within a Markov chain Monte Carlo scheme, but at the expense of requiring the simulation of

additional variables. As an alternative, we may use a Metropolis-Hastings algorithm for $\boldsymbol{\lambda}$ only. To obtain a Markov chain with good mixing properties, we sample the complete $\boldsymbol{\lambda}$ vector via a single Metropolis-Hastings step using a multivariate random walk based on a normal proposal.

Sampling the *complete* $\boldsymbol{\lambda}$ vector via a multivariate normal proposal requires a covariance matrix that closely mimics the dependence within the posterior distribution. Again, we suppose we have simple random sampling at phase I and a saturated log-linear model. (Case-control sampling and/or non-saturated models follow in a straightforward manner.) The phase II data $\boldsymbol{n}^{\mathrm{yxz}}$ provide information on the $y \times x \times z$ table, but these data are biased due to the phase I sampling. However, we can correct for the bias by creating expected counts $N^{\star}_{yxz} = \mathrm{E}[N_{yxz}] = N_{\cdots} \times \widehat{\mathrm{pr}}(Y = y, X = x, Z = z)$, with

$$\begin{aligned} \widehat{\mathrm{pr}}(Y = y, X = x, Z = z) &= \widehat{\mathrm{pr}}(X = x \mid Y = y, Z = z) \times \widehat{\mathrm{pr}}(Y = y, Z = z) \\ &= \frac{n_{yxz}}{n_{y\cdot z}} \times \frac{N_{y\cdot z}}{N_{\cdots}} \end{aligned}$$

to give $N^{\star}_{yxz} = n_{yxz} N_{y\cdot z} / n_{y\cdot z}$. A saturated log-linear model is fitted to the $N^{\star}_{yxz}$ data using maximum likelihood estimation, and the variance-covariance matrix of the proposal for $\boldsymbol{\lambda}$ is taken as a constant, c, times the asymptotic variance-covariance matrix of the maximum likelihood estimate. The constant is chosen to achieve acceptance rates of approximately 25–30% (Roberts et al., 1997). The above scheme captures the dependence in the posterior and has proved to be well-behaved Markov chains in the examples we have worked on.

3.2.4 Auxiliary Variable Sampling Scheme

For either phase I sampling scheme, the posterior may be explicitly calculated, up to proportionality, but would require a summation over all possible realizations of $\boldsymbol{N}^{\mathrm{yxz}}$, which in usual applications will be a vast space. However, the $\boldsymbol{N}^{\mathrm{yxz}}$ may be introduced as auxiliary variables within a Markov chain Monte Carlo scheme. We describe such a scheme for simple random sampling at phase I. A Markov chain is constructed which cycles, with

Metropolis-Hastings steps (Section 2.5.7), between the conditional distributions:

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{N}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{yxz}}) &\propto \mathrm{pr}(\boldsymbol{N}^{\mathrm{yxz}}|N_{\cdots}, \boldsymbol{\lambda})\mathrm{I}(\boldsymbol{N}^{\mathrm{yxz}} \in \mathcal{S}_{yxz}) \times \pi(\boldsymbol{\lambda}) \\ \mathrm{pr}(\boldsymbol{N}^{\mathrm{yxz}}|\boldsymbol{N}^{\mathrm{y\cdot z}}, \boldsymbol{n}^{\mathrm{yxz}}, \boldsymbol{n}^{\mathrm{y\cdot z}}, \boldsymbol{\lambda}) &\propto \mathrm{pr}(\boldsymbol{N}^{\mathrm{yxz}}|\boldsymbol{\lambda})\mathrm{I}(\boldsymbol{N}^{\mathrm{yxz}} \in \mathcal{S}_{yxz}) \\ &\quad \times \mathrm{pr}(\boldsymbol{n}^{\mathrm{yxz}}|\boldsymbol{N}^{\mathrm{yxz}}, \boldsymbol{n}^{\mathrm{y\cdot z}}). \end{aligned} \tag{3.18}$$

The conditional distribution, (3.18), can be normalized so that direct sampling could be performed, but this will rarely be computationally feasible because of the large sample space which must be enumerated to normalize the distribution. Hence, we use a technique adapted from Diaconis and Sturmfels (1998) and described in Section 2.5.7. Given the sampling scheme in which the $N_{y\cdot z}$ margin is fixed, we sample within the columns of the phase I data (with the data laid out as in Table 2.2). For each value of y and z (*i.e.* for each column), two distinct values of x are randomly sampled from $1, ..., m_x$, denoted by x' and x''. Then, the proposal $\widetilde{\boldsymbol{N}}^{\mathrm{yxz}}$ differs from $\boldsymbol{N}^{\mathrm{yxz}}$ in only two entries, with $\widetilde{N}_{yx'z} = N_{yx'z} - t$ and $\widetilde{N}_{yx''z} = N_{yx''z} + t$ and t sampled at random from $\{1, 2, \ldots, T\}$ for fixed $T > 0$. This proposal clearly respects the column margin, but we must ensure that the constraints encoded in (3.11) are satisfied. If $N_{yxz} \notin \mathcal{S}_{yxz}$, we immediately reject the proposed values, and retain the current value. The acceptance probability is the minimum of 1 and the ratio of (3.18) evaluated at the proposed point, $\widetilde{\boldsymbol{N}}^{\mathrm{yxz}}$, to the current point, $\boldsymbol{N}^{\mathrm{yxz}}$. The integer T is chosen according to the size of the population, and may vary according to the size of the marginal for each y and z. All values of T retain irreducibility of the Markov chain, but efficiency will depend on a choice that produces both movement around the sample space, and acceptance probabilities that are not too small.

In the examples considered in this thesis, the direct sampling scheme described in Section 3.2.3 proved more efficient, so that is the method we adopt in the examples described below. In Appendix A.3, we provide results from analyzing the Wilms tumour data using the auxiliary variable sampling scheme to detail the superiority of the direct sampling scheme approach.

3.3 Equivalence of Bayesian Two-phase Approaches Under Case-control and Simple Random Sampling at Phase I

We want to show that the Bayesian two-phase approach under case-control sampling at phase I is equivalent to the Bayesian two-phase approach under simple random sampling at phase I. We can approach this in two ways. One avenue is to extend the results of Seaman and Richardson (2004) and show the equivalence of the marginal $\boldsymbol{\beta}$ posterior distributions using carefully chosen prior distributions. Our idea was assigning Dirichlet priors to the exposure-confounder probabilities (or gamma priors on the exposure-confounder $\boldsymbol{\lambda}$ parameters) such that these parameters could then be integrated from the model. Seaman and Richardson (2004) considered the case where only exposure variables were observed for each subject in their approach and the proof of the result requires integrating out nuisance parameters. On the way to proving the result for the two-phase design, we first attempted to extend the main result of their paper to the case where exposure and confounder variables are observed for each subject. We were only able to integrate out one set of nuisance parameters, rather than both that would be needed in the proof of the equivalence result. However, this showed that we were able to at least reduce the dimensionality of the problem, and indicated that in the extension to the two-phase design, perhaps one of x or z could be integrated from the model.

Another possibility is to prove an equivalent result to that shown in Breslow and Holubkov (1997a) and show that the inference for the logistic regression parameters are the same whether case-control or simple random sampling is assumed at phase I. To prove this result, we need to show that the conditional distributions for $\boldsymbol{\beta}$ are equivalent under the two sampling schemes. When we attempted to show this, the conditional distribution for $\boldsymbol{\beta}$ under case-control sampling at phase I is the same as that under simple random sampling at phase I, except that it contains an additional term that arises due to the conditioning on the number of cases and controls, $N_{1\cdot\cdot}$ and $N_{0\cdot\cdot}$, respectively. To date, we have been unsuccessful proving the equivalence using either approach, and no obvious counterexample exists. As we discuss in Chapter 6, we plan to pursue these proofs further and we also plan to consider prospective sampling at phase I.

3.4 Simulated Data

3.4.1 A Simulation Study

In this section, we compare our proposed method with NPML using simulated data and simple random sampling at phase I. There are clearly many possible scenarios to investigate and we choose to consider binary x and z, with a fixed phase I sample size. We generated the data in the same spirit as the Wilms tumour data set, but for simplicity ignored stage information while retaining central (x) and institutional (z) histology. The observed proportions of individuals with unfavourable central and institutional histology, as well as the odds ratio characterizing the dependence between x and z, were used to simulate the data. The parameters used in the simulation were: $\beta_0 = -2.16, \beta_x = 1.59, \beta_z = 0.15, \beta_{xz} = 0.17$, $\text{pr}(X = 1) = 0.11, \text{pr}(Z = 1) = 0.1$, with an xz odds ratio of 120. The total number of subjects in the phase I population for the simulation study is 5000. Table 3.2 displays the full data from which the phase II data are sampled. Fitting a saturated logistic regression model to the full data gives the estimates $\beta_0 = -2.21, \beta_x = 1.37, \beta_z = -0.03$, and $\beta_{xz} = 0.64$, which we use to compare different estimators. We report three simulations with varying phase II sample sizes, denoted by "small", "medium" and "large". For the "small" scenario, we fix the phase II sample sizes as 125 each of favourable and unfavourable institutional histology non-cases, and 50 each of favourable and unfavourable institutional histology cases for a total of 350 individuals. For the "medium" scenario, we fix the phase II sample sizes as 250 each of favourable and unfavourable institutional histology non-cases, and 100 each of favourable and unfavourable institutional histology cases totaling 700 individuals, whereas for the "large" scenario, we fix the phase II sample sizes as 500 favourable institutional histology non-cases, 200 favourable institutional histology cases, and all unfavourable institutional histology individuals for a total of $1,184$ individuals.

Two different sets of prior distributions for $\boldsymbol{\beta}$ were evaluated. The first used a normal prior distribution with a large variance for β_0, and more informative but realistic priors for β_x, β_z, and β_{xz}. Specifically, we assigned independent normal prior distributions with variance $\log(5)/1.96$, which corresponds to a 95% prior interval of $(-5, 5)$ for each of the log odds ratios, *i.e.* $(0.01, 148)$ for the odds ratios. The second set of priors were normal

Table 3.2: Number of Wilms tumour cases and non-cases by institutional histology (IH) and central histology (CH) for the simulation study: complete data from which the phase II data are sampled for the simulation study. In a two-phase design, the internal cells are unobserved.

	Favourable IH		Unfavourable IH		
	Non-cases	Cases	Non-cases	Cases	
Favourable CH	3930	430	94	10	
Unfavourable CH	109	47	212	168	
Total	4039	477	306	178	5000

distributions with large variance for each of the $\boldsymbol{\beta}$ parameters ("flat priors"). For the remainder of the $\boldsymbol{\lambda}$ nuisance parameters, we followed the procedure outlined in Section 3.2.2. We ran the chains for a total of 450,000 iterations, where the first 50,000 were used for tuning.

In each simulation, individuals were randomly sampled from the population to obtain the phase II data, and a total of 500 data sets were generated. For each sampled data set, we analysed the data using weighted and non-parametric maximum likelihood, as well as the Bayesian two-phase method. Since we are fitting a saturated logistic regression model, the pseudo- and non-parametric maximum likelihood estimates are identical, as shown in Breslow and Holubkov (1997a). In the event of a zero count in the phase II data, the maximum likelihood estimates converge to boundary values at $\pm\infty$ so that the fitted probabilities can equal zero. The bias of these estimators is not a well-defined quantity, hence estimation of it is similarly not well-defined. In these situations, we remove the infinite estimates to obtain empirical estimates of bias and only fit the Bayesian method to the data. In the case of "small" phase II sample sizes, there were 30 such data sets. The bias, variance and mean squared error (MSE) were calculated for each method and are shown in Appendix A.1. In addition, Figure 3.1 shows violin plots comparing the results from NPML and the two Bayesian two-phase analyses using small, medium and large phase II sample sizes.

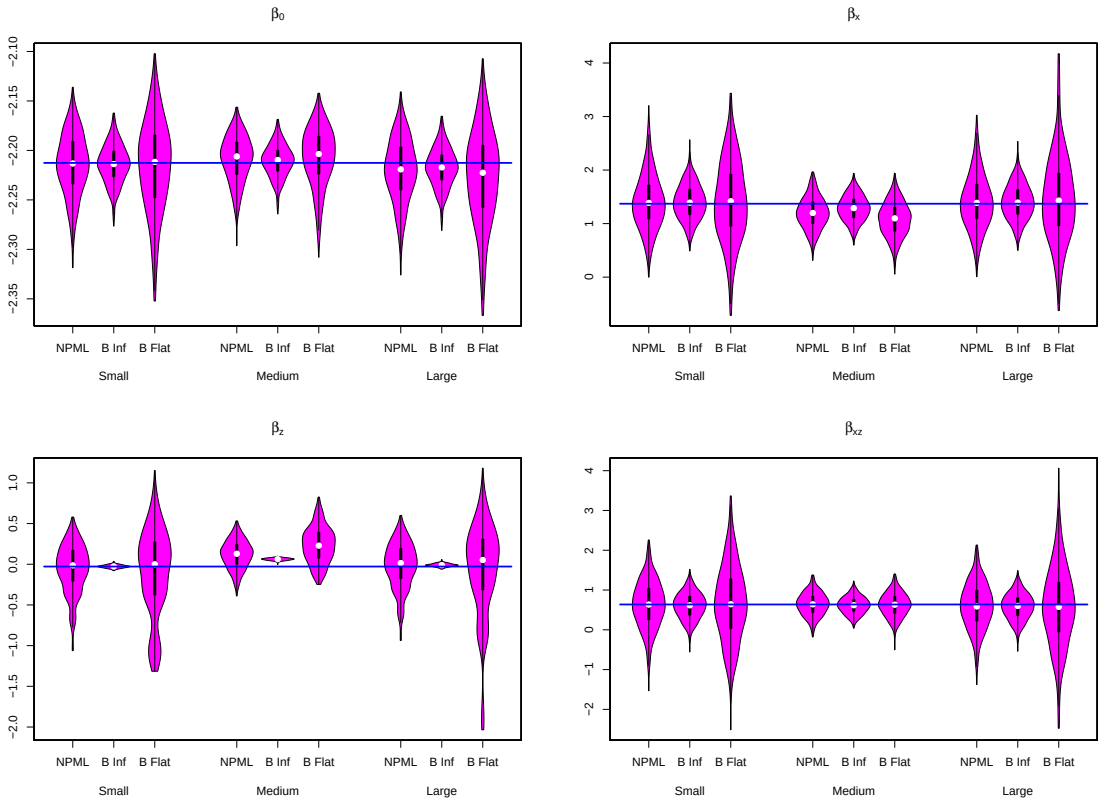


Figure 3.1: Simulation study results: violin plots comparing NPML and two Bayesian two-phase analyses using small, medium and large phase II sample sizes with informative (B Inf) and flat priors (B Flat). The blue lines indicate the true value of the parameters.

For the small sample sizes, there are two sets of Bayesian results, including and excluding the 30 data sets with zero counts. For the weighted likelihood method, there were an additional 47 data sets where the MLE estimates converge to boundary values at $\pm\infty$, and so these were also removed in the calculation of the MSE and its decomposition into the bias and variance. Not surprisingly, when the zero count cases are included, the Bayes results with flat priors perform far more poorly than the Bayes results with informative priors, particularly for β_{xz}, in terms of the precision. (We emphasize that we do not recommend the use of flat priors in practice for small phase II counts.) Even including the 30 zero count cases, the Bayes approach with informative priors outperforms the likelihood approaches (in terms of MSE) for all parameters but β_{xz}. When the zero count cases are excluded, both Bayes approaches have lower MSE than the likelihood approaches (apart from β_0 for the flat prior case), with the informative priors case being far superior (2 to 3 times smaller for β_0, β_x and β_z, and roughly 10 times smaller for β_{xz}).

Turning to the medium sample sizes, the informative priors approach has the smallest MSE for all parameters, with the flat priors approach being the next smallest for all parameters but β_0, though the flat priors results are close to those of the three likelihood approaches.

Finally, for the large sample sizes, the three likelihood approaches have slightly lower MSE than the Bayes flat prior results, with the Bayes informative priors analysis giving the lowest MSE for all parameters but β_z.

For all three sample sizes, the Bayesian analysis with informative priors has lower variance than the flat priors case, however the pattern is less clear in terms of bias. For β_x and β_z, the flat priors approach has lower bias, while for β_0 and β_{xz}, the informative priors approach has lower bias. Compared to the likelihood approaches, the Bayes estimates were slightly more biased, however the variances were generally much smaller, resulting in smaller MSEs.

3.4.2 Sparse Data Example

To illustrate the potential advantages of the Bayesian approach, we present an analysis of a single dataset for which the NPML approach fails. The population data were generated in the same way as for the population of the simulation study described in Section 3.4.1, where the total number of subjects in the phase I population for the simulation is 1500. Table 3.3 displays the unobserved population data, while Table 3.4 displays the phase II data. All favourable IH cases and samples of the favourable IH non-cases and unfavourable IH cases and non-cases were used as the phase II sample. We picked a realization of the data in which a zero count resulted in the phase II data. Due to the zero count in the upper right corner of this table, the NPML estimate is not available, although since we are fitting a saturated logistic regression model, the pseudo- and non-parametric maximum likelihood estimates are identical, as in the simulation study of Section 3.4.1.

Table 3.3: Number of Wilms tumour cases and non-cases by institutional histology (IH) and central histology (CH) in the simulated data: full data.

	Favourable IH		Unfavourable IH		
	Non-cases	Cases	Non-cases	Cases	
Favourable CH	1171	128	25	3	
Unfavourable CH	24	19	75	55	
Total	1195	147	100	58	1500

A normal prior distribution with a large variance was used for β_0, while more informative but realistic priors were used for β_x, β_z and β_{xz}. Specifically, we assigned independent normal prior distributions with variance $\log(5)/1.96$, which corresponds to a 95% prior interval of $(-5, 5)$ for each of the log odds ratios. We ran the chain for a total of 200,000 iterations, where the first 50,000 were used to tune the Markov chain, with a constant of $c = 0.77$ giving an acceptance rate of approximately 30%. This analysis took 9 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM.

Results from the complete data given in Table 3.3 are shown in the second column of

Table 3.4: Number of Wilms tumour cases and non-cases by institutional histology (IH) and central histology (CH) in the simulated data: phase II data.

	Favourable IH		Unfavourable IH		
	Non-cases	Cases	Non-cases	Cases	
Favourable CH	295	128	11	0	
Unfavourable CH	5	19	39	50	
Total	300	147	50	50	547

Table 3.5, while results from fitting the phase II data alone are shown in the third column. The parameter estimates and 95% credible intervals for the Bayesian two-phase approach are displayed in column 6 of Table 3.5. Analysis of the phase II data only yields spurious results for β_z and β_{xz}, due to the zero count. Further, there is not much of an improvement in these estimates when the weighted likelihood or pseudo-likelihood approach is used (results displayed in columns 4 and 5 of Table 3.5). In all three cases, the MLE estimates converge to boundary values at $\pm\infty$, however because of machine precision limitations, they appear to converge to large finite values. Appendix A.2 displays the trace plots for this example, with the posterior samples thinned by 100. Based on $75,000$ samples from the posterior distribution, the effective number of independent samples is between $2,000$ and $5,000$ for the seven $\boldsymbol{\lambda}$ parameters. Convergence was assessed through visual inspection of trace plots by running three chains with different starting values.

We compare the results of the Bayesian two-phase analysis to those with fitting a logistic regression model to the complete data. The posterior means and 95% credible intervals obtained from the Bayesian analysis for the log odds parameters, $\boldsymbol{\beta}$, and the nuisance log-linear parameters, λ^X, λ^Z and λ^{XZ}, are displayed in Figure 3.2, along with summaries from fitting a logistic model to the full phase I data. The Bayes results agree well with the full data estimates, with the exception of β_z and β_{xz}. Figure 3.3 displays some aspects of the dependence in the posterior via scatterplot representations of the bivariate posterior marginal distributions, which shows strong negative dependence between β_z and β_{xz}. The

Table 3.5: Estimates and 95% intervals from five analyses of the sparse data example. NPML results are unavailable due to the zero count in the phase II data. True values: $\beta_0 = -2.16$, $\beta_x = 1.59$, $\beta_z = 0.15$, $\beta_{xz} = 0.17$.

	Complete Data	Phase II Sample	WL	PL	Bayesian Two-Phase
β_0	-2.21	-0.83	-2.22	-2.21	-2.21
	(-2.40, -2.03)	(-1.04, -0.63)	(-2.40,-2.03)	(-2.39, -2.03)	(-2.39, -2.03)
β_x	1.98	2.17	2.17	2.17	1.76
	(1.35, 2.61)	(1.16, 3.18)	(1.16,3.18)	(1.16,3.18)	(1.02, 2.55)
β_z	0.09	-14.7	-9.69	-9.89	-0.33
	(-1.12, 1.30)	(-874.8, 845.3)	(-9.96,-9.41)	(-126.3,106.5)	(-1.39, 0.67)
β_{xz}	-0.17	13.6	9.44	9.64	0.45
	(-1.57, 1.23)	(-846.4, 873.7)	(8.34,10.54)	(-106.8,126.0)	(-0.69, 1.59)

zero entry is modeled by $\exp(\beta_z)$ and so the latter estimate is low, while the "large" entry of 50 in the phase II data (see Table 3.4) is modeled by $\exp(\beta_z + \beta_{xz})$ and so the interaction parameter, β_{xz}, is correspondingly overestimated. In general, the Bayesian credible intervals are wide but contain the true values of each parameter. Hence, Bayesian inference is reliable in this example and may be useful in other sparse data situations.

3.5 Examples

3.5.1 Wilms Tumour Data

We return to the Wilms tumour data introduced in Section 2.8.1 in which simple random sampling is assumed at phase I. Recall that the phase II data was obtained by selecting all the Wilms tumour cases, all unfavourable institutional histology non-cases and approximately 10% of the favourable institutional histology non-cases. The full population and phase II data are provided in Tables 2.3 and 2.4. Using a disease model that includes main effects for stage and unfavourable central histology, as well as their interaction, we compare the NPML approach with the Bayesian two-phase approach.

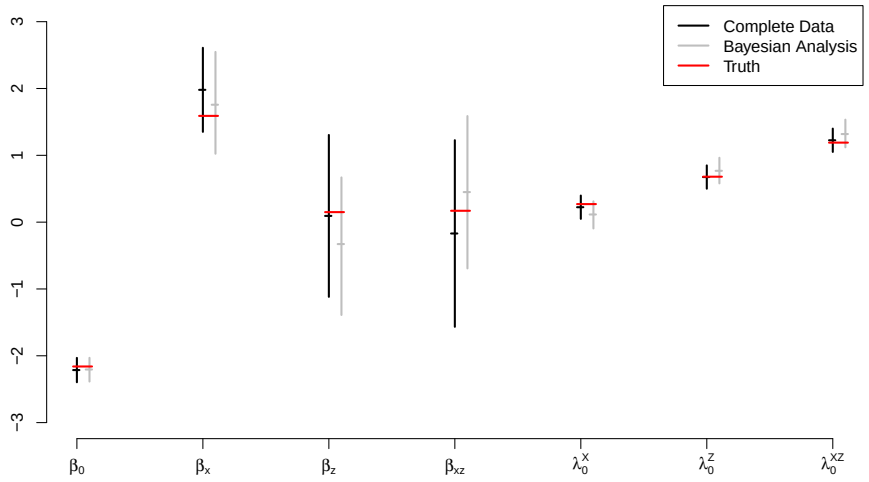


Figure 3.2: Comparison of estimates and 95% intervals of log odds ratios $\boldsymbol{\beta}$ and log-linear parameters $\boldsymbol{\lambda}$, for the sparse data example.

The Bayesian approach assigned independent zero mean normal prior distributions on $\boldsymbol{\beta}$ with large variances and the approach outlined in Section 3.2.2 for the remaining $\boldsymbol{\lambda}$ parameters; the details are contained in Appendix A.3.

The first 100,000 iterations were used to tune the Markov chain, with a constant of $c = 0.32$ giving an acceptance rate of approximately 30%. We then ran the chain for a further 500,000 iterations (though convergence was achieved at around 200,000 iterations). The 600,000 iterations took 50 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM. Appendix A.3 displays the trace plots for the $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ parameters, which displayed no obvious convergence problems. The effective sample sizes for the 23 $\boldsymbol{\lambda}$ parameters are between $1,600$ and $5,300$, based on $300,000$ samples from the posterior distribution. Convergence was assessed through visual inspection of trace plots by running three chains with different starting values.

Results from the complete data are shown in the second column of Table 3.6. The parameter estimates and 95% intervals for the NPML and Bayesian approaches are displayed

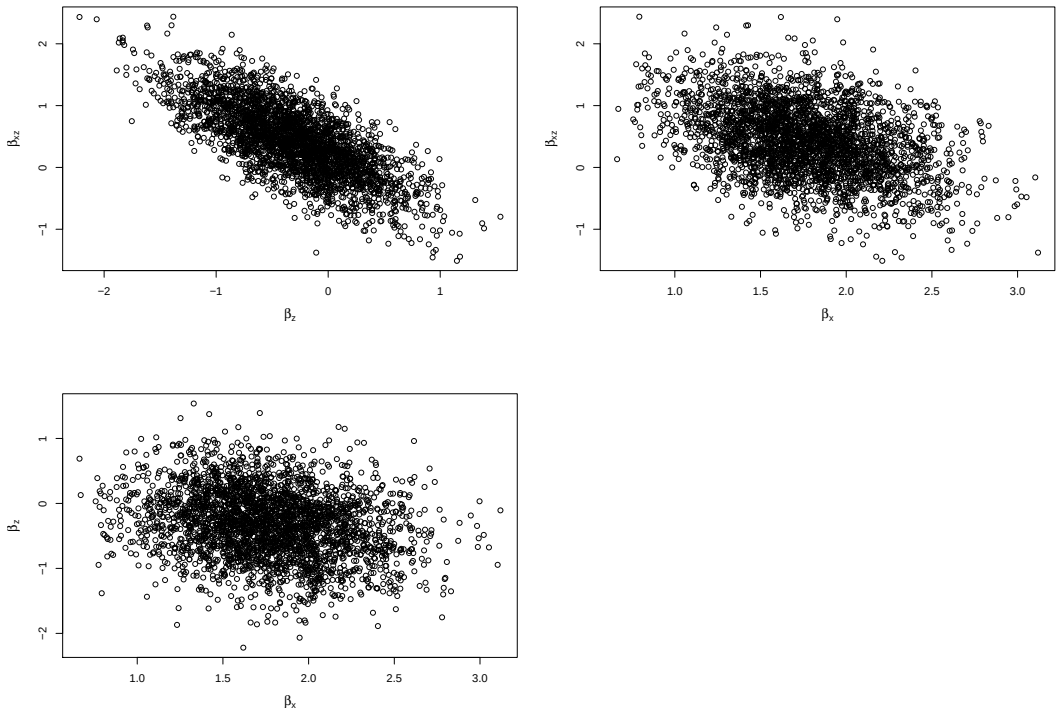


Figure 3.3: Scatterplots of three bivariate posterior distributions in the simulated data example.

Table 3.6: Disease model point and 95% interval estimates for the Wilms tumour relapse data for various methods.

	Complete Data	Phase II Sample	NPML	Bayesian Two-Phase
β_0	-2.71 (-2.92, -2.50)	-0.35 (-0.61, -0.08)	-2.58 (-2.83, -2.32)	-2.58 (-2.84, -2.34)
Stage II	0.77 (0.48, 1.05)	0.55 (0.17, 0.93)	0.55 (0.17, 0.94)	0.56 (0.17, 0.95)
Stage III	0.77 (0.48, 1.07)	0.41 (0.02, 0.79)	0.49 (0.09, 0.88)	0.49 (0.10, 0.88)
Stage IV	1.05 (0.71, 1.39)	0.58 (0.12, 1.03)	0.96 (0.47, 1.46)	0.99 (0.50, 1.47)
UH*	1.31 (0.82, 1.80)	-0.63 (-1.16, -0.09)	1.26 (0.70, 1.82)	1.25 (0.67, 1.80)
Stage II:UH	0.15 (-0.49, 0.78)	0.28 (-0.43, 0.99)	0.29 (-0.47, 1.05)	0.30 (-0.45, 1.09)
Stage III:UH	0.59 (-0.03, 1.21)	0.70 (0.01, 1.39)	0.73 (0.01, 1.46)	0.74 (0.01, 1.50)
Stage IV:UH	1.26 (0.50, 2.02)	1.67 (0.79, 2.55)	1.43 (0.51, 2.36)	1.42 (0.49, 2.39)

*: Unfavourable central histology

in columns 4 and 5 of Table 3.6. The NPML and Bayesian two-phase results are in accordance with the Bayesian credible intervals tending to be very slightly narrower. Compared to the results using the full data, the NPML and Bayesian two-phase approaches slightly underestimate the main effects, while slightly overestimating the interaction terms. Analysis of the phase II data only (column 2) results in extreme bias due to the biased sampling. In particular, the unfavourable central histology parameter estimate from an analysis of the phase II data only results in a negative estimate of $-0.63(-1.16, -0.09)$ compared to an estimate of $1.31(0.82, 1.80)$ from the full data analysis. In addition, there is a large loss of precision, as we would expect. In this example, the NPML and Bayes approach give essentially identical inference.

3.5.2 Case-Control Sampling at Phase I Example

As outlined in Section 3.2, the log-linear framework that we use allows for simple random or case-control sampling at phase I. In this section, we provide an example in which case-control sampling was carried out. The data consist of preliminary information on age, sex, smoking and disease status from a case-control study of residential exposure to radon and

lung cancer (Wacholder and Weinberg, 1994). Random digit dialing was used to identify controls younger than 65 years, while those over 65 were sampled at random from Health Care Financing Administration lists. All cases and a subset of controls were selected for a telephone interview. The controls were selected using a randomized recruitment process where the probabilities of recruitment varied by age and sex. Individuals were classified into one of four smoking categories, depending on their smoking behaviour 10 years prior to the interview: never smoker, ex-smoker, light smoker (fewer than 20 cigarettes per day), or heavy smoker (20 or more cigarettes per day). Age and sex data were available for $12,695$ individuals, which comprises the phase I data, shown in Table 3.7. Smoking status was only known for a much smaller subsample of $4,423$ individuals, which gives the phase II data shown in Table 3.8.

Table 3.7: Phase I data, $N_{y\cdot z}$, for the case-control example.

Age Category	Female		Male	
(years)	Controls	Cases	Controls	Cases
40–59	2738	67	2589	100
60–64	468	45	437	97
65–69	1147	74	1010	123
70–74	1119	55	903	131
75–79	838	40	637	77

Following Wacholder and Weinberg (1994), the disease model included main effects for age and smoking. We performed four analyses of these data: weighted likelihood, pseudo-likelihood, non-parametric maximum likelihood (NPML) and a Bayesian two-phase analysis.

In our notation, y is an indicator variable for lung cancer, z is a categorical variable with levels corresponding to the age-by-sex groups and x is a categorical variable representing smoking behavior. For illustration, we fit a log-linear model using all two-way interactions between age, sex and smoking. We specified independent zero mean normal distributions with large variances for $\boldsymbol{\beta}$. For the remainder of the $\boldsymbol{\lambda}$ nuisance parameters, we followed the

Table 3.8: Phase II data, n_{yxz}, for the case-control example.

		Smoking Status							
		Never		Ex-smoker		Light		Heavy	
Sex	Age Category	Control	Case	Control	Case	Control	Case	Control	Case
	40–59	320	7	26	3	31	15	37	34
	60–64	135	6	10	0	11	10	17	22
Female	65-69	228	7	26	6	12	12	18	34
	70–74	214	6	18	3	12	11	16	28
	75–79	156	8	11	7	11	8	0	7
	40–59	385	6	127	2	47	6	121	73
	60–64	210	2	99	7	22	7	67	65
Male	65–69	240	7	136	18	31	10	75	63
	70–74	289	8	164	25	26	16	62	54
	75–79	195	7	119	17	11	11	25	25

procedure outlined in Section 3.2.2, based on the Knuiman and Speed (1988) and Dellaportas and Forster (1999) formulation. Recall that since we are in the case-control sampling at phase I situation, we cannot estimate λ_y^Y.

Table 3.9 gives the results from the four analyses. For the Bayesian analysis, $150,000$ iterations were run, where the first 50,000 were used to tune the Markov chain with a constant of $c = 0.20$ giving an acceptance rate of approximately 30%. The $150,000$ iterations took 57 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM. Appendix A.4 contains trace plots for the $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ parameters. Convergence appears to be achieved at around $50,000$ iterations. Based on $50,000$ samples from the posterior distribution, the effective number of independent samples is between 200 and 700 for the 34 $\boldsymbol{\lambda}$ parameters. Convergence was assessed through visual inspection of trace plots by running three chains with different starting values.

As expected, for these data with large counts and the use of non-informative priors,

Table 3.9: Estimates and 95% intervals from four analyses of the case-control example investigating the association between lung cancer and smoking; WL = weighted likelihood, PL = pseudo-likelihood, NPML = non-parametric maximum likelihood.

	WL	PL	NPML	Bayesian Two-Phase
Intercept	-5.69 (-6.00, -5.38)	-5.71 (-6.02, -5.41)	-5.73 (-6.03, -5.42)	–
Age 60–64	1.61 (1.32, 1.89)	1.65 (1.37, 1.93)	1.65 (1.36, 1.93)	1.63 (1.35, 1.91)
Age 65–69	1.24 (0.98, 1.50)	1.25 (0.99, 1.51)	1.25 (0.99, 1.51)	1.23 (0.98, 1.49)
Age 70–74	1.41 (1.14, 1.68)	1.46 (1.19, 1.73)	1.47 (1.20, 1.73)	1.43 (1.18, 1.70)
Age 75–79	1.70 (1.39, 2.01)	1.72 (1.41, 2.02)	1.72 (1.41, 2.03)	1.66 (1.37, 1.96)
Ex-smoking	1.76 (1.43, 2.09)	1.75 (1.41, 2.08)	1.77 (1.43, 2.10)	1.76 (1.43, 2.09)
Light smoking	3.04 (2.68, 3.39)	3.06 (2.71, 3.41)	3.07 (2.72, 3.42)	3.07 (2.73, 3.41)
Heavy smoking	3.86 (3.57, 4.16)	3.88 (3.59, 4.17)	3.90 (3.61, 4.19)	3.90 (3.62, 4.20)

there is strong agreement between the likelihood-based and Bayesian analyses, with the Bayesian credible intervals tending to be slightly narrower. Hence, this example provides an instance where the Bayesian method provides reliable estimates and 95% intervals in the phase I case-control sampling situation.

3.5.3 $2 \times 2 \times 2 \times 3 \times 7$ *Contingency Table*

One criticism of our approach is that the number of parameters required to fit the log-linear model (3.8) explodes when m_x or m_z is large, and hence may affect the performance of the Bayesian approach. This example concerns a higher-order table to illustrate the performance of the method under a larger parameter space with respect to $\boldsymbol{\lambda}$, which demonstrates the method's ability to cope in such situations.

The data come from a 1981 survey of employees of a large national corporation and have previously been analyzed by Fowlkes, Freeny, and Landwehr (1988). The aim of the survey was to relate job satisfaction to simple demographic variables as well as to variables that measure job characteristics, such as perceived stress. Table A.5 shows the number of craft employees cross-classified by age, gender, race and region. There are a total of $9,949$ individuals in the data set. We artificially create a two-phase data set and

at phase I, we assume we have information on employees' job satisfaction ($y = 0/1$ for not satisfied/satisfied) and region ($z = 1, \ldots, 7$). The phase I counts are displayed in Table 3.10. At phase II, 100 controls and 200 cases were sampled from each region and information on age, race, and sex were obtained. The phase II data are shown in Table A.6.

As described in Fowlkes et al. (1988), the disease model fit to these data had main effects for age, race, sex and region, as well as an interaction between sex and race; hence, the disease model has 12 parameters. For illustration, we fit the model using all two-way interactions between age, race, sex and region, which gives an additional 39 coefficients. Let A represent age, R race, S sex and Z region. Again, a multivariate normal prior with large variances was assigned for $\boldsymbol{\beta}$ and the Dellaportas and Forster (1999) prescription described in Section 3.2.2 was used for the remaining $\boldsymbol{\lambda}$ nuisance parameters.

In Table 3.11, we report the estimates for $\boldsymbol{\beta}$ based on the complete data, and NPML and the Bayesian approach applied to the two-phase data. In the Bayesian approach, the first 50,000 iterations were used to tune the Markov chain, with a constant of $c = 0.17$ giving an acceptance rate of approximately 30%. We then ran the chain for a further 200,000 iterations. The 250,000 iterations took 138 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM. Trace plots for the $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ parameters are contained in Appendix A.5. Convergence appears to be acheived around $100,000$ iterations. There are four $\boldsymbol{\lambda}$ parameters which converge less well than the others, namely the two $\boldsymbol{\lambda}^{AR}$ and the two $\boldsymbol{\lambda}^{AS}$ parameters. The effective sample sizes for these four $\boldsymbol{\lambda}$ parameters are between 45 and 80, based on $100,000$ samples from the posterior. For the remaining 47 $\boldsymbol{\lambda}$ parameters, the effective number of independent samples is between 200 and $1,300$. Convergence was assessed through visual inspection of trace plots by running three chains with different starting values.

Again, we see strong correspondence between the NPML and Bayesian approaches, illustrating that the Bayesian method is able to cope with the relatively high dimension of the $\boldsymbol{\lambda}$ parameter space. While the Bayesian credible intervals tended to be slightly narrower than the NPML confidence intervals, both procedures provide essentially identical inference in this example. Further, we see that inference based on the two-phase data is quite close to that based on the complete data, showing the efficiency of the two-phase design.

Table 3.10: Phase I Data, $N_{y\cdot z}$, for the $2 \times 2 \times 2 \times 3 \times 7$ dimensional table example.

	Northwest ($Z = 1$)	Mid-Atlantic ($Z = 2$)	Southern ($Z = 3$)	Midwest ($Z = 4$)	Northwest ($Z = 5$)	Southwest ($Z = 6$)	Pacific ($Z = 7$)
Satisfied	1161	406	916	1240	1221	971	462
Not satisfied	738	166	514	749	711	482	209

Table 3.11: Estimates and 95% intervals for $2 \times 2 \times 2 \times 3 \times 7$ dimensional table example.

	Complete Data	NPML	Bayesian Two-Phase
β_0	-0.01 (-0.18, 0.17)	-0.02 (-0.36, 0.32)	-0.02 (-0.38, 0.32)
Age $35 - 44$	0.13 (0.03, 0.23)	0.20 (-0.02, 0.43)	0.20 (-0.02, 0.42)
Age > 44	0.36 (0.26, 0.46)	0.42 (0.19, 0.65)	0.42 (0.20, 0.65)
Region Mid-Atlantic	0.44 (0.23, 0.64)	0.42 (0.21, 0.62)	0.42 (0.22, 0.62)
Region Southern	0.18 (0.03, 0.32)	0.19 (0.04, 0.34)	0.19 (0.03, 0.34)
Region Midwest	0.09 (-0.04, 0.22)	0.06 (-0.07, 0.20)	0.06 (-0.07, 0.20)
Region Northwest	0.12 (-0.01, 0.26)	0.13 (-0.01, 0.28)	0.13 (-0.01, 0.27)
Region Southwest	0.29 (0.14, 0.43)	0.27 (0.12, 0.42)	0.28 (0.12, 0.43)
Region Pacific	0.41 (0.22, 0.60)	0.42 (0.23, 0.62)	0.44 (0.24, 0.63)
Male	0.49 (0.27, 0.71)	0.37 (-0.09, 0.83)	0.31 (-0.16, 0.77)
White Race	0.21 (0.04, 0.39)	0.10 (-0.29, 0.48)	0.11 (-0.31, 0.51)
Male:White Race	-0.38 (-0.62, -0.14)	-0.12 (-0.63, 0.40)	-0.06 (-0.58, 0.44)

3.6 Summary

In this chapter, we have described a Bayesian approach to the analysis of data arising from two-phase study designs. The proposed approach uses a log-linear model to model the disease-exposure-confounder relationship, and introduces the unobserved population disease-exposure-confounder counts as auxiliary variables to facilitate the derivation of the posterior distribution following Holubkov (1995). Due to the potentially large dimensionality of the confounder/exposure space, we follow the specification proposed by Knuiman and Speed (1988), and refined by Dellaportas and Forster (1999), where multivariate normal prior distributions are assigned to the collections of main effect and interaction terms in the log-linear model. Sampling from the posterior distribution proceeds via a Markov chain Monte Carlo scheme in which the complete vector of parameters from the log-linear model is updated via a single Metropolis-Hastings step using a multivariate random walk based on a normal proposal.

The data examples of Section 3.5 compare the Bayesian approach to the NPML approach proposed by Breslow and Holubkov (1997a), where the results agree reasonably well in all cases. Additionally, the results of the simulation study of Section 3.4 suggest that the Bayesian parameter estimates have lower mean squared error than those obtained from the likelihood approaches, however they also tend to have some finite-sample biases.

The Bayesian approach we have outlined requires a far greater level of model specification than the NPML approach, namely an exposure-confounder model. This is a potentially major drawback and in situations in which the data are numerous and there are no dependencies that need to be modeled (via the use of random effects, for example), we would recommend the NPML approach. However, for the situation in which the covariates are all discrete and the data are sparse, the Bayesian approach is beneficial. The $2 \times 2 \times 2 \times 3 \times 7$ contingency table example from Section 3.5 provided an example in which the dimensionality of the covariate space was quite large. If the log-linear model had been fit using all interactions between age, race, sex and region, 82 additional coefficients would have been required rather than the 39 required using only two-way interactions. In this example, we saw that the Bayesian method was able to handle the relatively high dimension of the $\boldsymbol{\lambda}$

parameter space, and gave essentially identical inference as the NPML approach. Hence, in situations in which x and/or z take on many values, it may be beneficial to avoid the saturated log-linear model (3.8), which will contain many parameters, in favour of a simpler form. Random effects models can also be used, as we demonstrate in the next chapter.

Despite the complexities associated with the full probability modeling approach described here, a motivation for this approach is that random effects may be introduced to account for dependencies in the data, which we describe in Chapter 4.

Chapter 4

BAYESIAN RANDOM EFFECTS ANALYSIS OF TWO-PHASE DATA

In Chapter 3, we described a Bayesian approach to the analysis of data arising from two-phase studies. One benefit of the Bayesian approach is that random effects may be easily introduced, and that is the extension we describe in this chapter. There are two ways in which we might approach the addition of random effects into our model. The first is modeling the exposure-confounder relationship using random effects to smooth the fitted cell probabilities in large contingency tables, particularly in the case of sparse data. We describe this approach in Section 4.2. We can also include random effects into the disease model to smooth across geographic areas, for example. This is discussed in Section 4.3. These methods are illustrated using the North Carolina infant mortality data example, described in Section 4.4.

4.1 The Likelihood

For both simple random and case-control sampling at phase I, the likelihood in the random effects case is of the same form as in the fixed effects case. Recall from Section 3.2.1 that in the case of simple random sampling at phase I, the likelihood is

$$\text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{\cdots}, \boldsymbol{\lambda}) \quad \propto \quad \text{pr}(\boldsymbol{N}^{\text{y·z}} | N_{\cdots}, \boldsymbol{\lambda}) \times \text{pr}(\boldsymbol{n}^{\text{yxz}} | \boldsymbol{n}^{\text{y·z}}, \boldsymbol{\lambda}),$$

where

$$\boldsymbol{N}^{\text{y·z}} | N_{\cdots}, \boldsymbol{\lambda} \sim \text{Multinomial}_{2 \times m_z} \left(N_{\cdots}, \{\text{pr}(Y = y, Z = z)\}_{y=0,1; 1 \le z \le m_z} \right),$$

and

$$\boldsymbol{n}^{\text{yxz}} | \boldsymbol{n}^{\text{y·z}}, \boldsymbol{\lambda} \sim_{\text{ind}} \text{Multinomial}_{m_x} \left(n_{y \cdot z}, \{\text{pr}(X = x | Z = z, Y = y)\}_{1 \le x \le m_x} \right), \quad y = 0, 1; 1 \le z \le m_z,$$

for $\text{pr}(Y = y, Z = z) = \sum_{x=1}^{m_x} r_{yxz}$ and $\text{pr}(X = x | Z = z, Y = y) = \frac{r_{yxz}}{\sum_{v=1}^{m_x} r_{yvz}}$.

In the case-control sampling at phase I situation, the likelihood is

$$\text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{0\cdot\cdot}, N_{1\cdot\cdot}, \boldsymbol{\lambda}) \quad \propto \quad \text{pr}(\boldsymbol{N}^{\text{y·z}} | N_{1\cdot\cdot}, N_{0\cdot\cdot}, \boldsymbol{\lambda}) \times \text{pr}(\boldsymbol{n}^{\text{yxz}} | \boldsymbol{n}^{\text{y·z}}, \boldsymbol{\lambda}),$$

where

$$\boldsymbol{N}^{\text{y·z}} | N_{1\cdot\cdot}, N_{0\cdot\cdot}, \boldsymbol{\lambda} \sim_{\text{ind}} \text{Multinomial}_{m_z} \left(N_{y\cdot\cdot}, \{\text{pr}(Z = z | Y = y)\}_{1 \le z \le m_z} \right), y = 0, 1,$$

and

$$\boldsymbol{n}^{\text{yxz}} | \boldsymbol{n}^{\text{y·z}}, \boldsymbol{\lambda} \sim_{\text{ind}} \text{Multinomial}_{m_x} \left(n_{y\cdot z}, \{\text{pr}(X = x | Z = z, Y = y)\}_{1 \le x \le m_x} \right), \\ y = 0, 1; 1 \le z \le m_z,$$

for $\text{pr}(Z = z | Y = y) = \sum_{x=1}^{m_x} r_{xz|y}$ and $\text{pr}(X = x | Z = z, Y = y) = \frac{r_{xz|y}}{\sum_{v=1}^{m_x} r_{vz|y}}$.

The probabilities r_{yxz} and $r_{xz|y}$ are calculated using the log-linear model that we specify, defined in terms of μ_{yxz}, the mean of the disease-exposure-confounder count in cell (y, x, z) of the $2 \times m_x \times m_z$ contingency table:

$$\begin{aligned} r_{yxz} &= \frac{\mu_{yxz}}{\sum_u \sum_v \sum_w \mu_{uvw}} \\ r_{xz|y} &= \frac{\mu_{yxz}}{\sum_v \sum_w \mu_{yvw}}. \end{aligned}$$

As described below, the particular form of the log-linear model changes depending on the kind of smoothing we would like to perform. We now discuss smoothing across the cells of the contingency table.

4.2 Random Exposure-Confounder Relationship

As discussed in Chapter 3, using a saturated log-linear model is not always desirable, particularly in situations in which x and/or z take on many values. For example, in the large contingency table example considered in Section 3.5.3, we fit a log-linear model using all two-way interactions between the exposure and confounder variables rather than a saturated exposure-confounder model. We may sometimes wish to model the x–z margin using random effects for smoothing purposes. Hence, another alternative to the saturated

exposure-confounder model is to introduce random effects into the log-linear model by modeling the exposure-confounder relationship using random effects:

$$\log(\mu_{yxz}) = \mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + V_{xz}^{XZ}, \tag{4.1}$$

with the usual sum-to-zero constraints imposed for identifiability. We use the notation V_{xz}^{XZ} to denote the random λ_{xz}^{XZ} terms. (Note that in addition to $\boldsymbol{\lambda}^{XZ}$, $\boldsymbol{\lambda}^{X}$, $\boldsymbol{\lambda}^{Z}$ and/or $\boldsymbol{\lambda}^{YXZ}$ can be modeled as random effects in an analogous manner.) The $V_{xz}^{XZ}, x = 1, \ldots, m_x - 1$,$z = 1, \ldots, m_z - 1$ are assigned independent normal distributions with mean zero and variance σ_λ^2. Note that since we place proper prior distributions on $\boldsymbol{V}^{XZ}$, imposing the sum-to-zero constraints is not strictly necessary. The sum-to-zero constraints are imposed on both margins of $\boldsymbol{V}^{XZ}$ so that the interpretations of the parameters are the same as in the frequentist analysis. This also improves convergence of the Markov chains. Letting $\tau_\lambda = \sigma_\lambda^{-2}$ denote the precision, we adopt a conjugate gamma hyperprior distribution with shape parameter a_λ and rate parameter b_λ for τ_λ:

$$\begin{aligned} \boldsymbol{V}_{xz}^{XZ} | \tau_\lambda &\sim N(0, \tau_\lambda^{-1}) \\ \tau_\lambda &\sim G(a_\lambda, b_\lambda), \end{aligned}$$

where a_λ and b_λ are suitably chosen constants based on the context. The choice of the gamma distribution as the hyperprior for the precision variable is convenient as the marginal distribution of $\boldsymbol{V}_{xz}^{XZ}$ is a Student's t-distribution with $2a_\lambda$ degrees of freedom, location zero, and scale b_λ/a_λ. Further, by specifying the degrees of freedom and the range within which the $\boldsymbol{V}_{xz}^{XZ}$ lie with probability q, we can easily compute the corresponding values for a_λ and b_λ according to (2.14). Details are provided in Section 2.6.2. As we will see in Section 4.4.4, the particular choice of a_λ and b_λ makes little difference on the posterior estimates for $\boldsymbol{\lambda}$ when the phase II sample sizes are large enough. However, with small phase II sample sizes, the posterior estimates for $\boldsymbol{\lambda}$ can be quite sensitive to the choice of hyperprior distribution parameters, and hence they should be chosen with care.

Including uncorrelated random effects in a log-linear model as in (4.1) does not affect the correspondence between $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ})$ and $\boldsymbol{\beta}$. However, the prior we assign to $\boldsymbol{V}^{XZ}$ does imply a certain correlation structure for the exposure-confounder log odds ratios

(Coull and Agresti, 2003). Let $\delta_{(y)xz}$ denote the exposure-confounder odds ratio for disease status y, and let $\boldsymbol{\delta}^{(y)XZ} = (\delta_{(y)22}, \ldots, \delta_{(y)m_x m_z})$. Then, in terms of the log-linear model (4.1),

$$\begin{aligned} \log \delta_{(y)xz} &= \log\left(\frac{\mu_{yxz}\mu_{y11}}{\mu_{yx1}\mu_{y1z}}\right) \\ &= V_{xz}^{XZ} + V_{11}^{XZ} - V_{x1}^{XZ} - V_{1z}^{XZ} \end{aligned}$$

for $x = 2, \ldots, m_x, z = 2, \ldots, m_z$, where

$$V_{m_x z}^{XZ} = -\sum_{x=1}^{m_x - 1} V_{xz}^{XZ}$$

$$V_{x m_z}^{XZ} = -\sum_{z=1}^{m_z - 1} V_{xz}^{XZ}.$$

Then, we have $\boldsymbol{\delta}^{(y)XZ} = \boldsymbol{B}\boldsymbol{V}^{XZ}$, where $\boldsymbol{B}$ is an $(m_x - 1)(m_z - 1)$ square transformation matrix. The independent normal priors on $\boldsymbol{V}^{XZ}$ induce a multivariate normal distribution on $\boldsymbol{\delta}^{(y)XZ}$ with mean zero and variance-covariance matrix $\tau_\lambda^{-1}\boldsymbol{B}\boldsymbol{B}^T$. For example, let $m_x = 3$ and $m_z = 3$. Then, $\boldsymbol{V}^{XZ} = (V_{11}^{XZ}, V_{21}^{XZ}, V_{12}^{XZ}, V_{22}^{XZ})$ and

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 2 & 1 & -2 & 1 \\ 2 & -2 & 1 & 1 \\ 4 & 2 & 2 & 1 \end{pmatrix}.$$

Then, the induced multivariate normal distribution on $\boldsymbol{\delta}^{(y)XZ}$ is

$$\boldsymbol{\delta}^{(y)XZ} \sim N_4\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tau_\lambda^{-1}\begin{pmatrix} 4 & 4 & 4 & 1 \\ 4 & 10 & 1 & 7 \\ 4 & 1 & 10 & 7 \\ 1 & 7 & 7 & 25 \end{pmatrix}\right).$$

As in the case of fixed $\boldsymbol{\lambda}^{XZ}$ terms, we would like to base our inference on the $\boldsymbol{\beta}$ parameters from the disease model (3.17). We can assign a multivariate normal prior distribution for $\boldsymbol{\beta}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, which induces a multivariate normal prior on $\boldsymbol{\Lambda}^Y$ with mean $\boldsymbol{C}^{-1}\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$. As discussed in Section 3.2.2, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are chosen based on the context. We assign $\boldsymbol{\lambda}^X$ and $\boldsymbol{\lambda}^Z$ multivariate normal priors following the Dellaportas and Forster (1999) prescription.

4.2.1 Computation

The $\boldsymbol{V}^{XZ}$ vector can be updated together with the $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^{Y}, \boldsymbol{\lambda}^{X}, \boldsymbol{\lambda}^{Z}, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ})$ vector, or in a separate Metropolis-Hastings step via a multivariate random walk. In the former case, the Markov chain cycles between the conditional distributions

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{V}^{XZ} | N_{...}) &\propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{...}, \boldsymbol{\lambda}, \boldsymbol{V}^{XZ}) \times \pi(\boldsymbol{\lambda}) \pi(\boldsymbol{V}^{XZ} | \tau_{\lambda}) \\ p(\tau_{\lambda} | \boldsymbol{V}^{XZ}) &\sim G\left(a_{\lambda} + \frac{(m_x - 1)(m_z - 1)}{2}, b_{\lambda} + \frac{1}{2} \sum_{x=1}^{m_x - 1} \sum_{z=1}^{m_z - 1} \left(V_{xz}^{XZ} \right)^2 \right), \end{aligned} \quad (4.2)$$

and in the latter case, the chain iterates between

$$\begin{aligned} p(\boldsymbol{\lambda} | \boldsymbol{V}^{XZ}, N_{...}) &\propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{...}, \boldsymbol{\lambda}, \boldsymbol{V}^{XZ}) \times \pi(\boldsymbol{\lambda}) \\ p(\boldsymbol{V}^{XZ} | \boldsymbol{\lambda}, N_{...}) &\propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}} | N_{...}, \boldsymbol{\lambda}, \boldsymbol{V}^{XZ}) \times \pi(\boldsymbol{V}^{XZ} | \tau_{\lambda}) \\ p(\tau_{\lambda} | \boldsymbol{V}^{XZ}) &\sim G\left(a_{\lambda} + \frac{(m_x - 1)(m_z - 1)}{2}, b_{\lambda} + \frac{1}{2} \sum_{x=1}^{m_x - 1} \sum_{z=1}^{m_z - 1} \left(V_{xz}^{XZ} \right)^2 \right). \end{aligned}$$

In the examples considered in this thesis, we found that updating $\boldsymbol{\lambda}$ and $\boldsymbol{V}^{XZ}$ separately allowed for better control of acceptance probabilities and movement around the sample space for both parameters.

We take the proposal distribution for $\boldsymbol{V}^{XZ}$ as normal, centered at the current value with variance-covariance matrix a constant, c_{λ}, times a diagonal matrix with entries τ_{λ}^{-1}. Sampling the $\boldsymbol{\lambda}$ vector proceeds exactly as described in Section 3.2.2. As before, the constants are chosen to achieve acceptance rates of approximately 30% (Roberts et al., 1997).

We now discuss the inclusion of random effects into the disease model.

4.3 Random Disease Model

Assume that the confounding variable z now represents areas, of which there are m_z. We begin by describing a non-spatial random effects disease model.

4.3.1 Non-spatial Random Effects

The disease model (3.17) can be extended to include independent random intercept terms for area via

$$\log \left[\frac{\text{pr}(Y = 1 \mid X = x, Z = z)}{\text{pr}(Y = 0 \mid X = x, Z = z)} \right] = \boldsymbol{W}\boldsymbol{\beta} + V_z,$$

where V_z are random effects without spatial structure. Without loss of generality, we suppose there is simple random sampling of the phase I data. (The scenario of case-control sampling at phase I follows in a straightforward fashion.) In terms of the log-linear parameterization, we have

$$\log(\mu_{yxz}) = \mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} \tag{4.3}$$

with the set of constraints

$$\begin{aligned} \sum_y \lambda_y^Y = \sum_x \lambda_x^X = \sum_z \lambda_z^Z = 0, \\ \sum_y \lambda_{yx}^{YX} = \sum_x \lambda_{yx}^{YX} = 0, \\ \sum_y \lambda_{yz}^{YZ} = 0, \\ \sum_x \lambda_{xz}^{XZ} = \sum_z \lambda_{xz}^{XZ} = 0, \end{aligned}$$

where we place no constraint on $\sum_z \lambda_{yz}^{YZ}$. As for $\boldsymbol{V}^{XZ}$, imposing the sum-to-zero constraints is not strictly necessary since we assign proper priors on $\boldsymbol{\lambda}^{YZ}$. We impose the sum-to-zero constraints on the y margin of $\boldsymbol{\lambda}^{YZ}$ only so that the interpretations of the parameters are the same as in the frequentist analysis. As before, this improves the convergence of the Markov chains.

As described in Section 3.1.1, the $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ parameters are directly related:

$$\begin{aligned} \beta_0 &= -2(\lambda_0^Y + \lambda_{01}^{YX}) \\ \beta_x &= -2(\lambda_{0x}^{YX} - \lambda_{01}^{YX}) \end{aligned}$$

for $x = 2, \ldots, m_x$ and

$$V_z = -2\lambda_{0z}^{YZ},$$

for $z = 1, \ldots, m_z$. Hence, we have the same relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ as in the case of no confounder variables, where $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\Lambda}^Y$, for $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX})$ and $\boldsymbol{C}$ the $m_x \times m_x$ invertible matrix given in (3.7). As in the independent data case, we perform computations for $\boldsymbol{\lambda}$ first, and then transform to $\boldsymbol{\beta}$ using $\boldsymbol{C}$.

Prior Specification

We specify prior distributions for $\boldsymbol{\Lambda}^Y$, $\boldsymbol{\lambda}^X$, $\boldsymbol{\lambda}^Z$, $\boldsymbol{\lambda}^{XZ}$, $\boldsymbol{V} = (V_1, \ldots, V_{m_z})$ and τ_v, the precision of the random effects, where we assume

$$\pi(\boldsymbol{\Lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{XZ}, \boldsymbol{V}|\tau_v) = \pi(\boldsymbol{\Lambda}^Y)\pi(\boldsymbol{\lambda}^X)\pi(\boldsymbol{\lambda}^Z)\pi(\boldsymbol{\lambda}^{XZ})\pi(\boldsymbol{V}|\tau_v).$$

As described in Chapter 3, we assume a multivariate normal prior distribution for $\boldsymbol{\beta}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are chosen based on the context. Then, the induced multivariate normal prior for $\boldsymbol{\Lambda}^Y$ has mean $\boldsymbol{C}^{-1}\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$.

Depending on the context, we may want to consider $\boldsymbol{\lambda}^{XZ}$ as fixed or random. If we view $\boldsymbol{\lambda}^X$, $\boldsymbol{\lambda}^Z$, and $\boldsymbol{\lambda}^{XZ}$ as fixed effects, then we can follow the approach of Dellaportas and Forster (1999) as described in Section 3.2.2. However, for those situations where we view $\boldsymbol{\lambda}^{XZ}$ as random, we follow the procedure outlined in Section 4.2 and assign $\boldsymbol{\lambda}^{XZ}$ independent normal prior distributions with gamma hyperpriors, and denote them $\boldsymbol{V}^{XZ}$. In this case, $\boldsymbol{\lambda}^X$ and $\boldsymbol{\lambda}^Z$ would be assigned the usual multivariate normal prior distributions of Dellaportas and Forster (1999).

For $\boldsymbol{V}$, we adopt a similar strategy as described for $\boldsymbol{V}^{XZ}$ in Section 4.2. We place independent normal prior distributions with zero mean and variance τ_v^{-1} on each V_z, $z = 1, \ldots, m_z$, and assign a gamma hyperprior distribution with shape parameter a_v and rate parameter b_v on the precision τ_v:

$$\begin{aligned} V_z|\tau_v &\sim N(0, \tau_v^{-1}) \\ \tau_v &\sim G(a_v, b_v), \end{aligned}$$

where, again, a_v and b_v are suitably chosen constants based on the context.

Computation

We construct a Markov chain and update the parameters using separate Metropolis-Hastings steps for $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{XZ})$ and $\boldsymbol{V}$, and a Gibbs step for τ_v. The Markov chain

cycles between the conditional distributions:

$$p(\boldsymbol{\lambda}|\boldsymbol{V}, N_{...}) \propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}}|N_{...}, \boldsymbol{\lambda}, \boldsymbol{V}) \times \pi(\boldsymbol{\lambda}) \tag{4.4}$$

$$p(\boldsymbol{V}|\boldsymbol{\lambda}, \tau_v, N_{...}) \propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}}|N_{...}, \boldsymbol{\lambda}, \boldsymbol{V}) \times \pi(\boldsymbol{V}|\tau_v)$$

$$p(\tau_v|\boldsymbol{V}) \sim G\left(a_v + \frac{m_z}{2}, b_v + \frac{1}{2}\sum_{i=1}^{m_z} V_i^2\right). \tag{4.5}$$

For $\boldsymbol{\lambda}^{XZ}$ fixed, we follow the procedure described in Section 3.2.3, where the Metropolis-Hastings step for $\boldsymbol{\lambda}$ is a multivariate normal random walk based on a normal proposal. When we consider $\boldsymbol{\lambda}^{XZ}$ to be random, we propose values for $\boldsymbol{V}^{XZ}$ using a normal distribution centered on the current value with variance-covariance matrix diagonal with entries $c_\lambda \tau_\lambda^{-1}$. As described in Section 4.2, the $\boldsymbol{V}^{XZ}$ parameters can also be updated simultaneously with $\boldsymbol{\lambda}$.

The Metropolis-Hastings step for $\boldsymbol{V}$ is performed via block updating (that is, updating all parameters at the same time), using a random walk based on a multivariate normal proposal centered around the current value of $\boldsymbol{V}$ with variance-covariance matrix taken to be a constant, c_v, times a diagonal matrix Σ with entries τ_v^{-1}. The constants c_λ and c_v are chosen to achieve acceptance rates of approximately 30% (Roberts et al., 1997). We sample τ_v from its full conditional distribution (4.5). If $\boldsymbol{\lambda}^{XZ}$ is treated as random, τ_λ is sampled from its full conditional distribution (4.2).

We now discuss the disease model in which we supplement the independent random effects with spatial terms.

4.3.2 Spatial Random Effects

In the extension to spatial random effects, the disease model (3.17) now includes two different random effect terms:

$$\log\left[\frac{\text{pr}(Y=1 \mid X=x, Z=z)}{\text{pr}(Y=0 \mid X=x, Z=z)}\right] = \boldsymbol{W}\boldsymbol{\beta} + U_z + V_z,$$

where U_z and V_z are random effects with and without spatial structure, respectively. In terms of the log-linear parameterization, we would fit the same model with an identical set of constraints as in the non-spatial model, given in (4.3). The correspondence between $\boldsymbol{\beta}$

and $\boldsymbol{\lambda}$ will remain the same, however and we now have

$$U_z + V_z = -2\lambda_{0z}^{YZ},$$

for $z = 1, \ldots, m_z$.

The prior distributions for $\boldsymbol{\lambda}$ and $\boldsymbol{V}$ are unchanged from the non-spatial model, and are as given in Section 4.3.1. Let $\boldsymbol{U} = (U_1, \ldots, U_{m_z})$. We assume that $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent. We adopt the intrinsic conditional autoregression (ICAR) model for $\boldsymbol{U}$ introduced by Besag et al. (1991) and described in Section 2.6:

$$\begin{aligned} \pi(\boldsymbol{U} \mid \tau_u) &\propto \tau_u^{(m_z-1)/2} \exp\left(-\frac{\tau_u}{2}\sum_{i\sim j}(U_i - U_j)^2\right) \\ &= \tau_u^{(m_z-1)/2} \exp\left(-\frac{\tau_u}{2}\boldsymbol{U}\boldsymbol{K}\boldsymbol{U}\right), \end{aligned}$$

where $i \sim j$ denotes all pairs of neighbouring areas i and j, and $\boldsymbol{K}$ denotes the matrix representing the neighbourhood structure of areas z. The definition of neighbours that we adopt is that based on contiguity. For the precision parameter τ_u, we adopt the conjugate gamma prior Gamma(a_u, b_u), where a_u and b_u are suitably chosen constants based on the context.

The updating of the parameters proceeds via a Markov chain, which cycles between the conditional distributions:

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{V}, \boldsymbol{U}, N_{\ldots}) &\propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}}|N_{\ldots}, \boldsymbol{\lambda}, \boldsymbol{V}, \boldsymbol{U}) \times \pi(\boldsymbol{\lambda}) \\ p(\boldsymbol{V}|\boldsymbol{\lambda}, \boldsymbol{U}, \tau_v, N_{\ldots}) &\propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}}|N_{\ldots}, \boldsymbol{\lambda}, \boldsymbol{V}, \boldsymbol{U}) \times \pi(\boldsymbol{V}|\tau_v) \\ p(\boldsymbol{U}|\boldsymbol{\lambda}, \boldsymbol{V}, \tau_u, N_{\ldots}) &\propto \text{pr}(\boldsymbol{N}^{\text{y·z}}, \boldsymbol{n}^{\text{y·z}}, \boldsymbol{n}^{\text{yxz}}|N_{\ldots}, \boldsymbol{\lambda}, \boldsymbol{V}, \boldsymbol{U}) \times \pi(\boldsymbol{U}|\tau_u) \\ p(\tau_v|\boldsymbol{V}) &\sim G\left(a_v + \frac{m_z}{2}, b_v + \frac{1}{2}\sum_{i=1}^{m_z} V_i^2\right) \qquad (4.6) \\ p(\tau_u|\boldsymbol{U}) &\sim G\left(a_u + \frac{m_z - 1}{2}, b_u + \frac{1}{2}\sum_{i\sim j}(U_i - U_j)^2\right) \qquad (4.7) \end{aligned}$$

using Metropolis-Hastings steps for $p(\boldsymbol{\lambda}|\boldsymbol{V}, \boldsymbol{U}, N_{\ldots})$, $p(\boldsymbol{V}|\boldsymbol{\lambda}, \boldsymbol{U}, \tau_v, N_{\ldots})$ and $p(\boldsymbol{U}|\boldsymbol{\lambda}, \boldsymbol{V}, \tau_u, N_{\ldots})$ (Tierney, 1994), and Gibbs steps for $p(\tau_v|\boldsymbol{V})$ and $p(\tau_u|\boldsymbol{U})$.

Updating the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{V}$ proceeds exactly as described in Section 4.3.1, where we can, again, view $\boldsymbol{\lambda}^{XZ}$ as fixed or random. Since single-site updating can have poor

convergence and mixing properties in spatial models, the Metropolis-Hastings step for $\boldsymbol{U}$ is performed via block-updating (Knorr-Held and Rue, 2002). We use an improper Gaussian Markov random field (GMRF) with mean the current value of $\boldsymbol{U}$ and precision matrix $c_u\boldsymbol{Q}$ as our proposal distribution, where $\boldsymbol{Q} = \tau_u\boldsymbol{K}$ and the tuning parameter c_u is selected to achieve acceptance rates of approximately 30% (Roberts et al., 1997). As described in Section 2.6.4, sampling from an improper GMRF is relatively straightforward using Algorithm 3.1 of Rue and Held (2005) for sampling from an improper GMRF with mean zero. Finally, we sample τ_v and τ_u from their full conditional distributions, given in (4.6) and (4.7), respectively. In the case where $\boldsymbol{\lambda}^{XZ}$ is treated as random, τ_λ is sampled from its full conditional distribution (4.2).

As we will see in the next chapter, there are situations where we may want to predict the population counts in each disease-exposure-confounder group. The predictions can be done within the MCMC scheme. Once the chains have achieved convergence, the updated vector of $\boldsymbol{\lambda}$, $\boldsymbol{V}$ and $\boldsymbol{U}$ parameters are used to calculate the disease-exposure-confounder cell probabilities, and the $\boldsymbol{N}^{\text{yxz}}$ are sampled from a multinomial distribution using the calculated cell probabilities. We can evaluate the predictions using the bias, variance and mean squared error (MSE). For S draws from the posterior distribution, we have

$$\begin{aligned}
\text{Bias}_{yxz} &= \left(\bar{\hat{N}}_{yxz} - N_{yxz}\right), \quad \bar{\hat{N}}_{yxz} = \frac{1}{S}\sum_{s=1}^{S}\hat{N}_{yxz}^{(s)}, \\
\text{Variance}_{yxz} &= \left(\frac{1}{S}\sum_{s=1}^{S}(\hat{N}_{yxz}^{(s)} - \bar{\hat{N}}_{yxz})^2\right), \\
\text{MSE} &= \sum_{y=0}^{1}\sum_{x=1}^{m_x}\sum_{z=1}^{m_z}\left(\text{Bias}_{yxz}^2 + \text{Variance}_{yxz}\right).
\end{aligned}$$

Estimates with small MSE will be generally preferred.

4.4 Examples

4.4.1 $2 \times 2 \times 2 \times 3 \times 7$ Contingency Table Example

We revisit the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example described in Section 3.5.3. Recall that the outcome of interest is job satisfaction, the confounder variable is region and the

exposure variables are age, race and sex. Because of the high-dimensionality of the $x \times z$ parameter space, we model the $\boldsymbol{\lambda}^{XZ}$ as random. As before, we fit a disease model with main effects for age, race, sex and region, as well as an interaction between sex and race. Let A denote the 3 age categories, R race, S sex and Z region. We assigned independent mean zero normal priors with large variances for β, and we used the Dellaportas and Forster (1999) formulation for $\boldsymbol{\lambda}^{A}, \boldsymbol{\lambda}^{R}, \boldsymbol{\lambda}^{S}, \boldsymbol{\lambda}^{Z}, \boldsymbol{\lambda}^{AR}, \boldsymbol{\lambda}^{AS}, \boldsymbol{\lambda}^{RS}$, and $\boldsymbol{\lambda}^{ARS}$. Details can be found in Appendix B.2.

For $\boldsymbol{V}^{XZ}$, we assigned each of the m_z elements a normal prior distribution with mean zero and precision τ_λ, where we assigned τ_λ a gamma hyperprior with shape parameter 0.5 and rate parameter 0.0005. Using this hyperprior, the (0.025, 0.975) quantiles for σ_λ are (0.014, 1.01). There were 66 $\boldsymbol{V}^{XZ}$ parameters to consider.

Table 4.1: Estimates and 95% intervals of the $\boldsymbol{\beta}$ parameters from three analyses of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example.

	Complete Data	NPML	Bayesian Two-Phase
β_0	-0.01 (-0.18, 0.17)	-0.02 (-0.36, 0.32)	-0.01 (-0.35, 0.34)
Age 35-44	0.13 (0.03, 0.23)	0.20 (-0.02, 0.43)	0.20 (-0.02, 0.43)
Age $>$ 44	0.36 (0.26, 0.46)	0.42 (0.19, 0.65)	0.42 (0.19, 0.65)
White Race	0.21 (0.04, 0.39)	0.10 (-0.29, 0.48)	0.11 (-0.28, 0.50)
Male	0.49 (0.27, 0.71)	0.37 (-0.09, 0.83)	0.36 (-0.10, 0.81)
Mid-Atlantic	0.44 (0.23, 0.64)	0.42 (0.21, 0.62)	0.43 (0.22, 0.63)
Southern	0.18 (0.03, 0.32)	0.19 (0.04, 0.34)	0.16 (0.01, 0.31)
Midwest	0.09 (-0.04, 0.22)	0.06 (-0.07, 0.20)	0.06 (-0.08, 0.19)
Northwest	0.12 (-0.01, 0.26)	0.13 (-0.01, 0.28)	0.12 (-0.02, 0.25)
Southwest	0.29 (0.14, 0.43)	0.27 (0.12, 0.42)	0.26 (0.12, 0.41)
Pacific	0.41 (0.22, 0.60)	0.42 (0.23, 0.62)	0.41 (0.22, 0.61)
Male:White Race	-0.38 (-0.62, -0.14)	-0.12 (-0.63, 0.40)	-0.11 (-0.62, 0.40)

Table 4.1 displays the estimates and 95% intervals for $\boldsymbol{\beta}$ based on the complete data,

and NPML and the Bayesian random effects approach applied to the two-phase data. In the Bayesian approach, we ran $300,000$ iterations, using the first $50,000$ to tune the chain. The $300,000$ iterations took 250 minutes to run on an AMD Opteron 2350 2GHz (four cores) processor with 4GB RAM. With constants $c = 0.24$ and $c_\lambda = 0.018$, we observed acceptance rates of 31% and 34% for $\boldsymbol{\lambda}$ and $\boldsymbol{V}^{XZ}$, respectively. Figures B.9 - B.19 display the trace plots for the $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{V}^{XZ}$ and τ_λ parameters. In each case, the samples are thinned by 100. Convergence was assessed through visual inspection of trace plots by running two chains with different starting values.

We see very strong agreement between the Bayesian and NPML approaches, with the Bayesian intervals tending to be very slightly narrower. Recall that in the fixed effects Bayesian approach, we fit a log-linear model which included all two-way interactions between age, race, sex and region. In the random effects Bayesian approach, however, the log-linear model included all two-way, three-way and four-way interactions between age, race, sex and region. (Note that the random effects Bayesian approach is not directly comparable to the fixed effects Bayesian approach because we have included higher order fixed effects interaction terms in the log-linear model of the random effects approach.) The results of the Bayesian random effects analysis have very good agreement with the NPML approach. Figure 4.1 compares the estimated $\boldsymbol{\lambda}^{XZ}$ to the estimates from the full data analysis (indicated by red crosses). In general, the Bayesian estimates agree well with the true values, where most of the 95% credible intervals contain the full data estimates. We also analyzed the data assuming a Gamma(1,0.026) prior for τ_λ. Figure B.8 compares the prior and posterior median and 95% intervals for σ_λ under the two assumed priors for τ_λ. Given that the posterior median is located in the left tail of the Gamma (1, 0.026) prior distribution, the Gamma(0.5, 0.0005) prior on τ_λ, which places more mass near zero, is the more appropriate prior for this analysis.

This example demonstrates that the Bayesian random effects approach can handle the high-dimensionality of the $x \times z$ parameter space, producing results that agree well with the NPML approach. Hence, this Bayesian hierarchical method offers another, more flexible, alternative to modeling a saturated log-linear model, while being only slightly more computationally demanding.

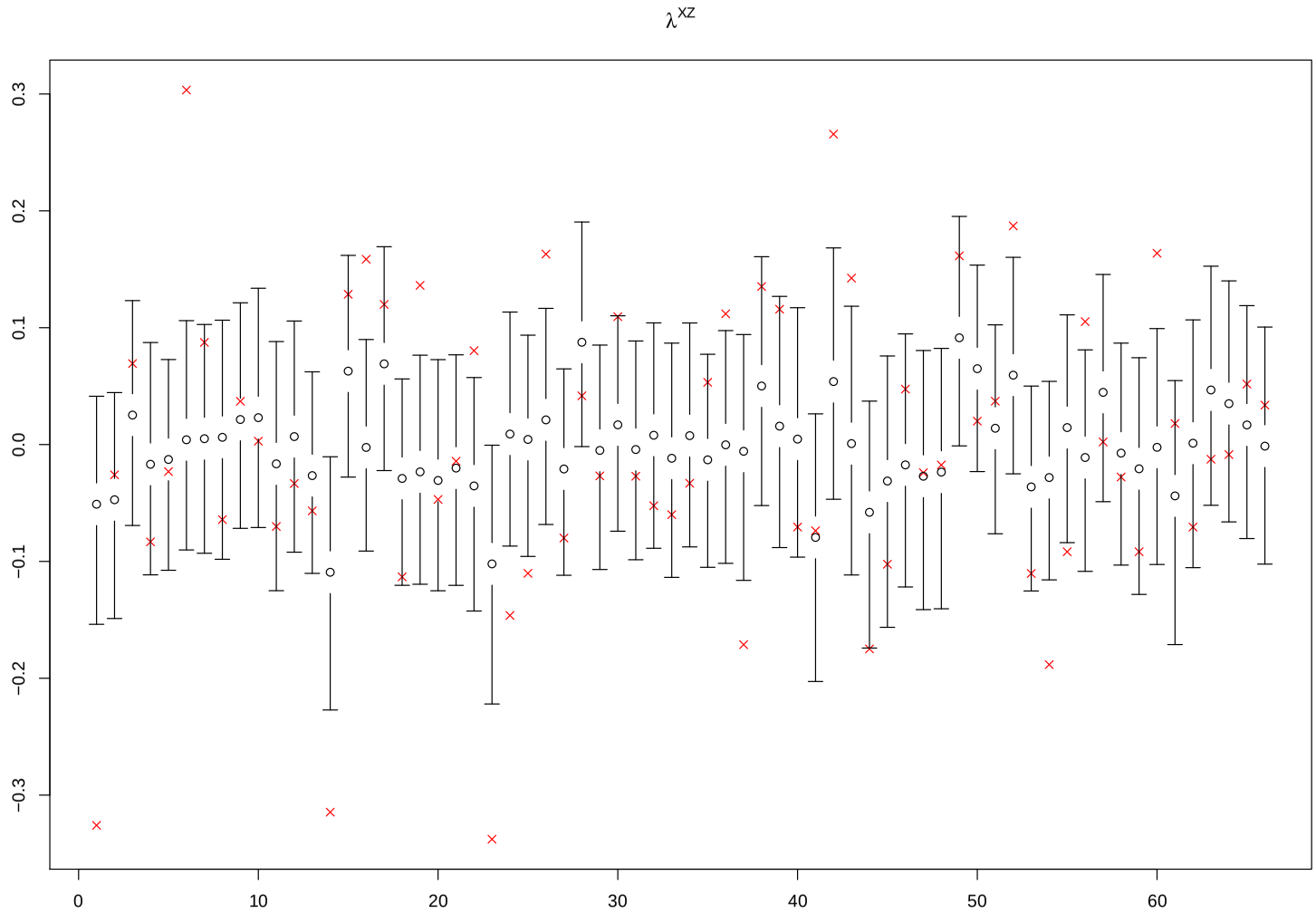


Figure 4.1: Comparing $\boldsymbol{\lambda}^{XZ}$ to the estimates from the full data analysis in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example. Red crosses indicate the true values.

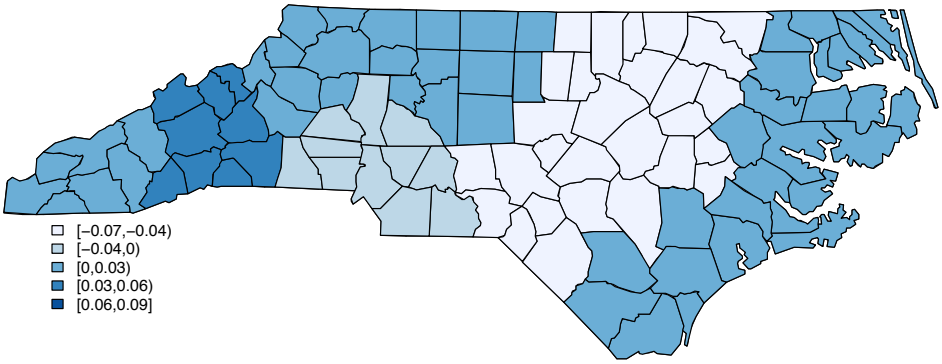
4.4.2 North Carolina Infant Mortality Example

We return to the North Carolina infant mortality data set described in Section 2.8.2. Recall that in this example, we are interested in estimating the association between infant mortality and birth weight, adjusting for race and gender, with particular interest in the possible effect modification by race on the infant mortality-birth weight association. The phase II data were obtained by sampling 200 non-cases and 50 cases from each of the 10 regions. The full population and the phase II data are provided in Tables 2.6 and 2.7, respectively. We perform a spatial analysis using a disease model that includes main effects for gender, race, low birth weight status, and an interaction term between race and low birth weight and we compare the Bayesian two-phase approach with the full individual-level data analysis.

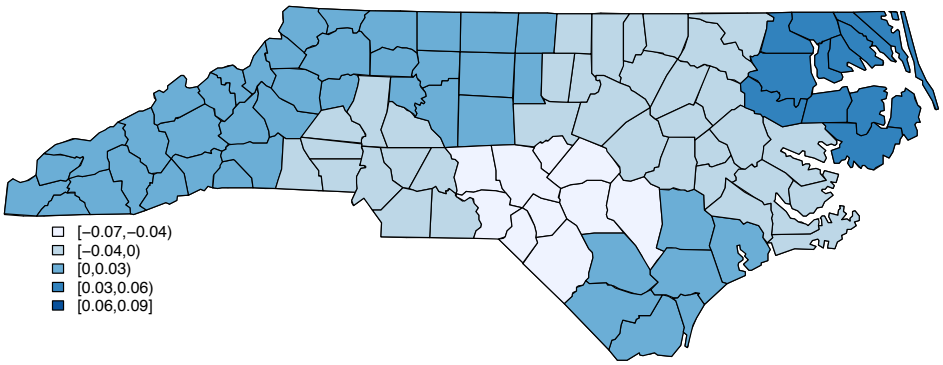
In the Bayesian approach, the log-linear model used included all two-way interactions between race, sex, birth weight and region. Let R denote race, S sex, W birth weight and Z region. We assigned independent zero mean normal prior distributions with large variances on $\boldsymbol{\beta}$ and the approach outlined in Section 4.3.2 for the remaining $\boldsymbol{\lambda}$, $\boldsymbol{V}$ and $\boldsymbol{U}$ parameters, where we have assumed the $\boldsymbol{\lambda}^{XZ}$ are fixed. Appendix B.3 provides the details. We assigned τ_v and τ_u gamma hyperprior distributions with shape parameter 1 and rate parameter 0.026.

In the Bayesian approach, we ran a total of $1,000,000$ iterations, where the first $50,000$ were used to tune the chain, and constants $c = 0.15$, $c_v = 0.002$ and $c_u = 220$ gave acceptance rates of 53%, 33%, and 30% for $\boldsymbol{\lambda}$, $\boldsymbol{V}$ and $\boldsymbol{U}$, respectively. The $1,000,000$ iterations took 824 minutes to run on a AMD Opteron 2350 2GHz (four cores) processor with 4GB RAM. Figures B.20 - B.26 display the trace plots for all parameters in the model. In all cases, the samples have been thinned by 100. Convergence was assessed through visual inspection of trace plots by running two chains with different starting values.

Parameter estimates and 95% intervals for the fixed effects disease model parameters are shown in Table 4.2. We compare the Bayesian random effects two-phase approach to the NPML approach using region as a fixed effect in the disease model since random effects are not available. (Breslow and Chatterjee did not make this available in the software provided to analyze two-phase data.) We notice fairly good agreement between the NPML

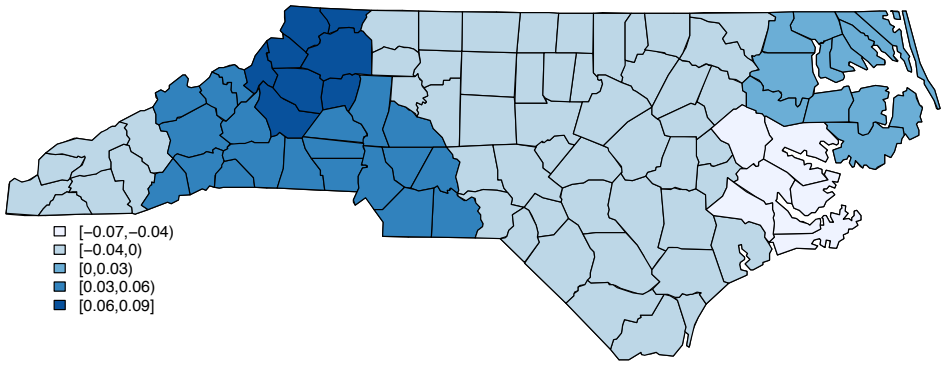


(a) Posterior mean estimates for $\boldsymbol{V}$ from the complete data analysis.

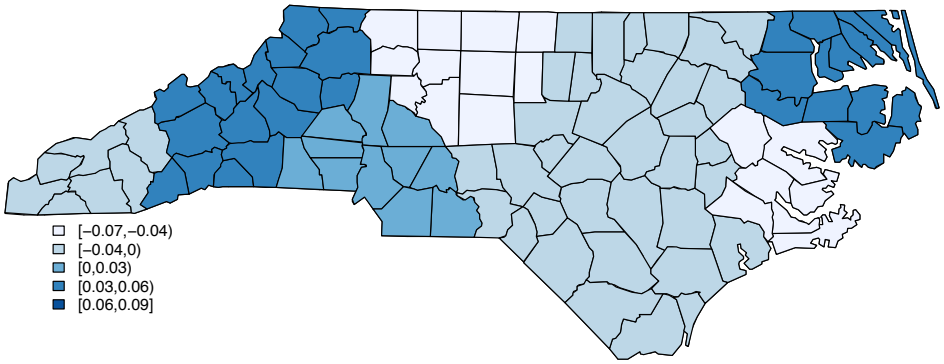


(b) Posterior mean estimates for $\boldsymbol{U}$ from the complete data analysis.

Figure 4.2: Posterior mean estimates of the log residual relative risks from the spatial analysis of the full North Carolina infant mortality data.



(a) Posterior mean estimates for $\boldsymbol{V}$.



(b) Posterior mean estimates for $\boldsymbol{U}$.

Figure 4.3: Posterior mean estimates of the log residual relative risks from the Bayesian two-phase random effects analysis of the North Carolina infant mortality data.

Table 4.2: Fixed effects estimates and 95% intervals from three analyses of the North Carolina infant mortality data example.

	Full Data	NPML	Bayesian Two-Phase
β_0	-6.28 (-6.38, -6.18)	-6.39 (-6.78, -5.99)	-6.24 (-6.49, -6.01)
Sex	0.35 (0.29, 0.40)	0.50 (0.25, 0.76)	0.50 (0.26, 0.74)
Race	0.35 (0.24, 0.46)	0.28 (-0.11, 0.68)	0.25 (-0.13, 0.64)
LBW	3.31 (3.24, 3.39)	3.08 (2.77, 3.39)	3.09 (2.79, 3.37)
Race:LBW	0.10 (-0.02, 0.23)	0.53 (-0.02, 1.08)	0.55 (0.002, 1.11)
τ_v	115.5 (36.3, 263.2)	–	76.7 (19.0, 197.9)
τ_u	81.3 (20.0, 208.9)	–	61.0 (10.0, 170.3)
σ_u	0.025	–	0.088 (0.037, 0.175)

and Bayesian approaches with the exception of the intercept. Since we include region as a fixed effect in the NPML approach, the interpretation of the intercept in this model is the log odds of infant mortality among normal birth weight female infants of white race for the baseline Southern mountains region. In the Bayesian and full data analyses, however, the interpretation of the intercept is the log odds of infant mortality among normal birth weight female infants of white race for a typical region. Compared to the full data analysis, the Bayesian approach slightly underestimates the main effects for race and low birth weight, while slightly overestimating the main effect for gender and the interaction term. In addition, the 95% interval of the main effect for race contains zero, though the corresponding interval for the interaction excludes zero.

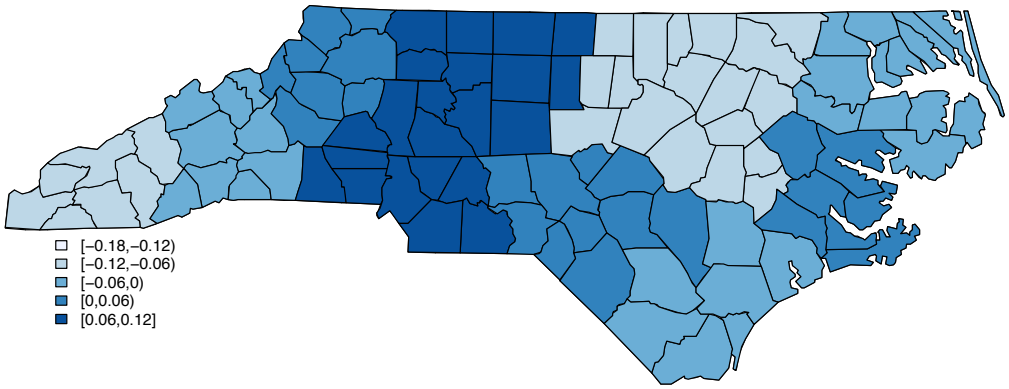
Figure 4.2 displays two maps for the complete data analysis, one for the unstructured random effects V_z and the other for the spatially structured random effects U_z. Figure 4.3 is the corresponding figure for the two-phase analysis. The scale is the same on all four plots. Comparing the four maps, the Bayesian spatial and non-spatial residual relative risk maps are comparable to the spatial and non-spatial residual relative risk maps from the full

data analysis, respectively. In particular, the posterior mean estimates for the unstructured and spatially structured random effects are close to zero for both analyses, indicating that there is not a great deal of residual variability in relative risk. The two-phase posterior mean estimates for σ_v and ω_u are 0.11 (0.07, 0.23) and 0.13 (0.08, 0.32), respectively. These estimates are not directly comparable however, as ω_u must be interpreted conditionally. We can obtain an unconditional estimate, however, by empirically calculating the standard deviation of the spatial random effect estimates, denoted by σ_u, which is 0.09 (0.04, 0.17). The values for σ_v and σ_u are comparable, consistent with what we observed in Figure 4.3. Compared to the full data analysis, the Bayesian two-phase analysis slightly overestimates the spatial and non-spatial standard deviations. Table B.4 displays the complete results for the random effects terms from the complete data and the Bayesian random effects analyses.

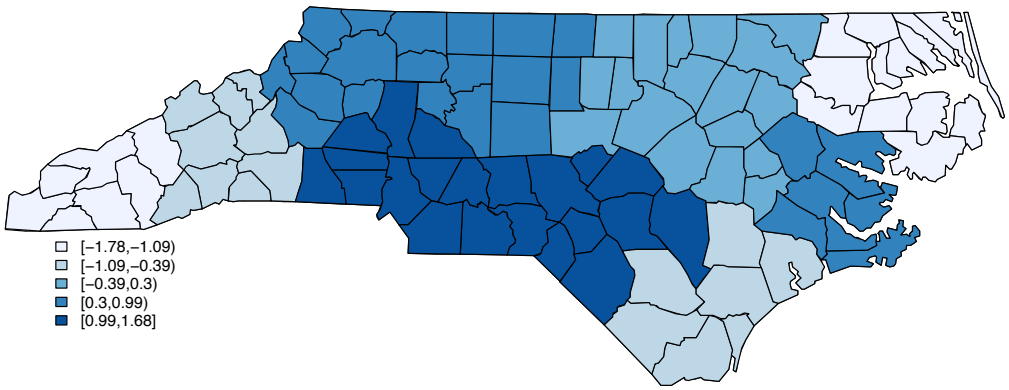
4.4.3 Simulated Data with Strong Spatial Dependence Example

To illustrate the potential benefits of fitting a spatial model over the simpler non-spatial model, we simulated a data set with strong spatial dependence. We generated the data using the geography of the North Carolina infant mortality data, and used low birth weight status as our exposure variable. We calculated the cell probabilities using the $\boldsymbol{\lambda}$ values from fitting a log-linear model to the full North Carolina data, and simulated non-spatial random effects from a normal distribution with mean zero and variance $\sigma_v^2 = 0.47^2$, and spatial random effects from an ICAR model with zero mean and conditional variance $\omega_u^2 = 0.69^2$. The total population size is $500,000$ and is displayed in Table B.5. The phase II data were obtained by sampling 30 deaths and 170 non-deaths from each region, and are shown in Table 4.3. Figure 4.4 displays the log residual relative risks from a spatial analysis of the complete data. From the maps, we notice the strong spatial component of the data, with the risks being especially high in the Southern foothills and the Southern heartland.

The two-phase data was analyzed using two different Bayesian approaches. The first approach used a non-spatial random effects disease model, while the second approach used a spatial random effects disease model. We compared the Bayesian approaches to the NPML approach using fixed effects for low birth weight status and region (as random



(a) Posterior mean estimates for $\boldsymbol{V}$ from the complete data analysis.



(b) Posterior mean estimates for $\boldsymbol{U}$ from the complete data analysis.

Figure 4.4: Posterior mean estimates for the log residual relative risks from the spatial analysis of the complete simulated data with strong spatial dependence. Note the difference in scale between the two maps.

Table 4.3: Number of live and dead infants by North Carolina region, and low birth weight status in the simulated data example with strong spatial dependence: phase II data.

		Normal Birth Weight	Low Birth Weight
Southern Mountains	Alive	156	14
	Dead	11	19
Central Mountains	Alive	158	12
	Dead	4	26
Northern Mountains	Alive	162	8
	Dead	7	23
Southern Foothills	Alive	153	17
	Dead	4	26
Northern Foothills	Alive	161	9
	Dead	9	21
Southern Heartland	Alive	158	12
	Dead	10	20
Southern Coast	Alive	148	22
	Dead	7	23
Northern Heartland	Alive	156	14
	Dead	8	22
Central Coast	Alive	156	14
	Dead	6	24
Northern Coast	Alive	160	10
	Dead	5	25

effects are unavailable with this approach), as well as to a spatial analysis of the complete data. Further, for each of the Bayesian two-phase analyses, the resulting estimates are used to predict $\boldsymbol{N}^{\text{yxz}}$, and we evaluate the accuracy of the predictions using the bias, variance and mean squared error (MSE).

For each of the Bayesian analyses, we follow the prior specification procedure outlined in Sections 4.3.1 and 4.3.2, respectively, where we treat $\boldsymbol{\lambda}^{XZ}$ as fixed. The $\boldsymbol{\beta}$ parameters were assigned independent mean zero normal prior distributions with large variances, and the Dellaportas and Forster (1999) formulation described in Section 3.2.2 was used for

the remainder of the $\boldsymbol{\lambda}$ nuisance parameters. We assigned τ_v and τ_u gamma hyperprior distributions with shape parameter 1 and rate parameter 0.026.

For the Bayesian non-spatial analysis, $50,000$ iterations were run to tune the chain, with constants $c = 0.20$ and $c_v = 5.0 \times 10^{-6}$ giving acceptance rates of 47% and 80% for $\boldsymbol{\lambda}$ and $\boldsymbol{V}$, respectively. The chain was run for a further $450,000$ iterations, and the total $500,000$ iterations took 145 minutes to run on a AMD Opteron 2350 2GHz (four cores) processor with 4GB RAM. For the Bayesian spatial analysis, $500,000$ iterations were run, where the first $50,000$ were used for tuning. Constants $c = 0.20$, $c_v = 5.0 \times 10^{-6}$ and $c_u = 1000$ gave acceptance rates of 48%, 79% and 11% for $\boldsymbol{\lambda}$, $\boldsymbol{V}$ and $\boldsymbol{U}$, respectively, where the $500,000$ iterations took 102 minutes to run on a AMD Opteron 2350 2GHz (four cores) processor with 4GB RAM.

Table 4.4: Fixed effects estimates and 95% intervals for four analyses of the simulated data example with strong spatial dependence.

	Full Data	NPML	Bayesian Two-Phase	
			Non-Spatial	Spatial
β_0	-6.08 (-6.74, -5.41)	-7.62 (-8.11, -7.13)	-6.12 (-6.41, -5.81)	-6.20 (-6.41, -5.97)
LBW	3.44 (3.38, 3.49)	3.69 (3.37, 4.02)	3.72 (3.41, 4.05)	3.71 (3.37, 4.03)
τ_v	35.0 (2.14, 127.1)	–	1.06 (0.38, 2.11)	30.2 (7.1, 81.4)
τ_u	0.45 (0.16, 0.95)	–	–	0.36 (0.12, 0.74)
σ_u	1.06	–	–	1.14 (1.04, 1.26)

Table 4.4 displays the estimates and 95% intervals for the $\boldsymbol{\beta}$ parameters. The $\boldsymbol{\beta}$ parameter estimate associated with low birth weight agrees well between the NPML and Bayesian approaches, which are slightly larger than the full data estimate. We notice that the NPML estimate for the intercept is quite different than that in the other approaches. As in the North Carolina infant mortality data example, this is because the interpretation of the intercept is different in this approach. Since we include region as a fixed effect in the NPML approach, the interpretation of the intercept in this model is the log odds of infant mortal-

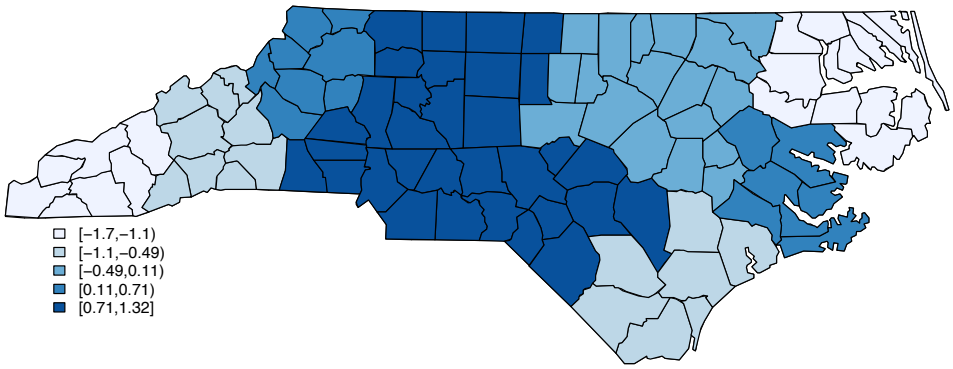
ity among normal birth weight infants for the baseline Southern mountains region. In the Bayesian and full data analyses, however, the interpretation of the intercept is the log odds of infant mortality among normal birth weight infants for a typical region.

Figure 4.5 displays the posterior mean estimates of the log residual relative risks. The posterior mean estimate for σ_v decreases drastically from 0.97 (0.69, 1.62) to 0.18 (0.11, 0.38) when the spatial random effect is added. As in the analysis of the full data, we see strong spatial dependence in the data, with the risks being particularly high in the Southern foothills and in the Southern heartland. We estimate that roughly 97.8% (97.4%, 98.2%) of the residual variability is spatial in nature. The posterior mean estimate for σ_u is 1.67 (1.16, 2.89), though this value is not directly comparable to σ_v. The empirical estimate for the standard deviation of the spatial random effects is 1.14 (1.04, 1.26). Table B.6 contains the full results for the random effects parameters.

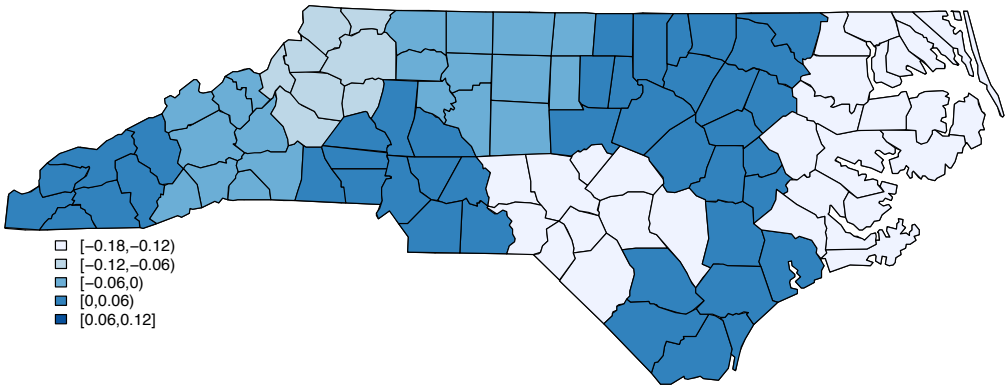
Table 4.5: Comparing bias, variance and mean squared error (MSE) for the predicted $\boldsymbol{N}^{\text{yxz}}$ values from the non-spatial and spatial Bayesian two-phase analyses of the simulated data example with strong spatial dependence.

	Bias ($\times 10^3$)	Variance ($\times 10^6$)	MSE ($\times 10^6$)
Non-Spatial	5.7	15.0	47.1
Spatial	5.2	16.7	43.3

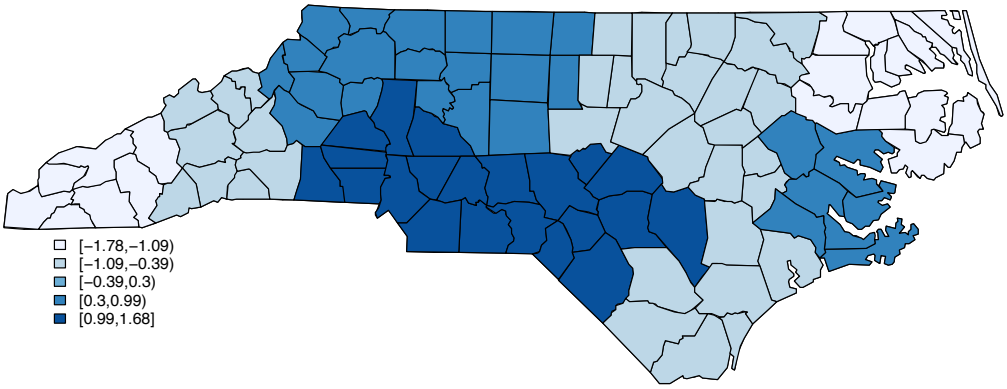
Figure B.27 displays the predicted $\boldsymbol{N}^{\text{yxz}}$ values with 95% intervals compared to the true population values (shown in red crosses) from the non-spatial Bayesian random effects analysis. In most cases, the predicted values are close to the truth and the 95% intervals contained the true values in all but four cases. The predicted $\boldsymbol{N}^{\text{yxz}}$ values from the spatial Bayesian random effects analysis are displayed in Figure B.28 along with the 95% intervals, where the true population values are indicated by red crosses. As in the non-spatial analysis, the predicted values are close to the true values, where only four intervals did not contain the true values. Table 4.5 comapres the bias, variance and MSE for these predictions from the spatial and non-spatial analyses. We emphasize that the MSE decomposition in this example is not over simulations, but over draws from the posterior distribution. The spatial



(a) Posterior mean estimates for $\boldsymbol{V}$ from the Bayesian non-spatial random effects analysis.



(b) Posterior mean estimates for $\boldsymbol{V}$ from the Bayesian spatial random effects analysis.



(c) Posterior mean estimates for $\boldsymbol{U}$ from the Bayesian spatial random effects analysis.

Figure 4.5: Posterior mean estimates for the log residual relative risks in the Bayesian two-phase random effects analysis of the simulated data with strong spatial dependence. Note the difference in scales between the three maps.

method produces slightly less biased estimates though the variance is slightly larger than for the non-spatial analysis, thus yielding a smaller MSE. Hence, analyzing the data using the slightly more complex Bayesian spatial approach yields better estimates than the non-spatial approach (measured in terms of the MSE).

Figures B.29 – B.32 display the trace plots for the $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{V}$, and τ_v parameters for the non-spatial analysis, while Figures B.33 – B.37 display the trace plots for the $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{V}$, $\boldsymbol{U}$, τ_v and τ_u parameters for the spatial analysis. In all cases, the samples have been thinned by 100. Convergence was assessed through visual inspection of trace plots by running two chains with different starting values.

4.4.4 Simulated Data with Strong Exposure-Confounder Relationship

The purpose of this example is to explore the effects on prediction of different treatments of the $\boldsymbol{\lambda}^{XZ}$ variable in a data set that has a strong exposure-confounder association. In particular, we perform three analyses of the data where we omit the $\boldsymbol{\lambda}^{XZ}$ variable from the log-linear model, treat $\boldsymbol{\lambda}^{XZ}$ as fixed and treat $\boldsymbol{\lambda}^{XZ}$ as random in the log-linear model.

We simulated a population using the North Carolina geography with 10 regions and low birth weight status as the exposure variable. We used the $\boldsymbol{\lambda}^{Y}, \boldsymbol{\lambda}^{X}, \boldsymbol{\lambda}^{Z}$, and $\boldsymbol{\lambda}^{YX}$ values from fitting a log-linear model to the full North Carolina infant mortality data, where the $\boldsymbol{\lambda}^{XZ}$ values were multiplied by 10 to create stronger (and statistically significant) exposure-confounder effects. The U_z and V_z values were taken to be the estimates from fitting a spatial model to the full data. The $\boldsymbol{\lambda}$, $\boldsymbol{V}$ and $\boldsymbol{U}$ values were used to calculate the cell probabilities of the $y \times x \times z$ table, and the data set was generated from sampling from a multinomial distribution with population size $500,000$. Table B.7 displays the full data. Three different phase II sample sizes were used (denoted "small", "medium" and "large"). To obtain the phase II data, we sample 10 deaths and 10 non-deaths from each region in the "small" case, all deaths and an equal number of non-deaths from each region in the "medium" case, and all deaths and 2000 non-deaths in the "large" case. Table 4.6 displays the phase II data for all three cases.

The two-phase data were analyzed using three different Bayesian random effects ap-

proaches, each using a random disease model with both spatial and non-spatial random effects. The first omits $\boldsymbol{\lambda}^{XZ}$ from the log-linear model, the second treats $\boldsymbol{\lambda}^{XZ}$ as a fixed effect, and the third treats $\boldsymbol{\lambda}^{XZ}$ as a random effect. For the first two analyses, we follow the prior specification procedure described in Section 3.2.2, whereas for the random effects analysis, we follow the procedure outlined in Section 4.2, in assigning prior distributions to $\boldsymbol{\lambda}^{XZ}$. In each analysis, the $\boldsymbol{\beta}$ parameters were assigned independent mean zero normal prior distributions with large variances, and the Dellaportas and Forster (1999) formulation described in Section 3.2.2 was used for $\boldsymbol{\lambda}^{X}$ and $\boldsymbol{\lambda}^{Z}$. Further, τ_v and τ_u were assigned gamma hyperprior distributions with shape parameter 1 and rate parameter 0.026, as in the examples considered in Sections 4.4.2 and 4.4.3.

In the analysis in which $\boldsymbol{\lambda}^{XZ}$ is considered fixed, we followed the Dellaportas and Forster (1999) prescription. For the scenario where we treat $\boldsymbol{\lambda}^{XZ}$ as random, we assign each component of $\boldsymbol{\lambda}^{XZ}$ a normal distribution with mean zero and precision τ_λ and we explore five different hyperprior distributions for τ_λ to examine the sensitivity of the results to the prior distributions. We assume *a priori* that the $\boldsymbol{\lambda}^{XZ}$ follow a Student's t-distribution with d degrees of freedom, where the $\boldsymbol{\lambda}^{XZ}$ fall within $\pm R$ with 95% probability. We can then use equation (2.14) to obtain a_λ and b_λ. The first hyperprior distribution we consider is a Gamma$(1, 0.026)$, which corresponds to $d = 2$ and $R = \log 2$. Given the wide range of the true values for λ^{XZ}, we also considered ranges of 1.5 and 2.0. For $R = 1.5$, we used $d = 1, 2$ which corresponds to a Gamma$(0.5, 0.0139)$ and Gamma$(1, 0.1215)$ hyperprior, respectively. For $R = 2.0$, we also considered $d = 1, 2$, which yields hyperprior distributions Gamma$(0.5, 0.0248)$ and Gamma$(1, 0.216)$, respectively. In the analyses where $\boldsymbol{\lambda}^{XZ}$ is omitted and is considered fixed, we ran the chains for $500,000$ iterations, where the first $50,000$ were used for tuning. In the analysis in which $\boldsymbol{\lambda}^{XZ}$ is treated as random, we ran the chains for $1,000,000$ iterations, with the first $50,000$ used for tuning.

We evaluate the performance of the Bayesian approaches by predicting the $\boldsymbol{N}^{\text{yxz}}$ values using the resulting estimates, and compare the results using the MSE and its decomposition into bias and variance. Table 4.7 compares the predictions from the five hyperprior distributions under the three different phase II sample sizes. For the medium and large phase II sample sizes, we notice that the results are fairly robust to the choice of hyperprior distri-

bution as the bias, variance and MSE are quite similar. This point is further illustrated in Figures B.39 and B.40, which shows the similarity between the posterior medians and 95% intervals for σ_λ under the five assumed prior distributions. For the small sample size case, however, we notice a substantial difference in both the bias and the variance between the different hyperprior distributions. We notice that the Gamma(1,0.216) hyperprior yields the smallest MSE, even though it produces nearly the most variable estimates, which is likely due to the wide prior range of ± 2 for $\boldsymbol{\lambda}^{XZ}$. A close second is the Gamma$(1, 0.1215)$ hyperprior, which assumes a slightly narrower prior on the random effects. This prior sacrifices some bias for much smaller variance. The remaining hyperpriors produce predicted values which have substantial bias and/or variance, resulting in large MSEs. The Gamma$(0.5, 0.0139)$, Gamma$(1, 0.216)$ and Gamma$(0.5, 0.0248)$ hyperpriors have the widest 95% intervals of the five priors, which perhaps explains the large variance estimates associated with their results. Figure B.38 compares the medians and 95% intervals for σ_λ under the five assumed prior distributions in the small sample size case. We notice the similarity in the posterior medians and intervals for the Gamma$(1, 0.1215)$ and Gamma$(1, 0.216)$ hyperpriors.

Figure 4.6 displays the posterior mean estimates for the $\boldsymbol{V}^{XZ}$ under the five hyperprior distributions, as well as the true values, and confirms the results of the table. The resulting posterior means for the hyperpriors where we assumed $d = 1$ often appear over smoothed. In general, the posterior means of $\boldsymbol{\lambda}^{XZ}$ under the Gamma$(0.5, 0.0139)$ and Gamma$(0.5, 0.0248)$ hyperpriors are much closer to zero than those from the remaining hyperpriors distributions, and consequently, are further from the true values. A similar pattern is seen for the posterior means from the Gamma$(1, 0.026)$ hyperprior, which is perhaps not too surprising given the restricted range assumed *a priori*.

Table 4.8 compares the predictions from the three Bayesian analyses in terms of the bias, variance and MSE. We use the results from the Bayesian random effects analysis assuming a Gamma(1,0.216) hyperprior distribution for τ_λ. For all three phase II sample sizes, the analysis omitting the $\boldsymbol{\lambda}^{XZ}$ parameters results in highly biased, high variance predictions. In the large phase II sample size case, the random $\boldsymbol{\lambda}^{XZ}$ analysis outperforms the fixed $\boldsymbol{\lambda}^{XZ}$ in terms of bias, variance and MSE. While the biases are comparable, the random effects analysis produces estimates with far smaller variance. In the medium phase II sample size

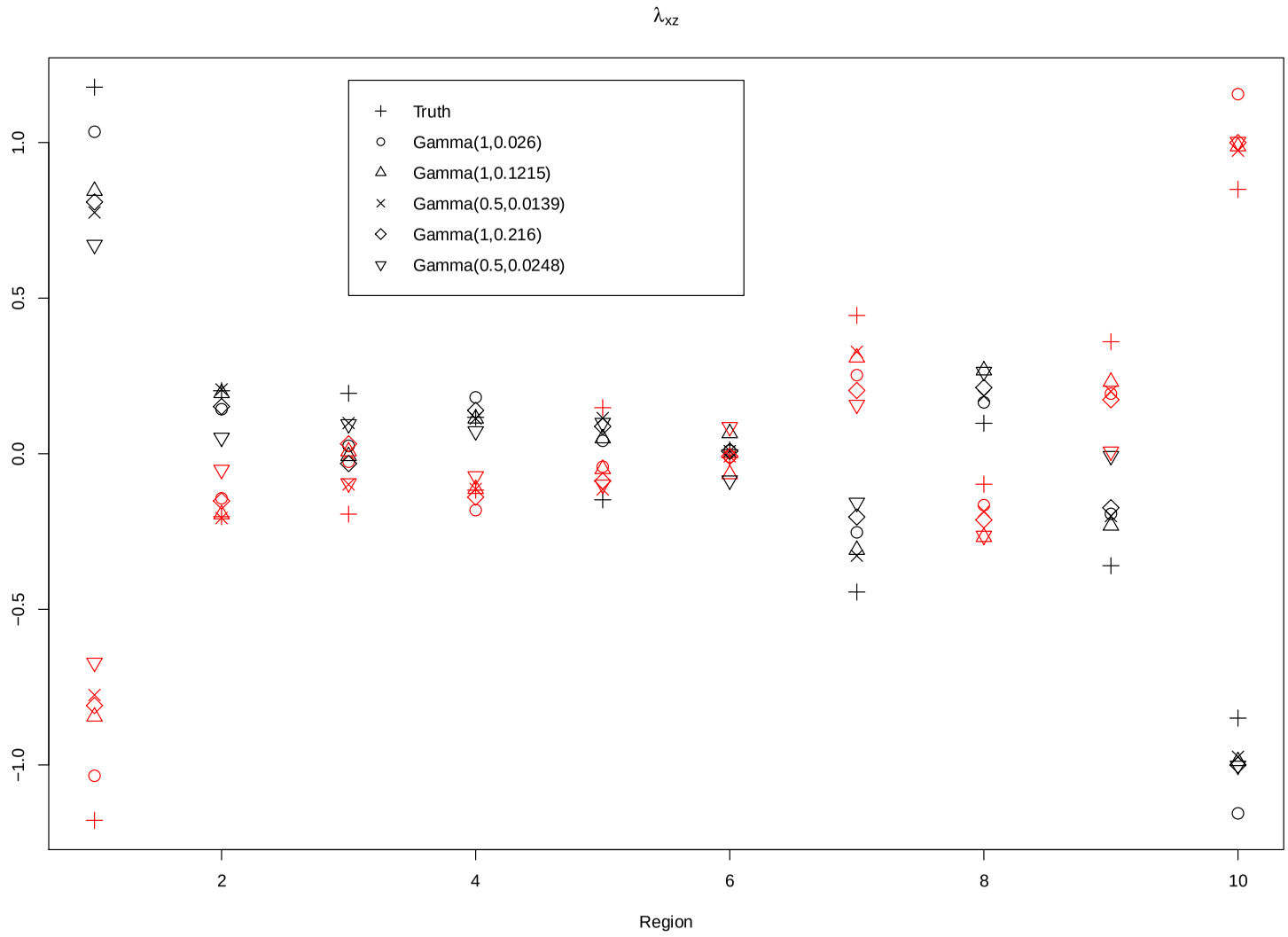


Figure 4.6: Comparing λ^{XZ} for the Bayesian two-phase random effects analyses using the five different hyperprior distributions with small phase II sample sizes for the simulated data example with strong exposure-confounder relationship. Black points indicate $X = 0$ and red points indicate $X = 1$.

scenario, the fixed and random $\boldsymbol{\lambda}^{XZ}$ analyses perform comparably, with the random effects analysis having slightly less bias, and slightly more variance than the fixed analysis. Finally, in the small phase II sample size case, the fixed analysis outperforms the random effects analysis with lower bias and variance, and consequently, MSE.

Figures B.41 - B.44 display some representative trace plots for the parameters in the scenario where $\boldsymbol{\lambda}^{XZ}$ is considered random with hyperprior Gamma(1, 0.1215). The trace plots for all analyses in the small phase II samples sizes case appear similar. Parameters converged without difficulty for the analyses involving the medium and large phase II sample sizes. Convergence was assessed through visual inspection of trace plots by running two chains with different starting values.

4.5 Summary

In this chapter, we have extended the work of Chapter 3 to allow for the inclusion of random effects. We described how spatial random effects along with independent normal random effects can be incorporated into the log-linear models of the previous chapter to control for residual spatial confounding. We adopted the spatial model introduced by Besag et al. (1991), which imposes non-parametric smoothing by assuming that the spatial effect in a particular area is similar to the mean of the spatial effects in close-by areas, with the strength of the similarity determined by the number of neighbours. In addition, we described how terms in the log-linear model, particularly the exposure-confounder terms $\boldsymbol{\lambda}^{XZ}$, can be modeled as random effects by treating the variance-covariance matrix of its prior distribution as unknown, and assigning it a conjugate hyperprior distribution. Sampling from the posterior distribution proceeded via a Markov chain Monte Carlo scheme using separate Metropolis-Hastings steps for each set of parameters.

We illustrated the methods of the chapter using the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example first presented in Section 3.5.3, the North Carolina infant mortality data, and two simulated data examples. Since there are no other methods that incorporate random effects when analysing two-phase data, we compared the results from the Bayesian random effects analyses to analyses of the complete data. For the examples considered, the results appeared reasonable. Further, the simulated data examples provides one scenario

highlighting the potential benefits of the Bayesian random effects analysis over the simpler fixed effects analysis described in Chapter 3. The inclusion of random effects into the log-linear model allowed for greater modeling flexibility than the fixed effects approach, and in general resulted in estimates with smaller MSE. This flexibility comes at the cost of longer computing times. Given the additional parameters to estimate, along with more iterations to converge, the computing times for the random effects models were noticeably longer than those for similar fixed effects models.

In Chapter 5, we retain the theme of predicting population counts, but we leave the two-phase study design scenario and move on to small area estimation in a developing world context.

Table 4.6: Number of live and dead infants by North Carolina region, and low birth weight status in the simulated data example with strong exposure-confounder relationship: phase II data. NBW=Normal birth weight, LBW=low birth weight.

		Small		Medium		Large	
		NBW	LBW	NBW	LBW	NBW	LBW
Southern Mountains	Alive	10	0	56	0	1984	14
	Dead	8	2	47	9	47	9
Central Mountains	Alive	10	0	187	9	1889	111
	Dead	3	7	56	140	56	140
Northern Mountains	Alive	10	0	120	8	1872	128
	Dead	0	10	45	83	45	83
Southern Foothills	Alive	9	1	847	59	1869	131
	Dead	3	7	274	632	274	632
Northern Foothills	Alive	10	0	660	80	1781	219
	Dead	4	6	139	601	139	601
Southern Heartland	Alive	9	1	880	70	1838	162
	Dead	2	8	256	694	256	694
Southern Coast	Alive	9	1	582	125	1632	368
	Dead	3	7	256	694	256	694
Northern Heartland	Alive	10	0	254	18	1867	133
	Dead	4	6	107	165	107	165
Central Coast	Alive	10	0	198	32	1689	311
	Dead	2	8	38	192	38	192
Northern Coast	Alive	6	4	155	52	1359	641
	Dead	0	10	13	194	13	194

Table 4.7: Comparing the bias, variance and mean squared error (MSE) for predicting $\boldsymbol{N}^{yxz}$ using the five different hyperprior distributions for $\boldsymbol{V}^{XZ}$ in the Bayesian two-phase analysis of the simulated data with strong exposure-confounder relationship under three different phase II sample sizes.

		Bias ($\times 10^3$)	Variance ($\times 10^4$)	MSE ($\times 10^6$)
	Gamma(1,0.026)	17.4	423.7	306.4
	Gamma(1,0.1215)	10.1	1751.3	120.1
Small	Gamma(0.5,0.0139)	12.6	3177.3	189.4
	Gamma(1,0.216)	7.9	3802.3	100.1
	Gamma(0.5,0.0248)	10.0	5036.0	151.0
	Gamma(1,0.026)	2.2	304.5	7.9
	Gamma(1,0.1215)	1.9	297.0	6.6
Medium	Gamma(0.5,0.0139)	2.0	286.4	6.9
	Gamma(1,0.216)	1.9	296.2	6.7
	Gamma(0.5,0.0248)	2.0	297.1	7.0
	Gamma(1,0.026)	0.9	179.0	2.6
	Gamma(1,0.1215)	0.9	189.0	2.7
Large	Gamma(0.5,0.0139)	0.9	175.1	2.5
	Gamma(1,0.216)	0.8	172.5	2.4
	Gamma(0.5,0.0248)	0.9	184.1	2.6

Table 4.8: Comparing the bias, variance and mean squared error (MSE) for predicting $\boldsymbol{N}^{yxz}$ for the three Bayesian two-phase analyses of the simulated data with strong exposure-confounder relationship under three different phase II sample sizes.

		Bias ($\times 10^3$)	Variance ($\times 10^4$)	MSE ($\times 10^6$)
Small	No $\boldsymbol{\lambda}^{XZ}$	12.7	1477.2	176.3
	Fixed $\boldsymbol{\lambda}^{XZ}$	5.0	2528.8	50.1
	Random $\boldsymbol{\lambda}^{XZ}$	7.9	3802.3	100.1
Medium	No $\boldsymbol{\lambda}^{XZ}$	10.0	209.0	101.8
	Fixed $\boldsymbol{\lambda}^{XZ}$	2.0	277.1	6.8
	Random $\boldsymbol{\lambda}^{XZ}$	1.9	296.2	6.7
Large	No $\boldsymbol{\lambda}^{XZ}$	11.5	50.1	132.7
	Fixed $\boldsymbol{\lambda}^{XZ}$	0.9	195.5	2.7
	Random $\boldsymbol{\lambda}^{XZ}$	0.8	172.5	2.4

Chapter 5

SMALL AREA ESTIMATION

In this chapter, we describe the integrated data collection and statistical analysis framework that we propose for improved mortality monitoring in areas without comprehensive vital records systems. In 2007, *The Lancet* published a series of articles titled "Who Counts?" (AbouZahr et al., 2007; Boerma and Stansfield, 2007; Hill et al., 2007; Horton, 2007; Mahapatra et al., 2007; Setel et al., 2007), which details the "scandal of invisibility" resulting from the lack of accurate, timely, full-coverage civil registration that affects much of the developing world. Civil registration systems are critical to providing vital statistics including birth rates by age of mother, mortality rates by sex, age and other demographic characteristics, and causes of death. This type of information is important for formulating good public health programs, developing regional, national, and global policies and implementing and evaluating public health actions. However, few countries can actually maintain ongoing civil registration systems and most countries in the developing world have little to no data from civil registration (Mathers et al., 2005). Hence, it is not possible to obtain useful vital statistics from these countries. The authors from *The Lancet* special series identify a need for representative data describing sex-, age-, and cause-specific mortality through time in small enough areas to be meaningful for local governance and health institutions.

On the way to full civil registration, AbouZahr et al. (2007) propose using censuses and survey-based approaches to obtain broad population data, as well as health and demographic surveillance systems (HDSS) sites coupled with verbal autopsy to obtain stratified mortality rates and causes of death. (Verbal autopsy is a term used to describe postmortem caregiver or family member interviews (Setel, 2011).) Surveillance systems provide detailed longitudinal information on a non-representative group of people. While this type of information is useful for studying changes in mortality over time, the non-representative nature of the data makes it difficult to generalize findings to the entire population. On the other

hand, survey samples provide representative data on a variety of population and health outcomes. However, most sample surveys do not revisit the same individuals, and hence cannot provide any longitudinal information making it difficult to measure the impacts of interventions on mortality. Hence, a monitoring system that combines the benefits of both of these data collection methods would provide useful indicators for large populations over prolonged periods of time, so that changes can be monitored and related to possible determinants, including interventions. On the way to such a monitoring system, we propose a sampling procedure which considers a snapshot of the population in which we do not include measurements taken over time. The general procedure is to use data from HDSS sites to inform survey sample sizes taken in neighbouring areas. Data from the HDSS sites and surrounding areas can then be used to obtain mortality predictions for the entire region. We begin by introducing the adopted notation in Section 5.1, followed by a description of the informed sampling procedure we are proposing in Section 5.2. We detail the analysis of the sampling data in Section 5.3 and conclude the chapter with a detailed simulation study.

5.1 Notation

Suppose the region we are interested in studying consists of I areas, where we assume that at least one area in this region is a HDSS site. Let k denote the number of HDSS sites, where $1 \leq k < I$. For convenience, areas $i = 1, \ldots, k$ will denote HDSS areas, and $i = k+1, \ldots, I$ will denote non-HDSS areas. Let $\boldsymbol{x}_i$ denote area-specific covariates such as socio-economic status (SES), $i = 1, \ldots, I$, and let j index the levels of demographic variables of interest, $j = 1, \ldots, J$. Our outcome of interest is denoted Y_{ij}, the unobserved number of deaths in area i and stratum j, and the number of individuals in area i and stratum j is denoted by N_{ij}. We assume that both N_{ij} and $\boldsymbol{x}_i$ are known quantities. The probability of death over a 5-year period in area i and stratum j is p_{ij}, and an estimate of this probability will be used to estimate Y_{ij}. Note that the exact time frame for the probability of death is not important for our purposes.

The survey sample corresponds to choosing n_{ij} individuals from area i and stratum j, of which y_{ij} are recorded as dying. We begin by describing the informed sampling procedure.

5.2 Informed Sampling

The goal of the informed sampling procedure is to obtain accurate and precise measurements for mortality rates to best predict the number of deaths in each area. All individuals from the HDSS sites are used to predict the number of deaths in the non-HDSS areas. Since we have information on all individuals living in the HDSS areas, we do not make any predictions for these areas. A total of $n_{\cdot\cdot}$ individuals are randomly sampled from the non-HDSS areas, where the sampling can proceed in several ways. For example, the areas and populations which have the highest predicted mortality rate can be oversampled to obtain as many events as possible, or we can oversample the areas with lowest predicted mortality rates to obtain more reliable estimates for those areas with fewer deaths. There are potential benefits to both schemes, but we chose to proceed with the former. Hence, the number of individuals sampled from each non-HDSS area is proportional to the predicted number of deaths based on the analysis of the HDSS data.

We fit a logistic regression model containing main effect terms for the area-specific covariates and demographic variables to the HDSS data only:

$$\text{logit } p_{ij} = \boldsymbol{x}_i\boldsymbol{\beta} + \gamma_j,$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, J$. Let $\widehat{\boldsymbol{\beta}}^{\star}$ and $\widehat{\gamma}_j^{\star}$ be the maximum likelihood estimates for $\boldsymbol{\beta}$ and γ_j, respectively, from the HDSS data. Then, the fitted probabilities are calculated via

$$\widehat{p}_{ij}^{\star} = \frac{\exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}^{\star} + \widehat{\gamma}_j^{\star})}{1 + \exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}^{\star} + \widehat{\gamma}_j^{\star})},$$

for $i = k+1, \ldots, I$ and $j = 1, \ldots, J$. The predicted numbers of deaths in the surrounding areas is $\widetilde{Y}_{ij} = N_{ij} \times \widehat{p}_{ij}^{\star}, i = k+1, \ldots, I$.

The sample sizes n_{ij} are selected proportionately to $\widetilde{Y}_{ij}$ so that areas and strata with more predicted deaths are sampled more heavily. In particular, we take

$$n_{ij} = n_{\cdot\cdot} \times \frac{\widetilde{Y}_{ij}}{\widetilde{Y}_{\cdot\cdot}},$$

for $i = k+1, \ldots, I, j = 1, \ldots, J$, where $\widetilde{Y}_{\cdot\cdot} = \sum_{i=k+1}^{I}\sum_{j=1}^{J}\widetilde{Y}_{ij}$ is the total number of predicted deaths in the non-HDSS areas.

Let W_{ij} denote the observed number of deaths in area i and strata j, where $W_{ij} = Y_{ij}$ for $i = 1, \ldots, k$, and $W_{ij} = y_{ij}$ for $i = k+1, \ldots, I$. As described in the next section, we model the W_{ij} as a function of the known demographic factors and area-level variables and use spatial smoothing to exploit similarities in risk among nearby areas.

5.3 Analysis

We assume a logistic regression model which includes main effect terms for the area-specific covariates and stratum effects, as well as random effects to account for unmeasured area-level covariates:

$$\text{logit } p_{ij} = \boldsymbol{x}_i\boldsymbol{\beta} + \gamma_j + V_i + U_i, \tag{5.1}$$

where V_i and U_i are non-spatial and spatial random effects, respectively, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Analogous to our approach in Chapter 4, we place independent normal prior distributions with zero mean and variance σ_v^2 on each V_i, while we adopt the ICAR model described in Section 2.6.3 for U_i:

$$U_i | U_j, j \in \delta_i \sim N\left(\overline{U}_i, \frac{\omega_u^2}{m_i}\right),$$

where δ_i is the set of neighbours of area i and m_i is the number of such neighbours.

We assign normal prior distributions to $\boldsymbol{\beta}$ and γ_j;

$$\begin{aligned} \boldsymbol{\beta} &\sim N\left(\mu_\beta, \Sigma_\beta\right) \\ \gamma_j &\sim N\left(\mu_j, \Sigma_j\right), \end{aligned}$$

where μ_β, μ_j, Σ_β and Σ_j, $j = 1, \ldots, J$, are chosen based on the context. We assign gamma hyperprior distributions with shape parameters a_v, a_u and rate parameters b_v, b_u to the precision variables $\tau_v = \sigma_v^{-2}$ and $\tau_u = \omega_u^{-2}$, respectively. The a_v, a_u and b_v, b_u parameters should be considered on a case-by-case basis.

Model (5.1) can be fitted in a straightforward manner using the integrated nested Laplace approximations approach of Rue et al. (2009), which is implemented in the `inla` package in `R`. Once we obtain estimates $\widehat{\boldsymbol{\beta}}, \widehat{\gamma}_j, \widehat{V}_i$, and $\widehat{U}_i$, we can calculate the fitted probabilities

$$\widehat{p}_{ij} = \frac{\exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_j + \widehat{V}_i + \widehat{U}_i)}{1 + \exp(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_j + \widehat{V}_i + \widehat{U}_i)},$$

which in turn can be used to estimate Y_{ij}:

$$\widehat{Y}_{ij} = y_{ij} + (N_{ij} - n_{ij}) \times \widehat{p}_{ij}, \tag{5.2}$$

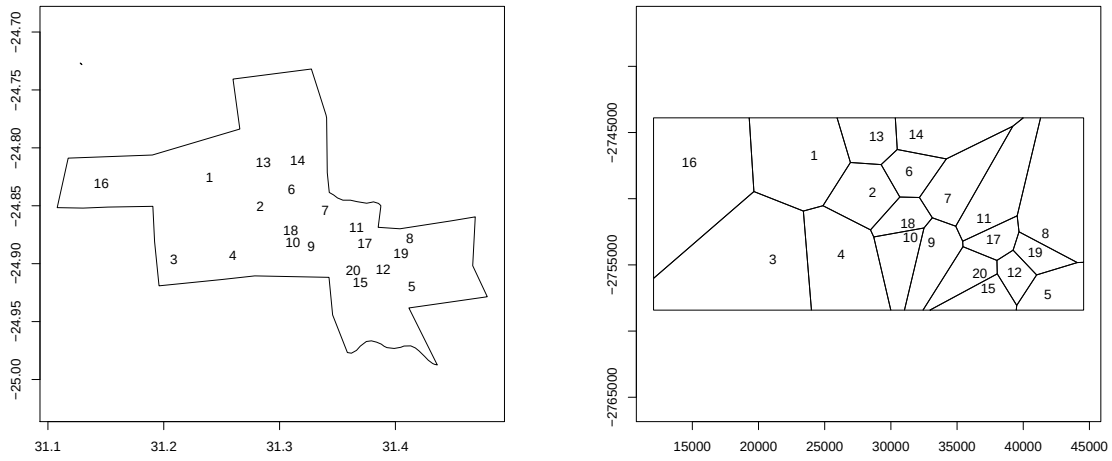
where $N_{ij} - n_{ij}$ is the number of unsampled individuals in village i and stratum j.

We now provide a detailed discussion of a simulation study performed to compare and contrast various sampling designs and models.

5.4 Simulation Study

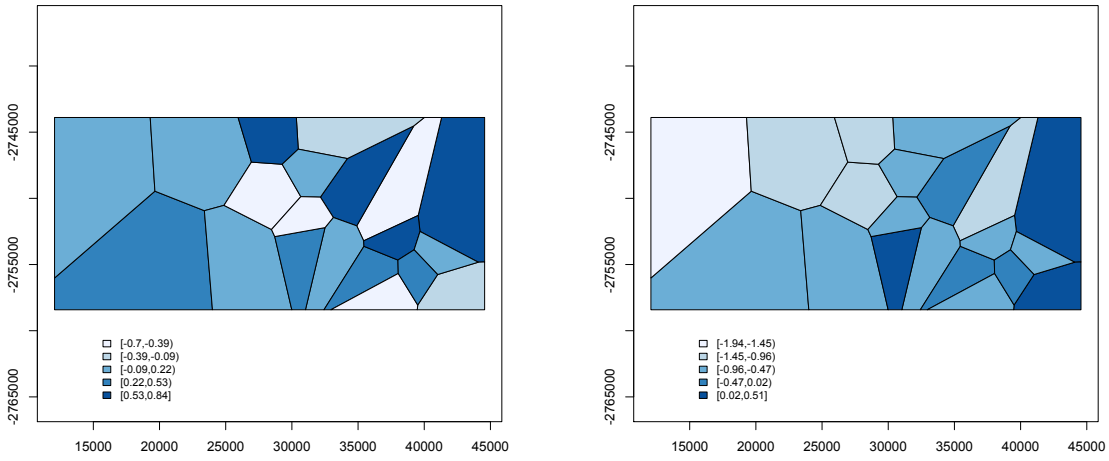
The study region that we create is based on the Agincourt HDSS site in South Africa (Kahn et al., 2007). We take 20 of the Agincourt villages for which we know the area centroids, shown in Figure 5.1(a). Because most of the villages are relatively isolated and do not share boundaries with nearby villages, it is not straightforward to define village neighbours for this region. There are several possible definitions of neighbours that we could use, including defining villages to be neighbours if their centroids fall within a certain distance of each other. We form a Dirichlet tesselation, which creates a set of tiles each associated with an area centroid. The tiles are the set of points that are closer to that area centroid than any other (Denison and Holmes, 2001). We then define neighbours as those villages whose tiles share an edge. Figure 5.1(b) shows the Dirichlet tesselation of the 20 villages in our study region.

We wish to estimate sex-by-age specific mortality rates in children under 5 years of age. We dichotomize age into less than 1 year, and 1-4 years old. Hence, $j = 1, \ldots, 4$ indexes the four sex and age categories. We assume a population of size $N = 28,000$, where an equal number are assigned to each village. Hence, there are $N_i = 1,400$ children in each village, of which 700 are girls and 700 are boys, with 150 of each sex being less than 1 years old, and 550 of each sex being 1-4 years old. We assume there are two area level covariates, socioeconomic status (SES) denoted by x_{1i} and population density denoted by x_{2i}, $i = 1, \ldots, 20$, and they are generated independently from a Unif(0,1) distribution. The non-spatial random effects are generated from a normal distribution with mean zero and variance $\sigma_v^2 = 0.22$, while the spatial random effects were generated from an ICAR model with zero mean and conditional variance $\omega_u^2 = 0.48$. Figure 5.2 maps the simulated spatial



(a) (b)

Figure 5.1: (a) Map of the 20 village centroids used in the simulation study within the Agincourt region outline. (b) The 20 villages of the simulation study, with Dirichlet tesselation defining neighbourhood structure.



(a) (b)

Figure 5.2: Maps of the simulated log residual relative risks for the Agincourt simulation study: (a) V_i, (b) U_i.

and non-spatial random effects. The spatial dependence is apparent, with higher risk in the eastern villages, and much lower risk in the western villages.

Based loosely on the real values from the Agincourt HDSS site, we assume the probability of death for young girls (*i.e.* less than 1 years old) is 0.050, for young boys is 0.105, for older girls (*i.e.* 1-4 years old) is 0.031, and for older boys is 0.066. We also assume that the odds of death is reduced by two-thirds for a one unit increase in SES, while the odds of death doubles for a one unit increase in population density. Hence, we use the following parameter values in the simulation:

$$\begin{aligned}
\gamma_1 &= \text{logit}(0.050) \\
\gamma_2 &= \text{logit}(0.105) \\
\gamma_3 &= \text{logit}(0.031) \\
\gamma_4 &= \text{logit}(0.066) \\
\beta_1 &= -1.1
\end{aligned}$$

$$\beta_2 \;=\; 0.7.$$

Combining all elements of the model, we generate deaths, Y_{ij}, from village i and stratum j by randomly drawing from a Binomial distribution with probabilities given by

$$\text{logit}(p_{ij}) = \beta_1 x_{1i} + \beta_2 x_{2i} + \gamma_j + V_i + U_i.$$

This resulted in a total of $1,230$ deaths. Figure 5.3 displays the predicted probabilities of death in the 20 villages for each age-sex stratum. We see clearly the higher probabilities of death in boys, and for young children. The total number of deaths ranged from 6 to 191 across villages.

Three villages are selected to be HDSS sites, within which extensive information is collected. The choice of villages is determined by the values of the two area-level covariates to ensure variation in these values for the HDSS sites, so the effects of these variables are not estimated to be zero. We selected the villages with both large SES and large population density, small SES and small population density, followed by a randomly sampled third village. We assign independent normal prior distributions with large variances for $\gamma_1, \gamma_2, \gamma_3, \gamma_4, \beta_1$ and β_2, and a Gamma$(1, 0.140)$ hyperprior distribution is used for the precision parameter τ_v while a Gamma$(1, 0.20/0.42)$ hyperprior is used for τ_u. We provide details on the choice of priors in Appendix C.1.

We compare four different sampling methods to show the informed sampling method we propose samples more events and produces better predictions for the number of deaths in each area stratified by age and sex (measured in terms of the mean squared error). The total sample size for each sampling design is $5,200$ children. This number was chosen to ensure that all sampling methods described below sample the same number of children.

1. **One-stage sampling from five villages:** Individuals are randomly sampled from a small number of villages. In particular, we sample five villages at random and sample $1,040$ children from each village.

2. **One-stage sampling from all villages:** We sample an equal number of children from all 20 villages (260 from each village).

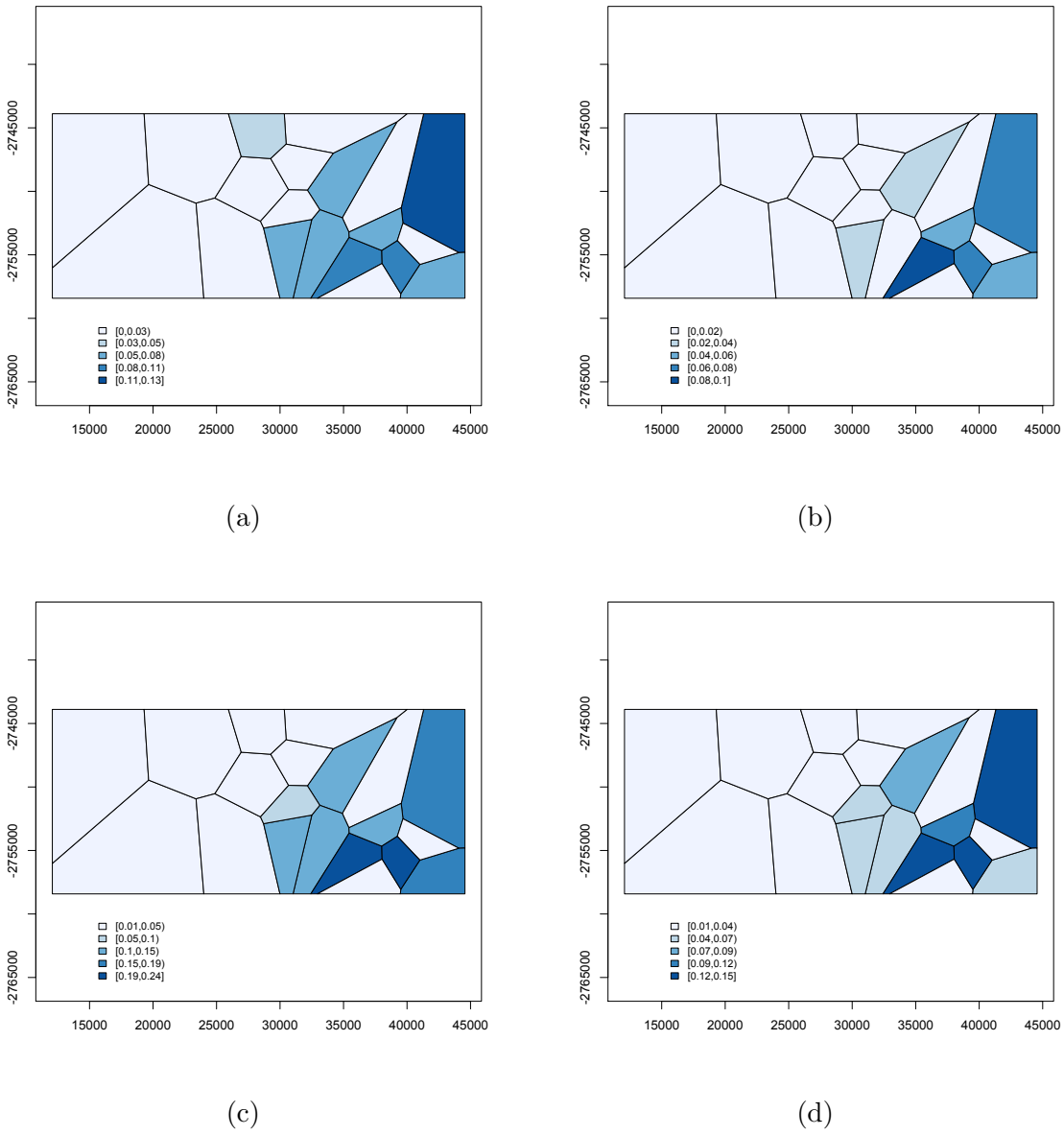


Figure 5.3: Probabilities of death for the simulated data by region in: (a) young girls, (b) older girls, (c) young boys, and (d) older boys. Note the difference in scale for each of the maps.

3. **HDSS with random sampling:** We select all children from the HDSS villages, and a random sample of $1,000$ children from the 17 remaining non-HDSS villages, which results in 59 children being sampled from each remaining village.

4. **HDSS with informative sampling:** We employ the informative sampling procedure where all children from the HDSS villages are sampled, and a total of $1,000$ children are sampled from the remaining non-HDSS villages in the manner described in Section 5.2.

Within each of these sampling schemes, various models may be fitted to the data to obtain the predicted village-age-sex specific death counts, $\widehat{Y}_{ij}$. We fit three additional models to model (5.1) to show that the spatial random effects model provides more reliable estimates.

I **Naïve Model:** The baseline model estimates a single probability as the overall empirical risk, $\widehat{p} = y_{\cdot\cdot}/n_{\cdot\cdot}$. The predicted number of deaths in each village and age-sex stratum is then (5.2) with $\widehat{p}_{ij} = \widehat{p}$ for $i = 1, \ldots, 20, j = 1, \ldots, 4$.

II **Age & Sex Model:** The age and sex model estimates four probabilities as the empirical risks, and assumes the risk is the same in all villages, $\widehat{p}_j = y_{\cdot j}/n_{\cdot j}$. The predicted number of deaths in each village and age-sex strata is then (5.2) with $\widehat{p}_{ij} = \widehat{p}_j$ for $i = 1, \ldots, 20$.

III **Logistic Regression Covariate Model:** We fit a logistic regression model to all villages sampled and estimate stratum and village effects:

$$\text{logit } p_{ij} = x_{1i}\boldsymbol{\beta}_1 + x_{2i}\boldsymbol{\beta}_2 + \gamma_j, \tag{5.3}$$

where $i = 1, \ldots, 20, j = 1, ..., 4$. Once we have estimates $\widehat{\gamma}_j$, $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, we can obtain the fitted probabilities:

$$\widehat{p}_{ij} = \frac{\exp(x_{1i}\widehat{\boldsymbol{\beta}}_1 + x_{2i}\widehat{\boldsymbol{\beta}}_2 + \widehat{\gamma}_j)}{1 + \exp(x_{1i}\widehat{\boldsymbol{\beta}}_1 + x_{2i}\widehat{\boldsymbol{\beta}}_2 + \widehat{\gamma}_j)},$$

which may be used in (5.2).

IV **Logistic Regression Random Effects Covariate Model:** The logistic regression random effects model requires all villages to have sampled data and estimates stratum and village-level effects as in (5.3) and in addition, introduces random effects to account for unmeasured village-level covariates. Specifically, we assume model (5.1) and proceed as described in Section 5.3.

For $s = 1, \ldots, S$ simulations, we subsample from our generated population using each of the four sampling designs described, and fit models I – IV to the sampled data. We denote the estimated number of deaths in survey village i and stratum j at simulation s by $\widehat{Y}_{ij}^{(s)}$:

$$\widehat{Y}_{ij}^{(s)} = y_{ij}^{(s)} + (N_{ij} - n_{ij}) \times \widehat{p}_{ij}^{(s)},$$

where the $\widehat{p}_{ij}^{(s)}$ are obtained from fitting models I-IV. Then, the mean squared error (MSE) is given by

$$\begin{aligned}
\text{MSE} &= \sum_{i=1}^{20}\sum_{j=1}^{4}\left(\frac{1}{S}\sum_{s=1}^{S}(Y_{ij} - \widehat{Y}_{ij}^{(s)})^2\right) \\
&= \sum_{i=1}^{20}\sum_{j=1}^{4}(\overline{\widehat{Y}}_{ij} - Y_{ij})^2 + \sum_{i=1}^{20}\sum_{j=1}^{4}\left(\frac{1}{S}\sum_{s=1}^{S}(\widehat{Y}_{ij}^{(s)} - \overline{\widehat{Y}}_{ij})^2\right) \\
&= \sum_{i=1}^{20}\sum_{j=1}^{4}\left(\text{Bias}(\widehat{Y}_{ij})^2 + \text{Var}(\widehat{Y}_{ij})\right),
\end{aligned}$$

where

$$\overline{\widehat{Y}}_{ij} = \frac{1}{S}\sum_{s=1}^{S}\widehat{Y}_{ij}^{(s)}$$

is the mean of the predicted counts over simulations in village i and stratum j. For interpretation of the results, we also evaluate the MSE associated with estimating the probabilities p_{ij} of death in village i and stratum j.

Table 5.1 displays the bias, variance and MSE, as well as the number of sampled deaths, based on 100 simulations for each combination of sampling strategy and analytical model. It is clear that the informed sampling design combined with a spatial logistic regression model outperforms all other combinations, when performance is measured in terms of the MSE for Y. Additionally, the informative sampling design captures more actual deaths than any of the other sampling schemes. The one-stage sampling schemes capture the

Table 5.1: Results from 100 simulations; 1,230 deaths. Model I=Naïve model; Model II=Age & sex model; Model III=Logistic regression covariate model; Model IV=Logistic regression covariate model with non-spatial and spatial random effects. It is not possible to fit the spatial model to the one-stage sampling from five villages plan since there are data from 5 villages only.

Sampling Design	Model	# Deaths	Y Based Estimates			p Based Estimates		
			Bias	Variance	MSE	Bias	Variance	MSE
One-stage	I	229	126.1	5179.8	21069.2	0.467	0.019	0.237
Sampling from	II	229	117.2	4854.3	18585.9	0.427	0.028	0.211
Five Villages	III	229	103.4	41508.2	52192.7	0.340	0.316	0.432
	IV	229	–	–	–	–	–	–
One-stage	I	226	154.5	217.3	24093.4	0.468	0.0003	0.219
Sampling from	II	226	158.9	1723.4	26961.2	0.533	0.045	0.329
All Villages	III	226	140.9	466.9	20330.6	0.343	0.005	0.123
	IV	226	127.3	1046.4	17239.9	0.110	0.014	0.026
HDSS with	I	479	219.6	45.3	48253.7	0.593	0.0001	0.352
Random	II	479	271.1	115.9	73584.7	1.224	0.002	1.500
Sampling	III	479	175.7	84.6	30964.9	0.541	0.0003	0.293
	IV	479	31.8	2649.1	3663.2	0.128	0.023	0.039
HDSS with	I	492	228.8	52.9	52422.8	0.607	0.0001	0.369
Informative	II	492	264.3	120.0	69955.9	1.114	0.002	1.242
Sampling	III	492	169.4	105.3	28791.2	0.519	0.001	0.270
	IV	492	27.9	2135.4	2913.2	0.127	0.018	0.034

least amount of deaths, where the scheme sampling from five randomly sampled villages captured 229 deaths and the scheme sampling from all villages captured 226 deaths. The sampling schemes that sample all individuals from the HDSS villages plus a sample from the remaining villages capture more than double the number of deaths from either of the one-stage sampling designs. They capture 447 deaths from the HDSS sites, plus an additional 32 deaths for random sampling and 45 deaths for informative sampling for a total of 479 and 492 deaths, respectively. In addition, when we increase the size of the survey sample to 5,000 for the HDSS plus survey sample schemes (so the total sample size is 9,200), the MSE based on the probabilities of death p_{ij} are the smallest for the informative sampling

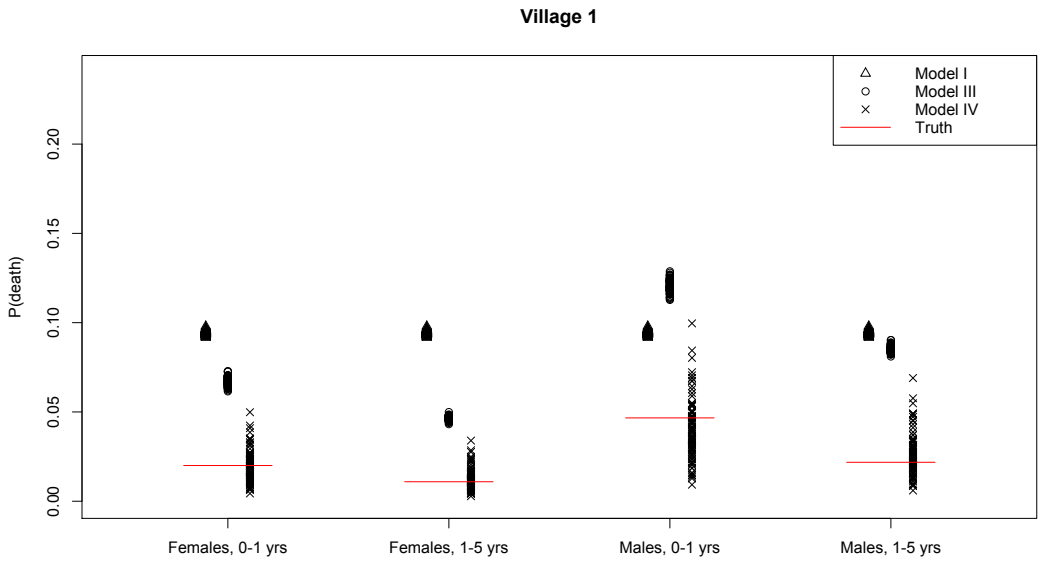


Figure 5.4: Comparing the bias and variance for models I, III and IV in village 1 for the informed sampling design across 100 simulations. Model I=Naïve model; Model II=Age & sex model; Model III=Logistic regression covariate model; Model IV=Logistic regression covariate model with non-spatial and spatial random effects.

using a spatial logistic regression model. The results are provided in Appendix C.1.

The sampling scheme which samples from five randomly sampled villages produces estimates which have both high bias and high variance, resulting in large MSEs for each of the analytical models. In addition, we notice that the three models which do not include random effects terms for village produce highly biased estimates, resulting in worse performance than the spatial model (measured in terms of the MSE) for each of the remaining sampling schemes. In addition, the sampling schemes involving sampling all individuals from the HDSS villages plus a sample from the remaining villages combined with fitting a spatial model to the data results in MSEs that are an order of magnitude less than any of the other combinations.

Figure 5.4 compares the bias and the variance for the informed sampling design for the naïve model, the logistic regression model and the logistic regression model including spatial

and non-spatial random effects in village 1. While the estimates from the random effects model clearly have more variance, they also have much less bias, resulting in smaller MSEs.

Part of the success of the informed sampling design reflects the careful choice of the HDSS villages so that they contain substantial variation in terms of the two village-level covariates. In fact, the HDSS villages selected happened to include the two villages with the greatest number of deaths. When we randomly select three villages to be the HDSS sites, the results still favour the informed sampling approach combined with fitting the sampled data with a spatial logistic regression model. We still see more deaths sampled using the informed sampling design, as well as estimates with far less bias and smallest MSE when fitting the sampled data using a spatial logistic regression model. Results are shown in Appendix C.1. In practice, the choice of HDSS villages may be made due to feasibility and cost of setting up a HDSS site, or HDSS sites may already be set up in a region (for example, as in Agincourt, South Africa).

5.5 Summary

We have proposed a mortality monitoring system, which combines highly informative data, such as that produced by HDSS sites, and survey sampling data from surrounding sites to obtain more sampled deaths, as well as more reliable estimates of counts of deaths. The informed sampling design that we propose fits a mortality model to data from HDSS sites only, which becomes the basis for predicting the number of deaths in the surrounding areas. We use these predictions to inform the survey sample of the remaining villages and we sample more heavily from populations with higher predicted counts of death. The sampled deaths are then modeled as functions of known demographic factors, area-level characteristics, and spatial smoothing is used in order to tune the model to each area.

We simulated a data set based loosely on data from the Agincourt HDSS site in South Africa, using 20 villages from that region as our study area. Using this data set, we evaluated the performance of four different sampling schemes and four different analytical models in terms of the number of deaths sampled, bias, variance, and MSE. The informed sampling scheme which fits the sampled data with a logistic regression model including spatial random effects resulted in estimated counts of death with the lowest MSE. Moreover, of the

four different sampling schemes, the informed sampling design captured substantially more deaths than the other three sampling schemes. Verbal autopsy methods can be applied to all or a fraction of these deaths to assign cause, which can be used to construct cause-specific distributions of deaths. Monitoring these distributions over time provides insight into the impact of interventions on specific causes of death. In addition, sampling more deaths allows for more accurate estimates of cause-specific mortality rates. This simulation study showed the benefits of the proposed informed sampling scheme combined with a spatial logistic regression model. Therefore, there is reason to further pursue the informed sampling design. However, as we discuss in the next chapter, more realistic simulations studies will need to be done to approximate real life situations more faithfully.

Chapter 6

DISCUSSION

6.1 Conclusions

In this thesis, we described an alternative approach to the current frequentist methods for the analysis of data arising from two-phase studies. The Bayesian method that we proposed uses a log-linear model to model the disease-exposure-confounder relationship, and assigns prior distributions to reduced sets of main effect and interaction terms in the log-linear model. A Markov chain Monte Carlo scheme was used for sampling from the posterior distribution using a single Metropolis-Hastings step. In the examples considered, the Bayesian and non-parametric maximum likelihood (NPML) approaches gave nearly identical inference. A simulated data example demonstrated the superiority of the Bayesian approach over NPML in the case of sparse data. Further, a simulation study showed that the Bayesian parameter estimates resulted in smaller mean squared errors than those obtained from the weighted, pseudo- and non-parametric maximum likelihood approaches.

We extended this model to include random effects in the log-linear model to perform various types of smoothing. In the first instance, random effects can be included in the log-linear model to smooth the cell probabilities in large contingency tables, particularly in the case of sparse data. In this case, we instead modeled the exposure-confounder interaction terms using independent mean zero normal prior distributions, and assigned a gamma hyperprior distribution to the precision parameter. A simulated example demonstrated the potential benefits of a random effects log-linear model over a fixed effects log-linear model.

In the second instance, we include spatial and non-spatial random effects in the log-linear model to control for residual spatial confounding. An intrinsic conditional autoregressive prior is assigned for the spatial random effects, which smooths spatial effects to the mean of the spatial effects in close-by areas. With this model, the nature of the spatial dependency is defined by the neighbourhood structure. The neighbour definition that we adopted is

that based on contiguity. We analysed the North Carolina infant mortality data using spatial and non-spatial random effects and found that the results agreed well with those from performing a similar analysis on the individual-level data. A simulated example was used to illustrate the improved reliability of estimates from the spatial random effects model (over the simpler non-spatial random effects model) in the case of spatially correlated data.

A major drawback of the Bayesian approach we have outlined is that it requires a joint exposure-confounder model. Hence, in situations where the data are numerous and there is no desire to include random effects in the model, we would recommend the NPML approach. However, the Bayesian approach is beneficial in sparse data situations in which the covariates are all discrete, where the NPML methodology breaks down. In this case, informative priors must be used in practice.

The use of the log-linear model is another potential limitation of the Bayesian approach since it will contain many parameters in situations where x and/or z take on many values. In this case, it may be preferable to use an unsaturated or random effects log-linear model. The ability to include random effects into the log-linear model in the Bayesian approach allows for much greater modeling flexibility than the current frequentist approaches.

Another limitation of the Bayesian approach arises when we are attempting to fit the wrong model. Although not explored in the thesis, the results of Scott and Wild (2002) provide insight into how the Bayesian approach might perform when the logistic or logistic-normal models that we use do not hold. Scott and Wild (1986) compared the robustness under model misspecification of survey weighted and semi-parametric maximum likelihood (ML) approaches to fitting logistic regression models with case-control data. The authors concluded that although semi-parametric ML estimation is more efficient when the model is true, survey weighting is more robust since it is the only approach that leads to consistent estimates in the presence of model misspecification. More recently, however, Scott and Wild (2002) argued that this view is not always justified. In their paper, the authors compared the performance of semi-parametric and survey weighted ML approaches fitting a linear logistic model when the true underlying model is quadratic, and found that the survey weighted ML approach gave a better approximation to the true slope only for those individuals at higher risk, whereas the semi-parametric ML approach was better for individuals at moderate risk

(which comprised roughly 95% of the population in their example). In particular, Scott and Wild (2002) suggest using the survey weighted method in situations where interest lies particularly in high risk individuals; when interest lies in more typical individuals from the population, the semi-parametric ML method will have smaller bias and will be more efficient. These results suggest that the Bayesian approach may produce reliable results in situations where we are interested in individuals at moderate risk. In those situations where we are interested in high risk individuals, the results from the Bayesian approach should be interpreted with caution.

In our approach to small area estimation in the context of the developing world, we proposed a mortality monitoring system, which combines detailed data from health and demographic surveillance systems (HDSS) sites with sample survey data from surrounding areas. The proposed approach provides a stepping stone to methods that incorporate the longitudinal data available from the HDSS sites. A preliminary simulation study illustrated the ability of the informed sampling design to capture more deaths compared to one-stage random sampling designs. Further, it showed that the informed sampling design combined with a spatial analysis of the sampled data results in estimates with the lowest MSE.

6.2 Future Work

The equivalence of the Bayesian approaches in the case of simple random and case-control sampling at phase I will continue to be explored. In particular, since the simple random sampling approach considered by Scott and Wild (1991, 1997) is actually a prospective sample (where individuals are sampled within the strata of the confounder variable), we hope to show the equivalence of the Bayesian approach under prospective sampling and case-control sampling at phase I, as was done in the frequentist paradigm. Furthermore, we hope to continue exploration of possible prior distributions that can be assigned in such a way that the exposure-confounder model can be integrated out asymptotically, as is done in Seaman and Richardson (2004) in a case-control setting. We also plan to conduct a simulation study to numerically compare the posterior distributions under case-control and simple random sampling at phase I.

In addition, we aim to reduce the computational intensity of the Bayesian approaches,

particularly the random effects approach, as the current computing times can be excessive. One possible approach is to use the hyperblock updating approach described in Knorr-Held and Rue (2002) in which all parameters and hyperparameters are updated in a single block. The authors demonstrated superior mixing and thus smaller simulation error for parameter estimates in the hyperblock approach.

Our approach only considers the case when the covariates are all discrete. The extension to the continuous or the mixed continuous and discrete covariates situations is not straightforward, though the approaches of Sinha, Mukherjee, and Ghosh (2004) and Dunson and Xing (2009) offer possible avenues for addressing this deficiency. In particular, the authors suggest assuming a Dirichlet process with support on the distribution of the continuous covariates. Normal prior distributions can be used on the other covariates, as is done in Sinha et al. (2004). In either case, the estimation of all parameters can still be done via a Markov chain Monte Carlo scheme.

We would like to extend the Bayesian approach to the analysis of two-phase data to a related design, the aggregate data design. Prentice and Sheppard (1995) consider a design in which detailed exposure and confounding factor data are available on a random sample of individuals from each area. The analysis method they propose involves the use of estimating equations for relative rate parameters. As we saw with two-phase studies, the current approach suffers from poor performance in small sub-samples (Guthrie and Sheppard, 2001) and difficulties in incorporating random effects. Hence, there is strong motivation for a non-GEE Bayesian approach.

For the small area estimation problem, we wish to extend the methodology to include time-varying covariates in the logistic model fit to the HDSS data to get the full benefit of the data from the HDSS sites. Hence, additional simulation studies will need to be performed to further assess the performance of the proposed informed sampling design. These should approximate a wider variety of real situations using additional covariates, different geographies and varying mortality rates. In addition, various spatial dependencies, and perhaps neighbour definitions, could be explored. These simulation studies can be informed using other existing HDSS sites in low- and middle-income countries, as well as historical datasets at the national level. Also, it may be worth exploring a survey sampling

scheme in which areas with the lowest predicted mortality rates are oversampled to obtain more reliable estimates for those areas with fewer deaths. Furthermore, a mixture of both sampling schemes could be explored (that is, a survey sampling scheme which combines oversampling areas with the highest predicted mortality rates and oversampling areas with lowest predicted mortality rates) to combine the benefits of both sampling schemes. The performances of these proposed schemes could be compared to that of the currently adopted scheme.

BIBLIOGRAPHY

C. AbouZahr, J. Cleland, F. Coullare, S.B. Macfarlane, F.C. Notzon, P. Setel, S. Szreter, R.N. Anderson, A. Bawah, A.P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, T. Evans, X.C. Figueroa, C.K. George, L. Gollogly, R. Gonzalez, D.R. Grzebien, K. Hill, Z. Huang, T.H. Hull, M. Inoue, R. Jakob, P. Jha, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, A.D. Lopez, D.M. Fat, M. Merialdi, L. Mikkelsen, J.K. Nien, C. Rao, K. Rao, O. Sankoh, K. Shibuya, N. Soleman, S. Stout, V. Tangcharoensathien, P.J. van der Maas, F. Wu, G. Yang, and S. Zhang. The way forward. *The Lancet*, 370:1791–1799, 2007.

A. Agresti. *Analysis of Ordinal Categorical Data*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2010.

A. Agresti and D.B. Hitchcock. Bayesian inference for categorical data analysis. *Statistical Methods and Applications: Journal of the Italian Statistical Society*, 14:297–330, 2005.

D. Ashby, J.L. Hutton, and M.A. McGee. Simple Bayesian analyses for case-control studies in cancer epidemiology. *Statistician*, 42:385–397, 1993.

L. Bernardinelli, C. Pascutto, N.G. Best, and W.R. Gilks. Disease mapping with errors in covariates. *Statistics in Medicine*, 16:741–752, 1997.

J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24:179–195, 1975.

J. Besag and P.J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55:25–37, 1993.

J. Besag and C. Kooperberg. On conditional and intrinsic auto-regressions. *Biometrika*, 82: 733–746, 1995.

J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, 43:1–59, 1991.

N. Best, L Waller, A. Thomas, E. Conlon, and R. Arnold. Bayesian models for spatially correlated disease and exposure data. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Sixth Valencia international meeting on Bayesian statistics*, pages 131–156, London, 1999. Oxford University Press.

N. Best, S. Richardson, and A. Thomson. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59, 2005.

J.T. Boerma and S.K. Stansfield. Health Statistics 1 Health statistics now: are we making the right investments? *Tuberculosis*, 369:779–786, 2007.

N. Breslow and N.E. Day. *Statistical Methods in Cancer Research, Volume 1- The Analysis of Case-Control Studies*. Scientific Publications No. 32. Lyon: International Agency for Research on Cancer, 1980.

N.E. Breslow and K.C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75:11–20, 1988.

N.E. Breslow and N. Chatterjee. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, 48:457–468, 1999.

N.E. Breslow and R. Holubkov. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B*, 59:447–461, 1997a.

N.E. Breslow and R. Holubkov. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine*, 16:103–116, 1997b.

S.P.J. Byrne and A.P. Dawid. Retrospective-prospective symmetry for the Bayesian analysis of case-control studies. *Biometrika*, 2012. To appear.

C.X. Chen, T. Lumley, and J. Wakefield. The use of sampling weights in Bayesian hierarchical models for small area estimation. Technical Report 583, Department of Statistics, University of Washington, 2011.

D.G. Clayton and J. Kaldor. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–682, 1987.

W.G. Cochran. *Sampling Techniques, Third Edition.* Wiley, New York, 1977.

P. Congdon and P. Lloyd. Estimating small area diabetes prevalence in the US using behavioral risk factor surveillance system. *Journal of Data Science*, 8:235–252, 2010.

B.A. Coull and A. Agresti. Generalized log-linear models with random effects, with application to smoothing contingency tables. *Statistical Modelling*, 3:251–271, 2003.

N. Cressie and N.H. Chan. Spatial modelling of regional variables. *Journal of the American Statistical Association*, 84:393–401, 1989.

P. Damien, J.C. Wakefield, and S.G. Walker. Gibbs sampling for bayesian nonconjugate and hierarchical models using auxiliary variables. *Journal of the Royal Statistical Society, Series B*, 61:331–344, 1999.

A.P. Dawid and S.L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposible graphical models. *The Annals of Statistics*, 21:1272–1317, 1993.

P. Dellaportas and J.J. Forster. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86:615–633, 1999.

D.G.T. Denison and C.C. Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57:143–149, 2001.

P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26:363–397, 1998.

P.J. Diggle, J.A. Tawn, and R.A. Moyeed. Model-based geostatistics (with discussion). *Applied Statistics*, 47:299–350, 1998.

D.B. Dunson and C. Xing. Non-parametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104:1042–1051, 2009.

P. Elliott, J.C. Wakefield, N.G. Best, and D.J. Briggs, editors. *Spatial Epidemiology: Methods and Applications*. Oxford University Press, 2000.

V.L. Ernster. Nested case-control studies. *Preventive Medicine*, 23:587–590, 1994.

M. Evans and G.H. Jang. Weak informativity and the information in one prior relative to another. *Statistical Science*, 26:423–535, 2011.

R.E. Fay, III and R.A. Herriot. Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74:269–277, 1979.

W.D. Flanders and S. Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10:739–747, 1991.

Y. Fong, H. Rue, and J. Wakefield. Bayesian inference for generalized linear mixed models. *Biostatistics*, 11:397–412, 2010.

E.B. Fowlkes, A.E. Freeny, and J.M. Landwehr. Evaluating logistic models for large contingency tables. *Journal of the American Statistical Association*, 83:611–622, 1988.

F. Galindo-Garre, J.K. Vermunt, and W.P. Bergsma. Bayesian posterior estimation of logit parameters with small samples. *Sociological Methods and Research*, 33:88–117, 2004.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall, New York, 2004.

A. Gelman, A. Jakulin, M.G. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2: 1360–1383, 2008.

M. Ghosh, L. Zhang, and B. Mukherjee. Equivalence of posteriors in the Bayesian analysis of the multinomial-Poisson transformation. *Metron-International Journal of Statistics*, LXIV:19–28, 2006.

M. Ghosh, J. Song, J.J. Forster, R. Mitra, and B. Mukherjee. On the equivalence of posterior inference based on retrospective and prospective likelihoods: application to a case-control study of colorectal cancer. *Statistics in Medicine*, 31:2196–2208, 2012.

I.J. Good. On the estimation of small frequencies in contingency tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 18:113–124, 1956.

I.J. Good. *The estimation of probabilities: an essay on modern Bayesian methods*. M.I.T. Press, Cambridge, MA, 1965.

I.J. Good. A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 29:399–431, 1967.

P. Gustafson, N.D. Le, and M. Vallée. A Bayesian approach to case-control studies with errors in covariables. *Biostatistics*, 3:229–243, 2002.

K.A. Guthrie and L. Sheppard. Overcoming biases and misconceptions in ecological studies. *Journal of the Royal Statistical Society, Series A*, 164:141–154, 2001.

M.S. Handock and M.L. Stein. A bayesian analysis of kriging. *Technometrics*, 35:403–410, 1993.

K. Hill, A.D. Lopez, K. Shibuya, P. Jha, C. AbouZahr, R.N. Anderson, A. Bawah, A.P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, J. Cleland, F. Coullare, T. Evans, X. Carrasco Figueroa, C.K. George, L. Gollogly, R. Gonzalez, D.R. Grzebien, Z. Huang, T.H. Hull, M. Inoue, R. Jakob, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, D.M. Fat, S. Macfarlane, P. Mahapatra, M. Merialdi, L. Mikkelsen, J.K. Nien, F.C. Notzon, C. Rao, K. Rao, O. Sankoh, P.W. Setel, N. Soleman, S. Stout, S. Szreter, V. Tangcharoensathien, P.J. van der Maas, F. Wu, G. Yang, S. Zhang, and M. Zhou. Interim measures for meeting needs for health sector data: births, deaths, and causes of death. *The Lancet*, 370:1726–35, 2007.

R. Holubkov. *Maximum likelihood estimation in two-stage case-control studies*. PhD thesis, University of Washington, Seattle, 1995.

R. Horton. Counting for health. *The Lancet*, 370:1526, 2007.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

D.A. Hsieh, C.F. Manski, and D. McFadden. Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association*, 80:651–662, 1985.

K. Kahn, S.M. Tollman, M.A. Collinson, S.J. Clark, R. Twine, B.D. Clark, M. Shabangu, F.X. Gmez-Oliv, O. Mokoena, and M.L. Garenne. Research into health, population and social transitions in rural south africa: Data and methods of the agincourt health and demographic surveillance system1. *Scandinavian Journal of Public Health*, 35:8–20, 2007.

E. Kass and L. Wasserman. A reference Bayesian test for nested hypothesis and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90: 928–934, 1995.

J.E. Kelsall and J.C. Wakefield. Discussion of "Bayesian models for spatially correlated disease and exposure data" by N. Best, L. Waller, A. Thomas, E. Conlon and R. Arnold. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Sixth Valencia international meeting on Bayesian statistics*, London, 1999. Oxford University Press.

R. King and S.P. Brooks. Prior induction in log-linear models for general contingency table analysis. *The Annals of Statistics*, 29:715–747, 2001.

L. Knorr-Held and H. Rue. On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29:597–614, 2002.

M.W. Knuiman and T.P. Speed. Incorporating prior knowledge into the analysis of contingency tables. *Biometrics*, 44:1061–1071, 1988.

T.D. Koepsell and N.S. Weiss. *Epidemiologic Methods: Studying the Occurrence of Illness*. Oxford University Press, 2003.

E.L. Korn and B.I. Graubard. *Analysis of Health Surveys.* John Wiley and Sons, New York, 1999.

T. Leonard. Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 37:23–37, 1975.

B.G. Leroux, X. Lei, and N. Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M.E. Halloran and D.A Berry, editors, *Statistical Models in Epidemiology. the Environment and Clinical Trials*, pages 179–192. Springer, New York, 1999.

D.V. Lindley. The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, 35:1622–1643, 1964.

T.S. Lumley. *Complex Surveys: A Guide to Analysis Using R (Wiley Series in Survey Methodology).* Wiley, 2010.

D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

P. Mahapatra, K. Shibuya, A.D. Lopez, F. Coullare, F.C. Notzon, C. Rao, and S. Szreter. Civil registration systems and vital statistics: successes and missed opportunities. *The Lancet*, 370:1653–1663, 2007.

R.J. Marshall. Bayesian analysis of case-control studies. *Statistics in Medicine*, 7:1223–1230, 1988.

H. Massam, J. Liu, and A. Dobra. A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics*, 37:3431–3467, 2009.

B. Matérn. *Spatial Variation, Second Edition.* Springer-Verlag, Berlin, 1986.

C.D. Mathers, D.M. Fat, M. Inoue, C. Rao, and A.D. Lopez. Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bulletin of the World Health Organization*, 83:171–177, 2005.

P. Müller and K. Roeder. A Bayesian semi-parametric model for case-control studies with errors in variables. *Biometrika*, 84:523–537, 1997.

P. Müller, G. Parmigiani, J. Schildkraut, and L. Tardella. A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics*, 55:858–866, 1999.

J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33:101–116, 1938.

M. Nurminen and P. Mutanen. Exact Bayesian analysis of two proportions. *Scand. J. Statist*, 14:67–77, 1987.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

R.L. Prentice and L. Sheppard. Aggregate data studies of disease risk factors. *Biometrika*, 82:113–25, 1995.

J.N.K. Rao. *Small Area Estimation.* John Wiley, New Jersey, 2003.

M. Reilly and M.S. Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82:299–314, 1995.

S. Richardson, C. Montfort, M. Green, G. Draper, and C. Muirhead. Spatial variation of natural radiation and childhood leukaemia incidence in great britain. *Statistics in Medicine*, 14:2487–2501, 1995.

G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

R.M. Royall. On finite population sampling theory under certain linear regression models. *Biometrika*, 57:377–387, 1970.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71:319–392, 2009.

C.-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, New York, 1992.

J. Schill, J. H. Jockel, K. Drescher, and J. Timm. Logistic analysis in case-control studies under validation sampling. *Biometrika*, 84:57–71, 1993.

A. Scott and C. Wild. Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 48:170–182, 1986.

A. Scott and C. Wild. On the robustness of weighted methods for fitting models to case-control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:207–219, 2002.

A.J. Scott and C.J. Wild. Fitting logistic regression in stratified case-control studies. *Biometrics*, 47:497–510, 1991.

A.J. Scott and C.J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 51:54–71, 1997.

S.R. Seaman and S. Richardson. Bayesian analysis of case-control studies with categorical covariates. *Biometrika*, 88:1073–1088, 2001.

S.R. Seaman and S. Richardson. Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika*, 91:15–25, 2004.

P. W. Setel. Verbal autopsy and global mortality statistics: if not now, then when? *Population Health Metrics*, 9:20, 2011.

P.W. Setel, S.B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout, and C. AbouZahr. A scandal of invisibility: making everyone count by counting everyone. *The Lancet*, 370: 1569–77, 2007.

B. Singh, G. Shukla, and D. Kundu. Spatio-temporal models in small-area estimation. *Survey Methodology*, 31:183–195, 2005.

S. Sinha, B. Mukherjee, and M. Ghosh. Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics*, 60:41–49, 2004.

A.-M. Staicu. On the equivalence of prospective and retrospective likelihood methods in case-control studies. *Biometrika*, 97:990–996, 2010.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22: 1701–1762, 1994.

S. Wacholder and C.R. Weinberg. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. *Biometrics*, 50:350–357, 1994.

S. Wacholder, D.T. Silverman, J.K. McLaughlin, and J.S. Mandel. Selection of controls in case-control studies: III. design options. *American Journal of Epidemiology*, 135:1042–1050, 1992.

J. Wakefield. *Bayesian and Frequentist Regression Analysis*. Springer, 2012.

J. Wakefield and S. Haneuse. Overcoming eological bias using the two-phase study design. *American Journal of Epidemiology*, 167:908–916, 2008.

J. Wakefield, S. Haneuse, A. Dobra, and E. Teeple. Bayes computation for ecological inference. *Statistics in Medicine*, 30:1381–1396, 2011.

J.C. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8: 158–183, 2007.

A.M. Walker. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*, 38:1025–1032, 1982.

L.A. Waller, B.P. Carlin, H. Xia, and A.E. Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617, 1996.

R. Weiss, R. Berk, W. Li, and M. Farrell-Ross. Death penalty charging in Los Angeles county: An illustrative data analysis using skeptical priors. *Sociological Methods and Research*, 28:91–115, 1999.

E.J. White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115:119–128, 1982.

A.S. Whittemore. Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society, Series B*, 59:589–602, 1997.

D.A. Williams. 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 31:949–952, 1975.

J. York, D. Madigan, I. Heuch, and R.T. Lie. Birth defects registered by double sampling: A Bayesian approach incorporating covariates and model uncertainty. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44:227–242, 1995.

M. Zelen and R.A. Parker. Case-control studies and Bayesian inference. *Statistics in Medicine*, 5:261–269, 1986.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

A.1 Details on the Simulation Study

As described in Section 3.3.1, two different prior distributions for $\boldsymbol{\beta}$ were evaluated. The induced multivariate normal prior distribution for $\boldsymbol{\Lambda}^Y = (\lambda_0^Y, \lambda_{00}^{YX}, \lambda_{00}^{YZ}, \lambda_{000}^{YXZ})$ has mean zero and variance-covariance matrix $\Lambda = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}_i\left(\boldsymbol{C}^{-1}\right)^T, i = 1, 2$, where

$$\boldsymbol{C} = \begin{pmatrix} -2 & -2 & -2 & -2 \\ 0 & 4 & 0 & 4 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & -8 \end{pmatrix} \tag{A.1}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 100^2 & 0 & 0 & 0 \\ 0 & \frac{\log(5)}{1.96} & 0 & 0 \\ 0 & 0 & \frac{\log(5)}{1.96} & 0 \\ 0 & 0 & 0 & \frac{\log(5)}{1.96} \end{pmatrix} \tag{A.2}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 100^2 & 0 & 0 & 0 \\ 0 & 100^2 & 0 & 0 \\ 0 & 0 & 100^2 & 0 \\ 0 & 0 & 0 & 100^2 \end{pmatrix},$$

where $\boldsymbol{\Sigma}_1$ corresponds to the informative priors, and $\boldsymbol{\Sigma}_2$ corresponds to the flat priors. For the remainder of the $\boldsymbol{\lambda}$ nuisance parameters, we assigned the following prior distributions:

$$\begin{aligned} \lambda_0^X &\sim N(0, 1) \\ \lambda_0^Z &\sim N(0, 1) \\ \lambda_{00}^{XZ} &\sim N(0, 1/4). \end{aligned}$$

Tables A.1-A.3 display the bias, variance and MSE for the five analyses of 500 simulated data sets and for each of the three phase II sample sizes.

Table A.1: Bias, variance and mean squared error (MSE) from five analyses of 500 simulated data sets using a small phase II sample size. There were 30 phase II data sets which contained a zero count.

Method	Parameter	Bias	Variance	MSE
	β_0	-7.76×10^{-4}	2.01×10^{-3}	2.01×10^{-3}
Weighted	β_x	-1.59×10^{-2}	4.88×10^{-1}	4.88×10^{-1}
Likelihood*	β_z	-7.84×10^{-2}	2.73×10^{-1}	2.79×10^{-1}
	β_{xz}	9.29×10^{-2}	8.11×10^{-1}	8.19×10^{-1}
	β_0	-3.27×10^{-3}	2.15×10^{-3}	2.16×10^{-3}
Maximum	β_x	4.37×10^{-2}	5.67×10^{-1}	5.69×10^{-1}
Likelihood**	β_z	-7.44×10^{-2}	2.76×10^{-1}	2.81×10^{-1}
	β_{xz}	3.13×10^{-2}	9.00×10^{-1}	9.01×10^{-1}
Bayesian	β_0	5.75×10^{-3}	8.64×10^{-4}	8.97×10^{-4}
Two-phase,	β_x	-2.71×10^{-1}	1.06×10^{-1}	1.79×10^{-1}
Informative	β_z	2.43×10^{-1}	5.68×10^{-2}	1.16×10^{-1}
Priors	β_{xz}	-4.91×10^{-3}	9.56×10^{-2}	9.57×10^{-2}
Bayesian	β_0	-1.41×10^{-2}	2.24×10^{-3}	2.44×10^{-3}
Two-phase,	β_x	1.00×10^{-1}	6.45×10^{-1}	6.55×10^{-1}
Flat	β_z	-3.88×10^{-2}	3.12×10^{-1}	3.13×10^{-1}
Priors	β_{xz}	-5.41×10^{-2}	1.02×10^{0}	1.03×10^{0}
Bayesian	β_0	5.82×10^{-3}	8.69×10^{-4}	9.03×10^{-4}
Two-phase,	β_x	-2.88×10^{-1}	9.95×10^{-2}	1.82×10^{-1}
Informative	β_z	2.63×10^{-1}	4.85×10^{-2}	1.18×10^{-1}
Priors*	β_{xz}	-9.51×10^{-3}	8.77×10^{-2}	8.78×10^{-2}
Bayesian	β_0	-1.37×10^{-2}	2.24×10^{-3}	2.42×10^{-3}
Two-phase,	β_x	4.27×10^{-2}	5.14×10^{-1}	5.15×10^{-1}
Flat	β_z	2.04×10^{-2}	2.10×10^{-1}	2.11×10^{-1}
Priors*	β_{xz}	-5.83×10^{-2}	7.76×10^{-1}	7.79×10^{-1}

*: Excludes 77 data sets where complete separation occurs.

**: Excludes 30 data sets with zero count.

Table A.2: Bias, variance and mean squared error (MSE) from five analyses of 500 simulated data sets using a medium phase II sample size.

Method	Parameter	Bias	Variance	MSE
	β_0	-1.52×10^{-3}	9.78×10^{-4}	9.80×10^{-4}
Weighted	β_x	3.38×10^{-2}	2.61×10^{-1}	2.62×10^{-1}
Likelihood	β_z	-2.80×10^{-2}	8.77×10^{-2}	8.85×10^{-2}
	β_{xz}	-3.82×10^{-3}	3.58×10^{-1}	3.58×10^{-1}
	β_0	-1.52×10^{-3}	9.78×10^{-4}	9.80×10^{-4}
Maximum	β_x	3.38×10^{-2}	2.61×10^{-1}	2.62×10^{-1}
Likelihood	β_z	-2.80×10^{-2}	8.77×10^{-2}	8.85×10^{-2}
	β_{xz}	-3.82×10^{-3}	3.58×10^{-1}	3.58×10^{-1}
Bayesian	β_0	-3.91×10^{-3}	5.47×10^{-4}	5.63×10^{-4}
Two-phase,	β_x	-1.71×10^{-1}	8.62×10^{-2}	1.15×10^{-1}
Informative	β_z	1.46×10^{-1}	2.96×10^{-2}	5.09×10^{-2}
Priors	β_{xz}	5.96×10^{-3}	7.99×10^{-2}	7.99×10^{-2}
Bayesian	β_0	-7.33×10^{-3}	1.00×10^{-3}	1.06×10^{-3}
Two-phase,	β_x	3.62×10^{-2}	2.52×10^{-1}	2.53×10^{-1}
Flat	β_z	8.73×10^{-3}	7.84×10^{-2}	7.85×10^{-2}
Priors	β_{xz}	-4.07×10^{-2}	3.40×10^{-1}	3.42×10^{-1}

Table A.3: Bias, variance and mean squared error (MSE) from five analyses of 500 simulated data sets using a large phase II sample size.

Method	Parameter	Bias	Variance	MSE
	β_0	-9.97×10^{-4}	3.75×10^{-4}	3.76×10^{-4}
Weighted	β_x	3.53×10^{-2}	1.05×10^{-1}	1.06×10^{-1}
Likelihood	β_z	9.97×10^{-4}	3.75×10^{-4}	3.76×10^{-4}
	β_{xz}	-3.53×10^{-2}	1.05×10^{-1}	1.06×10^{-1}
	β_0	-9.97×10^{-4}	3.75×10^{-4}	3.76×10^{-4}
Maximum	β_x	3.53×10^{-2}	1.05×10^{-1}	1.06×10^{-1}
Likelihood	β_z	9.97×10^{-4}	3.75×10^{-4}	3.76×10^{-4}
	β_{xz}	-3.53×10^{-2}	1.05×10^{-1}	1.06×10^{-1}
Bayesian	β_0	2.08×10^{-3}	2.58×10^{-4}	2.64×10^{-4}
Two-phase,	β_x	-8.17×10^{-2}	5.55×10^{-2}	6.21×10^{-2}
Informative	β_z	8.96×10^{-2}	2.22×10^{-4}	8.25×10^{-3}
Priors	β_{xz}	-1.93×10^{-2}	4.35×10^{-2}	4.39×10^{-2}
Bayesian	β_0	-4.50×10^{-3}	3.78×10^{-4}	3.98×10^{-4}
Two-phase,	β_x	3.81×10^{-2}	1.04×10^{-1}	1.05×10^{-1}
Flat	β_z	1.79×10^{-2}	3.72×10^{-4}	6.92×10^{-4}
Priors	β_{xz}	-5.27×10^{-2}	1.04×10^{-1}	1.07×10^{-1}

A.2 Sparse Data Example

In the sparse data example of Section 3.4.2, the informative prior distribution on $\boldsymbol{\beta}$ induces a multivariate normal prior distribution on $\boldsymbol{\Lambda}^Y = (\lambda_0^Y, \lambda_{00}^{YX}, \lambda_{00}^{YZ}, \lambda_{000}^{YXZ})$ that has mean zero and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$, where $\boldsymbol{C}$ and $\boldsymbol{\Sigma}$ are given by (A.1) and (A.2), respectively.

Figures A.1 and A.2 give the trace plots for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, where every 100th sample has been taken in all cases.

A.3 Wilms Tumour Example

For the Wilms tumour example described in Section 3.5.1, the uninformative prior on $\boldsymbol{\beta}$ induces a multivariate normal distribution on $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX})$ with mean zero and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$, where $\boldsymbol{C}$ and $\boldsymbol{\Sigma}$ are the 8×8 matrices given by

$$\begin{aligned}
\boldsymbol{C} &= \begin{pmatrix} -2 & -2 & 0 & 0 & \cdots & 0 \\ 0 & 2 & -2 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 2 & 0 & \cdots & 0 & -2 \\ 0 & 4 & 2 & 2 & \cdots & 2 \end{pmatrix} \\
\boldsymbol{\Sigma} &= \begin{pmatrix} 100^2 & & 0 \\ & \ddots & \\ 0 & & 100^2 \end{pmatrix}.
\end{aligned}$$

For $\boldsymbol{\lambda}^X = (\lambda_1^X, \ldots, \lambda_7^X)$, $\boldsymbol{\lambda}^Z = (\lambda_0^Z)$ and $\boldsymbol{\lambda}^{XZ} = (\lambda_{10}^{XZ}, \ldots, \lambda_{70}^{XZ})$, we assigned:

$$\begin{aligned}
\boldsymbol{\lambda}^X &\sim N_7\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 7 & & -1 \\ & \ddots & \\ -1 & & 7 \end{pmatrix}\right) \\
\lambda_0^Z &\sim N(0,1) \\
\boldsymbol{\lambda}^{XZ} &\sim N_7\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 55 & & -9 \\ & \ddots & \\ -9 & & 55 \end{pmatrix}\right),
\end{aligned}$$

following the Dellaportas and Forster (1999) prescription.

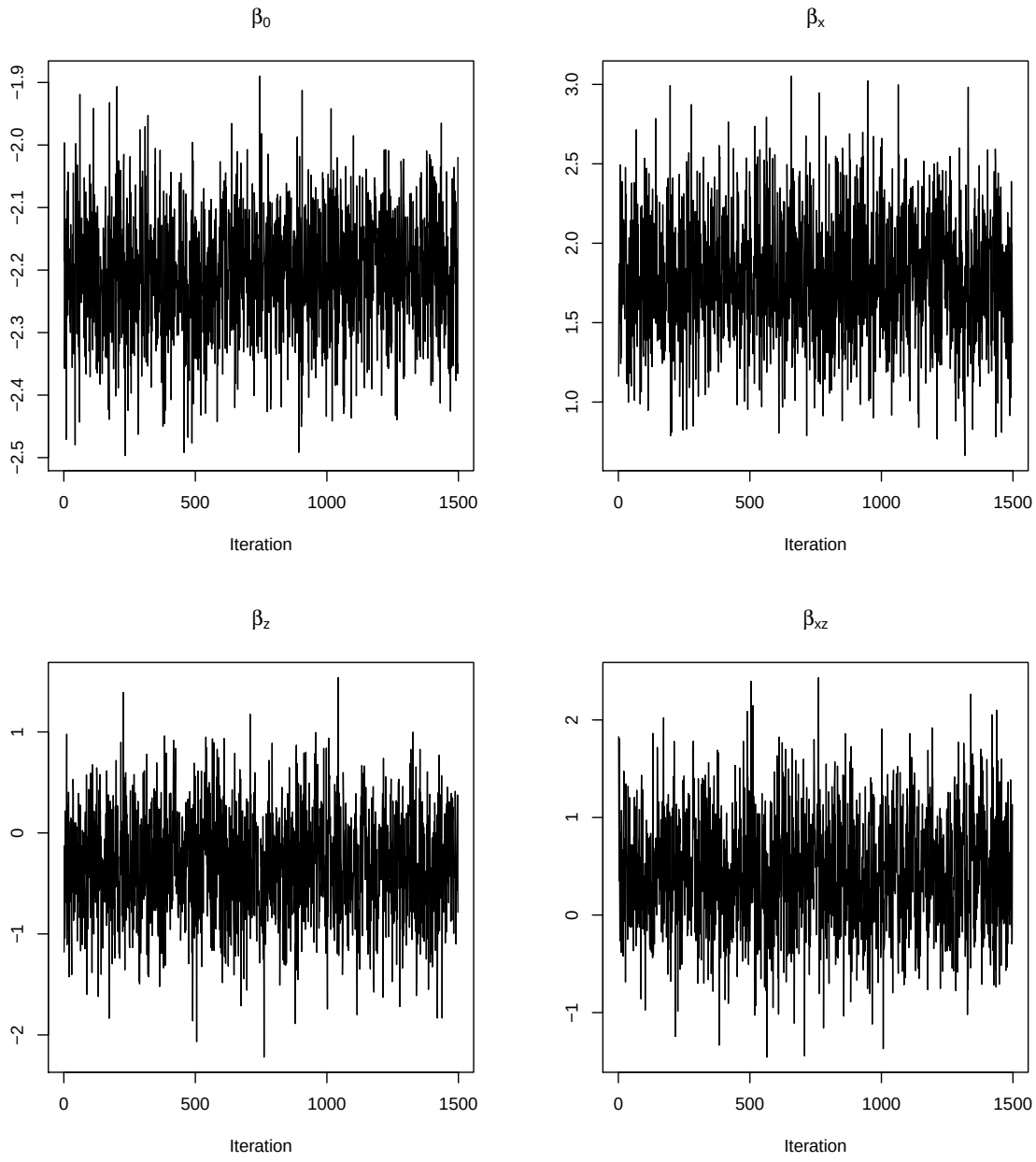


Figure A.1: Trace plots for the $\boldsymbol{\beta}$ parameters in the sparse data example.

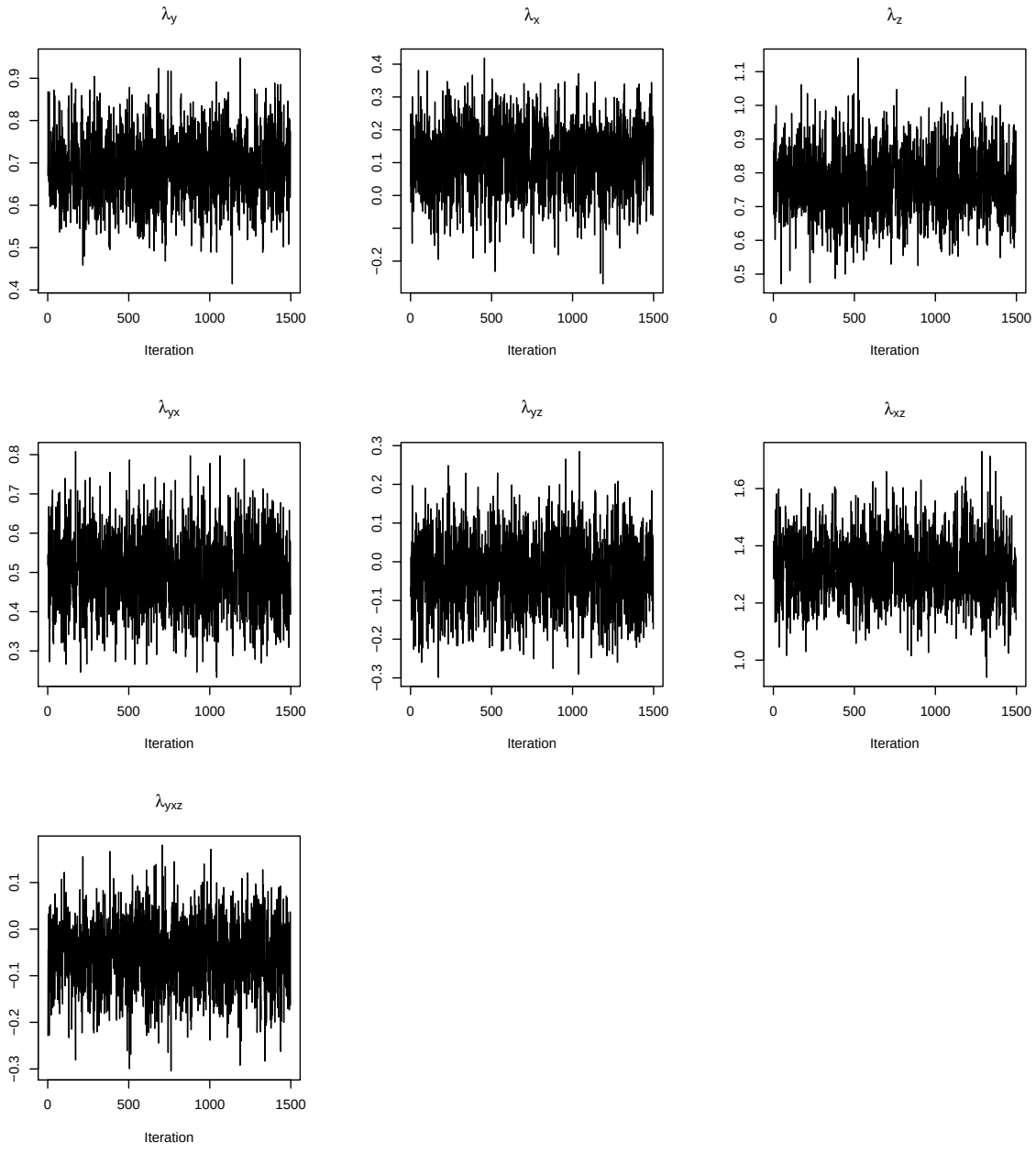


Figure A.2: Trace plots for the $\boldsymbol{\lambda}$ parameters in the sparse data example.

Figures A.3–A.5 give trace plots (with every 100th sample being taken) of the $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ parameters.

Table A.4 displays the estimates and 95% intervals for the Bayesian two-phase analysis of the Wilms tumour data using the auxiliary variable sampling scheme described in Section 3.2.4. We ran the chain for $500,000$ iterations, which took 79 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM. This is 29 minutes longer than the direct sampling scheme approach run for the same number of iterations on the same machine. The results generally agree quite well with the results from the direct sampling approach, however the estimates are more biased than those from the direct sampling approach. The estimates and 95% intervals from the two methods coincide after running $4,000,000$ iterations in each case, where the point estimates are identical to the results from the direct scheme shown in Table A.4.

Table A.4: Disease model point and interval estimates for the Wilms tumour relapse data comparing the auxiliary variable sampling scheme to the direct sampling scheme.

	Complete Data	Direct Scheme	Auxiliary Variable Scheme
β_0	-2.71 (-2.92, -2.50)	-2.56 (-2.81, -2.33)	-2.58 (-2.84, -2.34)
Stage II	0.77 (0.48, 1.05)	0.52 (0.18, 0.86)	0.56 (0.17, 0.95)
Stage III	0.77 (0.48, 1.07)	0.46 (0.08, 0.83)	0.49 (0.10, 0.88)
Stage IV	1.05 (0.71, 1.39)	0.95 (0.48, 1.41)	0.99 (0.50, 1.47)
UH*	1.31 (0.82, 1.80)	1.22 (0.66, 1.79)	1.25 (0.67, 1.80)
Stage II:UH	0.15 (-0.49, 0.78)	0.34 (-0.41, 1.11)	0.30 (-0.45, 1.09)
Stage III:UH	0.59 (-0.03, 1.21)	0.78 (0.06, 1.50)	0.74 (0.01, 1.50)
Stage IV:UH	1.26 (0.50, 2.02)	1.46 (0.55, 2.37)	1.42 (0.49, 2.39)

*: Unfavorable central histology

Figures A.6–A.8 give trace plots of the $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ parameters from the auxiliary variable sampling scheme, where the samples have been thinned by 100 in all cases. While the $\boldsymbol{\lambda}$ parameters seem to have converged, the convergence of the $\boldsymbol{N}^{\text{yxz}}$ is less apparent after $500,000$ iterations. The difficulty in getting the $\boldsymbol{N}^{\text{yxz}}$ to converge affect the entire MCMC algorithm, which makes the auxiliary variable sampling scheme less efficient than the direct

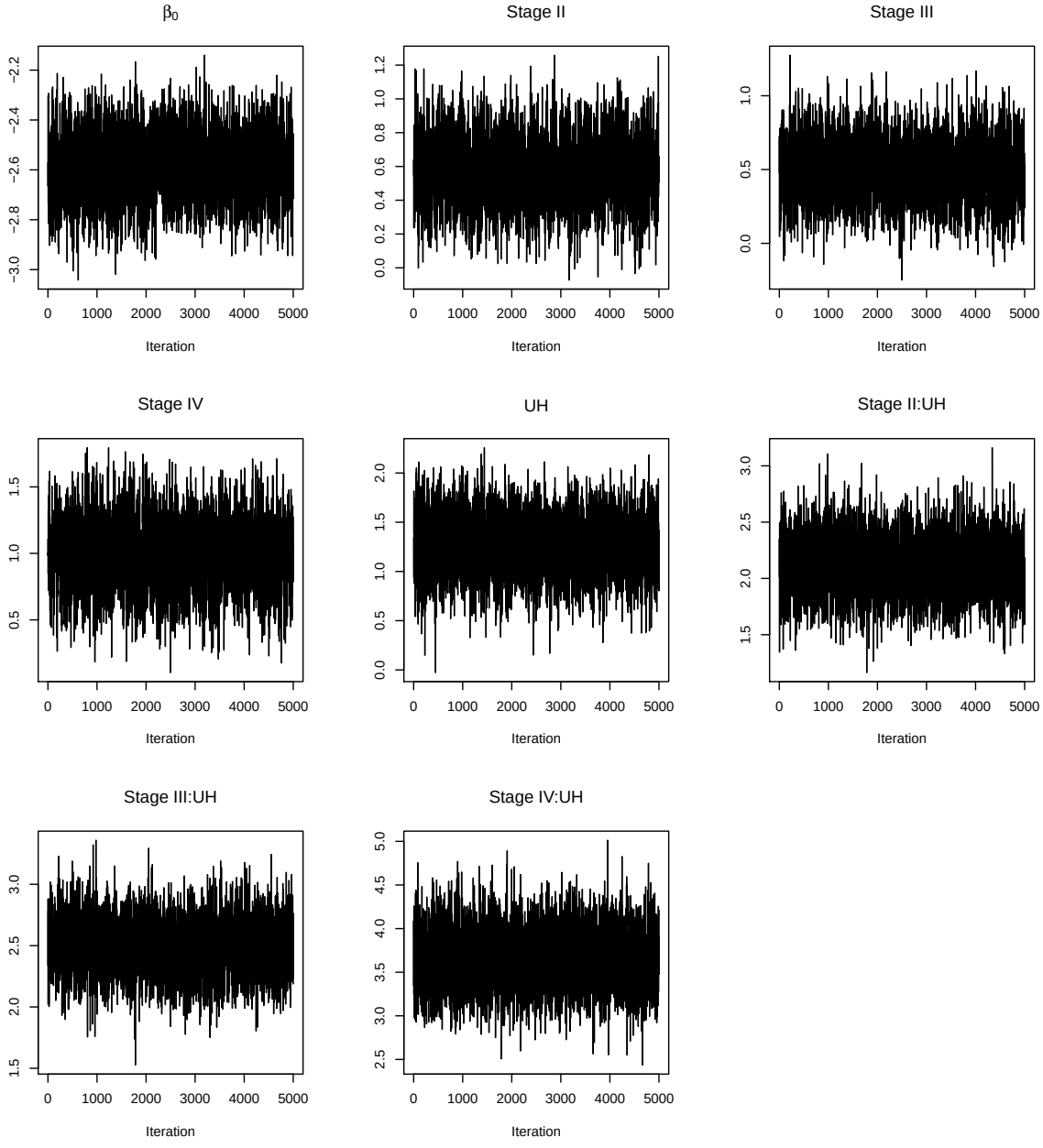


Figure A.3: Trace plots for the $\boldsymbol{\beta}$ parameters in the Wilms tumour data example.

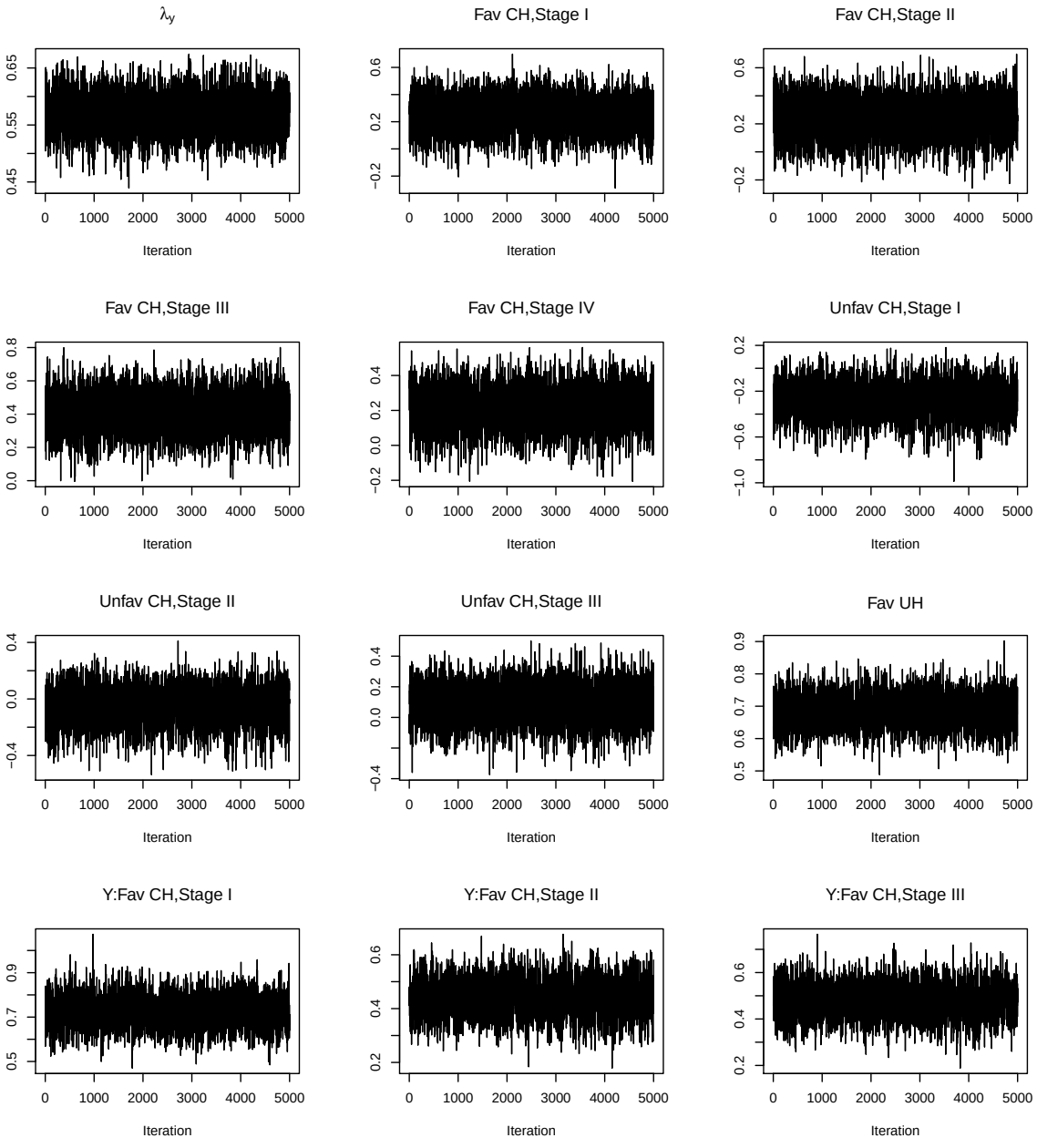


Figure A.4: Trace plots for the $\boldsymbol{\lambda}$ parameters in the Wilms tumour data example.

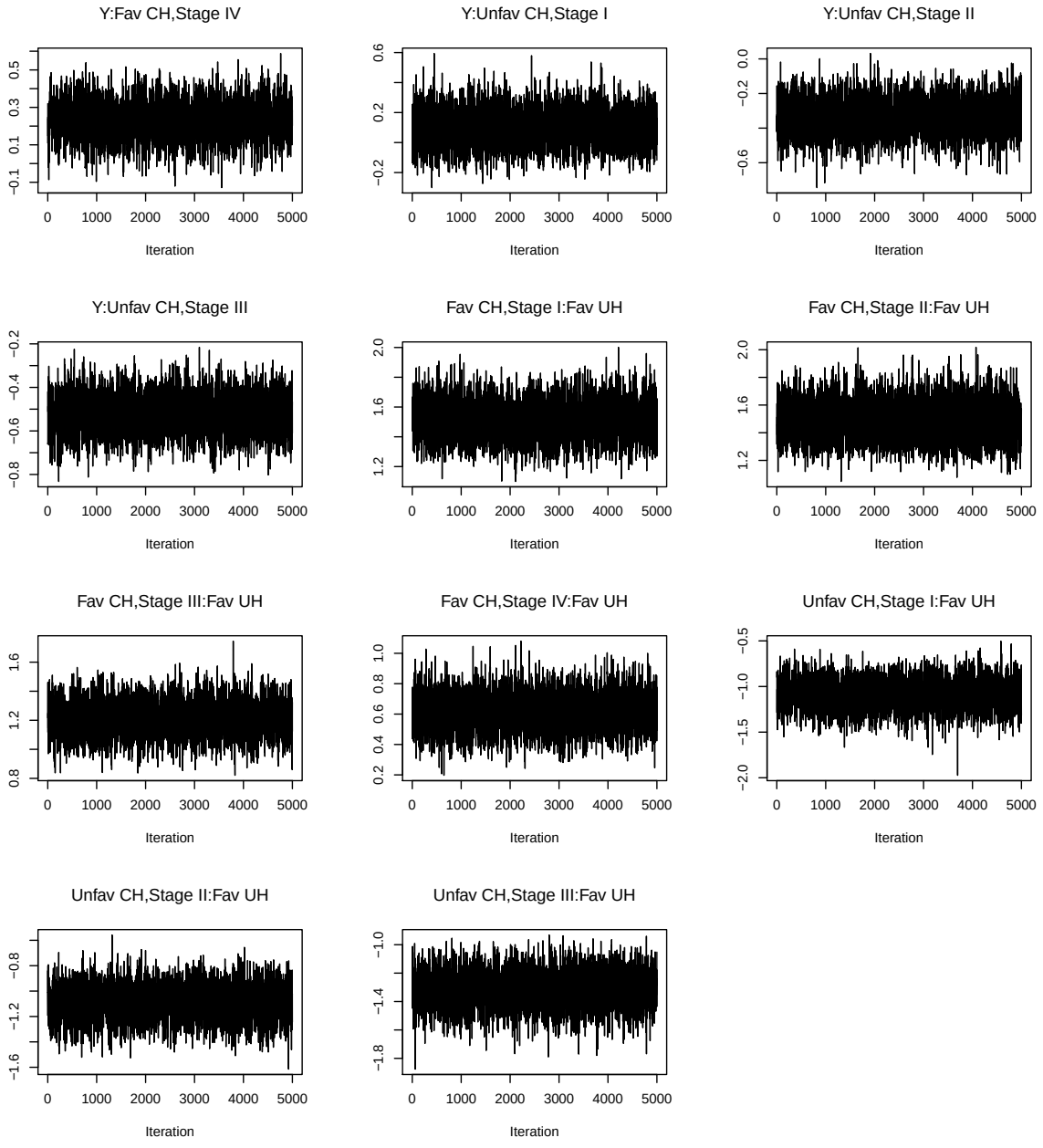


Figure A.5: Trace plots for the $\boldsymbol{\lambda}$ parameters in the Wilms tumour data example (continued).

scheme.

A.4 Case-Control Example

For the case-control example described in Section 3.5.2, the following prior distributions were assigned

$$
\begin{aligned}
\boldsymbol{\lambda}_A &\sim N_4\left(\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, 5\left(I_4 - \frac{1}{5}J_4\right)\right)\\
\boldsymbol{\lambda}_S &\sim N(0,1)\\
\boldsymbol{\lambda}_X &\sim N_3\left(\begin{pmatrix}0\\ 0 \\ 0\end{pmatrix}, 4\left(I_3 - \frac{1}{4}J_3\right)\right)\\
\boldsymbol{\lambda}_{AS} &\sim N_4\left(\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \frac{5}{2}\left(I_4 - \frac{1}{5}J_4\right)\left(I_4 - \frac{1}{5}J_4\right)\right)\\
\boldsymbol{\lambda}_{AX} &\sim N_{12}\left(\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \left(5\left(I_4 - \frac{1}{5}J_4\right)\otimes\left(I_3 - \frac{1}{4}J_3\right)\right)\left(4\left(I_4 - \frac{1}{5}J_4\right)\otimes\left(I_3 - \frac{1}{4}J_3\right)\right)\right),\\
\boldsymbol{\lambda}_{SX} &\sim N_3\left(\begin{pmatrix}0\\ 0 \\ 0\end{pmatrix}, 2\left(I_3 - \frac{1}{4}J_3\right)\left(I_3 - \frac{1}{4}J_3\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}_{YA}, \boldsymbol{\lambda}_{YX}) &\sim N_7\left(\mathbf{0}, \boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T\right)\\
\boldsymbol{C} &= \begin{pmatrix}
-2 & -2 & 0 & 0 & 0 & 0 & 0\\
2 & 0 & -2 & 0 & 0 & 0 & 0\\
2 & 0 & 0 & -2 & 0 & 0 & 0\\
4 & 2 & 2 & 2 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 2 & -2 & 0\\
0 & 0 & 0 & 0 & 2 & 0 & -2\\
0 & 0 & 0 & 0 & 4 & 2 & 2
\end{pmatrix}\\
\boldsymbol{\Sigma} &= \begin{pmatrix}
100^2 & & 0\\
& \ddots & \\
0 & & 100^2
\end{pmatrix},
\end{aligned}
$$

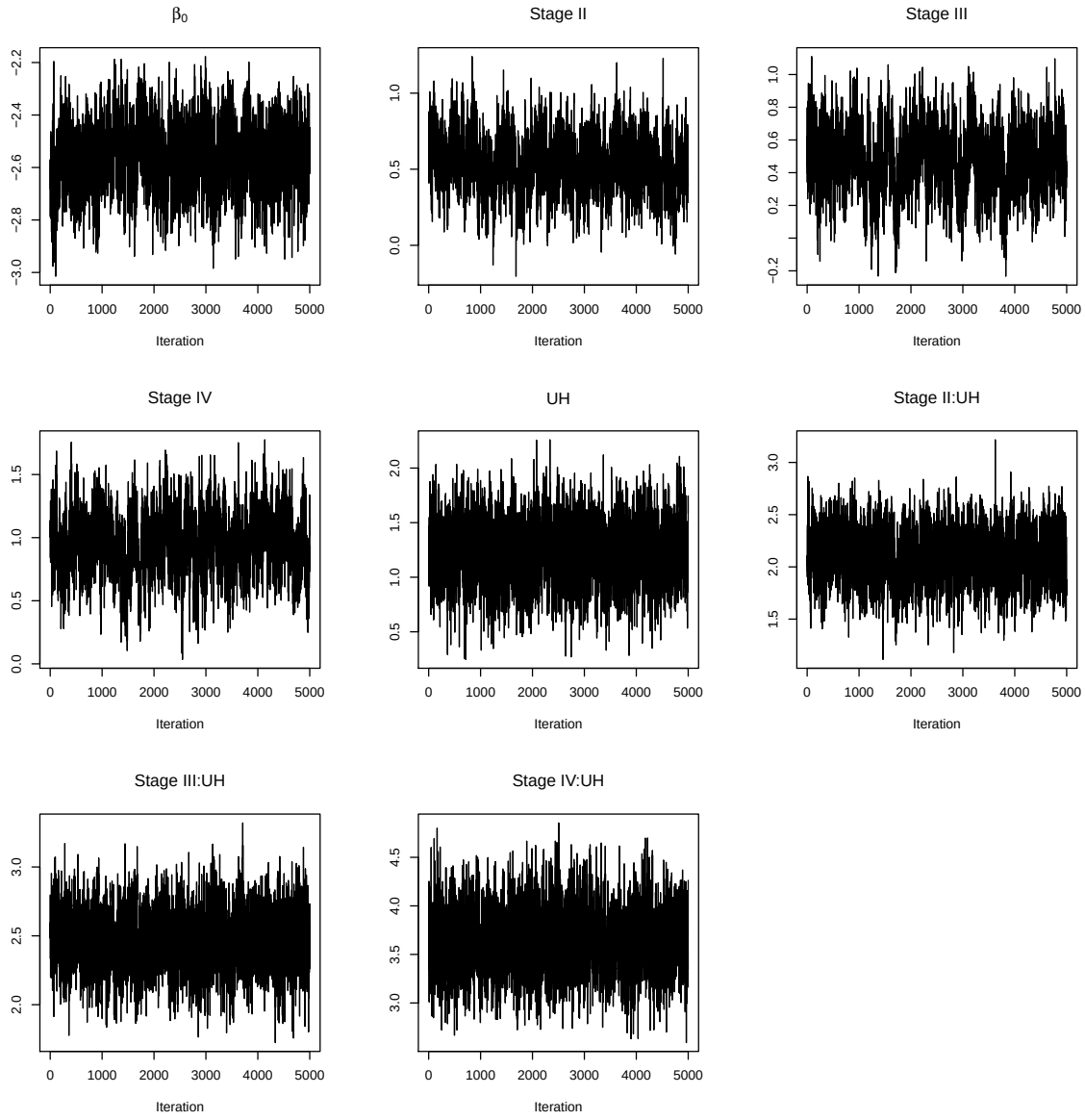


Figure A.6: Trace plots for the $\boldsymbol{\beta}$ parameters in the auxiliary variables sampling scheme analysis of the Wilms tumour data example.

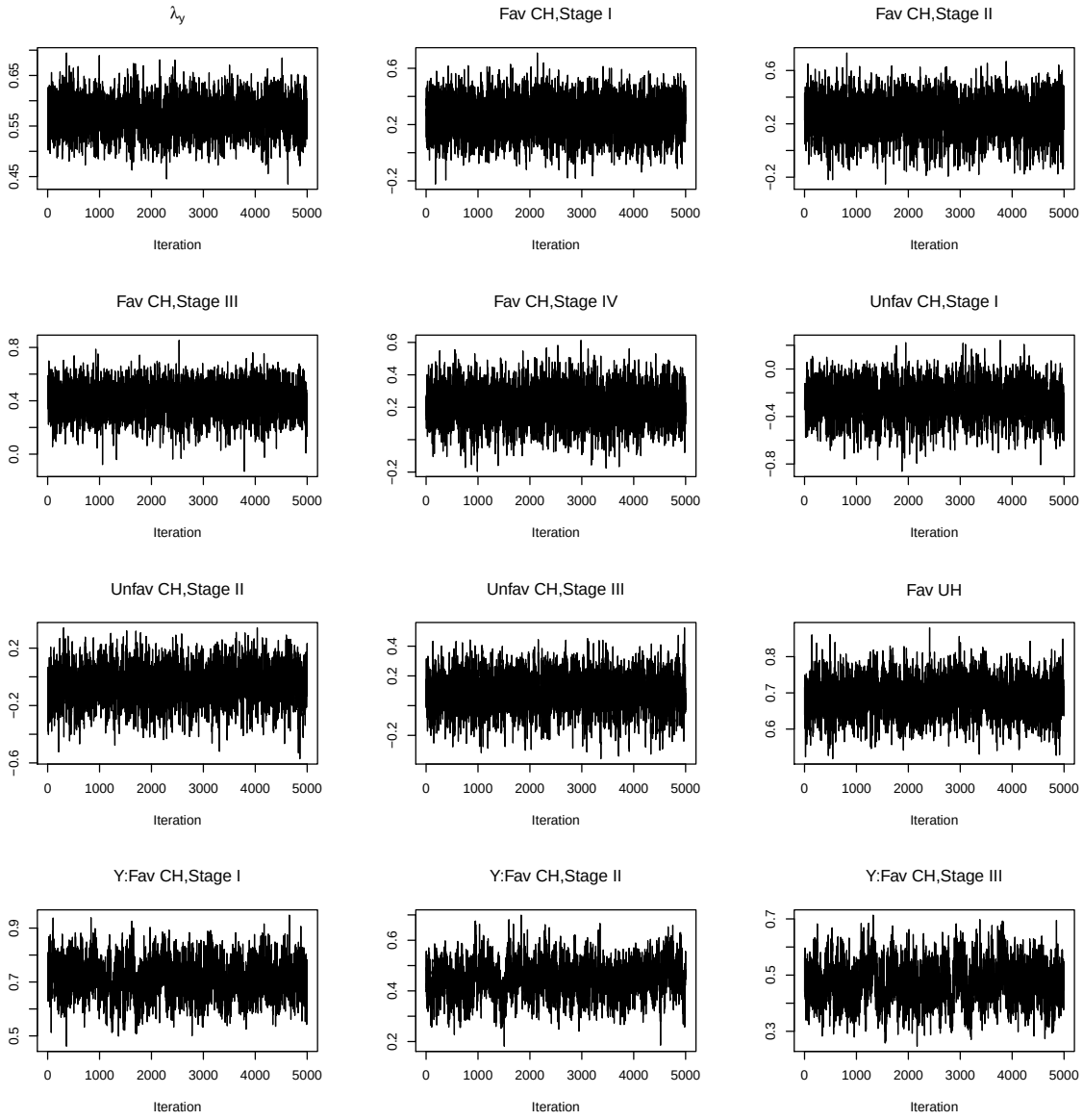


Figure A.7: Trace plots for the $\boldsymbol{\lambda}$ parameters in the auxiliary variables sampling scheme analysis of the Wilms tumour data example.

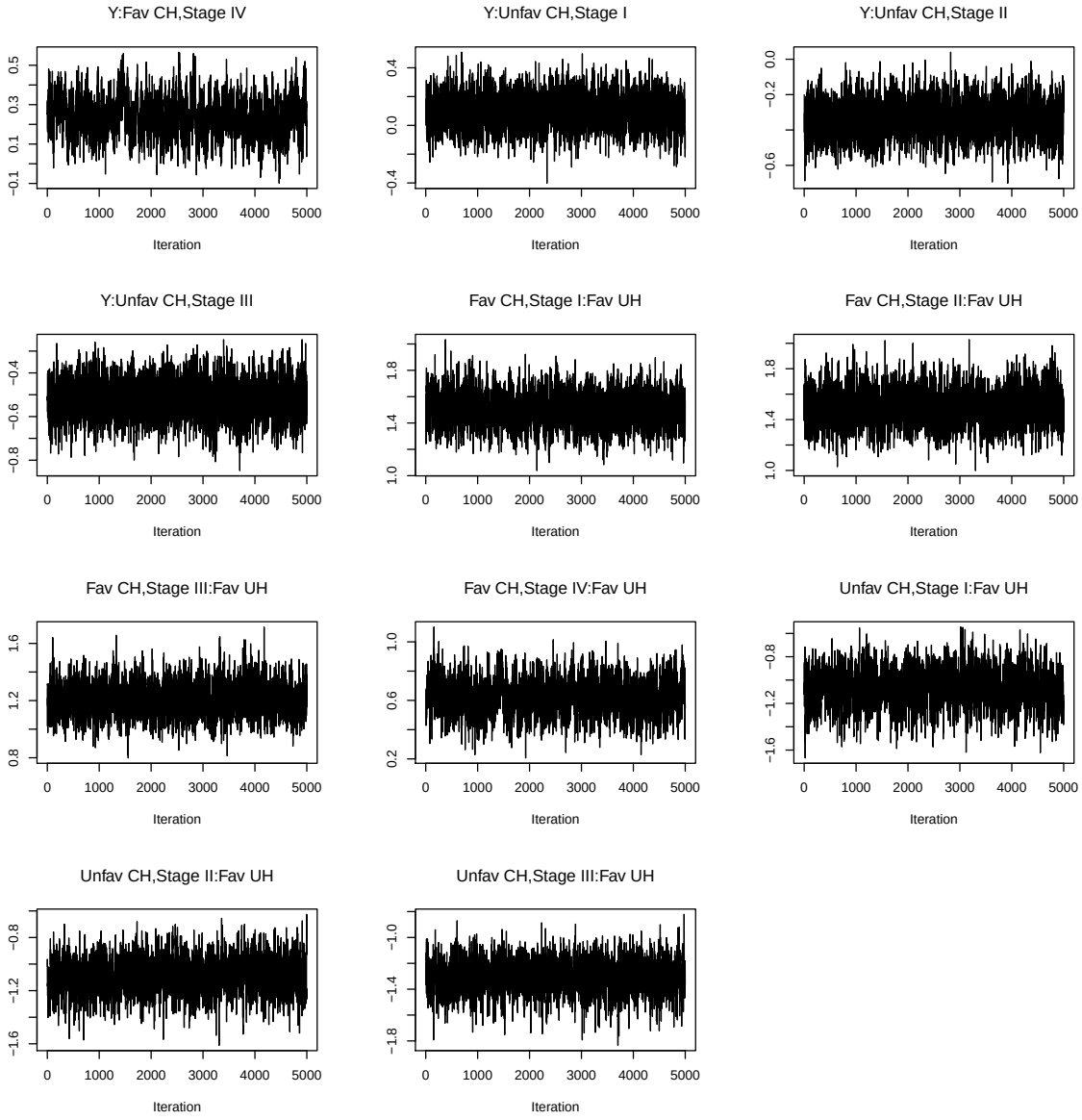


Figure A.8: Trace plots for the $\boldsymbol{\lambda}$ parameters in the auxiliary variables sampling scheme analysis of the Wilms tumour data example (continued).

where I_n is an $n \times n$ identity matrix, and J_n is an $n \times n$ matrix of ones.

Trace plots for the $\boldsymbol{\beta}$ parameters are given in Figure A.9, while trace plots for the 34 $\boldsymbol{\lambda}$ parameters are given in Figures A.10–A.12. In each case, we have taken every 100th sample from the Markov chain.

A.5 $2 \times 2 \times 2 \times 3 \times 7$ *Contingency Table Example*

Table A.5 displays the full data for the large contingency table example of Section 3.5.3, while Table A.6 displays the phase II data.

We assigned a multivariate normal prior with large variances for $\boldsymbol{\beta}$, which induces the following multivariate normal prior for $\boldsymbol{\Lambda}^Y$:

$$(\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YA}, \boldsymbol{\lambda}^{YR}, \boldsymbol{\lambda}^{YS}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{YRS}) \sim N_{12}\left(\mathbf{0}, \boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T\right)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 100^2 & & 0 \\ & \ddots & \\ 0 & & 100^2 \end{pmatrix}$$

$$\boldsymbol{C} = \begin{pmatrix}
-2 & -2 & 0 & -2 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & -2 \\
0 & 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \\
0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \\
0 & 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & -2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & -2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & -2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & -2 & 0 \\
0 & 0 & 0 & 0 & 0 & 4 & 2 & 2 & 2 & 2 & 2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -8
\end{pmatrix},$$

and, following the Dellaportas and Forster (1999) prescription, we used the following prior distributions for the remaining $\boldsymbol{\lambda}$ parameters:

$$\begin{aligned}
\lambda_A &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\right) \\
\lambda_R &\sim N(0,1) \\
\lambda_S &\sim N(0,1)
\end{aligned}$$

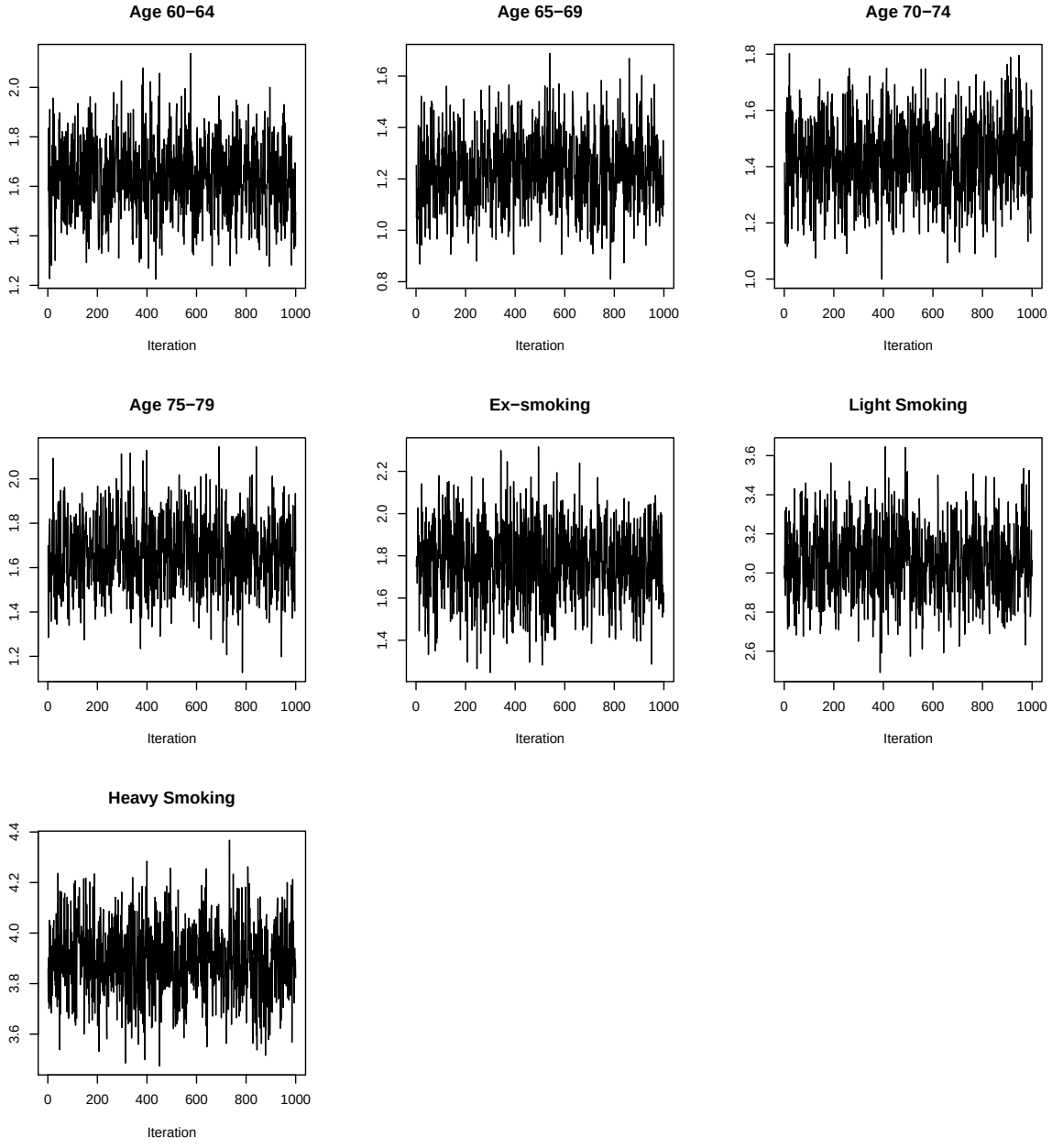


Figure A.9: Trace plots for the $\boldsymbol{\beta}$ parameters in the case-control example.

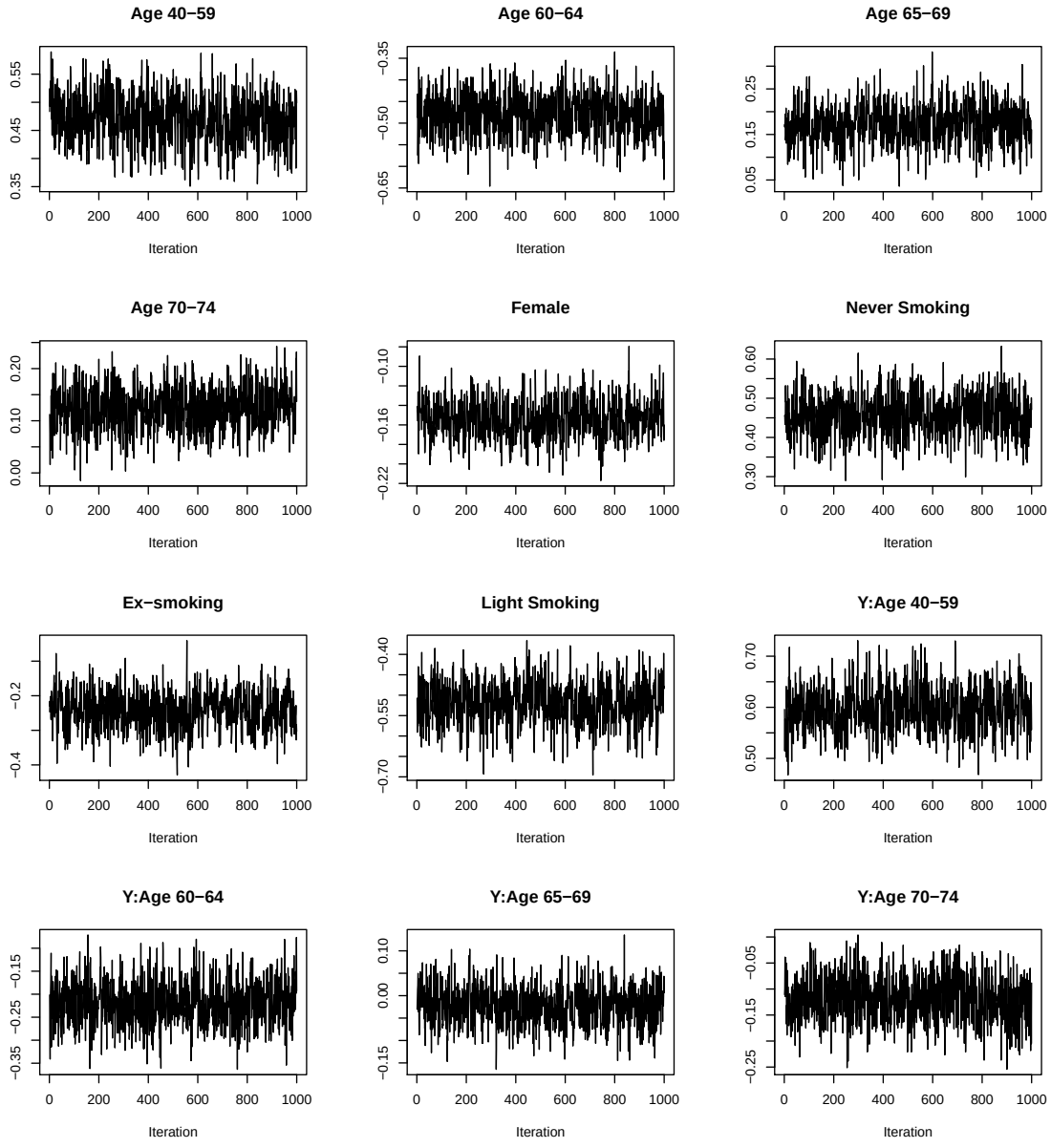


Figure A.10: Trace plots for the $\boldsymbol{\lambda}$ parameters in the case-control example.

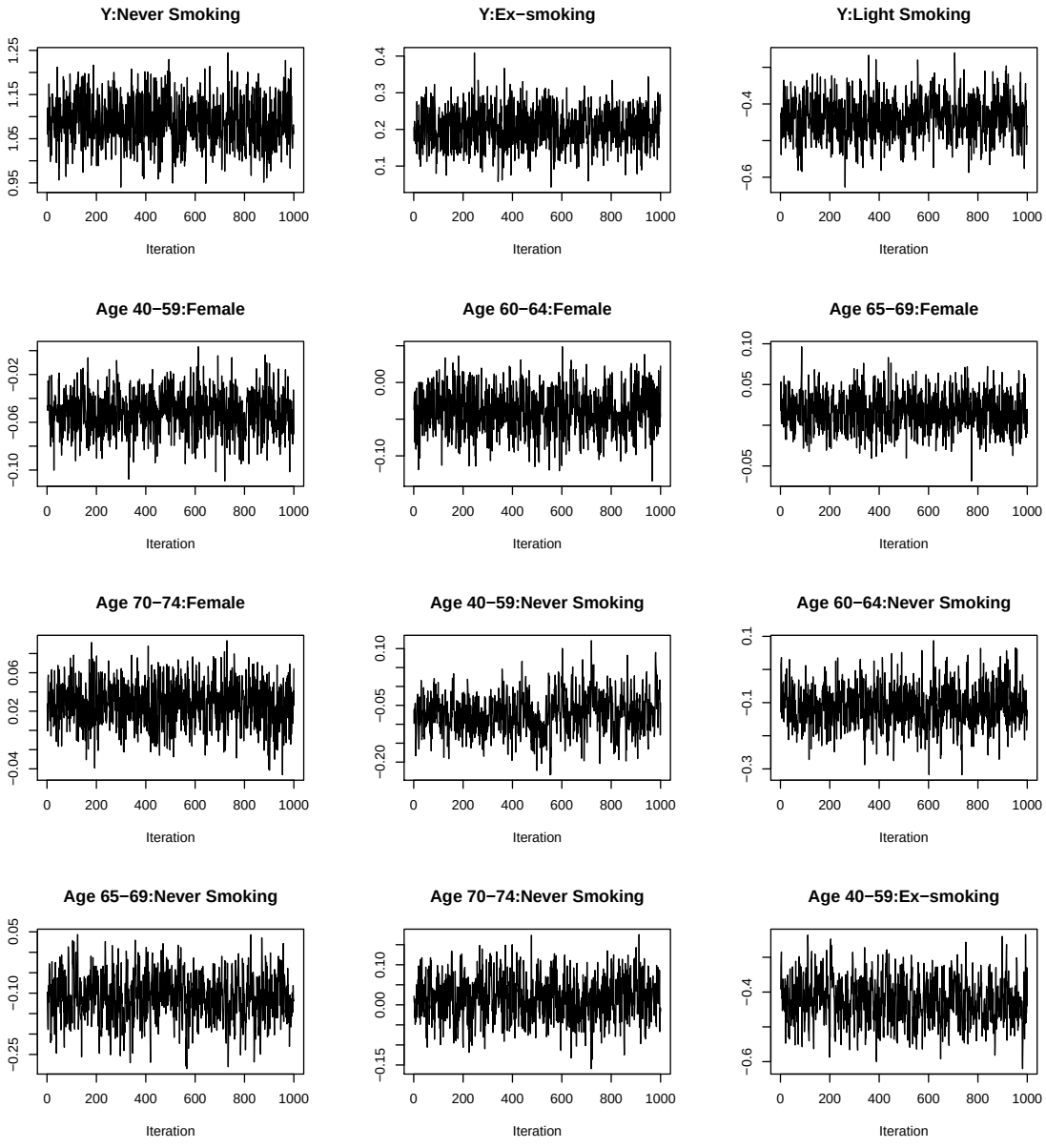


Figure A.11: Trace plots for the $\boldsymbol{\lambda}$ parameters in the case-control example (continued).

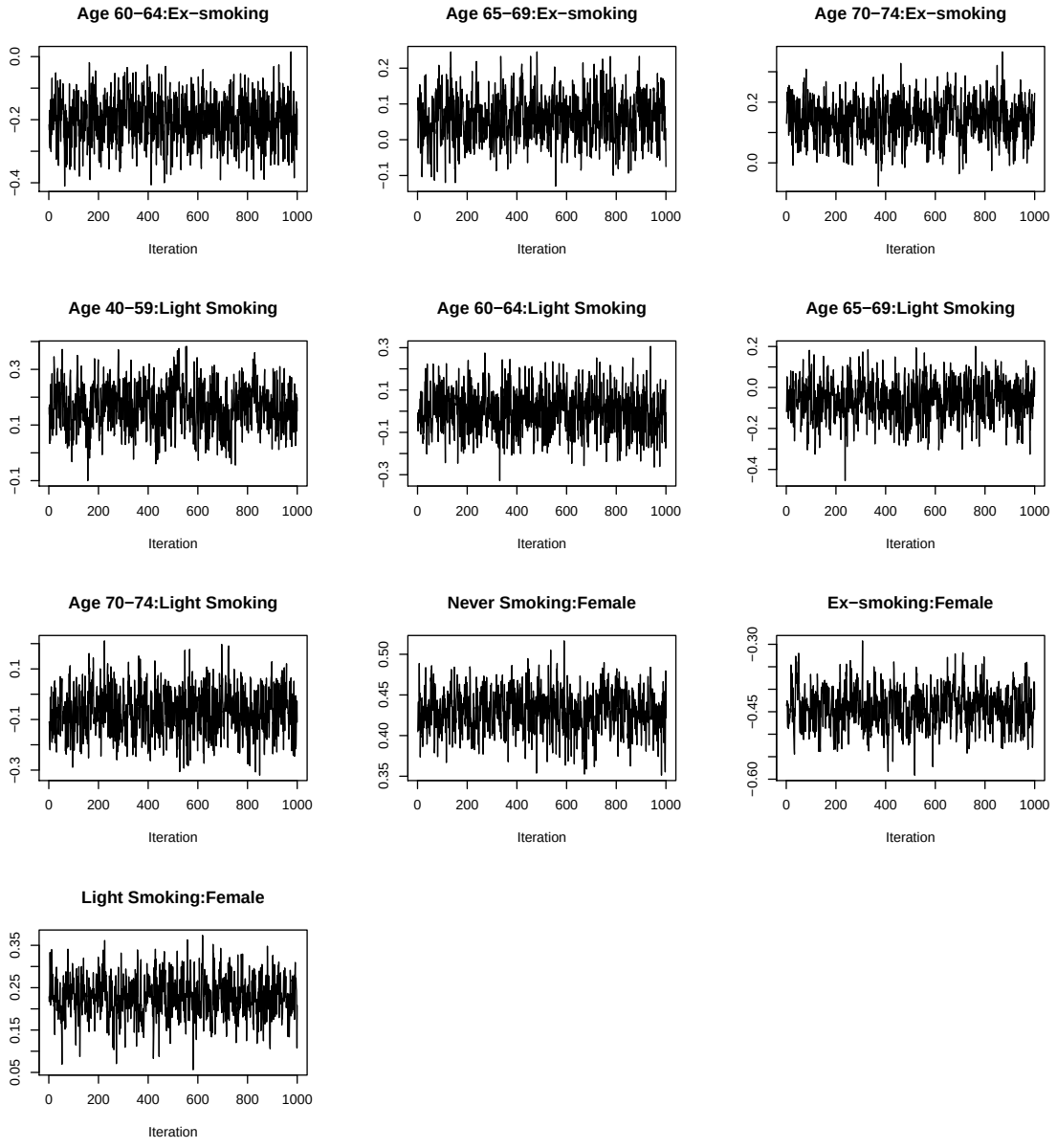


Figure A.12: Trace plots for the $\boldsymbol{\lambda}$ parameters in the case-control example (continued).

Table A.5: Population, N_{yxz}, for $2 \times 2 \times 2 \times 3 \times 7$ dimensional table example.

	White						Other					
	< 35		$35 - -44$		> 44		< 35		$35 - 44$		> 44	
Region	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Northeast												
Satisfied	288	60	224	35	337	70	38	19	32	22	21	15
Not satisfied	177	57	166	19	172	30	33	35	11	20	8	10
Mid-Atlantic												
Satisfied	90	19	96	12	124	17	18	13	7	0	9	1
Not satisfied	45	12	42	5	39	2	6	7	2	3	2	1
Southern												
Satisfied	226	88	189	44	156	70	45	47	18	13	11	9
Not satisfied	128	57	117	34	73	25	31	35	3	7	2	2
Midwest												
Satisfied	285	110	225	53	324	60	40	66	19	25	22	11
Not satisfied	179	93	141	24	140	47	25	56	11	19	2	12
Northwest												
Satisfied	270	176	215	80	269	110	36	25	9	11	16	4
Not satisfied	180	151	108	40	136	40	20	16	7	5	3	5
Southwest												
Satisfied	252	97	162	47	199	62	69	45	14	8	14	2
Not satisfied	126	61	72	27	93	24	27	36	7	4	5	0
Pacific												
Satisfied	119	62	66	20	67	25	45	22	15	10	8	6
Not satisfied	58	33	20	10	21	10	16	15	10	8	6	2

Table A.6: Phase II Data, n_{yxz}, for for $2 \times 2 \times 2 \times 3 \times 7$ dimensional table example.

	White						Other					
	< 35		$35 - 44$		> 44		< 35		$35 - 44$		> 44	
Region	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Northeast												
Satisfied	47	14	35	8	64	9	4	4	7	4	1	3
Not satisfied	32	7	20	3	17	3	3	8	2	2	1	2
Mid-Atlantic												
Satisfied	44	7	46	4	65	8	12	8	4	0	2	0
Not satisfied	26	10	23	5	22	2	4	2	1	2	2	1
Southern												
Satisfied	58	20	40	6	31	13	9	13	2	5	1	2
Not satisfied	21	14	18	8	17	7	5	6	0	2	2	0
Midwest												
Satisfied	39	17	43	6	63	10	5	5	4	6	2	0
Not satisfied	24	16	20	2	19	3	3	10	2	1	0	0
Northwest												
Satisfied	46	31	35	11	41	15	6	5	1	4	4	1
Not satisfied	29	20	16	5	19	6	1	0	1	0	0	3
Southwest												
Satisfied	46	19	39	9	48	9	11	8	6	2	3	0
Not satisfied	20	11	16	3	21	8	10	8	2	0	1	0
Pacific												
Satisfied	48	26	22	8	37	11	21	8	7	4	5	3
Not satisfied	28	17	8	4	9	5	8	8	6	4	2	1

$$
\begin{aligned}
\lambda_Z &\sim N_6\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 6 & & -1 \\ & \ddots & \\ -1 & & 6 \end{pmatrix}\right) \\
\lambda_{AR} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{6}\begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}\right) \\
\lambda_{AS} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{6}\begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}\right) \\
\lambda_{AZ} &\sim N_{12}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \left(3\left(I_2 - \frac{1}{3}J_2\right) \otimes \left(I_6 - \frac{1}{7}J_6\right)\right)\left(7\left(I_2 - \frac{1}{3}J_2\right) \otimes \left(I_6 - \frac{1}{7}J_6\right)\right)\right) \\
\lambda_{RS} &\sim N\left(0, \frac{1}{4}\right) \\
\lambda_{RZ} &\sim N_6\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \frac{7}{2}\left(I_6 - \frac{1}{7}J_6\right)\left(I_6 - \frac{1}{7}J_6\right)\right) \\
\lambda_{SZ} &\sim N_6\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \frac{7}{2}\left(I_6 - \frac{1}{7}J_6\right)\left(I_6 - \frac{1}{7}J_6\right)\right),
\end{aligned}
$$

where I_n is an $n \times n$ identity matrix, and J_n is an $n \times n$ matrix of ones.

Figure A.13 displays the trace plots for the 12 $\boldsymbol{\beta}$ parameters, while Figures A.14–A.17 display the trace plots for the 51 $\boldsymbol{\lambda}$ parameters. In each case, the samples are thinned by 100.

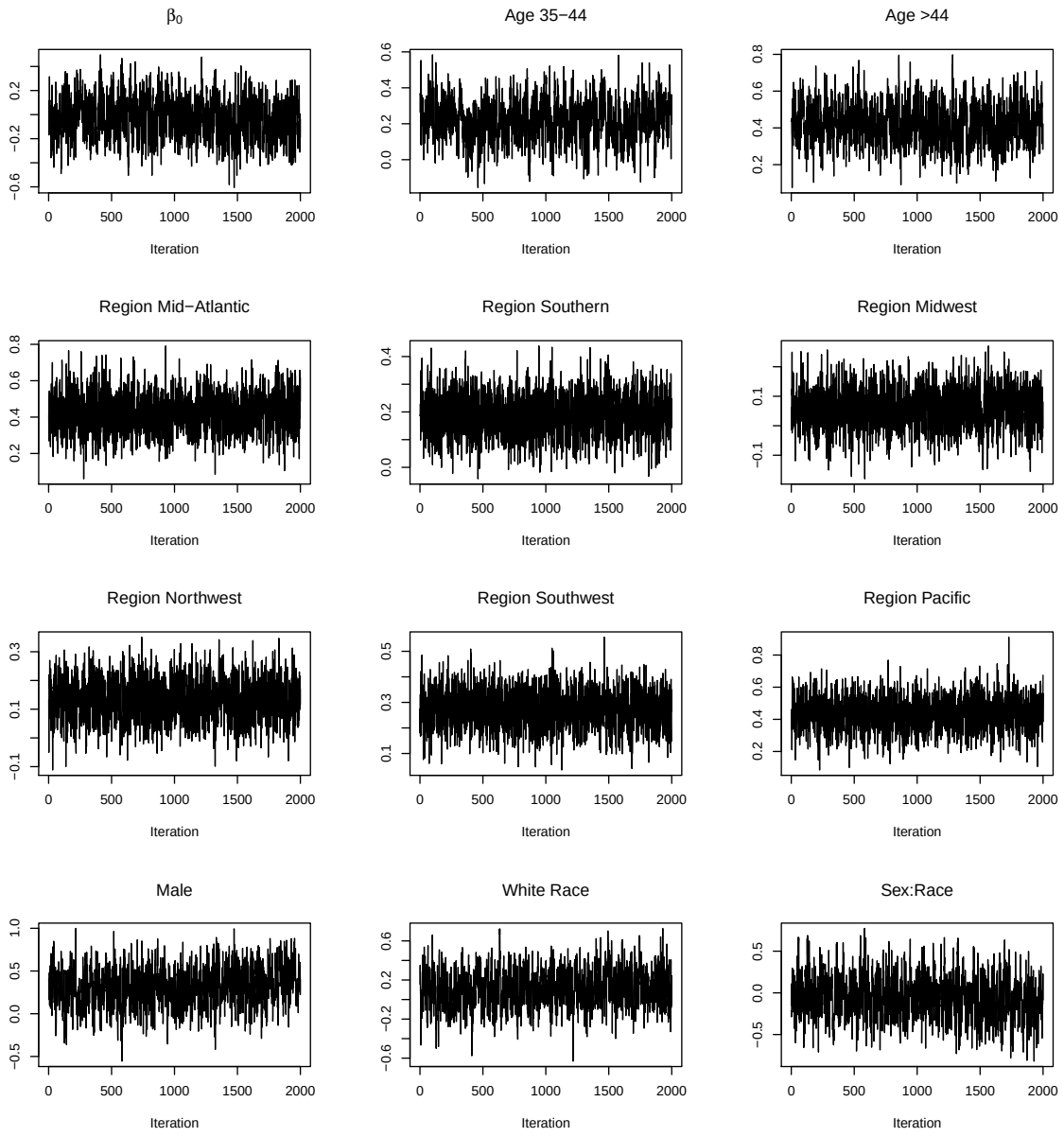


Figure A.13: Trace plots for the $\boldsymbol{\beta}$ parameters in the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example.

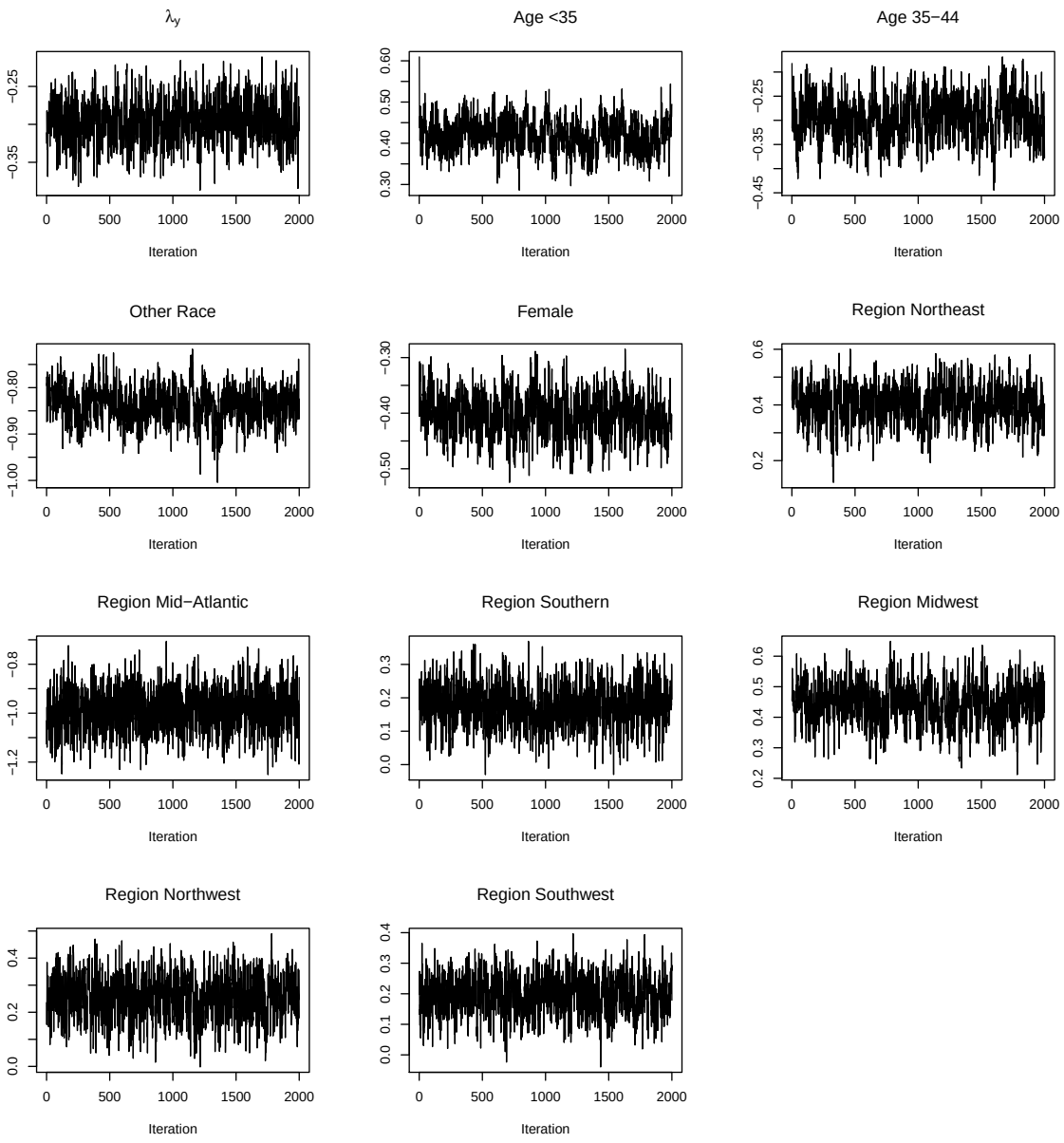


Figure A.14: Trace plots for the $\boldsymbol{\lambda}$ parameters in the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example.

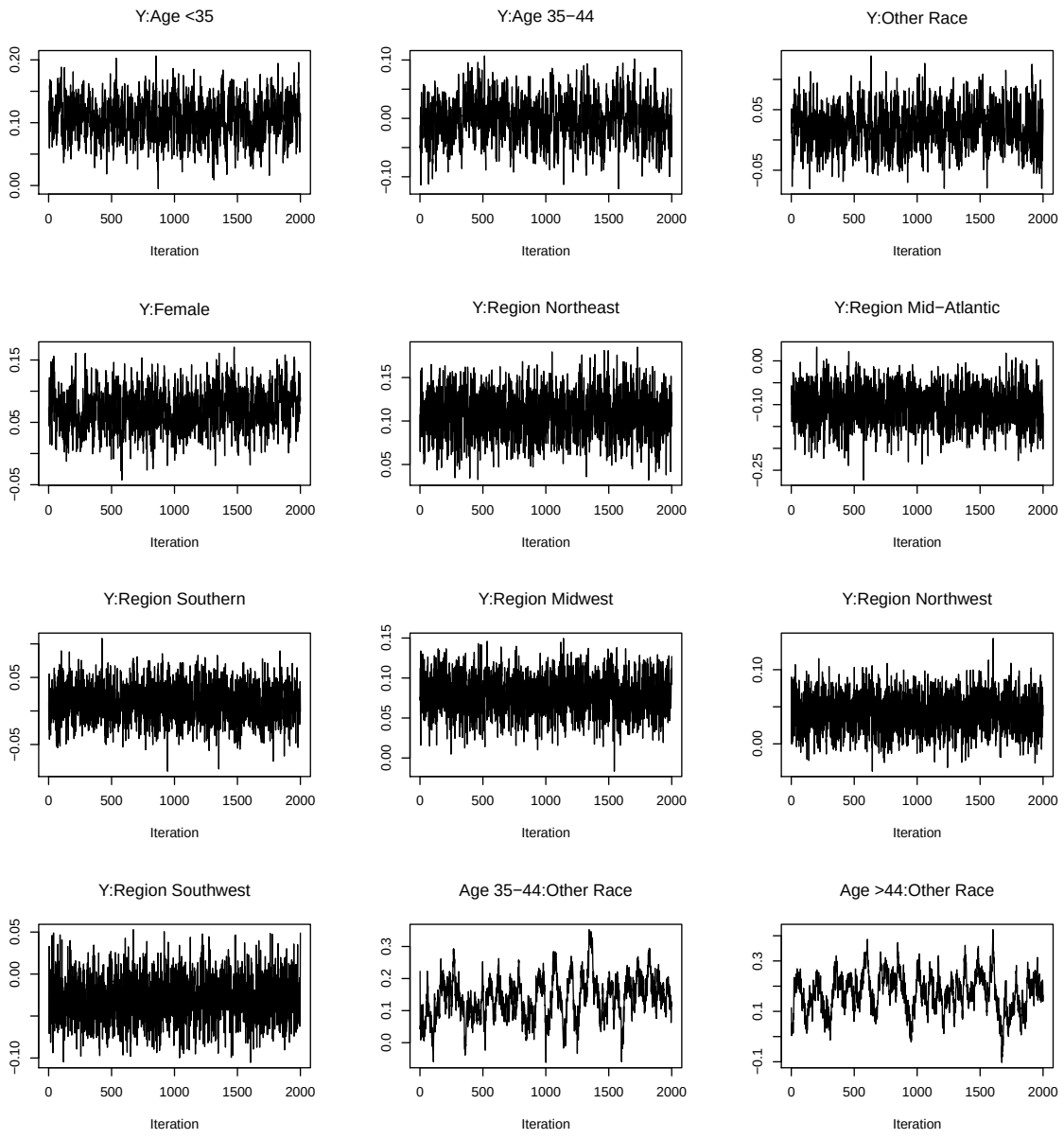


Figure A.15: Trace plots for the $\boldsymbol{\lambda}$ parameters in the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example (continued).

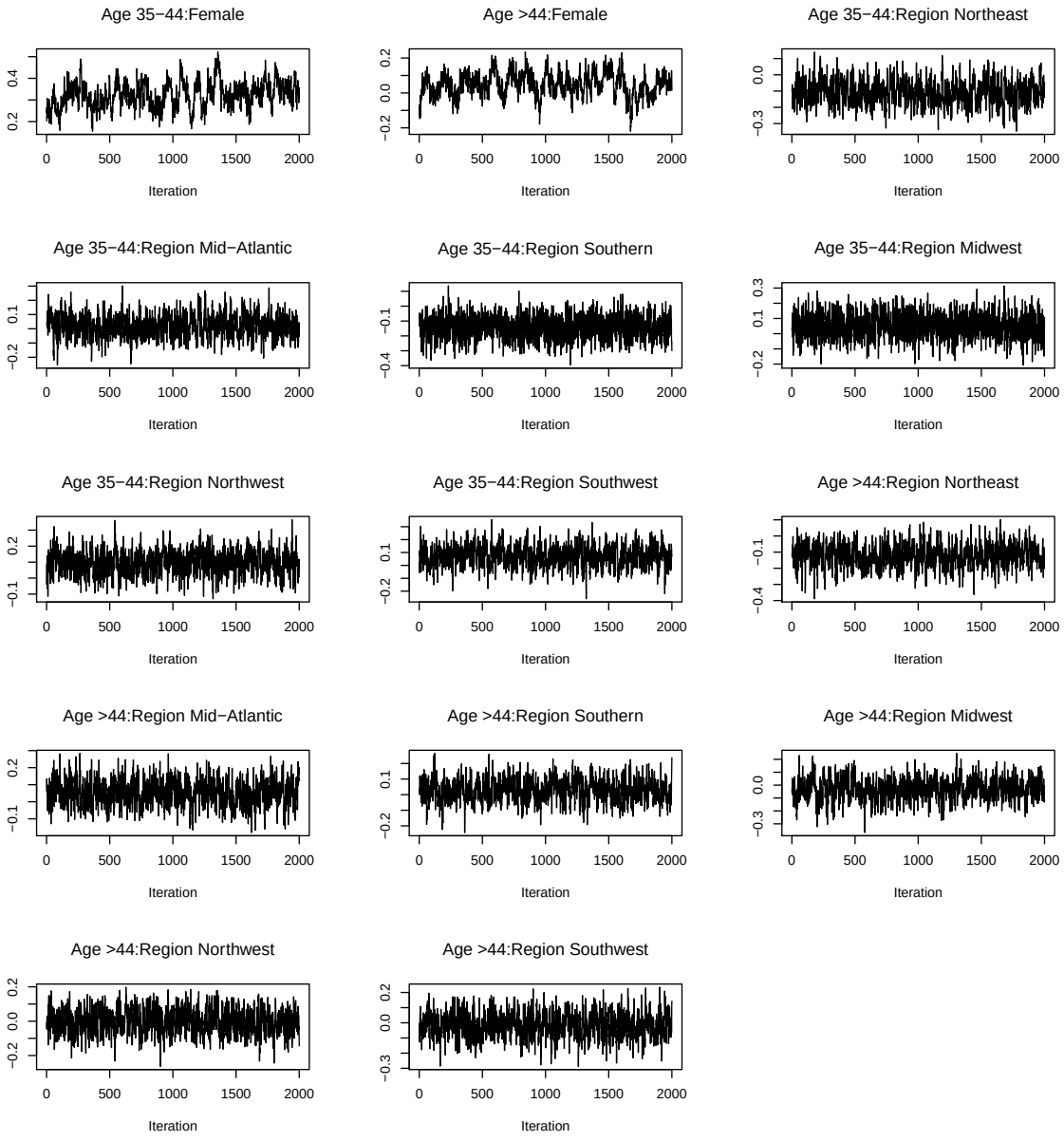


Figure A.16: Trace plots for the $\boldsymbol{\lambda}$ parameters in the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example (continued).

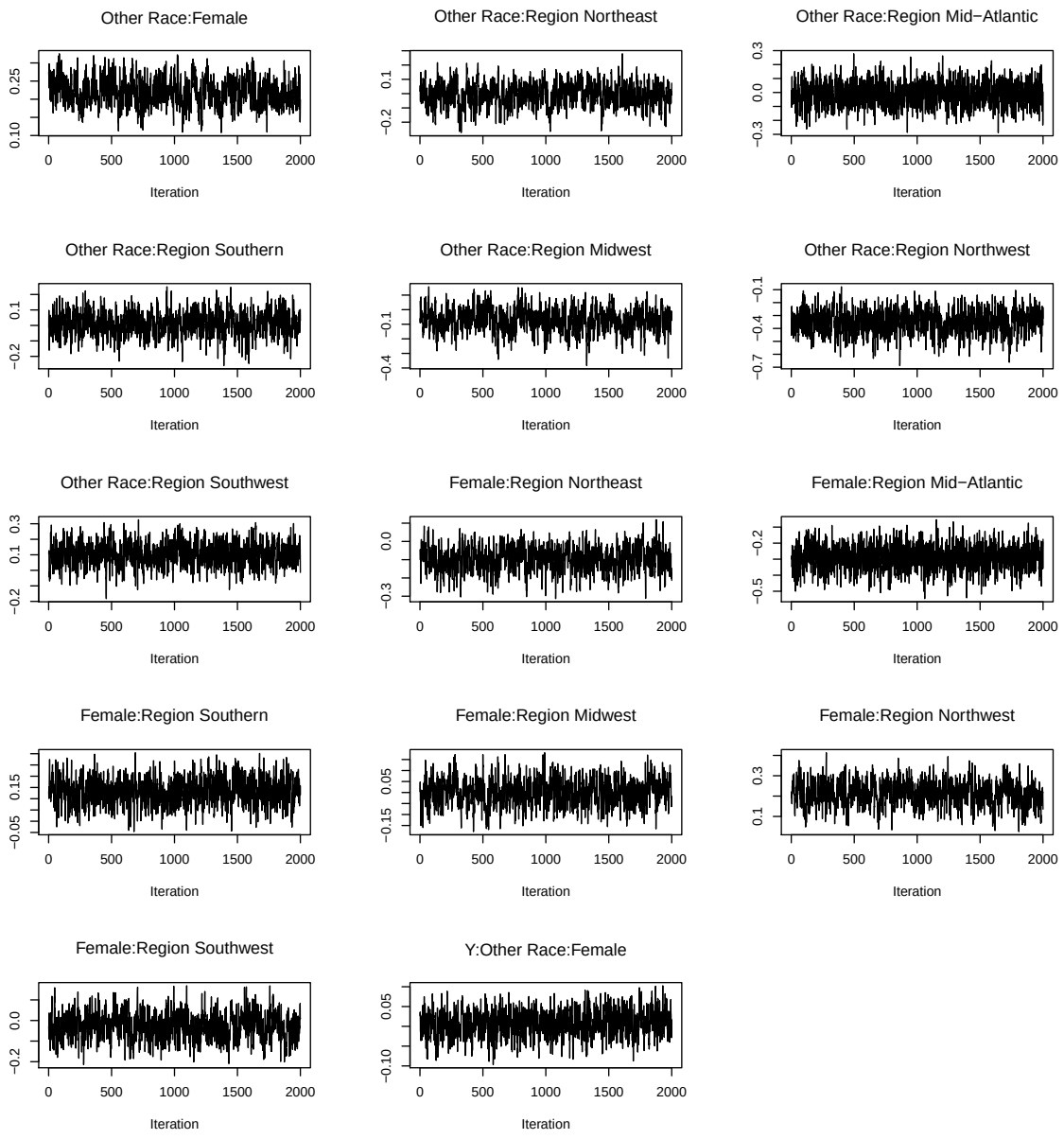


Figure A.17: Trace plots for the $\boldsymbol{\lambda}$ parameters in the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example (continued).

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

B.1 Verifying the Bayesian Spatial Analysis of Two-Phase Data

Since there are no frequentist methods which analyse two-phase data including random effects terms, we cannot compare the results obtained from the Bayesian two-phase analysis to those obtained from other methods. To verify that the method works, we instead artificially create a two-phase study using large phase II sample sizes so that a two-phase analysis will approximate an analysis on the full population. The full population data was given in Table 2.6, while the phase II data is given in Table B.1. Using a disease model that includes main effect terms for sex, race, and low birth weight status, as well as an interaction term between race and low birth weight, we compare the Bayesian two-phase approach with fitting a spatial model to the full data.

The log-linear model used included all two-way interactions between race, sex, birth weight and region. Let R denote race, S sex, W birth weight and Z region. The Bayesian approach assigned zero mean normal prior distributions on $\boldsymbol{\beta}$ with large variances and the approach outlined in Section 4.3.2, where we have assumed the $\boldsymbol{\lambda}^{XZ}$ are fixed. In particular, the induced multivariate normal prior on $\boldsymbol{\Lambda}^{Y} = (\boldsymbol{\lambda}^{Y}, \boldsymbol{\lambda}^{YS}, \boldsymbol{\lambda}^{YR}, \boldsymbol{\lambda}^{YW}, \boldsymbol{\lambda}^{YRW})$ has mean zero and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^{T}$, where $\boldsymbol{C}$ and $\boldsymbol{\Sigma}$ are the 5×5 matrices given by

$$\boldsymbol{C} = \begin{pmatrix} -2 & -2 & -2 & -2 & -2 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 4 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & -8 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 100^2 & & 0 \\ & \ddots & \\ 0 & & 100^2 \end{pmatrix}.$$

For the remaining $\boldsymbol{\lambda}$ nuisance parameters, we assigned

$$\begin{aligned}
\boldsymbol{\lambda}^S &\sim N(0,1) \\
\boldsymbol{\lambda}^R &\sim N(0,1) \\
\boldsymbol{\lambda}^W &\sim N(0,1) \\
\boldsymbol{\lambda}^Z &\sim N_9\left(\mathbf{0}, 10\left(I_9 - \frac{1}{10}J_9\right)\right) \\
\boldsymbol{\lambda}^{SR} &\sim N\left(0, \frac{1}{4}\right) \\
\boldsymbol{\lambda}^{SW} &\sim N\left(0, \frac{1}{4}\right) \\
\boldsymbol{\lambda}^{SZ} &\sim N_9\left(\mathbf{0}, 5\left(I_9 - \frac{1}{10}J_9\right)\left(I_9 - \frac{1}{10}J_9\right)\right) \\
\boldsymbol{\lambda}^{RW} &\sim N\left(0, \frac{1}{4}\right) \\
\boldsymbol{\lambda}^{RZ} &\sim N_9\left(\mathbf{0}, 5\left(I_9 - \frac{1}{10}J_9\right)\left(I_9 - \frac{1}{10}J_9\right)\right) \\
\boldsymbol{\lambda}^{WZ} &\sim N_9\left(\mathbf{0}, 5\left(I_9 - \frac{1}{10}J_9\right)\left(I_9 - \frac{1}{10}J_9\right)\right),
\end{aligned}$$

where I_n is an $n \times n$ identity matrix and J_n is an $n \times n$ matrix of ones. We assigned τ_v and τ_u gamma hyperprior distributions with shape parameter 1 and rate parameter 0.026.

The estimates for $\boldsymbol{\beta}$, U, V, τ_v and τ_u based on the complete data and the Bayesian approach are displayed in Tables B.2 and B.3. In the Bayesian approach, we ran a total of $500,000$ iterations, where the first $50,000$ were used to tune the chain, and constants $c_\lambda = 0.05$, $c_v = 0.002$ and $c_u = 190$ gave acceptance rates of 45%, 41%, and 32% for $\boldsymbol{\lambda}$, V and U, respectively. The $500,000$ iterations took 231 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM. We see very strong agreement between the Bayesian method and complete data analysis confirming the results of the Bayesian analysis. Figure B.1 displays the trace plots for the $\boldsymbol{\beta}$, $\log \tau_v$ and $\log \tau_u$ parameters, while Figures B.2 - B.5 display the trace plots for the $\boldsymbol{\lambda}$ parameters. Figures B.6 and B.6 display the trace plots for the $\boldsymbol{V}$ and $\boldsymbol{U}$ parameters, respectively. In each case, the samples are thinned by 100.

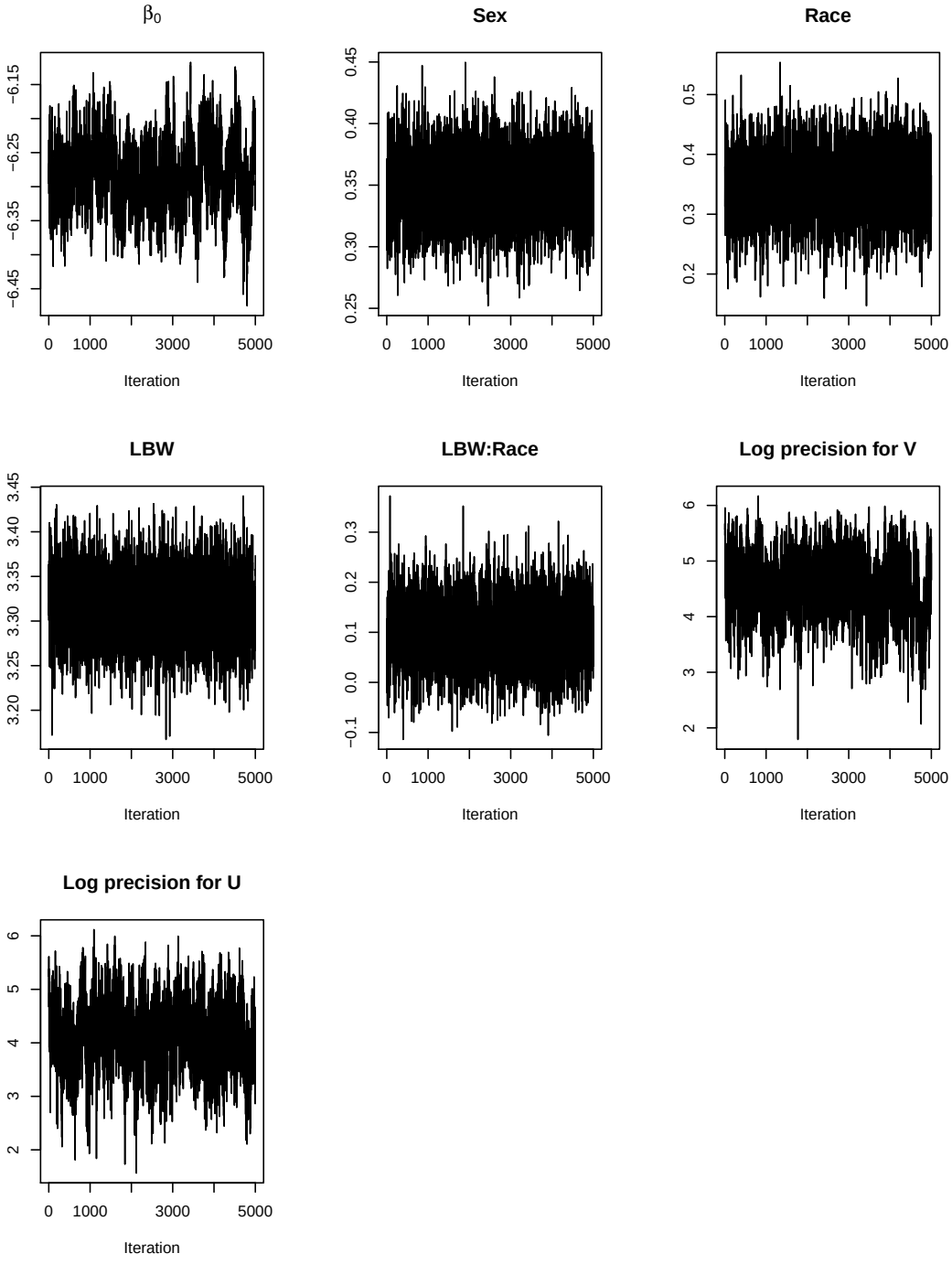


Figure B.1: Trace plots for the $\boldsymbol{\beta}$ parameters, $\log \tau_v$ and $\log \tau_u$ in the North Carolina infant mortality data example with large phase II sample sizes.

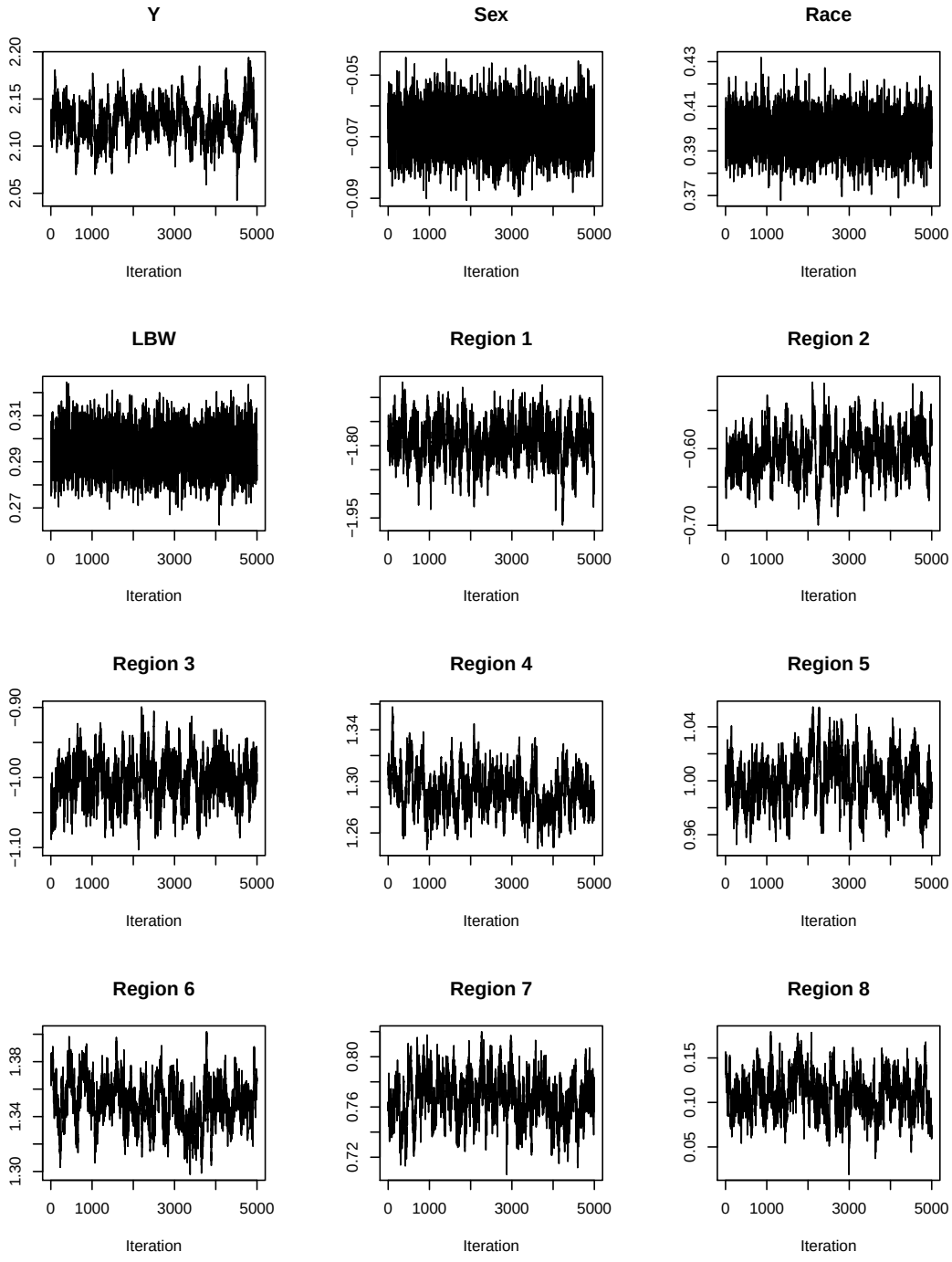


Figure B.2: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example with large phase II sample sizes (continued).

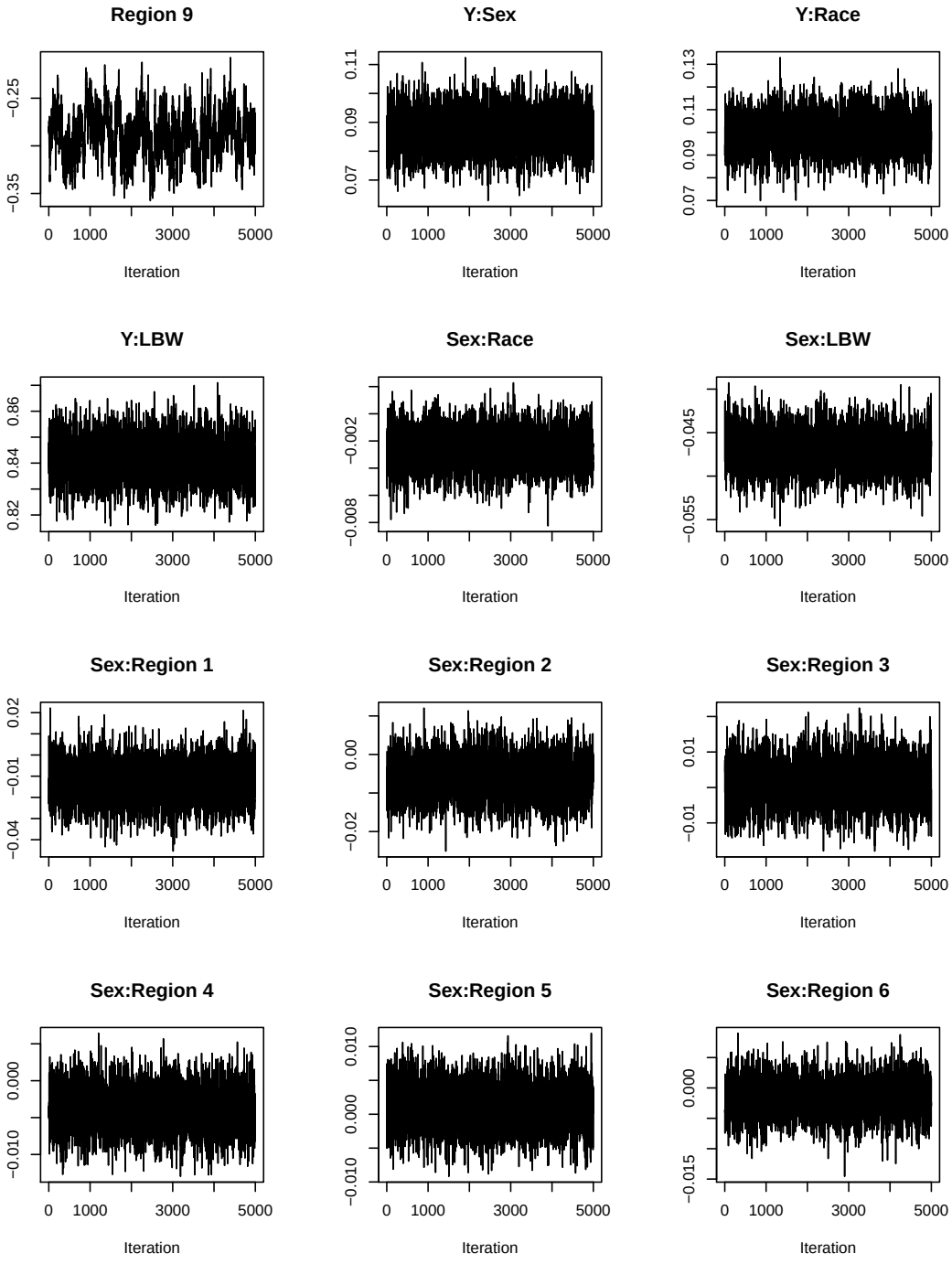


Figure B.3: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example with large phase II sample sizes (continued).

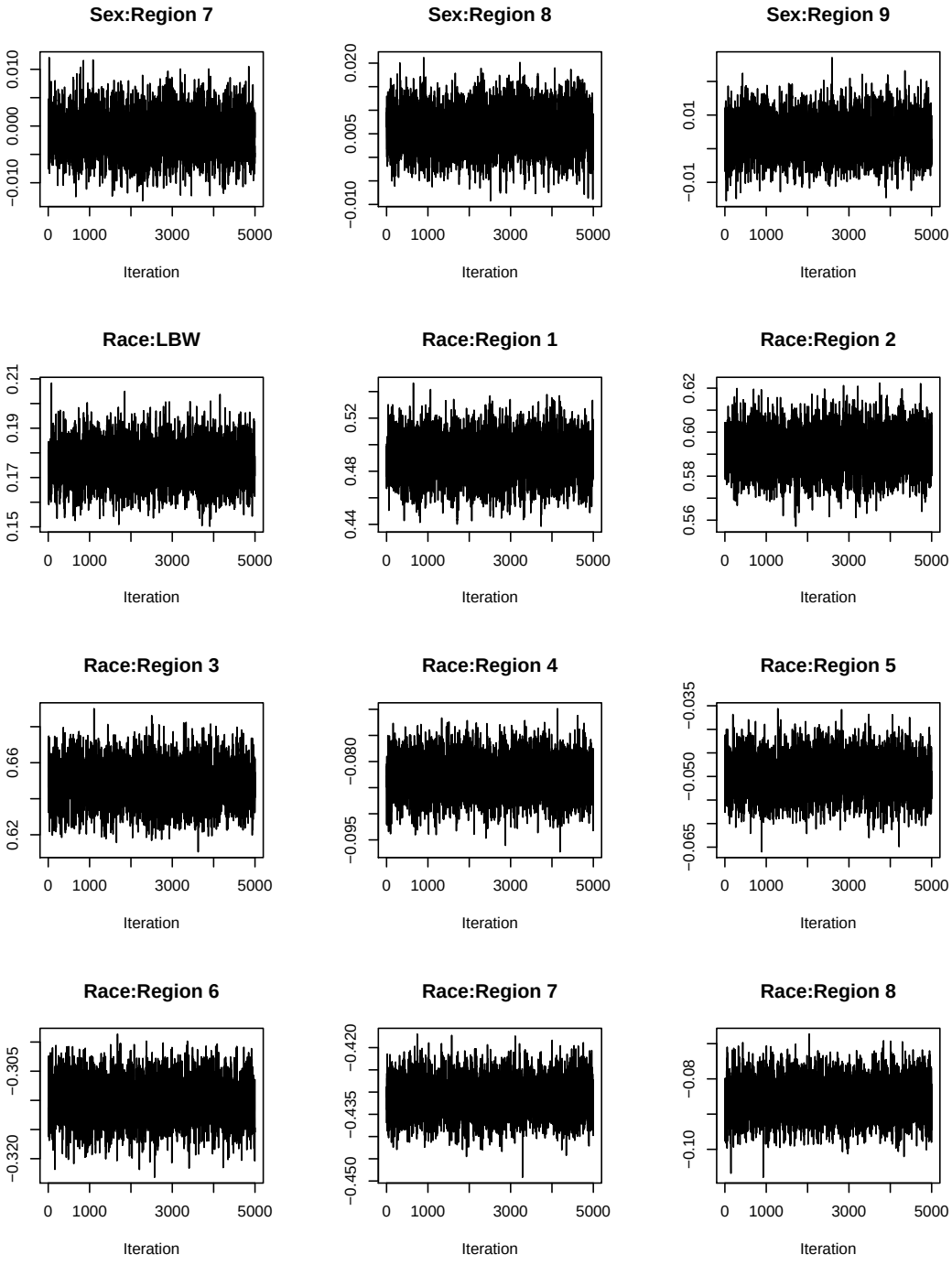


Figure B.4: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example with large phase II sample sizes (continued).

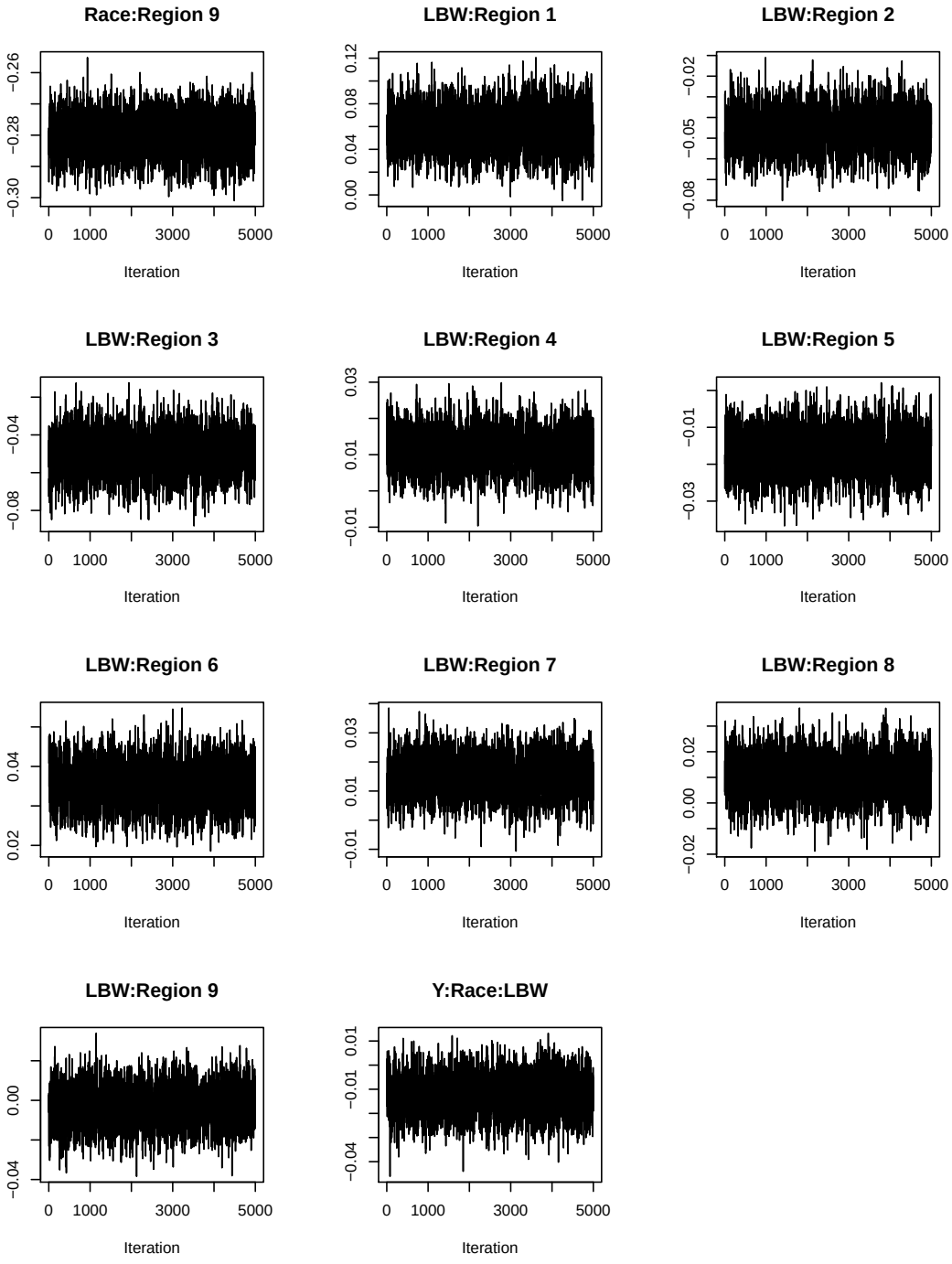


Figure B.5: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example with large phase II sample sizes (continued).

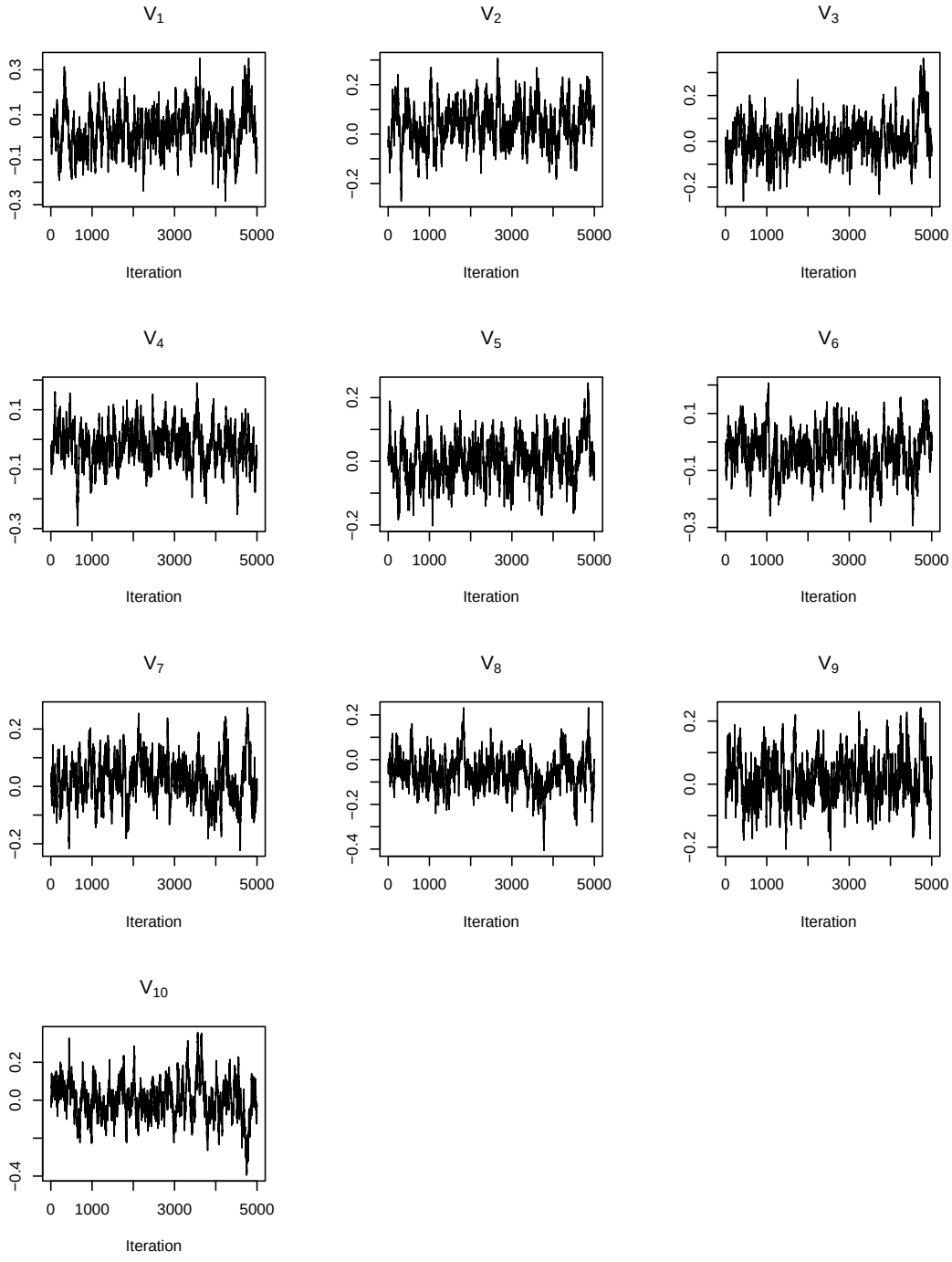


Figure B.6: Trace plots for the $\boldsymbol{V}$ parameters in the North Carolina infant mortality data example with large phase II sample sizes.

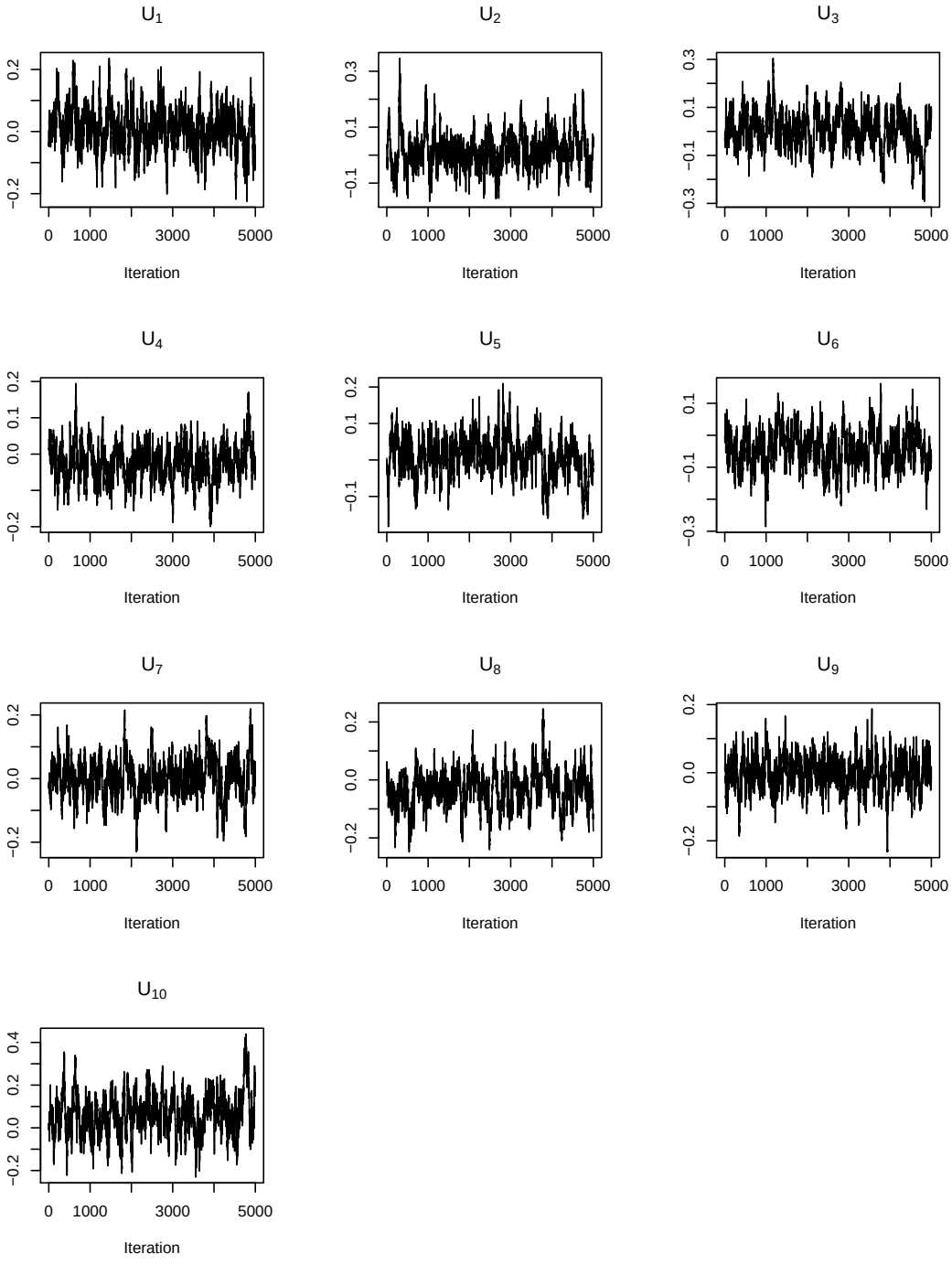


Figure B.7: Trace plots for the $\boldsymbol{V}$ parameters in the North Carolina infant mortality data example with large phase II sample sizes.

Table B.1: Number of live and dead infants by North Carolina region, race, sex and low birth weight status: phase II data.

		Normal Birth Weight				Low Birth Weight			
		Female		Male		Female		Male	
		White	Non-white	White	Non-white	White	Non-white	White	Non-white
Southern Mountains	Alive	4040	440	4379	469	322	29	295	26
	Dead	9	1	15	2	9	2	35	4
Central Mountains	Alive	14379	1188	15312	1302	1277	165	1211	166
	Dead	35	4	50	7	59	19	96	12
Northern Mountains	Alive	10365	780	10986	839	983	107	841	99
	Dead	35	2	40	2	49	4	54	7
Southern Foothills	Alive	54633	17251	58436	18436	4431	2684	3878	2251
	Dead	105	56	153	73	216	182	272	241
Northern Foothills	Alive	40230	11896	42393	12531	3361	1868	3044	1677
	Dead	77	28	114	35	187	157	216	202
Southern Heartland	Alive	48589	23906	51696	25295	3513	3696	3110	3195
	Dead	83	70	107	86	154	273	207	357
Southern Coast	Alive	22123	14120	23627	14829	1777	2129	1561	1834
	Dead	31	46	57	56	87	193	128	213
Northern Heartland	Alive	16923	5492	17858	5634	1384	834	1227	648
	Dead	32	6	43	22	72	50	89	71
Central Coast	Alive	8828	4038	9330	4268	672	734	560	570
	Dead	9	7	22	13	36	71	50	74
Northern Coast	Alive	3884	2833	4044	2811	304	438	300	386
	Dead	9	9	8	6	23	44	28	46

Table B.2: β estimates and 95% intervals for the North Carolina infant mortality data with large phase II sample sizes.

	Complete Data	Bayesian Two-Phase
β_0	-6.28 (-6.38, -6.18)	-6.28 (-6.39, -6.18)
Sex	0.35 (0.29, 0.40)	0.35 (0.29, 0.40)
Race	0.35 (0.24, 0.46)	0.35 (0.24, 0.45)
LBW	3.31 (3.24, 3.39)	3.32 (3.24, 3.39)
Race:LBW	0.10 (-0.02, 0.23)	0.10 (-0.02, 0.22)

Table B.3: Random effects estimates and 95% intervals for the North Carolina infant mortality data with large phase II sample sizes.

	Complete Data	Bayesian Two-Phase
V_1	0.02 (-0.14, 0.19)	0.04 (-0.15, 0.23)
V_2	0.04 (-0.10, 0.19)	0.05 (-0.10, 0.20)
V_3	0.01 (-0.15, 0.17)	0.02 (-0.12, 0.24)
V_4	-0.02 (-0.15, 0.10)	-0.02 (-0.15, 0.09)
V_5	0.001 (-0.12, 0.12)	0.01 (-0.13, 0.14)
V_6	-0.05 (-0.18, 0.09)	-0.04 (-0.18, 0.12)
V_7	0.01 (-0.13, 0.15)	0.02 (-0.12, 0.20)
V_8	-0.05 (-0.19, 0.08)	-0.06 (-0.24, 0.09)
V_9	0.02 (-0.12, 0.16)	0.02 (-0.12, 0.18)
V_{10}	0.02 (-0.15, 0.19)	0.0002 (-0.23, 0.23)
U_1	0.02 (-0.11, 0.16)	0.005 (-0.13, 0.13)
U_2	0.01 (-0.11, 0.13)	0.02 (-0.10, 0.16)
U_3	0.02 (-0.12, 0.16)	-0.002 (-0.19, 0.13)
U_4	-0.03 (-0.13, 0.07)	-0.02 (-0.13, 0.07)
U_5	0.01 (-0.09, 0.10)	0.003 (-0.12, 0.13)
U_6	-0.04 (-0.16, 0.07)	-0.04 (-0.16, 0.08)
U_7	0.01 (-0.10, 0.13)	0.005 (-0.14, 0.13)
U_8	-0.03 (-0.15, 0.07)	-0.03 (-0.15, 0.10)
U_9	-0.002 (-0.11, 0.10)	-0.004 (-0.11, 0.09)
U_{10}	0.03 (-0.13, 0.21)	0.07 (-0.11, 0.31)
τ_v	115.5 (36.3, 263.2)	106.2 (27.1, 247.2)
τ_u	81.3 (20.0, 208.9)	76.0 (18.2, 195.5)

B.2 $2 \times 2 \times 2 \times 3 \times 7$ *Contingency Table Example*

In the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table example of Chapter 4, the uninformative $\boldsymbol{\beta}$ prior induces the following prior on $\boldsymbol{\Lambda}^Y$:

$$(\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YA}, \boldsymbol{\lambda}^{YR}, \boldsymbol{\lambda}^{YS}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{YRS}) \sim N_{12}\left(\mathbf{0}, \boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T\right)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 100^2 & & 0 \\ & \ddots & \\ 0 & & 100^2 \end{pmatrix}$$

$$\boldsymbol{C} = \begin{pmatrix} -2 & -2 & 0 & -2 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 2 & 2 & 2 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -8 \end{pmatrix},$$

and, following the Dellaportas and Forster (1999) prescription, for the remaining fixed $\boldsymbol{\lambda}$ nuisance parameters, we assigned:

$$\begin{aligned}
\lambda_A &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\right) \\
\lambda_R &\sim N(0,1) \\
\lambda_S &\sim N(0,1) \\
\lambda_Z &\sim N_6\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 6 & & -1 \\ & \ddots & \\ -1 & & 6 \end{pmatrix}\right) \\
\lambda_{AR} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{6}\begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}\right) \\
\lambda_{AS} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{6}\begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}\right)
\end{aligned}$$

$$\begin{aligned}
\lambda_{RS} &\sim N\left(0, \frac{1}{4}\right) \\
\lambda_{ARS} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{3}{16}\left(I_2 - \frac{1}{3}J_2\right)\left(I_2 - \frac{1}{3}J_2\right)\left(I_2 - \frac{1}{3}J_2\right)\right)
\end{aligned}$$

where I_n is an $n \times n$ identity matrix, and J_n is an $n \times n$ matrix of ones.

Figure B.8 compares the prior and posterior median and 95% intervals for σ_λ under the two assumed priors for τ_λ.

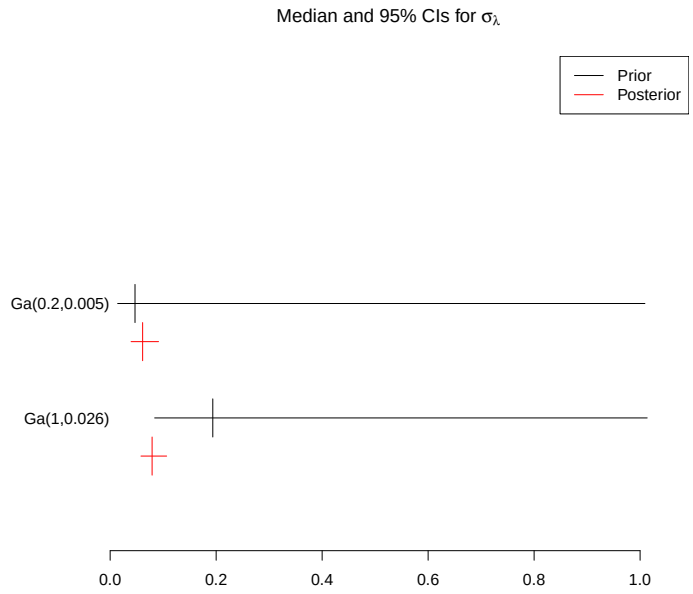


Figure B.8: Comparing the median and 95% intervals for σ_λ in the prior and the posterior distributions for two different priors of τ_λ in the random effects analysis of the $2\times2\times2\times3\times7$ contingency table.

Figures B.9 - B.19 display the trace plots for the $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{V}^{XZ}$ and τ_λ parameters. In each case, the samples are thinned by 100.

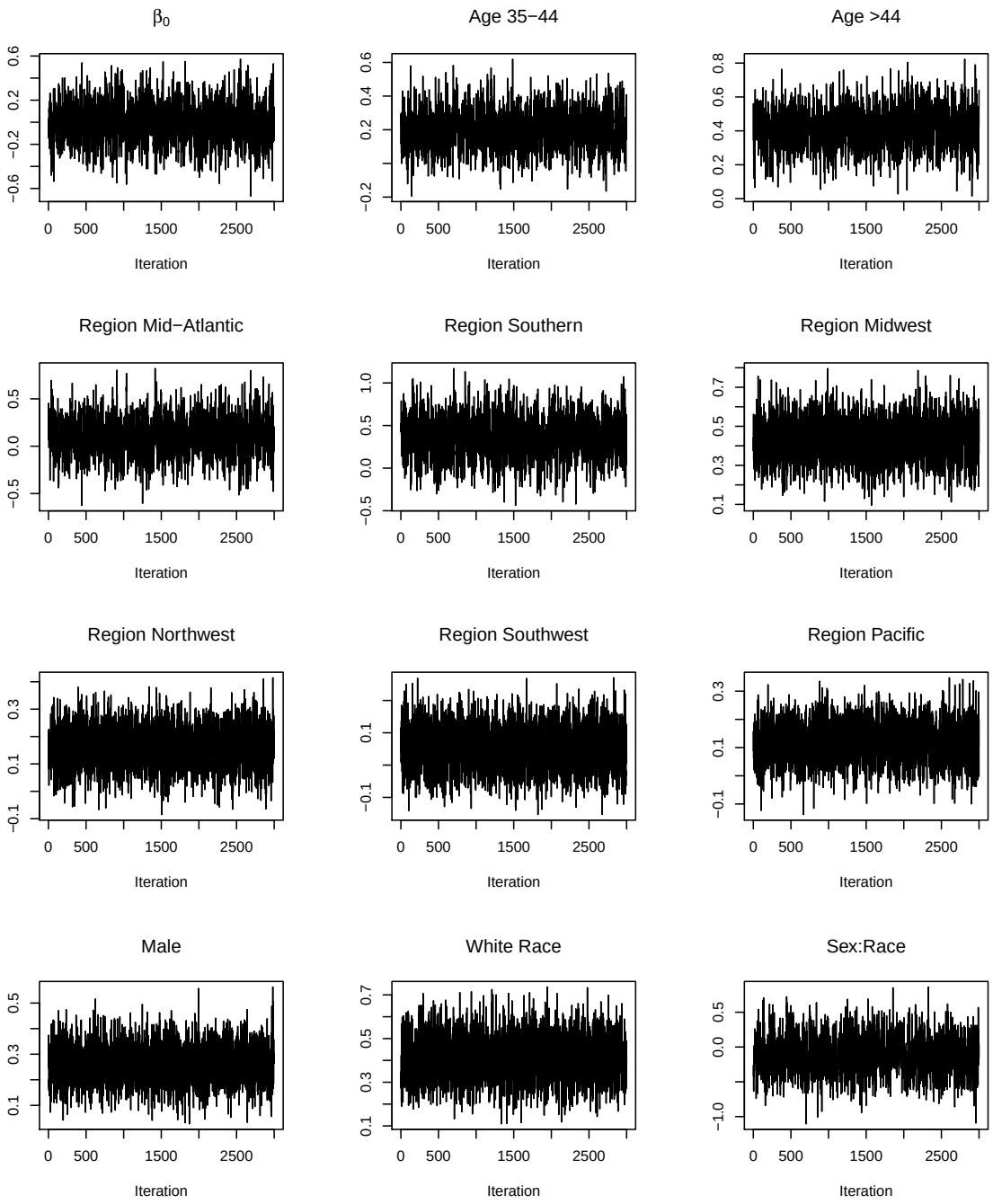


Figure B.9: Trace plots for the $\boldsymbol{\beta}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table.

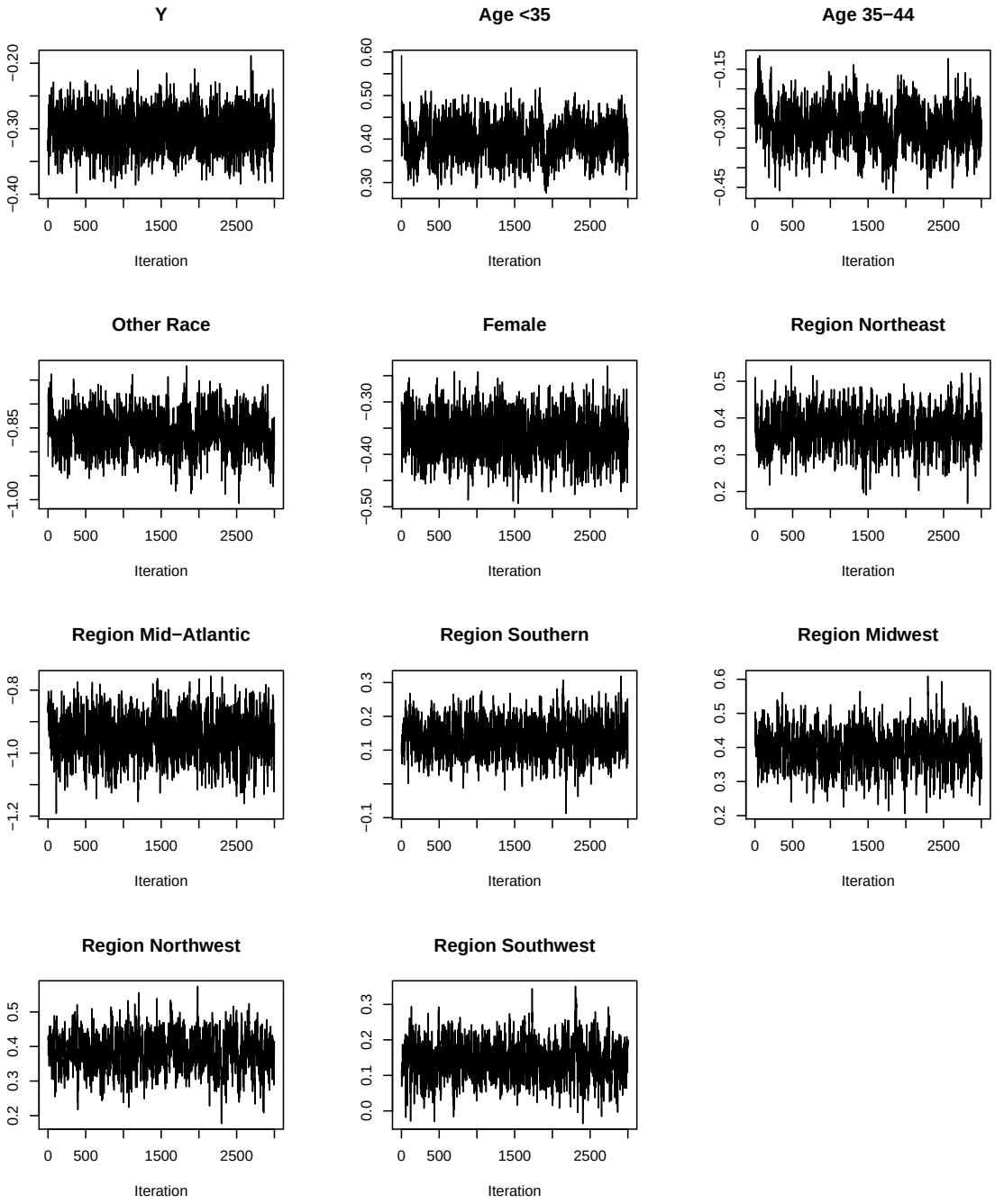


Figure B.10: Trace plots for the $\boldsymbol{\lambda}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table.

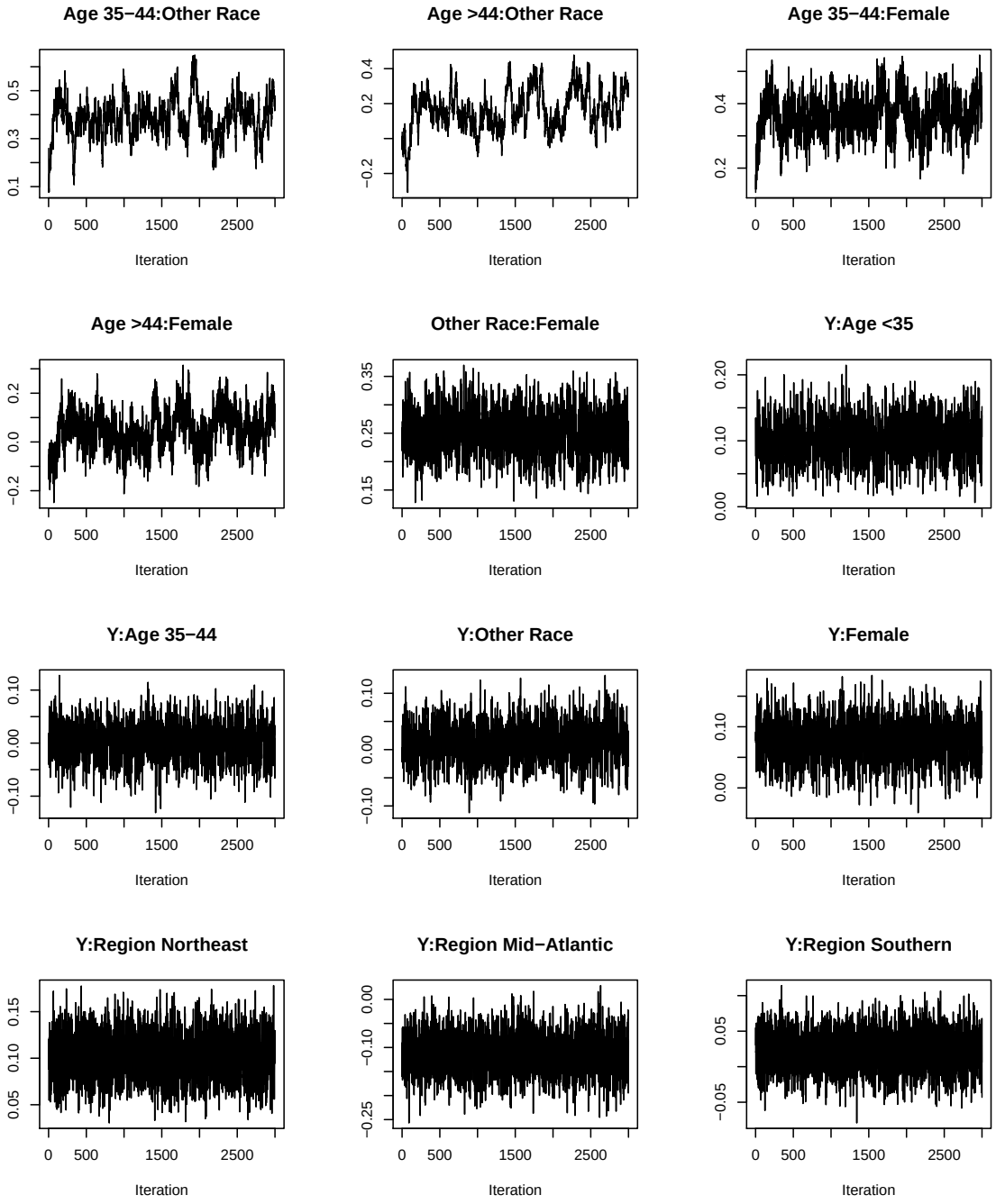


Figure B.11: Trace plots for the $\boldsymbol{\lambda}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table (continued).

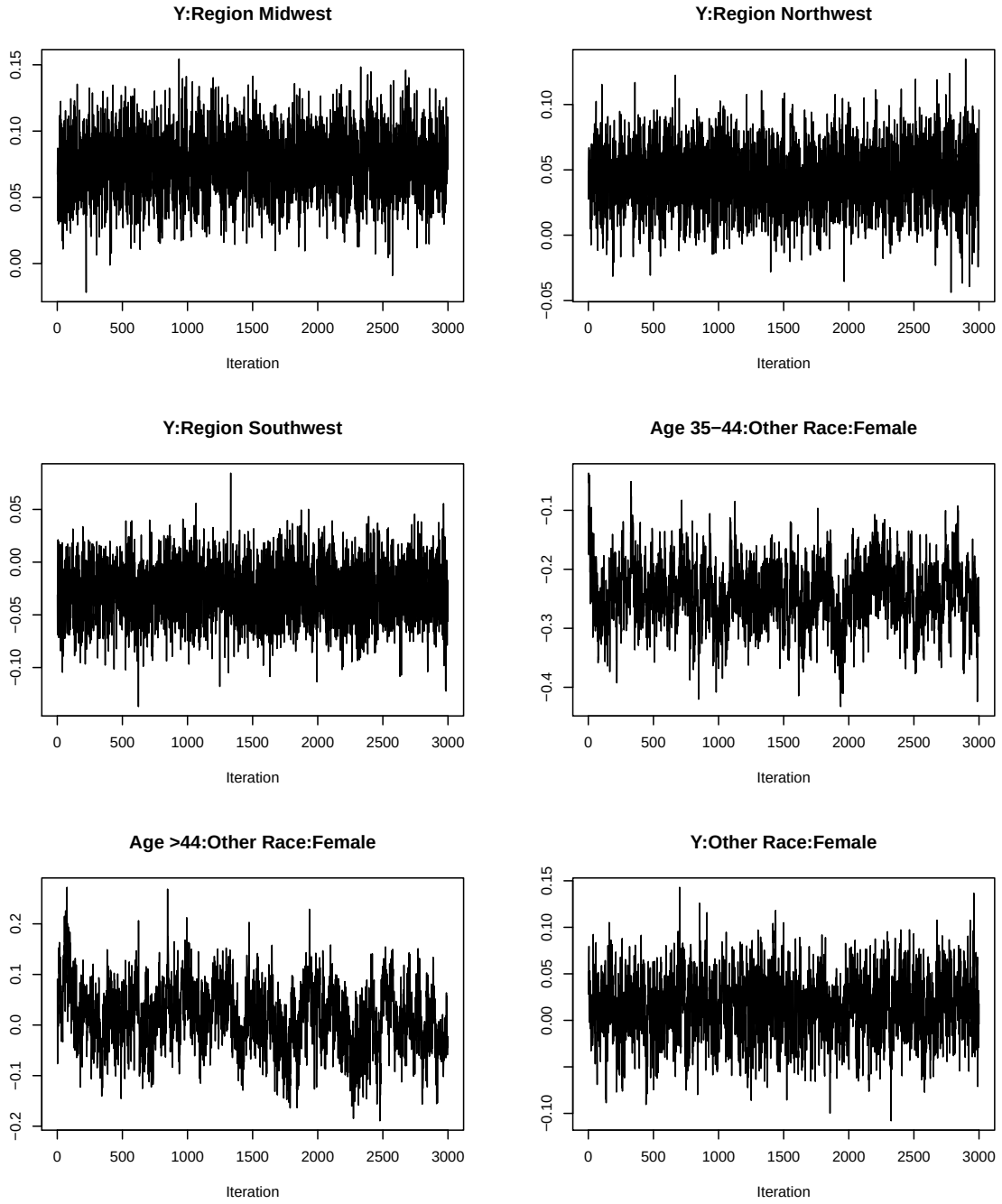


Figure B.12: Trace plots for the $\boldsymbol{\lambda}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table (continued).

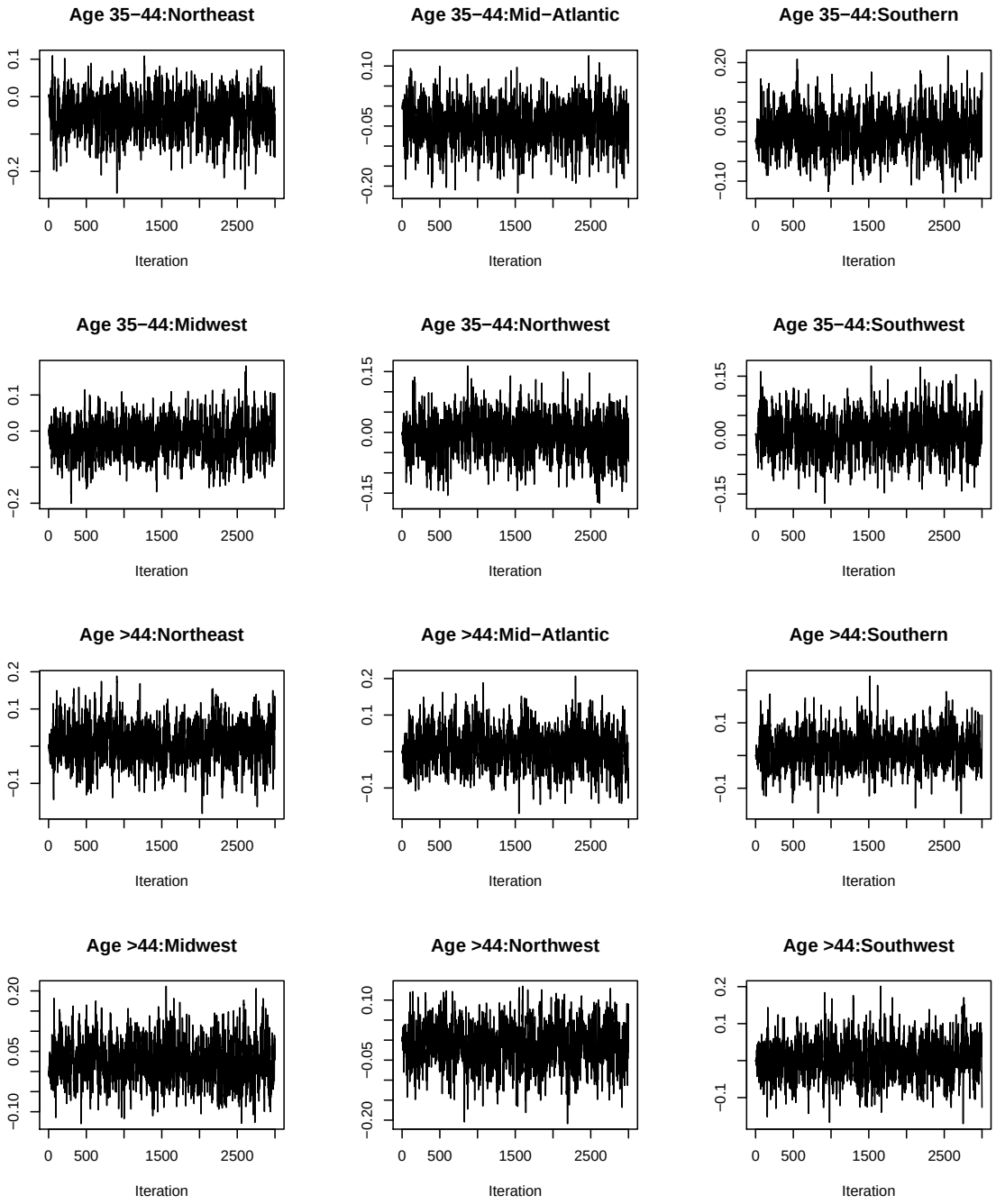


Figure B.13: Trace plots for the $\boldsymbol{V}^{XZ}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table.

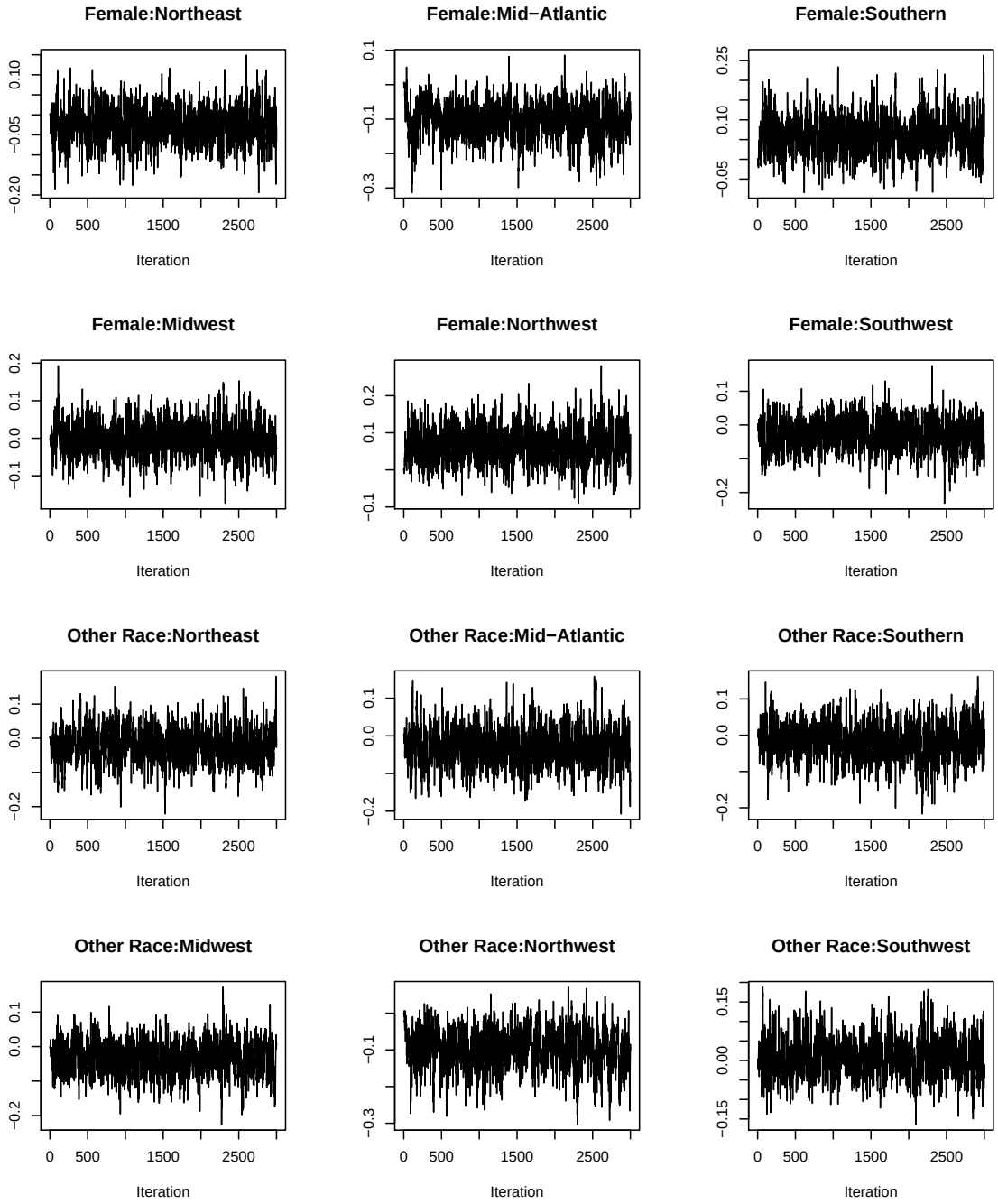


Figure B.14: Trace plots for the $\boldsymbol{V}^{XZ}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table (continued).

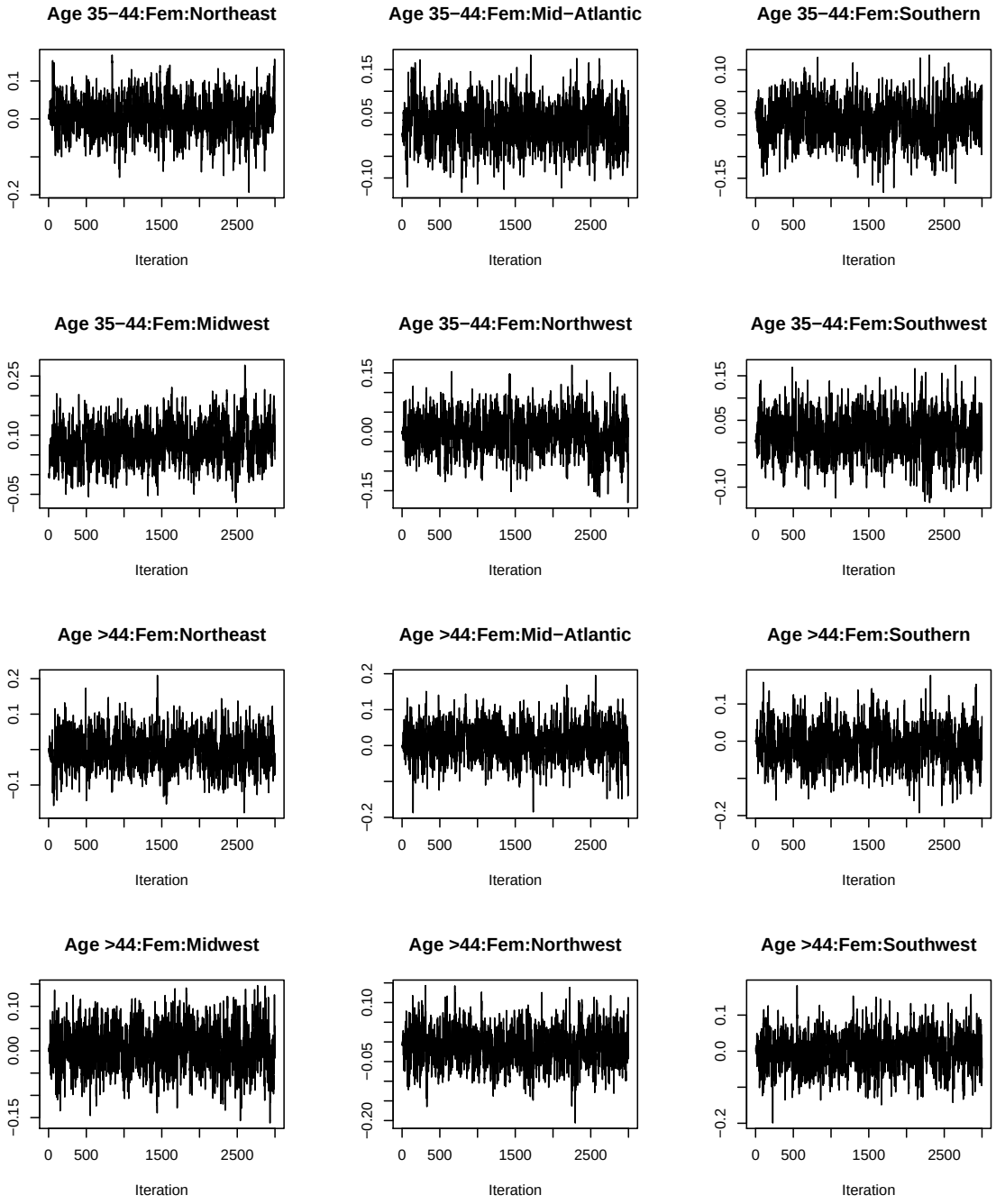


Figure B.15: Trace plots for the $\boldsymbol{V}^{XZ}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table (continued).

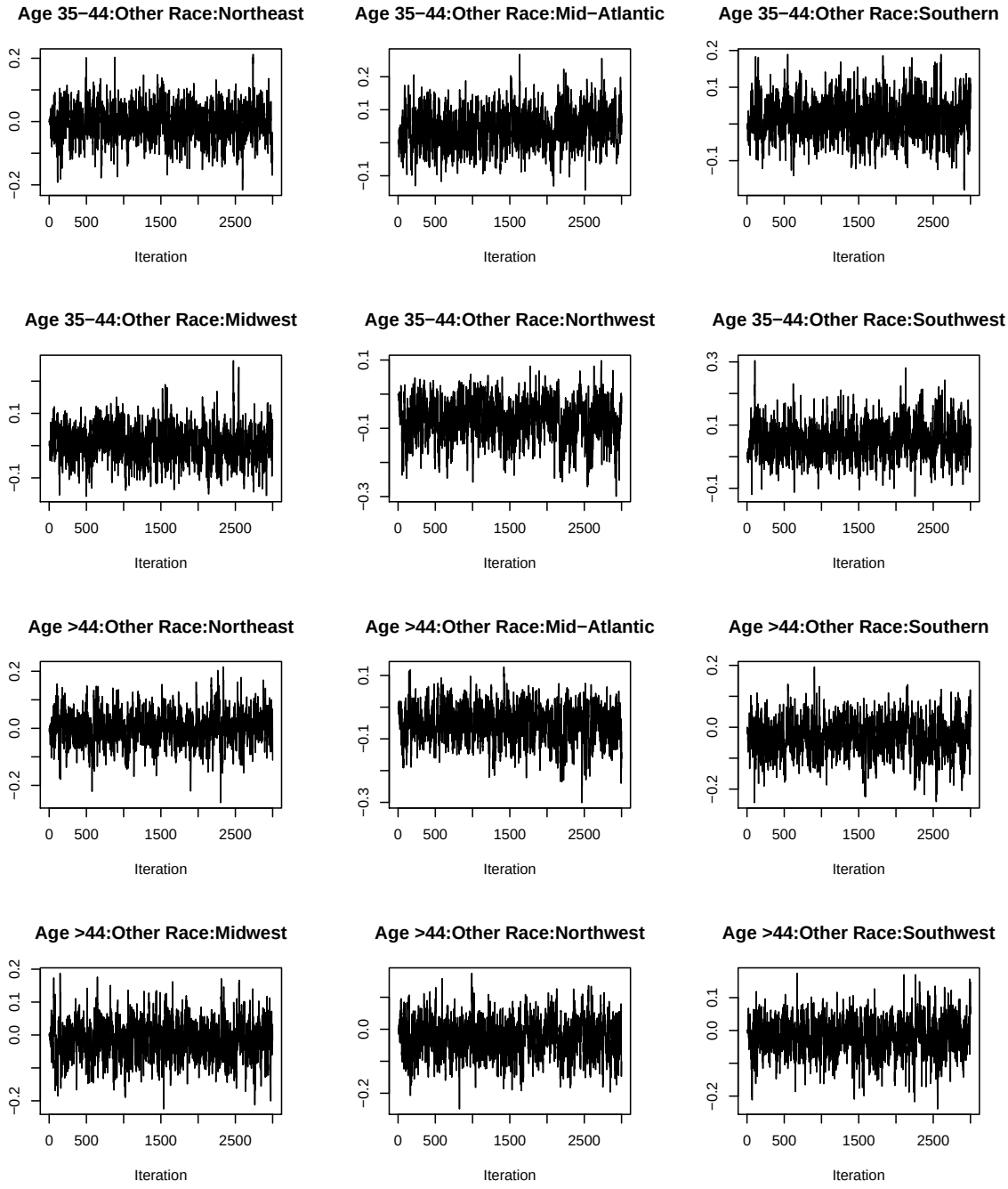


Figure B.16: Trace plots for the $\boldsymbol{V}^{XZ}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table (continued).

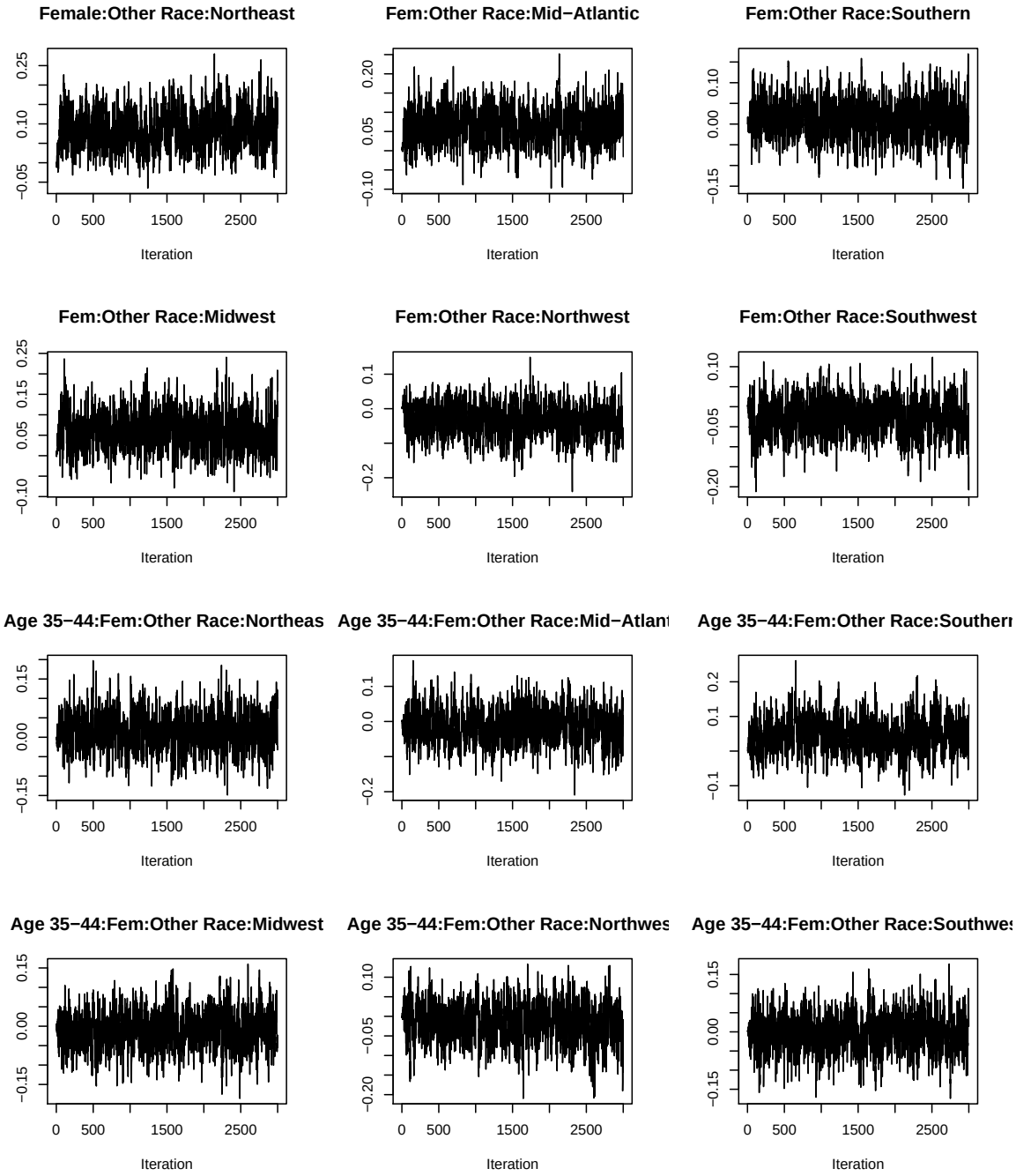


Figure B.17: Trace plots for the $\boldsymbol{V}^{XZ}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table (continued).

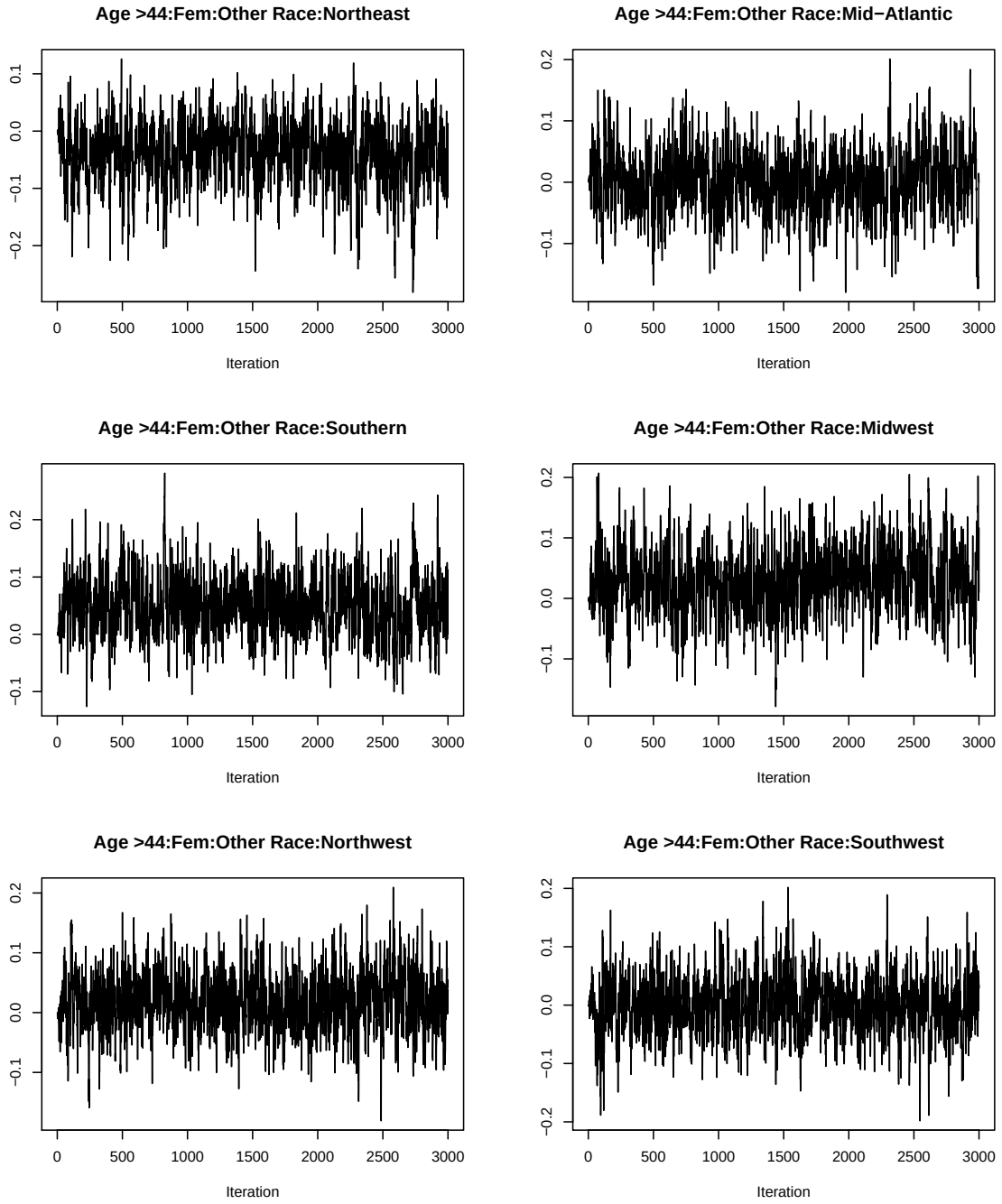


Figure B.18: Trace plots for the $\boldsymbol{V}^{XZ}$ parameters in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table (continued).

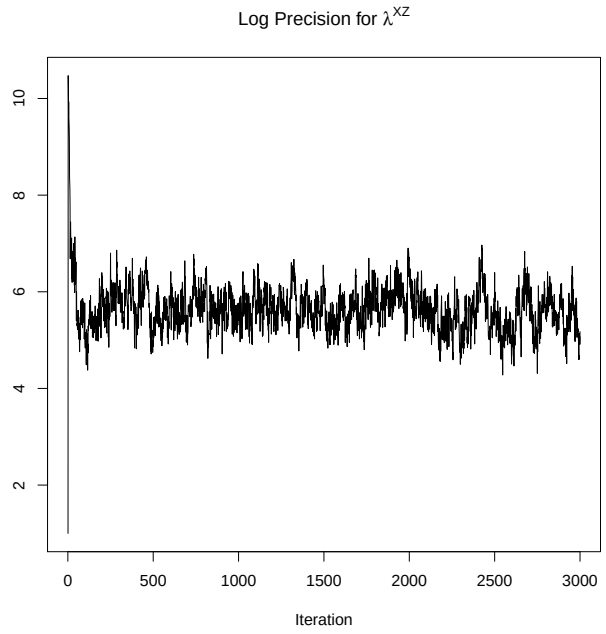


Figure B.19: Trace plots for $\log \tau_\lambda$ in the random effects analysis of the $2 \times 2 \times 2 \times 3 \times 7$ contingency table.

B.3 North Carolina Infant Mortality Data Example

For the North Carolina infant mortality data examined in Chapter 4 of the dissertation, we assigned zero mean normal prior distribution on $\boldsymbol{\beta}$ with large variances, which induces a multivariate normal prior on $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YS}, \boldsymbol{\lambda}^{YR}, \boldsymbol{\lambda}^{YW}, \boldsymbol{\lambda}^{YRW})$ that has mean zero and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$, where $\boldsymbol{C}$ and $\boldsymbol{\Sigma}$ are the 5×5 matrices given by

$$\begin{aligned}
\boldsymbol{C} &= \begin{pmatrix} -2 & -2 & -2 & -2 & -2 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 4 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & -8 \end{pmatrix} \\
\boldsymbol{\Sigma} &= \begin{pmatrix} 100^2 & & 0 \\ & \ddots & \\ 0 & & 100^2 \end{pmatrix}.
\end{aligned}$$

For the remaining $\boldsymbol{\lambda}$ nuisance parameters, we assigned

$$\begin{aligned}
\boldsymbol{\lambda}^S &\sim N(0,1) \\
\boldsymbol{\lambda}^R &\sim N(0,1) \\
\boldsymbol{\lambda}^W &\sim N(0,1) \\
\boldsymbol{\lambda}^Z &\sim N_9\left(\mathbf{0}, 10\left(I_9 - \frac{1}{10}J_9\right)\right) \\
\boldsymbol{\lambda}^{SR} &\sim N\left(0, \frac{1}{4}\right) \\
\boldsymbol{\lambda}^{SW} &\sim N\left(0, \frac{1}{4}\right) \\
\boldsymbol{\lambda}^{SZ} &\sim N_9\left(\mathbf{0}, 5\left(I_9 - \frac{1}{10}J_9\right)\left(I_9 - \frac{1}{10}J_9\right)\right) \\
\boldsymbol{\lambda}^{RW} &\sim N\left(0, \frac{1}{4}\right) \\
\boldsymbol{\lambda}^{RZ} &\sim N_9\left(\mathbf{0}, 5\left(I_9 - \frac{1}{10}J_9\right)\left(I_9 - \frac{1}{10}J_9\right)\right) \\
\boldsymbol{\lambda}^{WZ} &\sim N_9\left(\mathbf{0}, 5\left(I_9 - \frac{1}{10}J_9\right)\left(I_9 - \frac{1}{10}J_9\right)\right),
\end{aligned}$$

where I_n is an $n \times n$ identity matrix and J_n is an $n \times n$ matrix of ones.

Table B.4 displays the estimates and 95% intervals for the random effects terms from the complete data and the Bayesian random effects analyses.

Table B.4: Random effects estimates and 95% intervals for the North Carolina infant mortality data example.

	Full Data	Bayesian Two-Phase
V_1	0.02 (-0.14, 0.19)	-0.04 (-0.22, 0.15)
V_2	0.04 (-0.10, 0.19)	0.06 (-0.16, 0.30)
V_3	0.01 (-0.15, 0.17)	0.09 (-0.12, 0.34)
V_4	-0.02 (-0.15, 0.10)	0.05 (-0.16, 0.30)
V_5	0.001 (-0.12, 0.12)	-0.02 (-0.19, 0.13)
V_6	-0.05 (-0.18, 0.09)	-0.01 (-0.24, 0.22)
V_7	0.01 (-0.13, 0.15)	-0.01 (-0.24, 0.22)
V_8	-0.05 (-0.19, 0.08)	-0.03 (-0.28, 0.19)
V_9	0.02 (-0.12, 0.16)	-0.07 (-0.3, 0.11)
V_{10}	0.02 (-0.15, 0.19)	0.01 (-0.22, 0.22)
U_1	0.02 (-0.11, 0.16)	-0.02 (-0.21, 0.13)
U_2	0.01 (-0.11, 0.13)	0.03 (-0.14, 0.22)
U_3	0.02 (-0.12, 0.16)	0.06 (-0.11, 0.27)
U_4	-0.03 (-0.13, 0.07)	0.002 (-0.14, 0.21)
U_5	0.01 (-0.09, 0.10)	-0.04 (-0.18, 0.08)
U_6	-0.04 (-0.16, 0.07)	-0.002 (-0.14, 0.12)
U_7	0.01 (-0.10, 0.13)	-0.02 (-0.21, 0.17)
U_8	-0.03 (-0.15, 0.07)	-0.004 (-0.15, 0.14)
U_9	-0.002 (-0.11, 0.10)	-0.04 (-0.18, 0.09)
U_{10}	0.03 (-0.13, 0.21)	0.04 (-0.17, 0.29)

Figures B.20 - B.26 display the trace plots for all parameters in the model. In all cases, the samples have been thinned by 100.

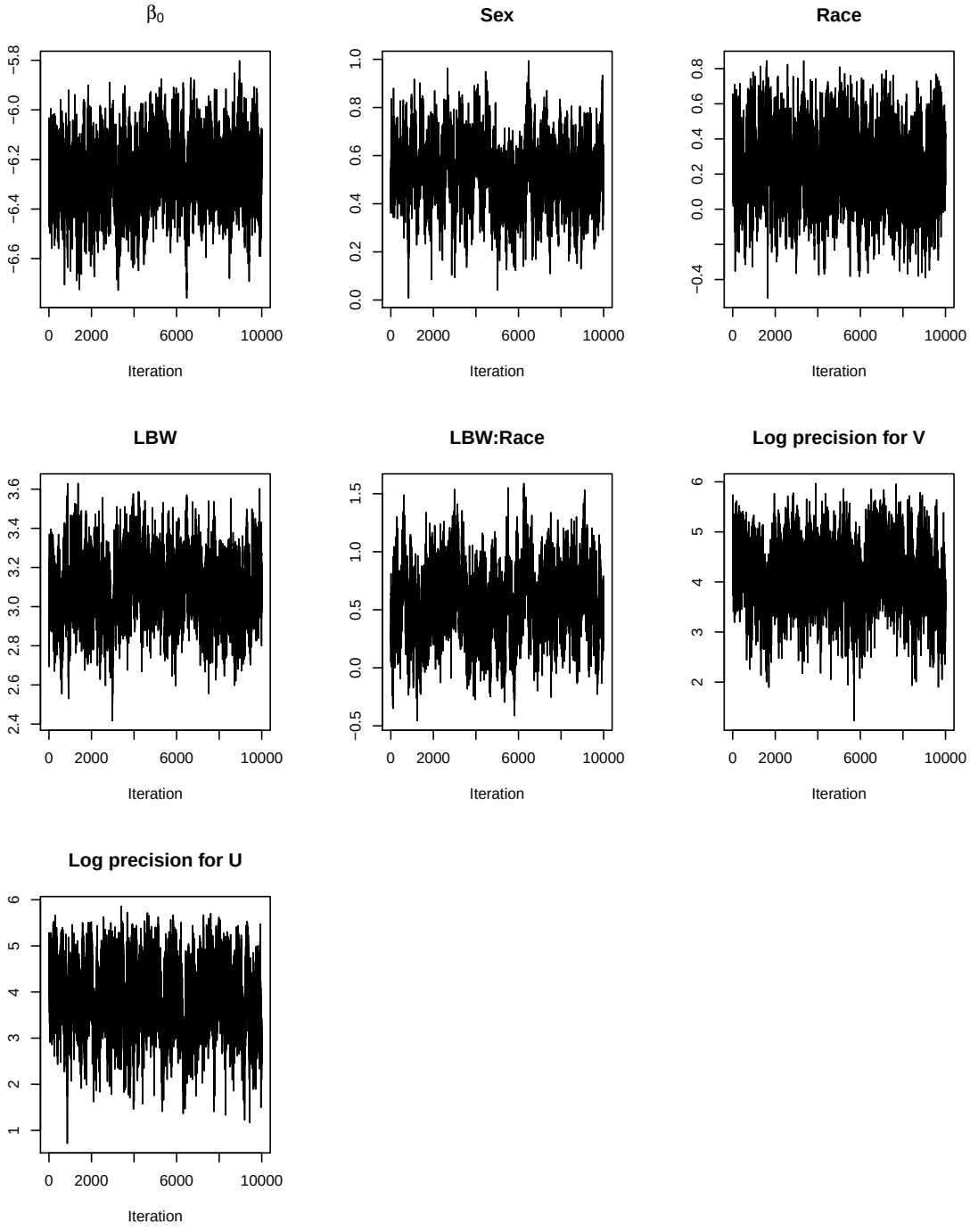


Figure B.20: Trace plots for the $\boldsymbol{\beta}$ parameters, $\log \tau_v$ and $\log \tau_u$ in the North Carolina infant mortality data example.

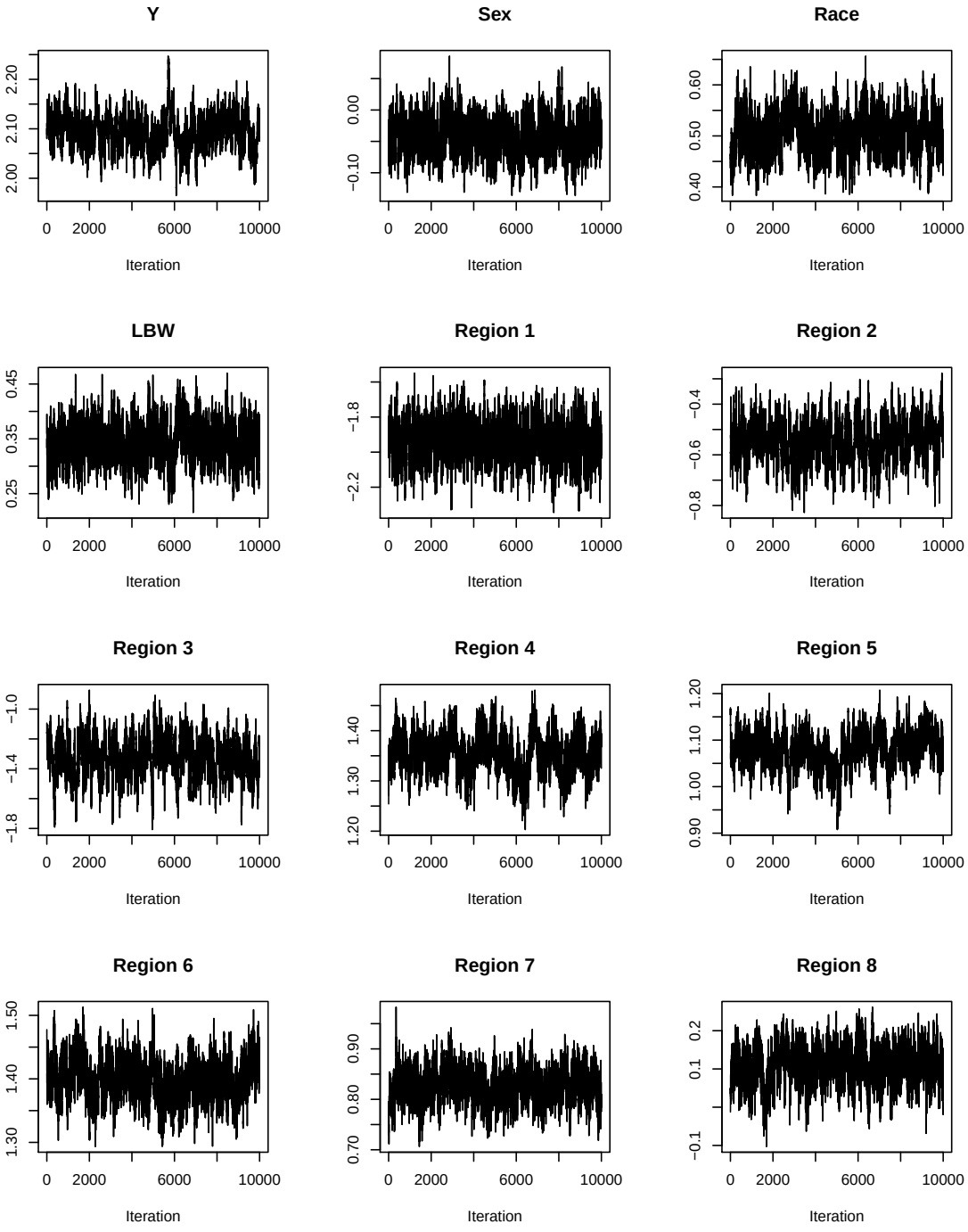


Figure B.21: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example (continued).

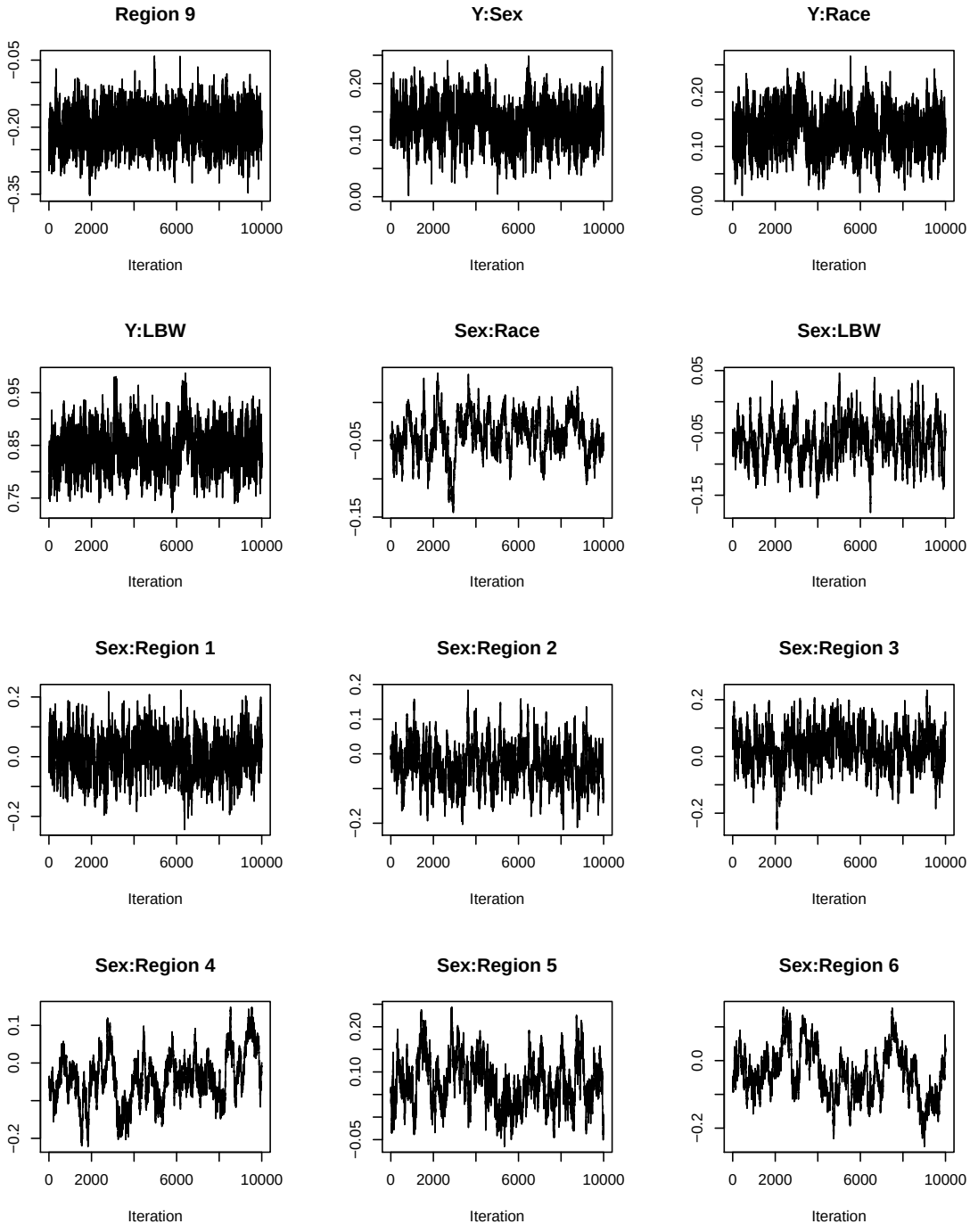


Figure B.22: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example (continued).

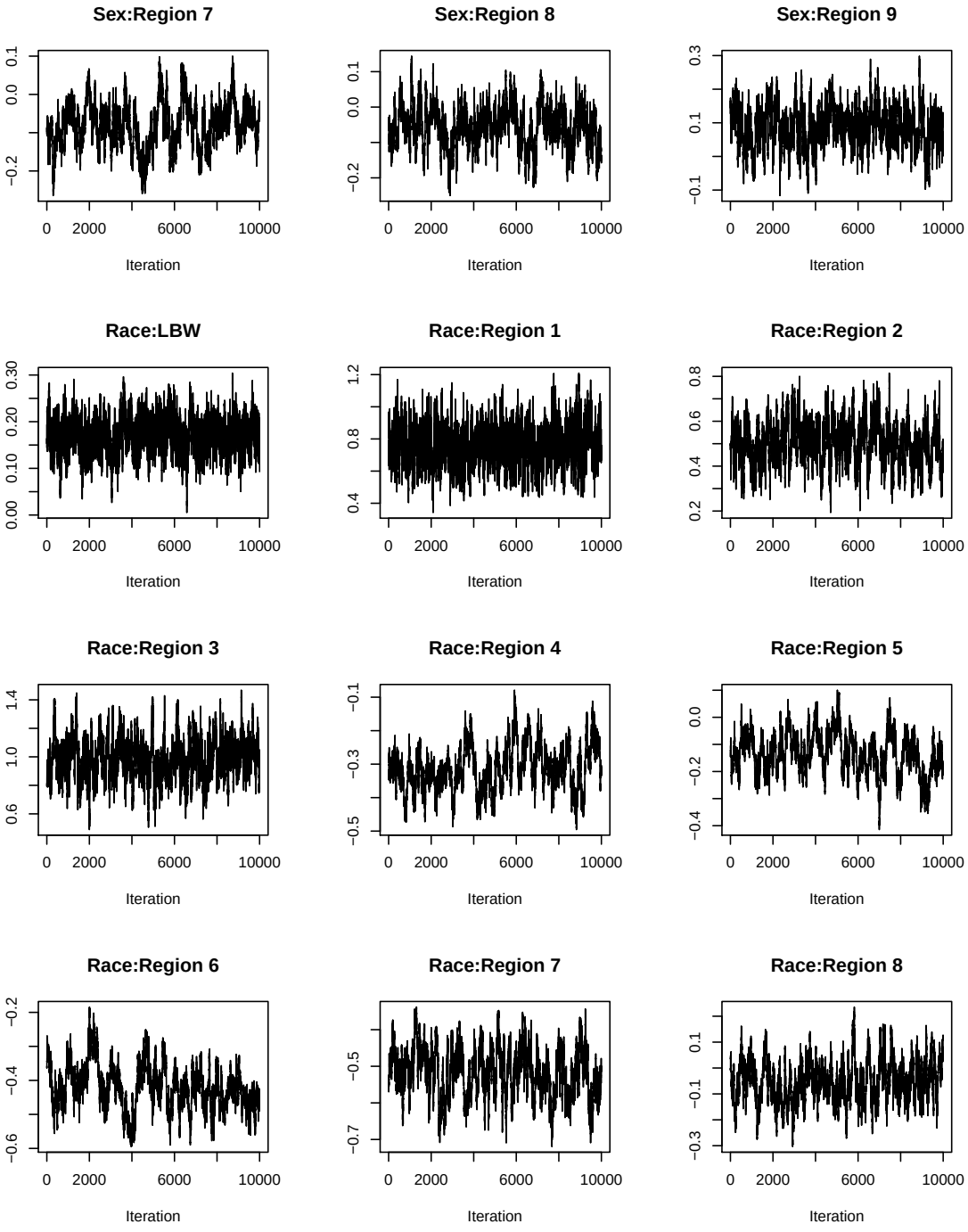


Figure B.23: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example (continued).

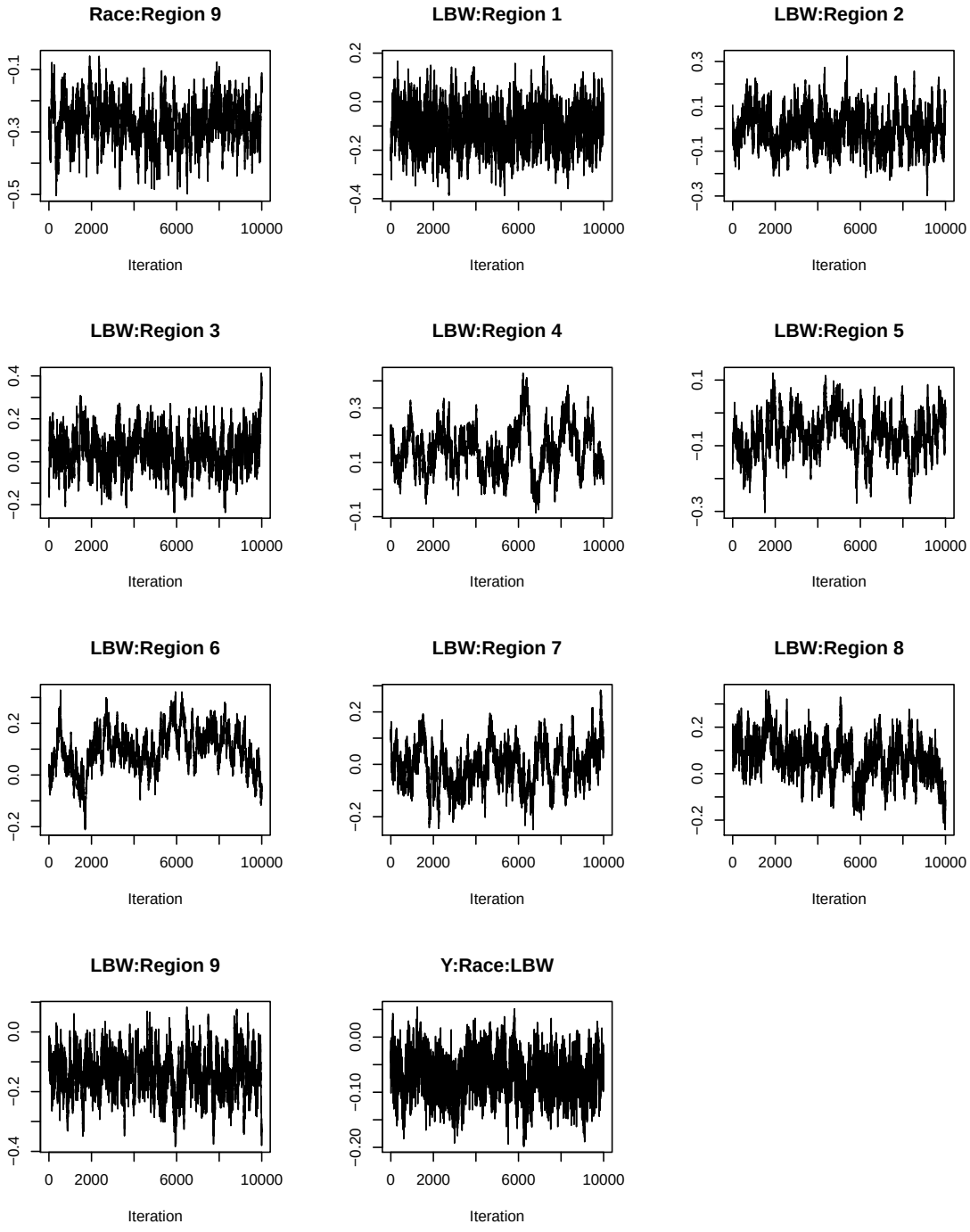


Figure B.24: Trace plots for the $\boldsymbol{\lambda}$ parameters in the North Carolina infant mortality data example (continued).

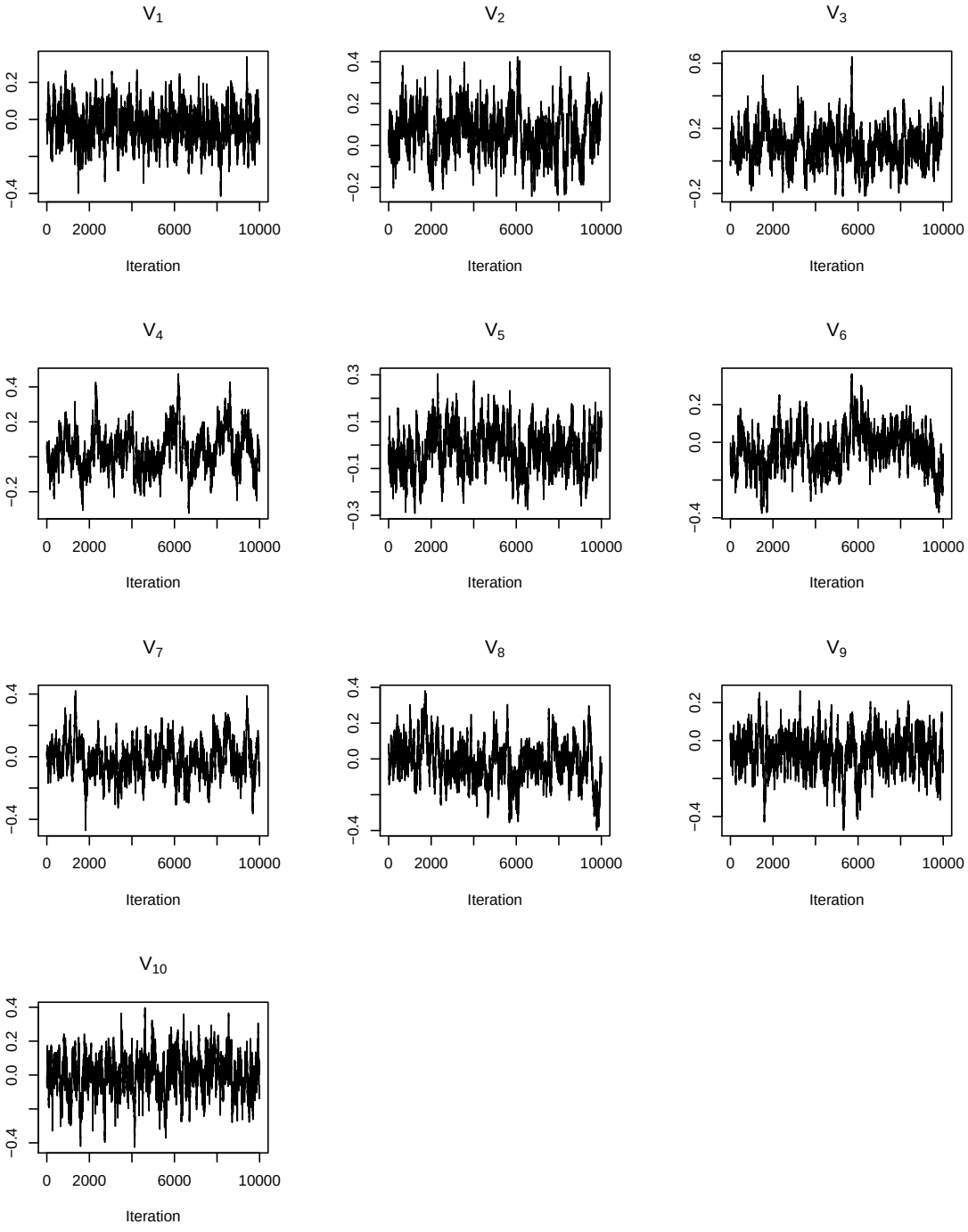


Figure B.25: Trace plots for the $\boldsymbol{V}$ parameters in the North Carolina infant mortality data example.

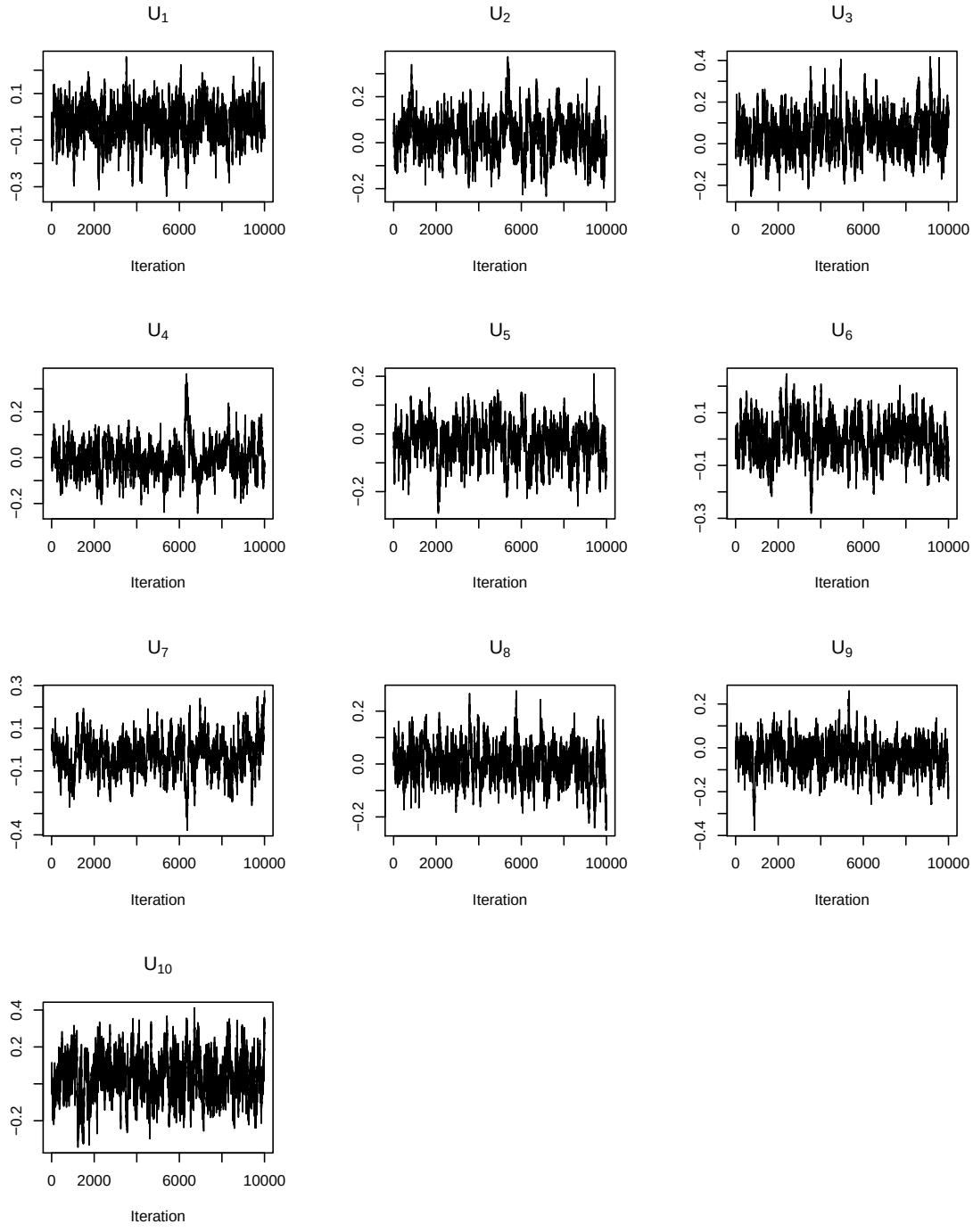


Figure B.26: Trace plots for the $\boldsymbol{V}$ parameters in the North Carolina infant mortality data example.

B.4 Simulated Data with Strong Spatial Dependence Example

Table B.5 displays the full data for the simulated data example with strong spatial dependence discussed in Chapter 4.

Table B.5: Number of live and dead infants by North Carolina region, and low birth weight status in the simulated data example with strong spatial dependence: full data.

		Normal Birth Weight	Low Birth Weight	
Southern Mountains	Alive	19499	1386	20885
	Dead	12	24	36
Central Mountains	Alive	35769	3107	38876
	Dead	50	154	204
Northern Mountains	Alive	15488	1299	16787
	Dead	54	149	203
Southern Foothills	Alive	53226	4674	57900
	Dead	735	2050	2785
Northern Foothills	Alive	66526	6193	72719
	Dead	279	898	1177
Southern Heartland	Alive	73871	6665	80536
	Dead	518	1386	1904
Southern Coast	Alive	93387	9506	102893
	Dead	113	319	432
Northern Heartland	Alive	44083	3955	48038
	Dead	84	214	298
Central Coast	Alive	17921	1798	19719
	Dead	80	208	288
Northern Coast	Alive	30925	3345	34270
	Dead	9	41	50

The non-informative prior distribution on $\boldsymbol{\beta}$ implies a multivariate normal prior with

mean zero and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^T$ for $\boldsymbol{\Lambda}^Y = (\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^{YX})$ where

$$\begin{aligned} \boldsymbol{C} &= \begin{pmatrix} -2 & -2 \\ 0 & 4 \end{pmatrix} \\ \boldsymbol{\Sigma} &= \begin{pmatrix} 100^2 & 0 \\ 0 & 100^2 \end{pmatrix}. \end{aligned}$$

and assigned

$$\begin{aligned} \boldsymbol{\lambda}^X &\sim N(0,1) \\ \boldsymbol{\lambda}^Z &\sim N_9\left(\mathbf{0}, 10\left(I_9 - \frac{1}{10}J_9\right)\right) \\ \boldsymbol{\lambda}^{XZ} &\sim N_9\left(\mathbf{0}, 5\left(I_9 - \frac{1}{10}J_9\right)\left(I_9 - \frac{1}{10}J_9\right)\right). \end{aligned}$$

Table B.6 displays the estimates and 95% intervals for the random effects parameters $\boldsymbol{V}$,$\boldsymbol{U}$,τ_v and τ_u.

Figures B.27 and B.28 display the predicted N^{yxz} values with 95% intervals compared to the true population values (shown in red crosses) for the non-spatial and spatial Bayesian random effects analyses.

Figures B.29-B.32 display the trace plots for the $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{V}$, and τ_v parameters for the non-spatial analysis, while figures B.33-B.37 display the trace plots for the $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{V}$, $\boldsymbol{U}$, τ_v and τ_u parameters for the spatial analysis. In each case, the samples have been thinned by 100.

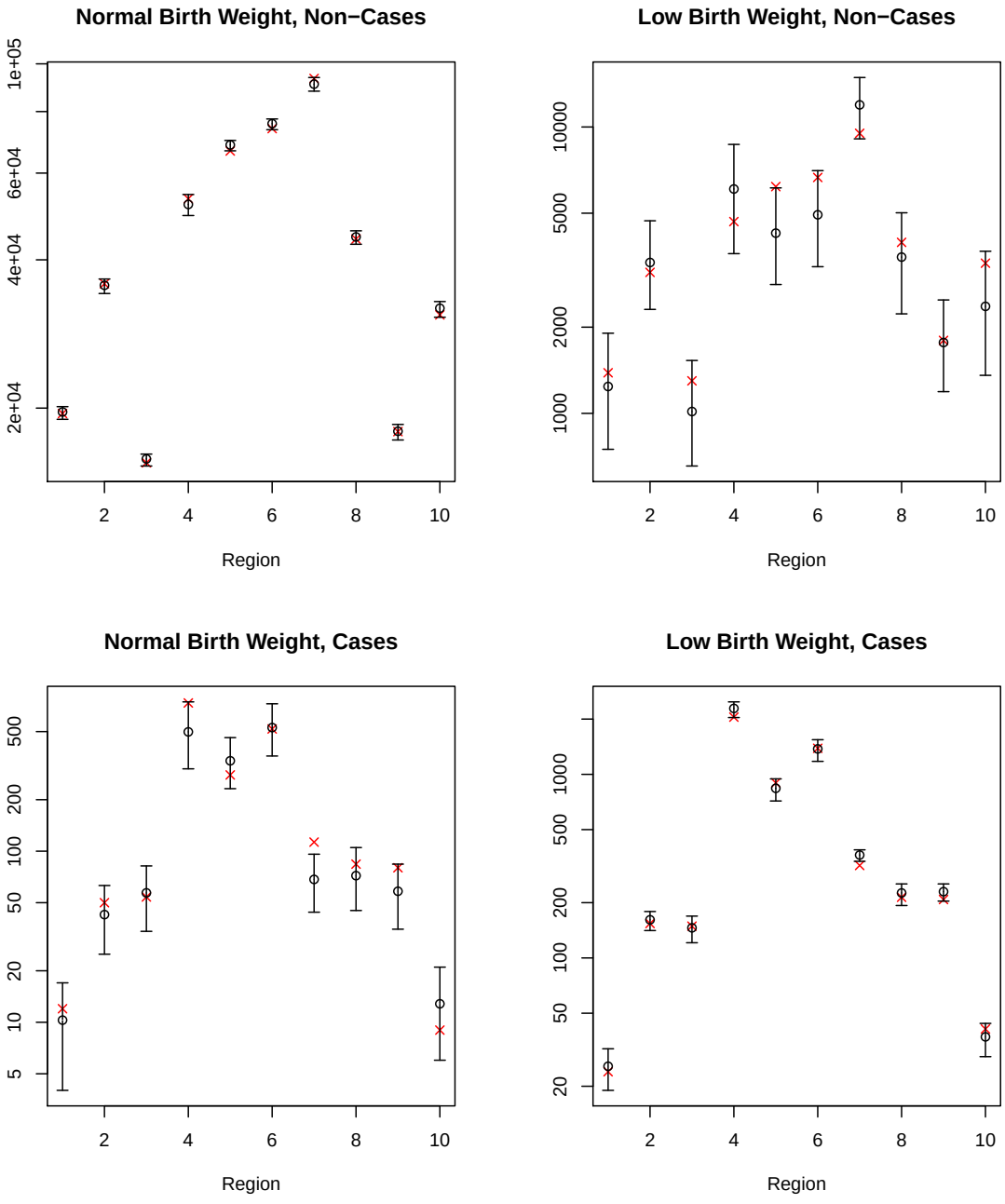


Figure B.27: Comparing the predicted N_{yxz} values from the non-spatial Bayesian two-phase analysis to the true values (shown in red crosses) for the simulated data example with strong spatial dependence.

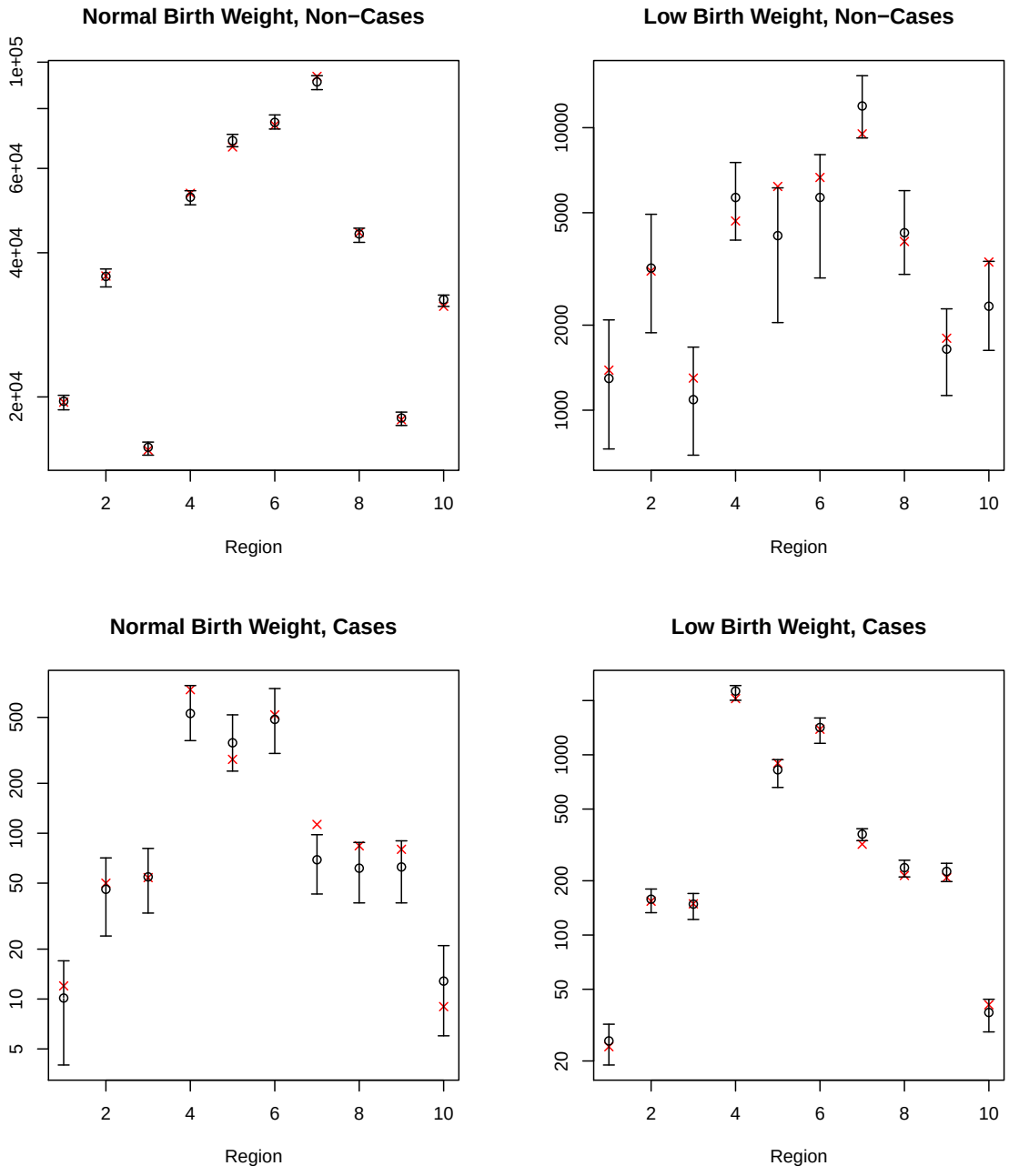


Figure B.28: Comparing the predicted N_{yxz} values from the spatial Bayesian two-phase analysis to the true values (shown in red crosses) for the simulated data example with strong spatial dependence.

Table B.6: Random effects estimates and 95% intervals for three analyses of the simulated data example with strong spatial dependence.

	Full Data	Bayesian Two-Phase	
		Non-Spatial	Spatial
V_1	-0.11 (-0.77, 0.28)	-1.40 (-1.87, -0.96)	0.01 (-0.40, 0.20)
V_2	-0.01 (-0.51, 0.46)	-0.64 (-1.01, -0.33)	-0.07 (-0.24, 0.23)
V_3	0.04 (-0.41, 0.59)	0.47 (0.14, 0.76)	-0.08 (-0.24, 0.09)
V_4	0.12 (-0.25, 0.80)	1.44 (0.96, 1.86)	0.11 (-0.10, 0.34)
V_5	0.06 (-0.33, 0.61)	0.79 (0.44, 1.10)	0.05 (-0.24, 0.25)
V_6	0.05 (-0.36, 0.62)	1.14 (0.89, 1.38)	-0.21 (-0.62, 0.49)
V_7	-0.05 (-0.62, 0.37)	-1.09 (-1.45, -0.85)	0.19 (0.02, 0.42)
V_8	-0.08 (-0.68, 0.30)	-0.33 (-0.74, 0.04)	0.03 (-0.29, 0.28)
V_9	0.02 (-0.43, 0.51)	0.36 (0.11, 0.69)	-0.11 (-0.28, 0.08)
V_{10}	-0.04 (-0.62, 0.41)	-1.72 (-2.12, -1.16)	-0.09 (-0.74, 0.37)
U_1	-1.23 (-1.71, -0.58)	–	-1.35 (-1.91, -0.73)
U_2	-0.39 (-0.84, 0.08)	–	-0.40 (-0.76, -0.10)
U_3	0.41 (-0.12, 0.84)	–	0.59 (0.14, 0.87)
U_4	1.68 (1.03, 2.05)	–	1.47 (1.19, 1.76)
U_5	0.62 (0.10, 0.99)	–	0.85 (0.57, 1.24)
U_6	1.03 (0.49, 1.42)	–	1.34 (0.89, 1.66)
U_7	-0.67 (-1.07, -0.13)	–	-1.19 (-1.40, -1.00)
U_8	-0.16 (-0.54, 0.41)	–	-0.43 (-0.82, -0.08)
U_9	0.51 (0.04, 0.94)	–	0.62 (0.26, 0.82)
U_{10}	-1.78 (-2.27, -1.20)	–	-1.52 (-2.11, -0.82)

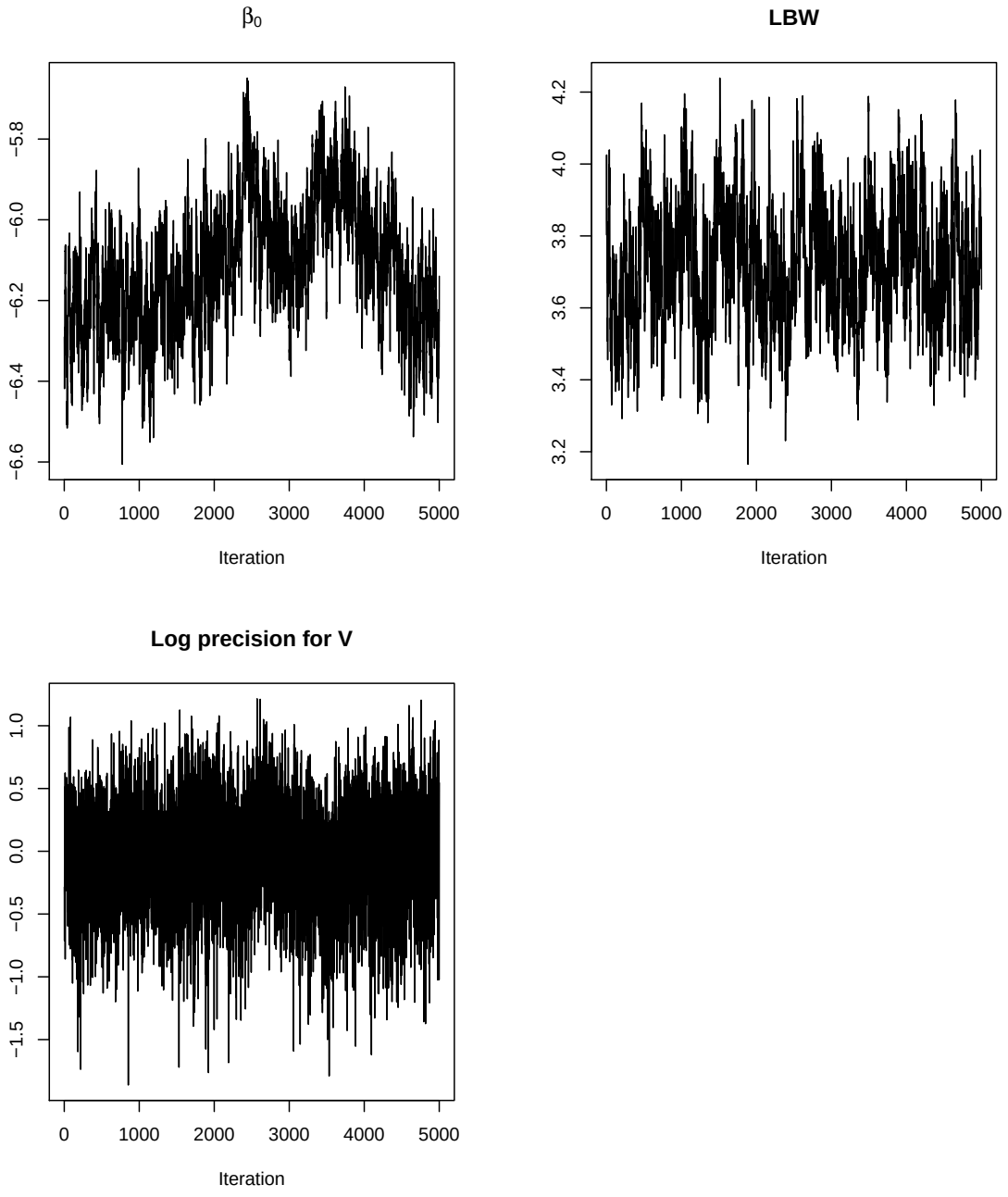


Figure B.29: Trace plots for the $\boldsymbol{\beta}$ parameters and $\log \tau_v$ in the non-spatial analysis of the simulated data example with strong spatial dependence.

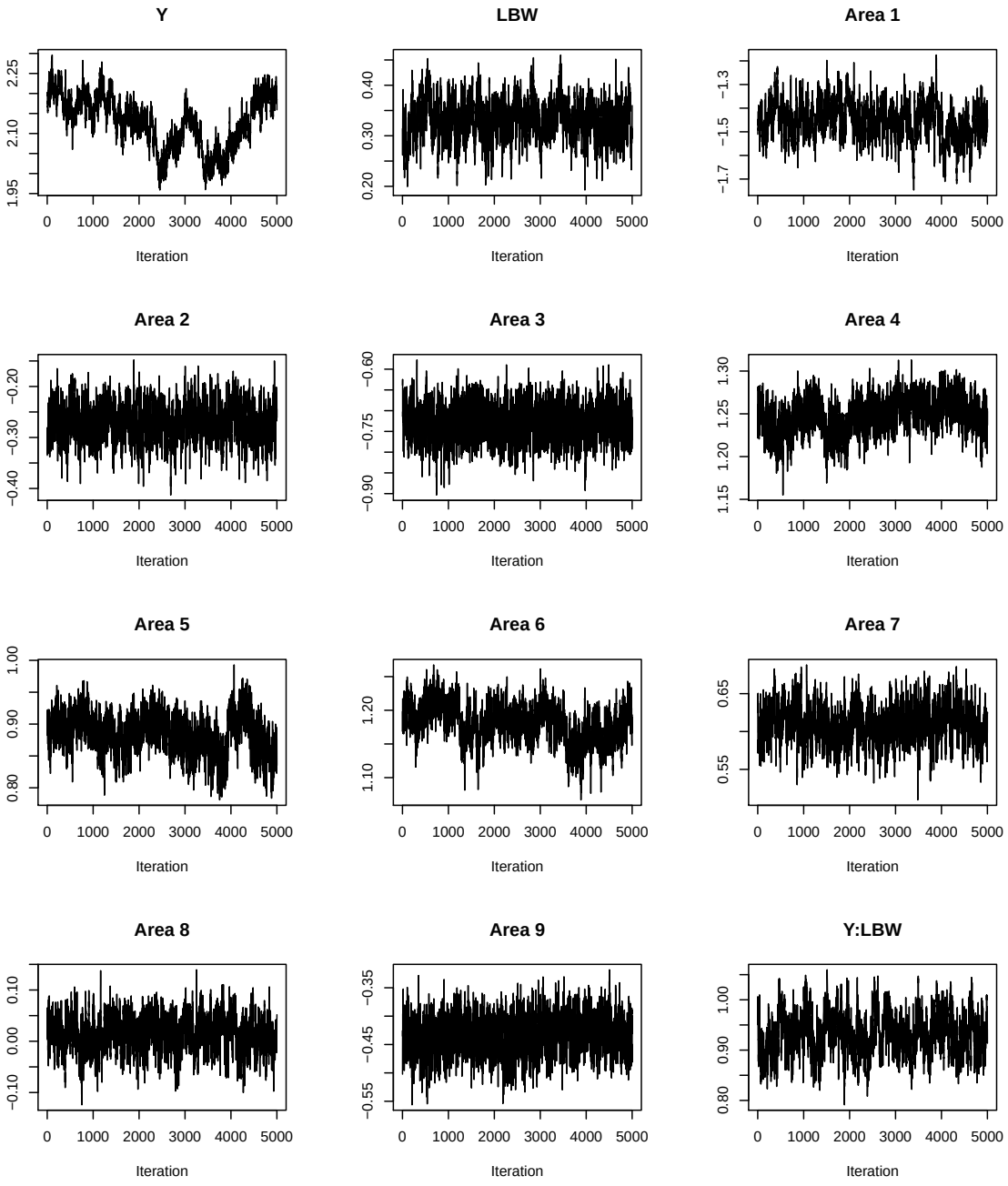


Figure B.30: Trace plots for the $\boldsymbol{\lambda}$ parameters in the non-spatial analysis of the simulated data example with strong spatial dependence.

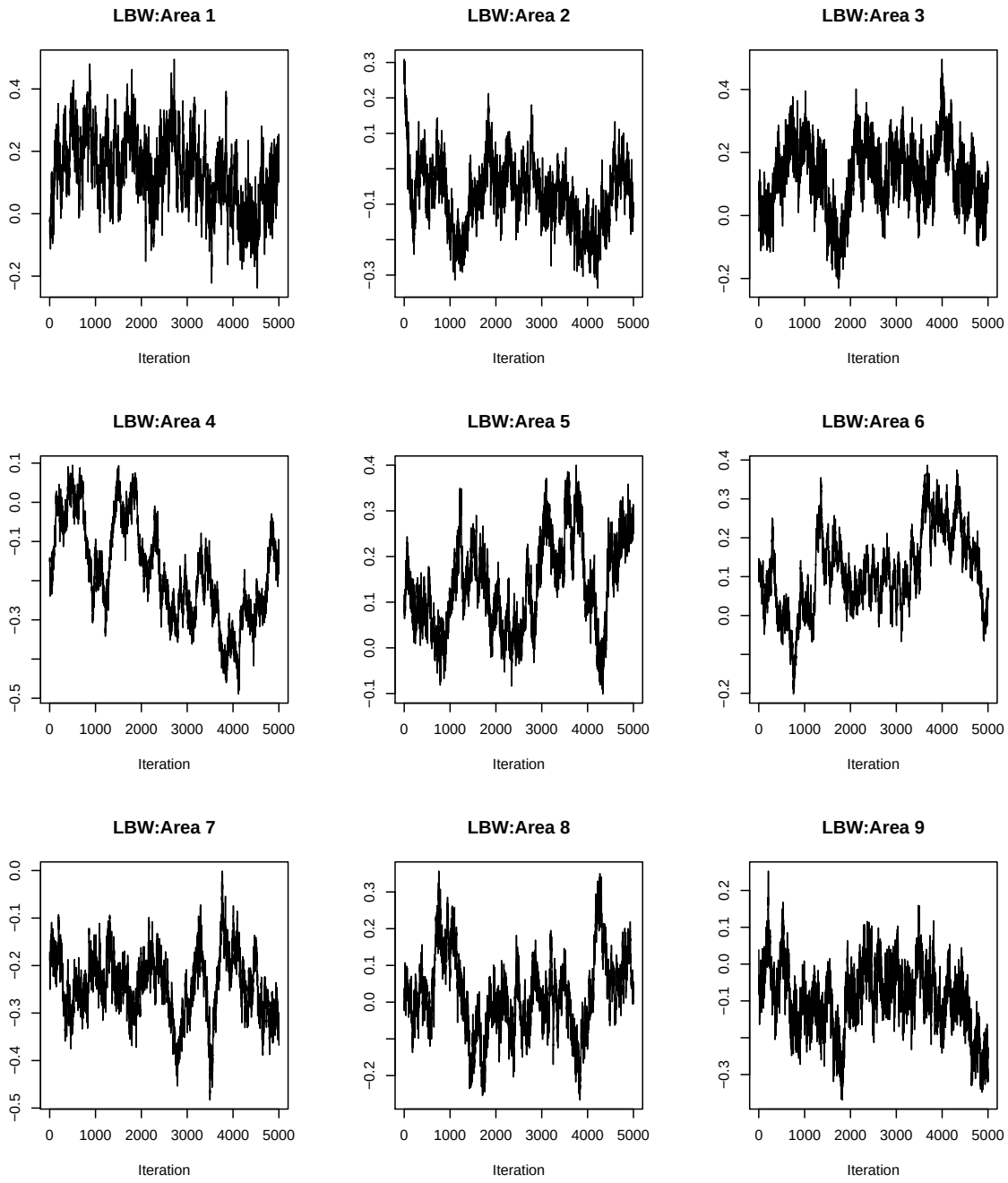


Figure B.31: Trace plots for the $\boldsymbol{\lambda}$ parameters in the non-spatial analysis of the simulated data example with strong spatial dependence (continued).

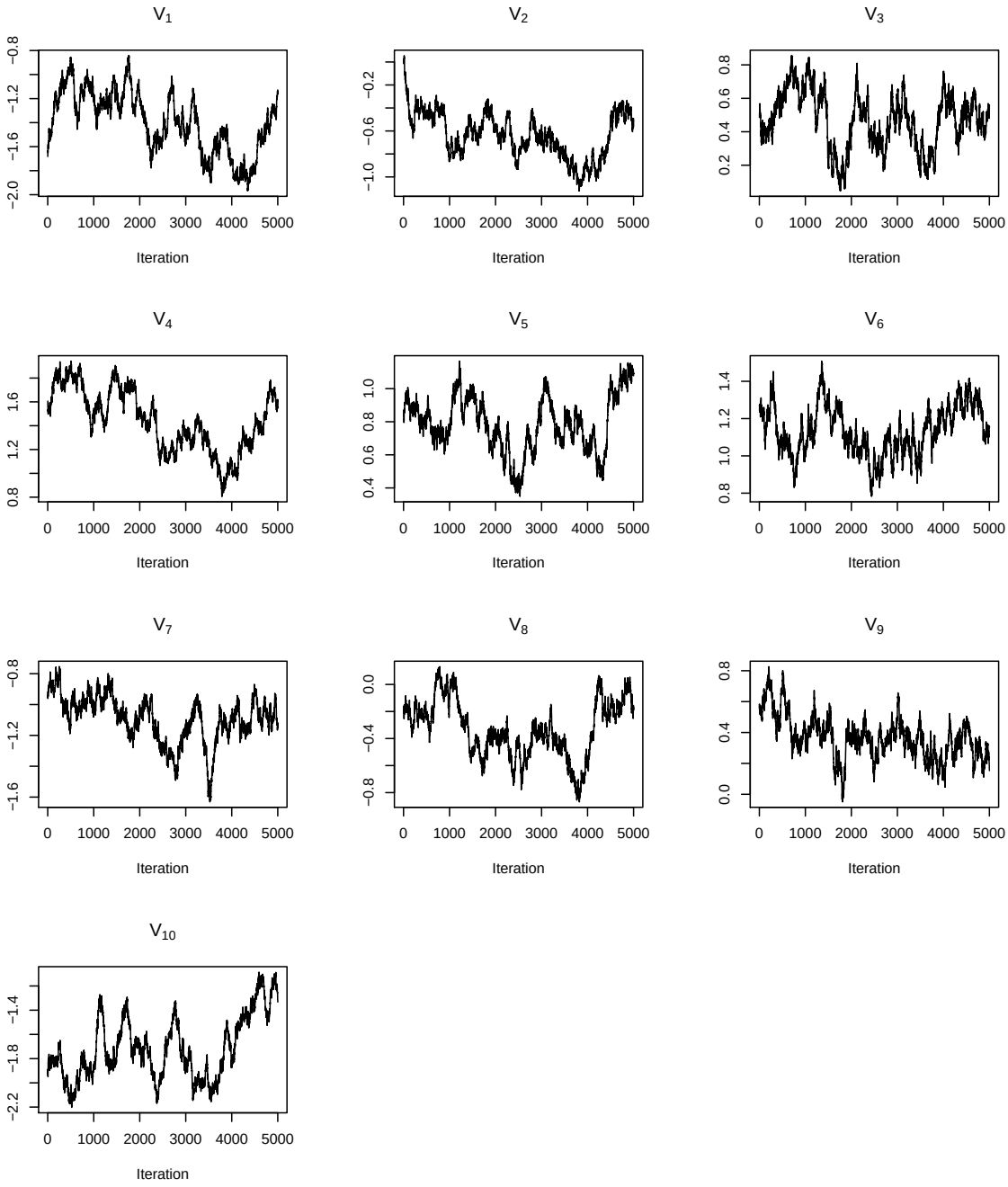


Figure B.32: Trace plots for the $\boldsymbol{V}$ parameters in the non-spatial analysis of the simulated data example with strong spatial dependence (continued).

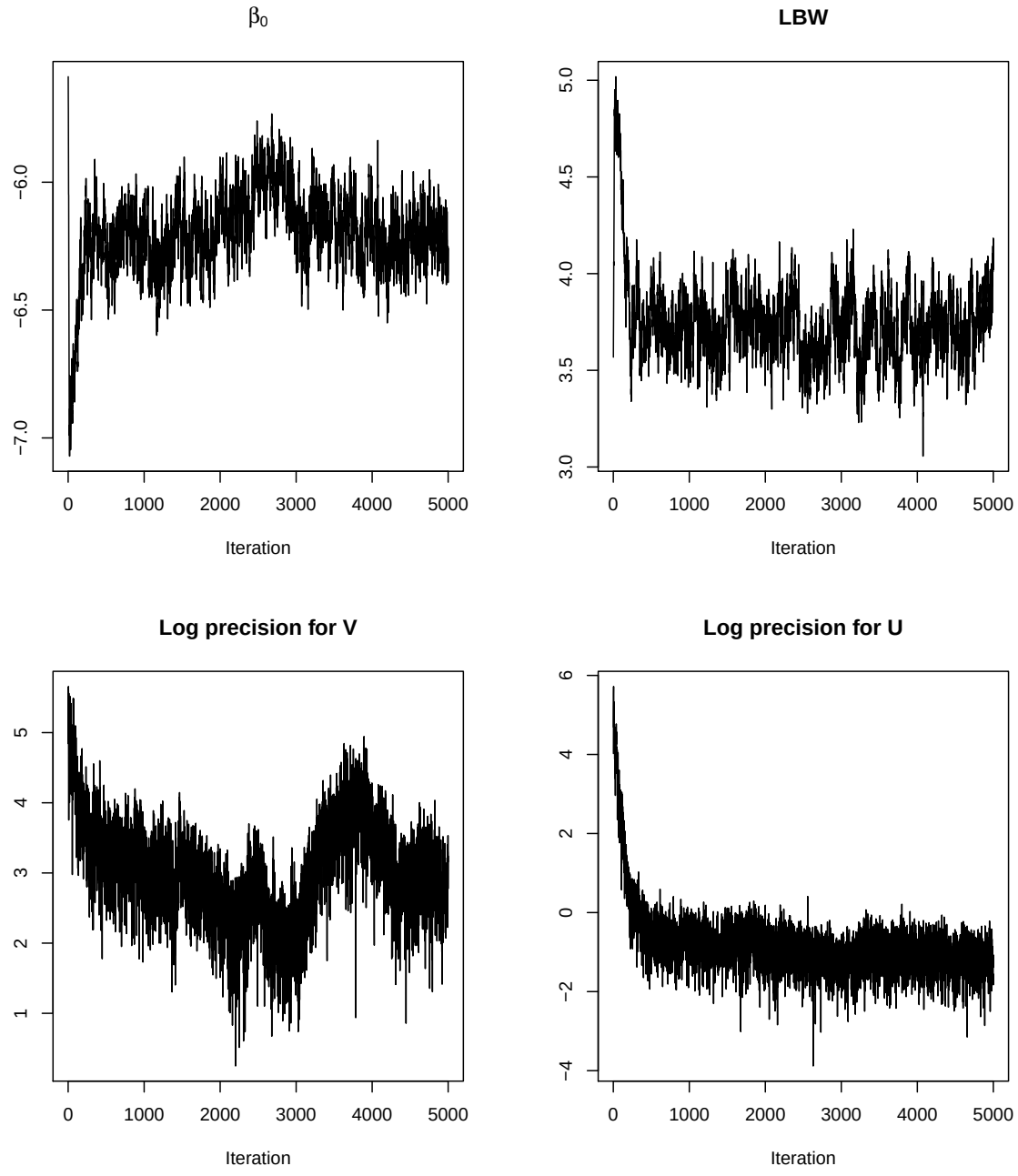


Figure B.33: Trace plots for the $\boldsymbol{\beta}$ parameters, as well as $\log \tau_v$ and $\log \tau_u$ in the spatial analysis of the simulated data example with strong spatial dependence.

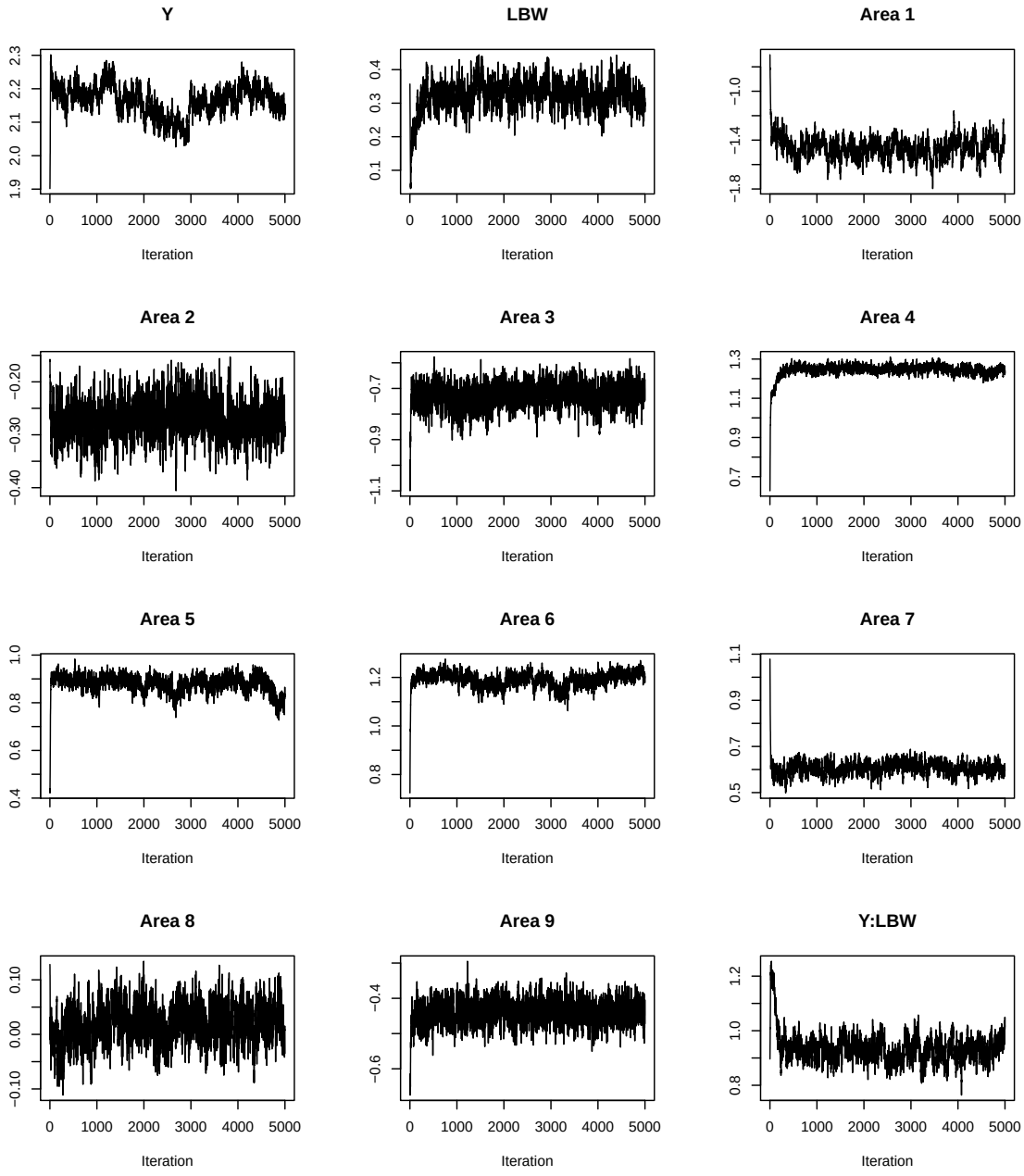


Figure B.34: Trace plots for the $\boldsymbol{\lambda}$ parameters in the spatial analysis of the simulated data example with strong spatial dependence.

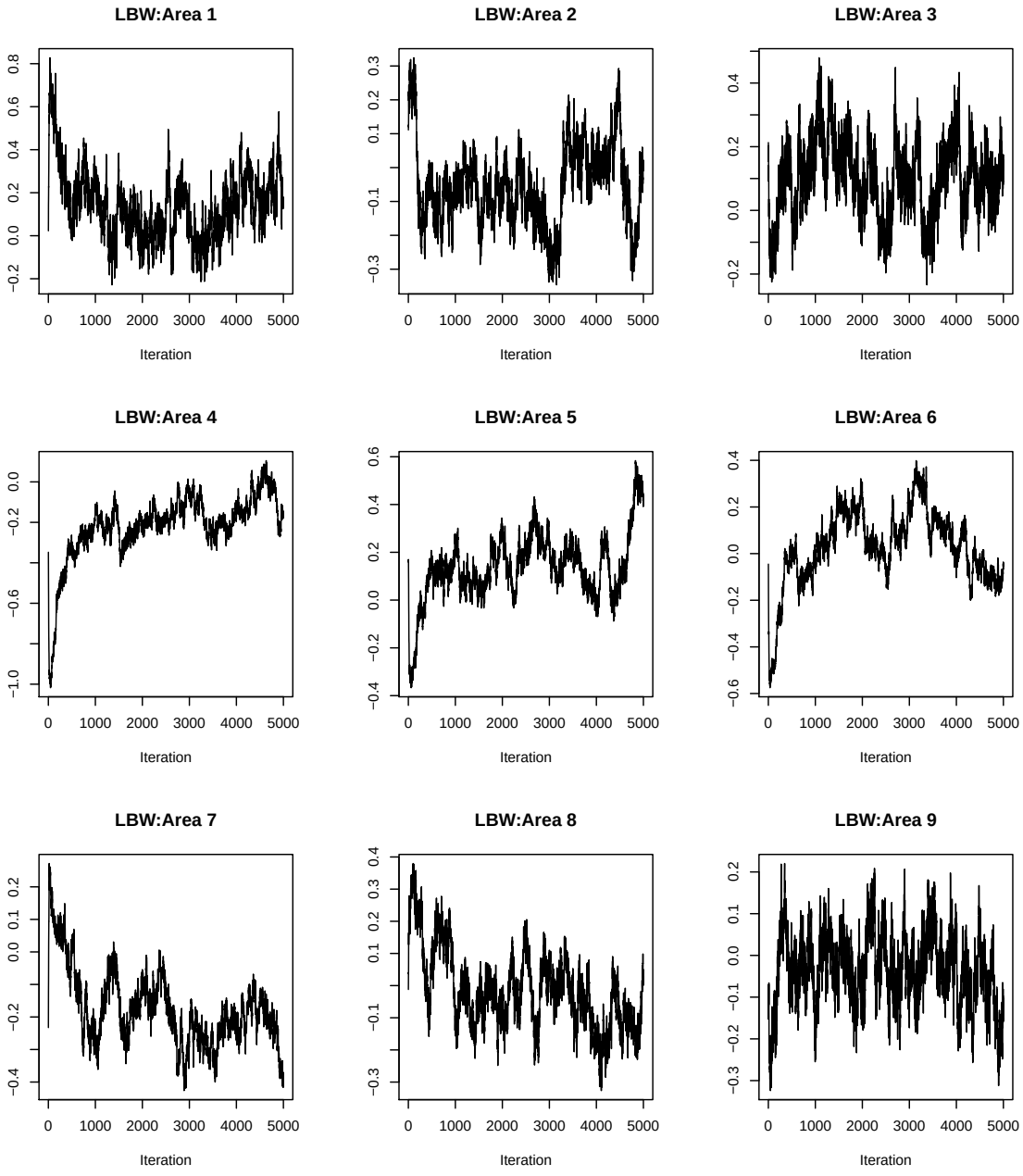


Figure B.35: Trace plots for the $\boldsymbol{\lambda}$ parameters in the spatial analysis of the simulated data example with strong spatial dependence (continued).

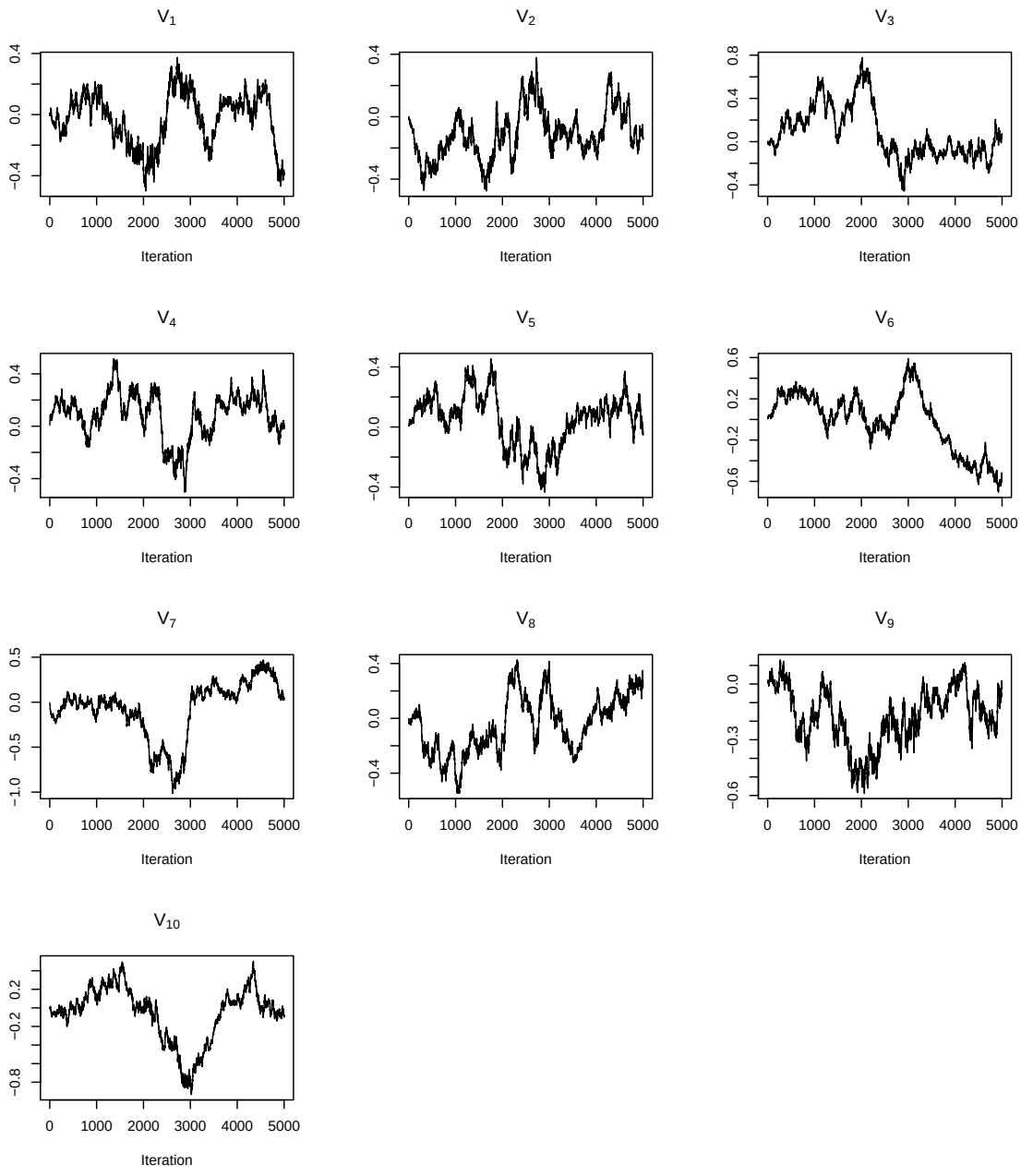


Figure B.36: Trace plots for the $\boldsymbol{V}$ parameters in the spatial analysis of the simulated data example with strong spatial dependence (continued).

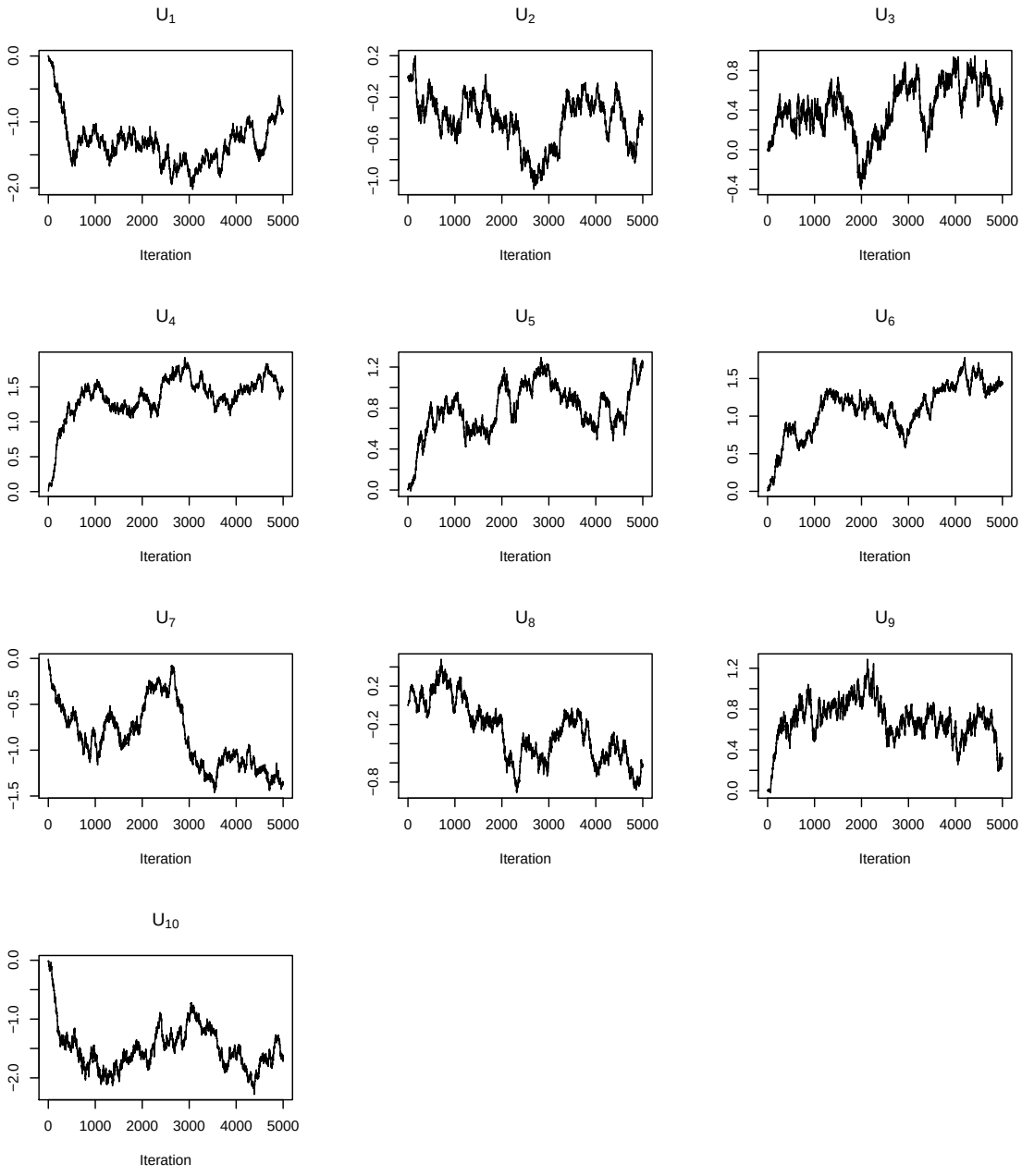


Figure B.37: Trace plots for the $\boldsymbol{U}$ parameters in the spatial analysis of the simulated data example with strong spatial dependence (continued).

B.5 Simulated Data with Strong Exposure-Confounder Relationship Example

Table B.7 displays the full data for the simulated data with strong exposure-confounder relationship discussed in Section 4.4.4.

Table B.7: Number of live and dead infants by North Carolina region, and low birth weight status in the simulated data example with strong exposure-confounder relationship: full data.

		Normal Birth Weight	Low Birth Weight	
Southern Mountains	Alive	20749	182	20931
	Dead	47	9	56
Central Mountains	Alive	27686	1653	29339
	Dead	56	140	196
Northern Mountains	Alive	20331	1256	21587
	Dead	45	83	128
Southern Foothills	Alive	119320	8536	127856
	Dead	274	632	906
Northern Foothills	Alive	66371	8062	74433
	Dead	139	601	740
Southern Heartland	Alive	107611	9498	117109
	Dead	256	694	950
Southern Coast	Alive	34071	7437	41508
	Dead	95	612	707
Northern Heartland	Alive	37646	2842	40488
	Dead	107	165	272
Central Coast	Alive	13248	2451	15699
	Dead	38	192	230
Northern Coast	Alive	4475	2183	6658
	Dead	13	194	207

The non-informative prior distribution on $\boldsymbol{\beta}$ implies a multivariate normal prior with

mean zero and variance-covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{C}^{-1}\right)^{T}$ to $\boldsymbol{\Lambda}^{Y}=(\boldsymbol{\lambda}^{Y},\boldsymbol{\lambda}^{YX})$ where

$$\begin{aligned}\boldsymbol{C} &= \begin{pmatrix} -2 & -2 \\ 0 & 4 \end{pmatrix} \\ \boldsymbol{\Sigma} &= \begin{pmatrix} 100^2 & 0 \\ 0 & 100^2 \end{pmatrix}.\end{aligned}$$

and assigned

$$\begin{aligned}\boldsymbol{\lambda}^{X} &\sim N(0,1) \\ \boldsymbol{\lambda}^{Z} &\sim N_9\left(\mathbf{0}, 10\left(I_9-\frac{1}{10}J_9\right)\right).\end{aligned}$$

In the analysis where $\boldsymbol{\lambda}^{XZ}$ is considered fixed, we assigned

$$\boldsymbol{\lambda}^{XZ}\sim N_9\left(\mathbf{0}, 5\left(I_9-\frac{1}{10}J_9\right)\left(I_9-\frac{1}{10}J_9\right)\right).$$

Figures B.38, B.39 and B.40 compare the prior and posterior medians and 95% intervals for σ_λ under the five assumed priors for τ_λ assuming small, medium and large phase II sample sizes, respectively.

Figures B.41 - B.44 display the trace plots for all parameters in the scenario where $\boldsymbol{\lambda}^{XZ}$ is considered random with hyperprior Gamma(1, 0.1215). The trace plots for all analyses in the small phase II samples sizes case appear similar. Parameters converged without difficulty for the analyses involving the medium and large phase II sample sizes.

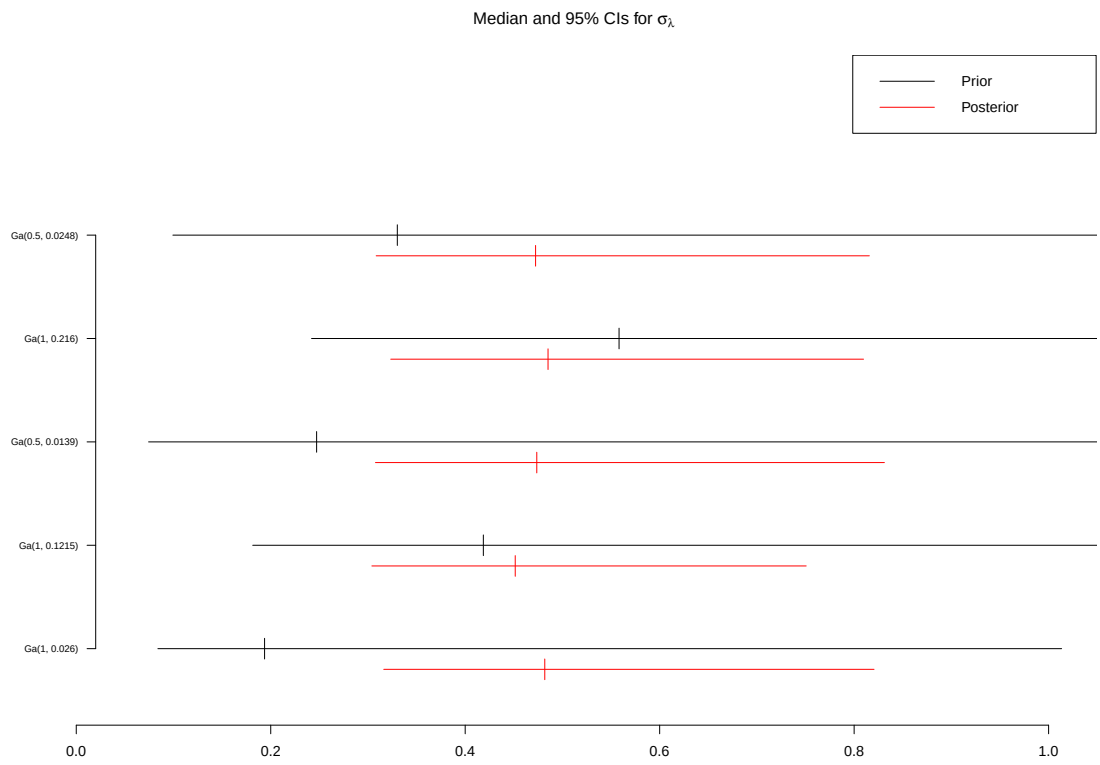


Figure B.38: Prior and posterior medians and 95% intervals for σ_λ under the five assumed priors for τ_λ assuming small phase II sample sizes.

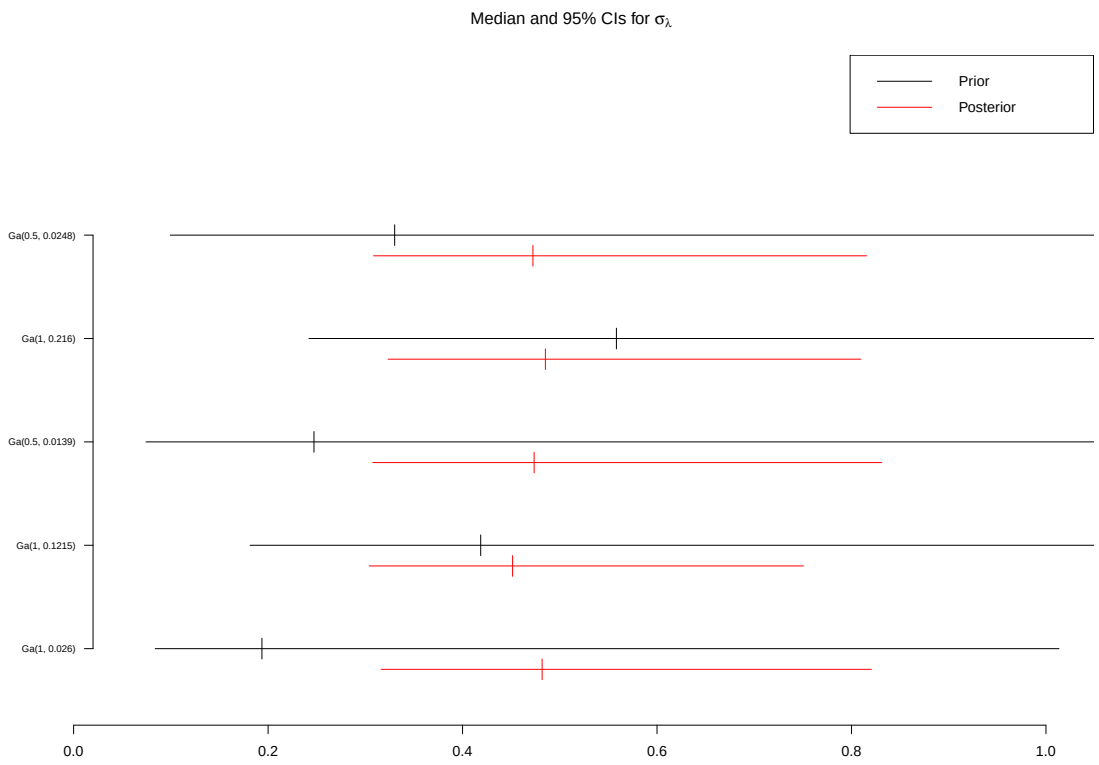


Figure B.39: Prior and posterior medians and 95% intervals for σ_λ under the five assumed priors for τ_λ assuming medium phase II sample sizes.

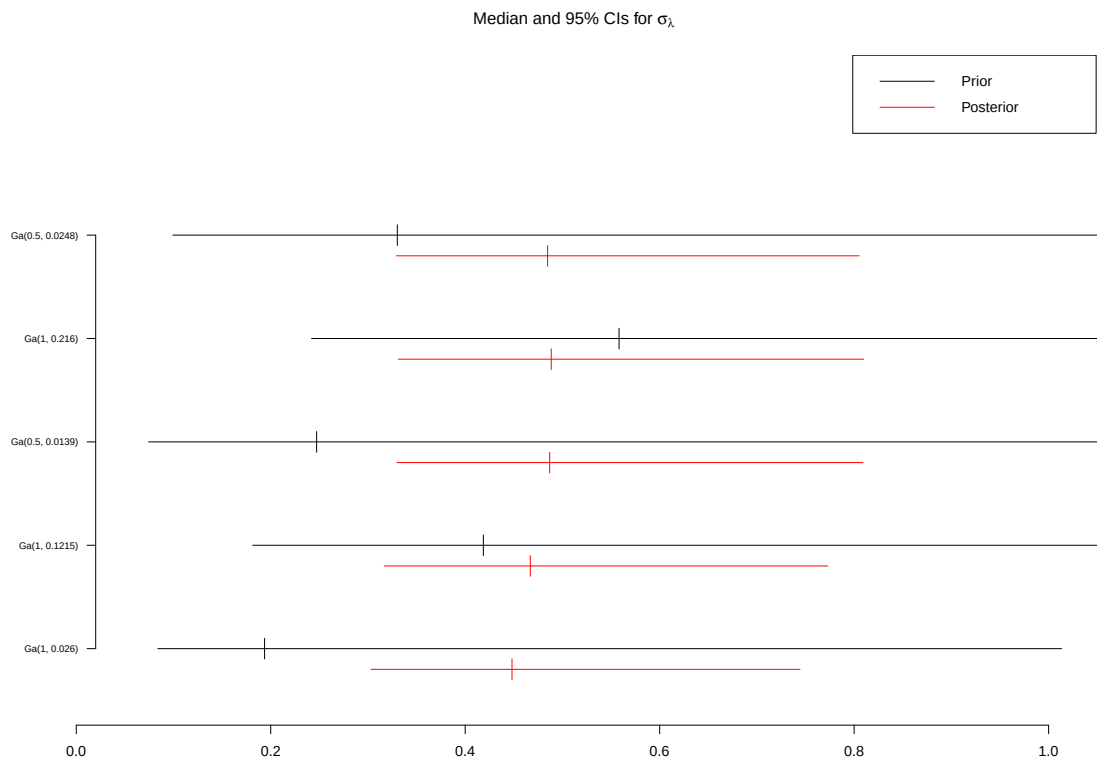


Figure B.40: Prior and posterior medians and 95% intervals for σ_λ under the five assumed priors for τ_λ assuming large phase II sample sizes.

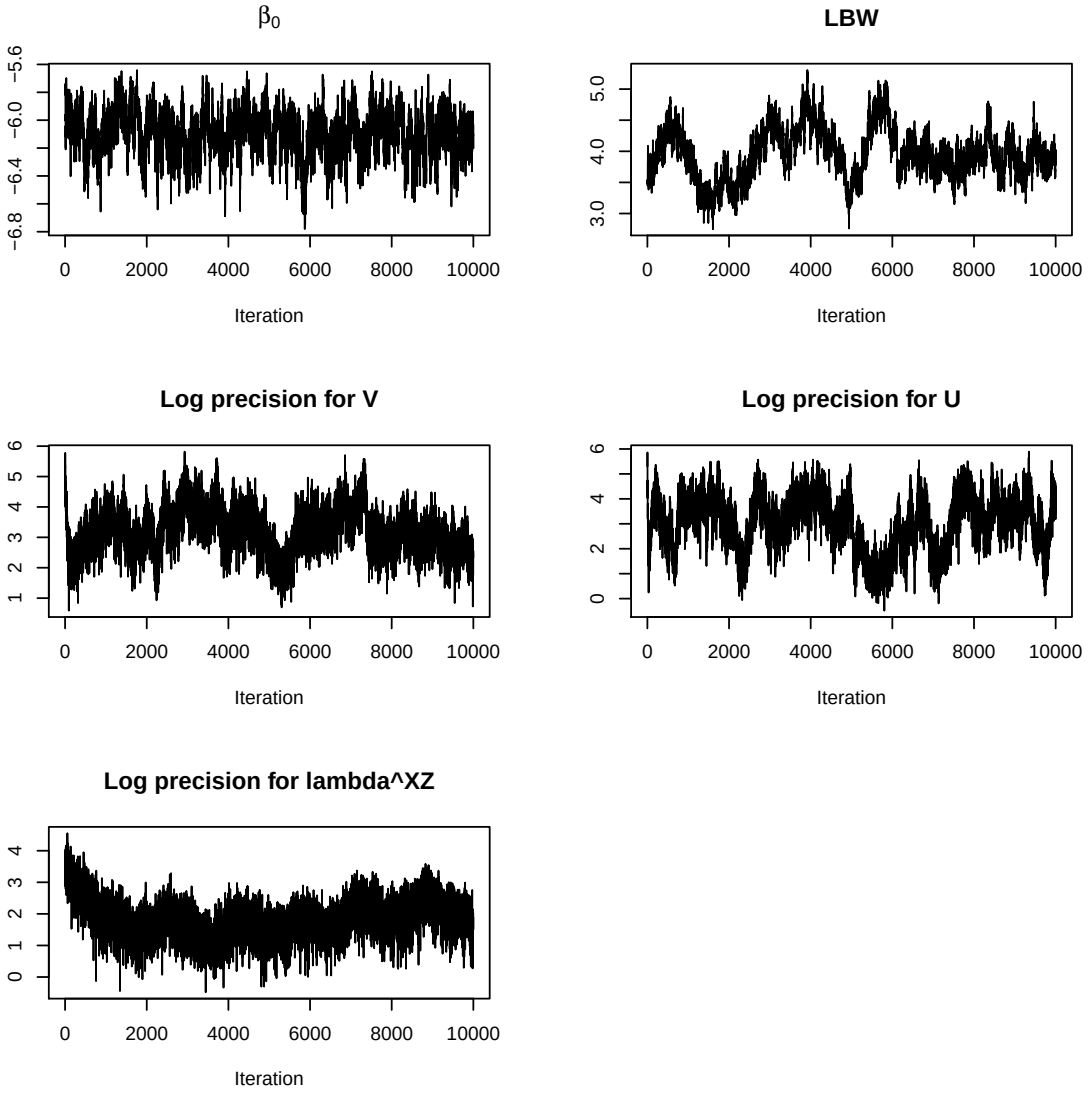


Figure B.41: Trace plots for the $\boldsymbol{\beta}$ parameters, as well as $\log \tau_v$, $\log \tau_u$ and $\log \tau_\lambda$ in the spatial analysis of the simulated data with strong exposure-confounder relationship treating $\boldsymbol{\lambda}^{XZ}$ as random for small phase II sample sizes.

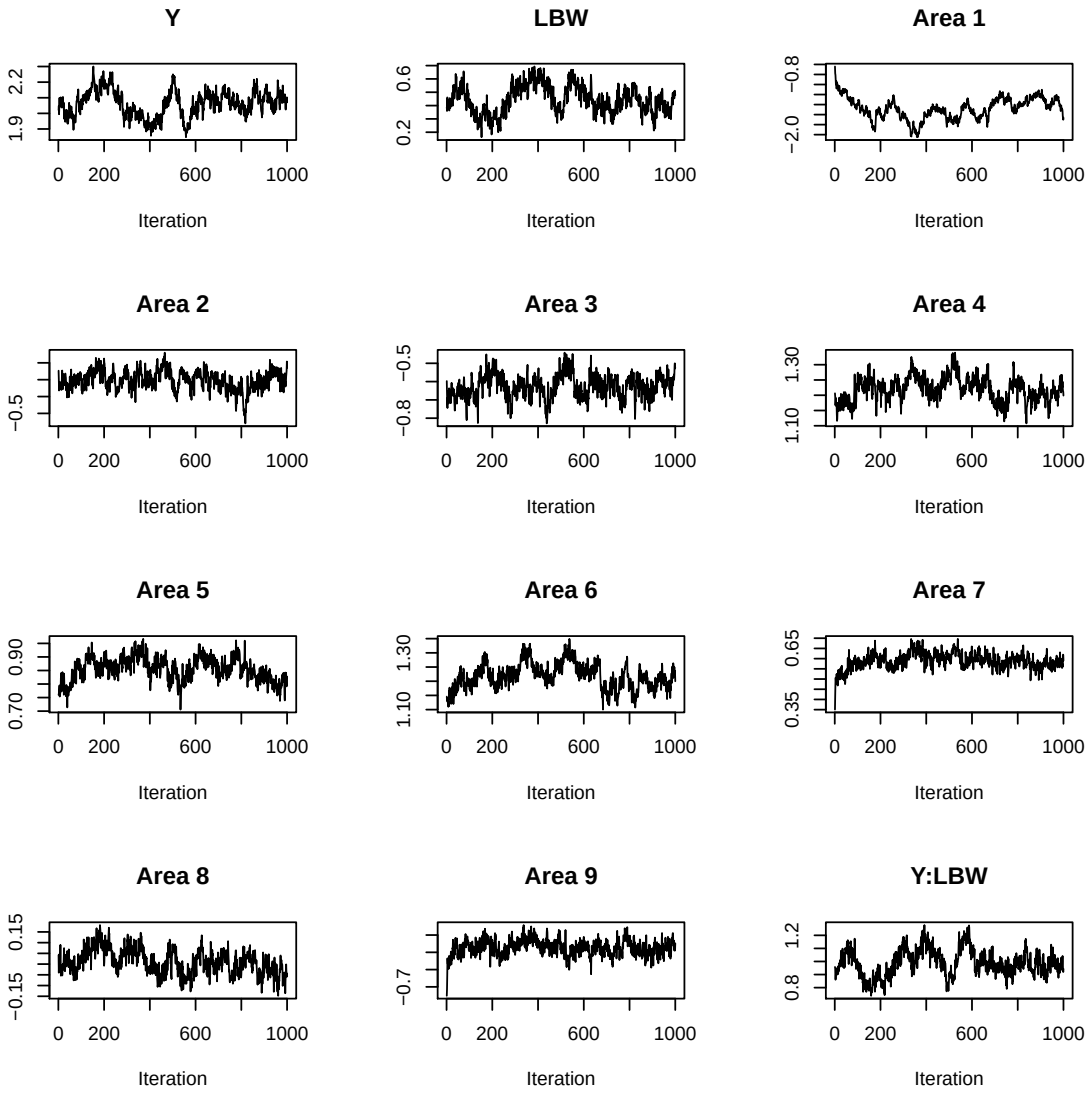


Figure B.42: Trace plots for the $\boldsymbol{\lambda}$ parameters in the spatial analysis of the simulated data with strong exposure-confounder relationship treating $\boldsymbol{\lambda}^{XZ}$ as random for small phase II sample sizes.

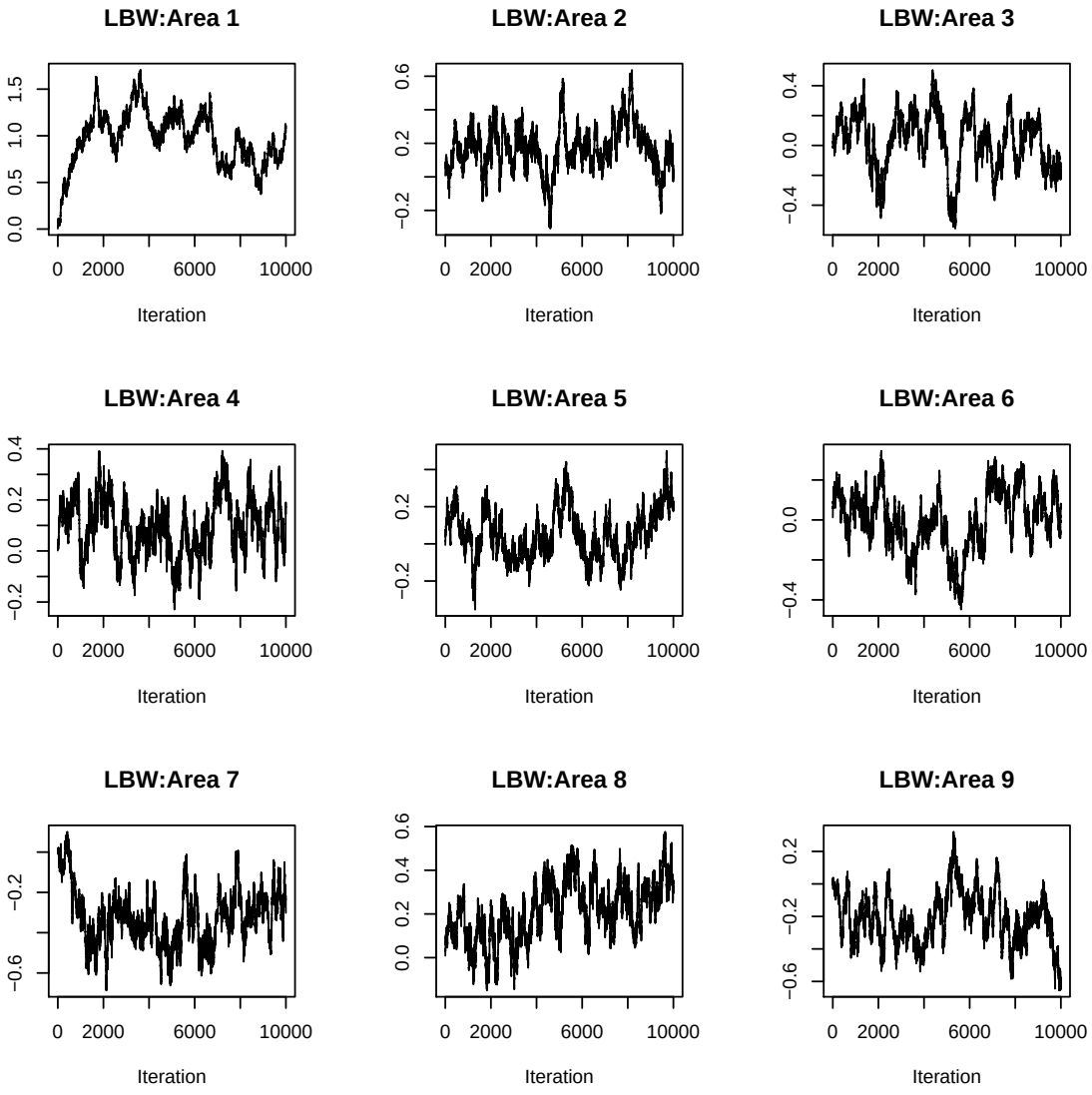


Figure B.43: Trace plots for the $\boldsymbol{\lambda}$ parameters in the spatial analysis of the simulated data with strong exposure-confounder relationship treating $\boldsymbol{\lambda}^{XZ}$ as random for small phase II sample sizes.

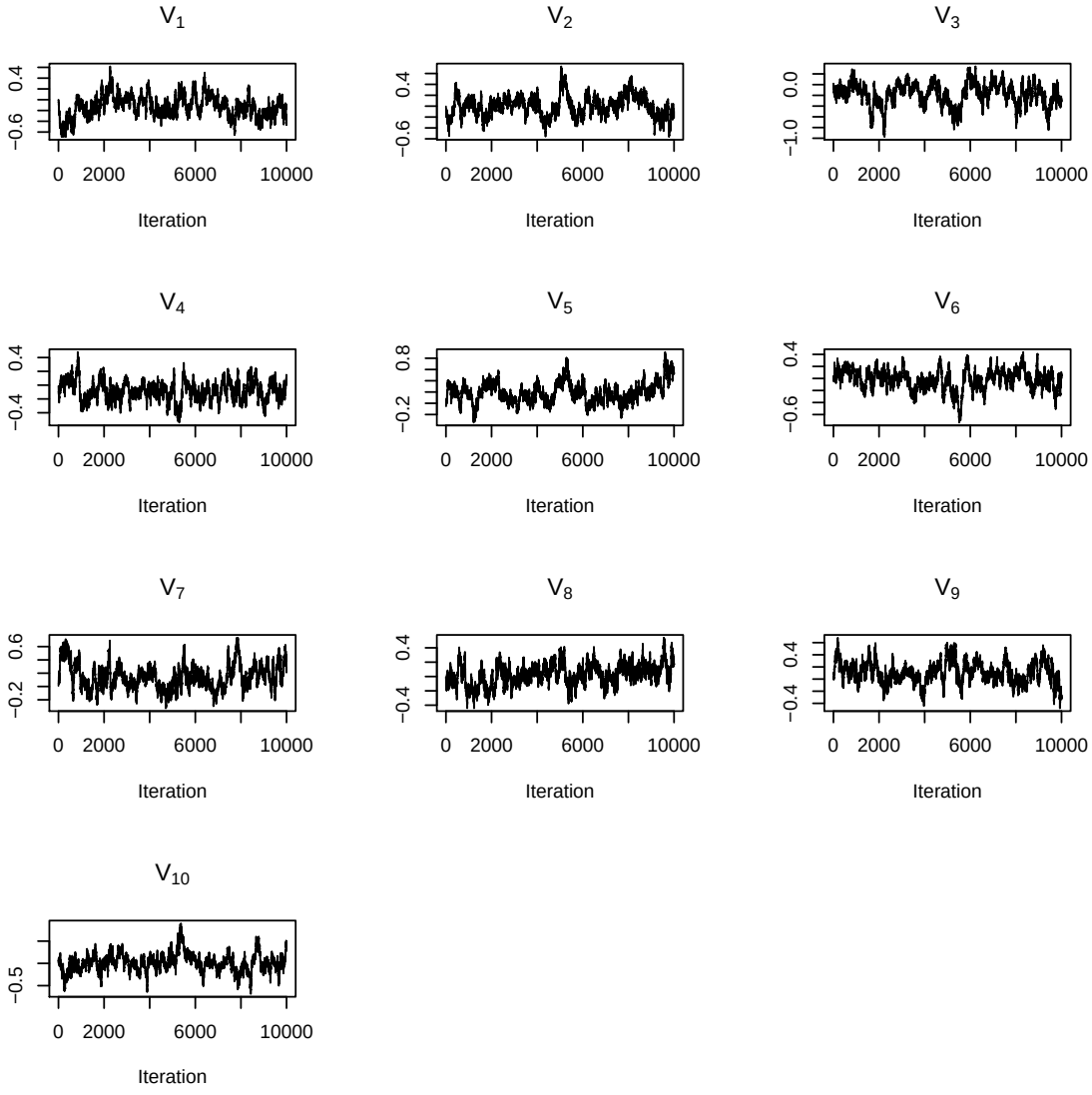


Figure B.44: Trace plots for the $\boldsymbol{V}$ parameters in the spatial analysis of the simulated data with strong exposure-confounder relationship treating $\boldsymbol{\lambda}^{XZ}$ as random for small phase II sample sizes.

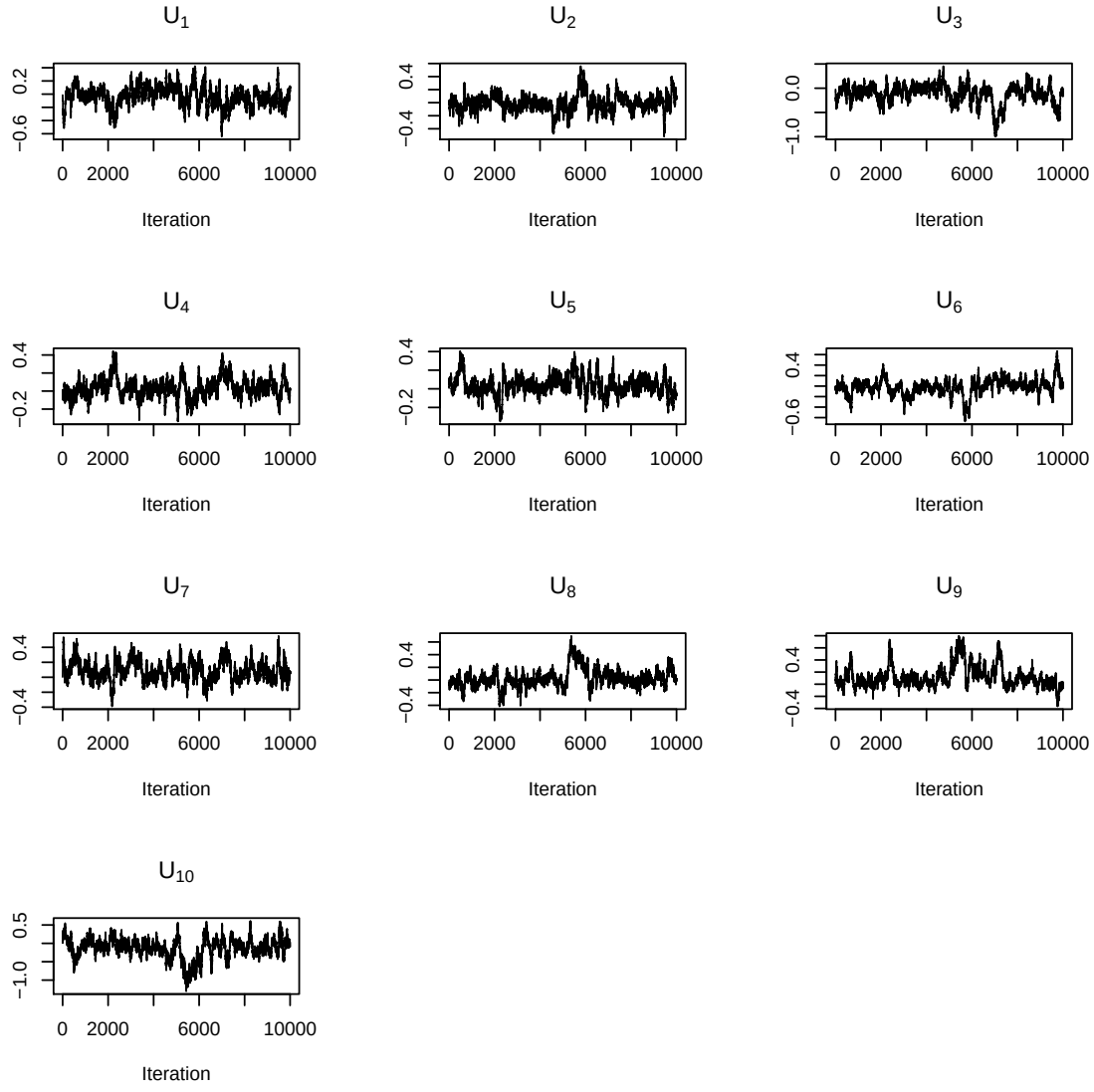


Figure B.45: Trace plots for the $\boldsymbol{U}$ parameters in the spatial analysis of the simulated data with strong exposure-confounder relationship treating $\boldsymbol{\lambda}^{XZ}$ as random for small phase II sample sizes.

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 5

C.1 Simulation Study

For the simulation study described in Section 5.4, we discuss the choice of hyperprior distributions for τ_v and τ_u. We follow the procedure outlined in Fong et al. (2010). We assume *a priori* that the residual relative risks lie in the interval $(0.2, 5)$ with probability 0.95 and assume $d = 2$ to give the Gamma$(1, 0.140)$ prior distribution for τ_v, using equation (2.14).

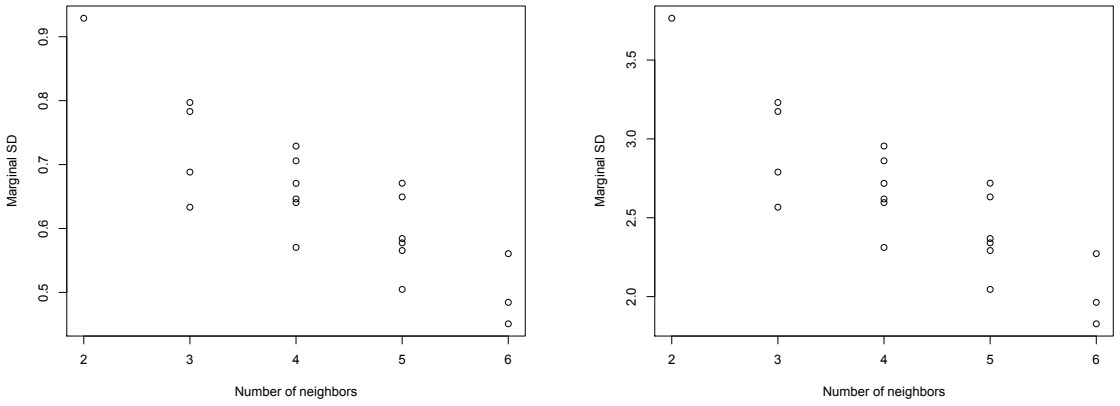
For τ_u, we begin by plotting the marginal variances assuming $\omega_u^2 = 1$, shown in Figure C.1(a). The mean of the marginal variances is 0.42, while the median of the inverse Gamma$(1, 0.140)$ assumed for τ_v is 0.20. Hence, we take the prior for τ_u to be Gamma$(1, 0.20/0.42)$. Figure C.1(b) displays the marginal variances under this prior. We notice that the variances are larger under this prior compared to assuming $\omega_u^2 = 1$, as desired. Figure C.2 maps simulations from the prior distributions of $\boldsymbol{V}$ and $\boldsymbol{U}$ for τ_v and τ_u set at the median of the prior distribution, and for τ_v and τ_u set at the 5% quantile of the prior distribution.

Table C.1 displays the bias, variance and MSE, as well as the number of sampled deaths, based on 100 simulations for each combination of sampling strategy and analytical model using larger survey sample sizes. In this case, we sample a total of $9,200$ individuals from each sampling design as follows:

1. **One-stage sampling from seven villages:** Individuals are randomly sampled from a seven randomly sampled villages and $1,314$ children are sampled from each village.

2. **One-stage sampling from all villages:** We sample an equal number of children from all 20 villages (460 from each village).

3. **HDSS with random sampling:** We select all children from the HDSS villages, and



(a) (b)

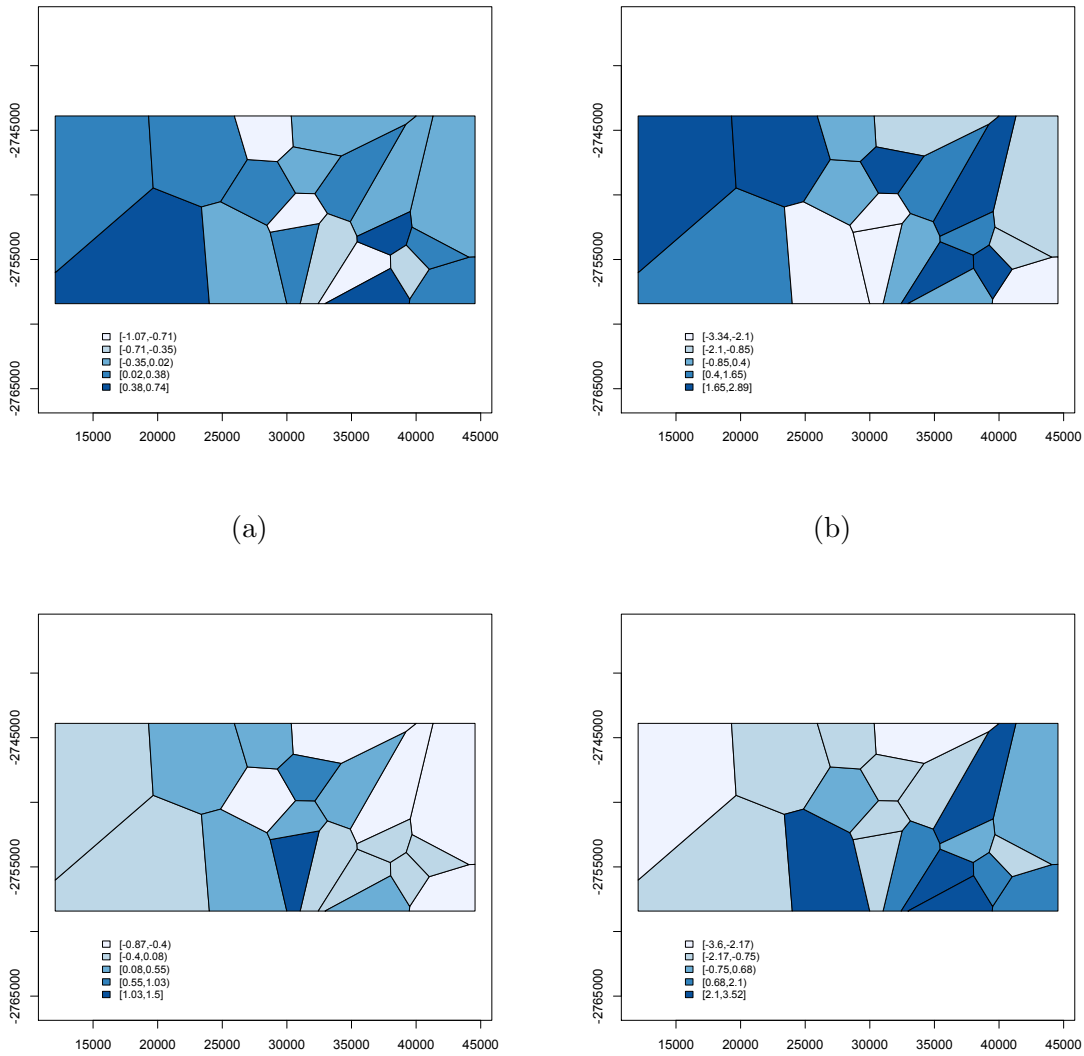
Figure C.1: Marginal standard deviations from an ICAR model with: (a) $\omega_u^2 = 1$, (b) $\tau_u \sim \text{Gamma}(1, 0.20/0.42)$.

a random sample of $5,000$ children from the 17 remaining non-HDSS villages, which results in 294 children being sampled from each remaining village.

4. **HDSS with informative sampling:** We employ the informative sampling procedure where all children from the HDSS villages are sampled, and a total of $5,000$ children are sampled from the remaining non-HDSS villages in the manner described in Section 5.2.

We sample far more deaths with the sampling schemes that sample all individuals from the HDSS villages plus a sample from the remaining villages than the one-stage sampling methods. Again, the MSE from the informed sampling design using a spatial logistic regression model produces results with substantially smaller MSE than any other sampling design and analytical model combination. In addition, the MSE based on the probabilities of death are also the smallest for the informative sampling using a spatial logistic regression model.

Table C.2 displays the bias, variance and MSE, as well as the number of sampled deaths, based on 100 simulations for each combination of sampling strategy and analytical model



(a) (b)

(c) (d)

Figure C.2: Simulations from the prior for: (a) V_i with τ_v set at the median of the prior, (b) V_i with τ_v set at the 5% quantile of the prior, (c) U_i with τ_u set at the median of the prior, (d) U_i with τ_u set at the 5% quantile of the prior.

Table C.1: Results from 100 simulations; 1,230 deaths. Model I=Naïve model; Model II=Age & sex model; Model III=Logistic regression covariate model; Model IV=Logistic regression covariate model with non-spatial and spatial random effects. It is not possible to fit the spatial model to the one-stage sampling from five villages plan since there are data from 5 villages only.

Sampling Design	Model	# Deaths	Y Based Estimates			p Based Estimates		
			Bias	Variance	MSE	Bias	Variance	MSE
One-Stage	I	385	107.1	6082.0	17554.6	0.470	0.012	0.233
Sampling from	II	385	100.6	5481.1	15598.3	0.427	0.018	0.200
Seven Villages	III	385	88.4	14005.4	21828.7	0.331	0.109	0.219
	IV	385	–	–	–	–	–	–
One-Stage	I	405	100.9	281.0	10454.0	0.467	0.000	0.218
Sampling from	II	405	102.7	602.3	11151.3	0.531	0.012	0.294
All Villages	III	405	75.1	444.4	6087.0	0.344	0.002	0.121
	IV	405	18.4	926.0	1265.1	0.101	0.007	0.018
HDSS with	I	611	124.8	154.0	15726.0	0.485	0.000	0.235
Random	II	611	162.6	271.4	26700.4	0.903	0.003	0.818
Sampling	III	611	98.5	216.0	9912.5	0.379	0.001	0.145
	IV	611	18.8	823.6	1175.8	0.095	0.006	0.015
HDSS with	I	672	144.9	152.7	21146.1	0.505	0.000	0.255
Informative	II	672	152.8	237.3	23583.2	0.729	0.002	0.534
Sampling	III	672	84.6	201.9	7356.1	0.378	0.001	0.144
	IV	672	17.3	637.2	935.9	0.102	0.005	0.015

using randomly selected villages as the HDSS villages. In this case, we again sample far more deaths with the sampling schemes that sample all individuals from the HDSS villages plus a sample from the remaining villages than the one-stage sampling methods. Again, the MSE from the informed sampling design using a spatial logistic regression model produces results with substantially smaller MSE than any other sampling design and analytical model combination.

Table C.2: Results from 100 simulations; 1,230 deaths. Model I=Naïve model; Model II=Age & sex model; Model III=Logistic regression covariate model; Model IV=Logistic regression covariate model with non-spatial and spatial random effects. It is not possible to fit the spatial model to the one-stage sampling from five villages plan since there are data from 5 villages only.

Sampling Design	Model	# Deaths	Y Based Estimates			p Based Estimates		
			Bias	Variance	MSE	Bias	Variance	MSE
One-Stage	I	229	126.1	5179.8	21069.2	0.467	0.019	0.237
Sampling from	II	229	117.2	4854.3	18585.9	0.427	0.028	0.211
Five Villages	III	229	103.4	41508.2	52192.7	0.340	0.316	0.432
	IV	229	–	–	–	–	–	–
One-Stage	I	226	154.5	217.3	24093.4	0.468	0.0003	0.219
Sampling from	II	226	158.9	1723.4	26961.2	0.533	0.045	0.329
All Villages	III	226	140.9	466.9	20330.6	0.343	0.005	0.123
	IV	226	127.3	1046.4	17239.9	0.110	0.014	0.026
HDSS with	I	282	129.7	60.6	16895.1	0.465	0.0001	0.216
Random	II	282	119.3	2090.9	16319.0	0.464	0.041	0.256
Sampling	III	282	123.2	415.4	15595.8	0.393	0.003	0.158
	IV	282	49.1	3249.6	5657.2	0.170	0.028	0.057
HDSS with	I	306	135.9	70.3	18541.4	0.469	0.0001	0.220
Informative	II	306	117.2	652.1	14384.5	0.429	0.012	0.196
Sampling	III	306	112.2	218.5	12806.5	0.375	0.001	0.142
	IV	306	58.6	2026.6	5464.1	0.206	0.017	0.059

VITA

Michelle Ross was born on April 22, 1983 in Montreal, Canada to Shirley and Brian Ross. In 2006 she graduated from McGill University, Montreal, Canada with a BSc (Honours) in Statistics. In 2007 she went on to earn an MSc in Statistics from also from McGill University.