**Distribution Agreement**

Signature:

____________________________ ______________

Zhiheng Xu Date

Bayesian Space-time Analysis in Carcinogenesis

By

Zhiheng Xu

Doctor of Philosophy

Biostatistics

Vicki Hertzberg
Advisor

Brani Vidakovic
Committee Member

Tianwei Yu
Committee Member

Lance Waller
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

Bayesian Space-time Analysis in Carcinogenesis

By

Zhiheng Xu

M.S., Georgia State University, 2004

M.S., Georgia Institute of Technology, 2002

B.S., Tianjin University, 1997

Advisor: Vicki Hertzberg, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

# Abstract

## Bayesian Space-time Analysis in Carcinogenesis

By Zhiheng Xu

Although the etiology of cancer remains under investigation, evidence has suggested that multiple events occur during carcinogenesis, the process of the transformation of normal cells into cancer cells. Statistical modeling of carcinogenesis has been used to study the cancer formation and cancer risk assessment. In this dissertation, I present three studies involving carcinogenesis models in estimating cancer mortality rates.

First, I develop a Bayesian Armitage-Doll multistage carcinogenesis model. The Armitage-Doll multistage model has been successfully employed in many carcinogenesis studies due to its simplicity in predicting cancer mortality rate. The model provides estimates of different numbers of stages for various types of cancer. This research is the first effort to use an alternative Bayesian approach in the Armitage-Doll multistage model. Different likelihoods and prior settings are discussed and sensitivity analysis and model assessment show that the Bayesian Armitage-Doll model fits the cancer mortality data well.

Second, two carcinogenesis models, the Armitage-Doll multistage model and the Moolgavkar-Venzon-Knudson Two-stage Clonal Expansion (TSCE) model, are used in updating the age-period-cohort (APC) model to target issues of lack of sound biological explanation and identifiability problems. I develop a Bayesian extended APC model where non-specific age effects are replaced by the hazard functions derived from multi-stage carcinogenesis models. The Bayesian extended APC model is applied to study colon cancer mortality rates in the US achieving high consistency between the estimated rates and observed rates for older age groups ($\geq 45$). In addition, model comparisons show that the Bayesian extended APC model can be used to replace the conventional APC model without increasing the deviance information criterion (DIC) values while providing a more sound biological meaning to the model.

Third, I further apply the Bayesian extended APC model to study the spatio-temporal variation in cancer mortality rates. Both the Armitage-Doll and TSCE carcinogenesis model are also used in the Area-APC model to replace the main age effect. The county level lung and colon cancer mortality data in Iowa are used as examples. The study shows the Bayesian extended AAPC model with area-cohort interaction and Armitage-Doll age effects achieved the lowest DIC values and good convergency among all models. The Bayesian extended AAPC model can be used to study spatial-temporal patterns of cancer mortality with strong biological prior beliefs in the age effects.

In summary, my dissertation focuses on developing carcinogenesis analytic approaches using Bayesian methods. The three studies show that carcinogenesis model can be used to study the relationship between cancer mortality rate and spatial and temporal effects from the underlying disease process.

# Bayesian Space-time Analysis in Carcinogenesis

By

Zhiheng Xu

M.S., Georgia State University, 2004

M.S., Georgia Institute of Technology, 2002

B.S., Tianjin University, 1997

Advisor: Vicki Hertzberg, PhD

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2011

## Acknowledgments

I would like to acknowledge and extend my heartfelt gratitude to the following persons who have helped me during my doctor study and provided enormous support and guidance toward the completion of this dissertation.

First and foremost I would like to thank my advisor Dr. Vicki Hertzberg for instilling in me the qualities of being a good scientist. I have been fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to hold my research to a high standard.

Second, I would like to thank my committee members Drs. Lance Waller, Tianwei Yu, and Brani Vidakovic for their insightful comments and constructive criticism at different stages of my research.

I owe special thanks to Dr. Michael Kutner, whose trust and encouragement have motivated me continuously to overcome setbacks and stay focused on my graduate study. I am also thankful Dr. Andre Rogatko who inspired me to pursue a Ph.D. in Biostatistics while I was working at Winship Cancer Institute. In addition, I acknowledge the financial support from Winship Cancer Institute during my graduate study.

I would like to give my gratitude to our department's students, faculty and staff. It is the best department you can ask for, talented students, knowledgeable faculty, and responsible staff. Thank you for providing me such a wonderful environment.

Finally I would like to thank my wife, who has given me both emotional and intellectual support during the whole process of dissertation research. She was always there to cheer me up and encourage me during the good and bad time. Without her, it is impossible to finish this project.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Cancer, the second most common cause of death in the United States, has become a critical public health concern since the last decade. The annual number of cancer deaths estimated by American Cancer Society is 292,540 for men and 269,800 for women which accounts for one quarter of total US mortality in 2009 [ACS, 2009]. Furthermore, among the three principle death causes in the United States, cancer mortality has consistently increased since 1950 while the mortalities from heart disease and cerebrovascular diseases have decreased more than 50%. The lifetime probability of developing cancer is 1 in 2 in men and 1 in 3 in women. The cancer statistics are dire, thus the need to improve cancer prevention and treatment is increasingly urgent. Fortunately, considerable amounts of money, effort, and resources have been allocated to cancer research from the federal and state governments and non-profit organizations since 1970s.

Quantitative methods have been used to model incidence, progression, and mortality of cancer. Much evidence has supported the assumption that multiple events are involved in carcinogenesis, the process that describes how normal cells are transformed into cancer cells. It is widely recognized that the multistage random process in carcinogenesis includes genetic changes and stochastic proliferation and differentiation of normal stem cells and genetically altered stem cells. Molecular biologists have discovered that a series of irreversible genetic changes have occurred on a single stem cell before developing into a tumor cell through stochastic proliferation and differentiation. In addition, the number of stages and pathways of the carcinogenesis process are significantly influenced by environmental factors underlying the individual. Among many mathematical models formulated in the structure of multistage stochastic process in carcinogenesis in the past 60 years, the Armitage-Doll model [Armitage and Doll, 1954] and the Moolgavkar-Venzon-Knudson Two-stage Clonal Expansion (TSCE) Carcinogenesis Model [Moolgavkar and Venzon, 1979, Moolgavkar and Knudson, 1981, Moolgavkar and Luebeck, 1990] have made influential impacts on cancer researchers, and have been widely applied to analyze cancer incidence and mortality rates.

In addition to multistage carcinogenesis model in analyzing cancer trends, a three-factor multiplicative model, age-period-cohort (APC) model, has gained much attention among statisticians and epidemiologists studying the separate effects and trends due to age, period and cohort for cancer incidence and mortality rates. Periods effects such as cancer screening technologies, have played important roles in cancer prevention and control. The occurrence of medical milestones such as mammograph has significantly reduced the breast cancer incidence and mortality in the past decades. Furthermore, cohort effects include both factors that occurred at the year of birth and those that affect disease rates related to year of birth. Therefore, period and cohort effects should be considered in the model along with age effect in fitting cancer incidence and mortality data. However, it is well known that there is a non-identifiability problem associated with all three factors (age, period and cohort) because of the exact linear relationship among them [Holford, 1983, 1991]. La Vecchia, et.al. [1998] also pointed out the random variation in the classical estimates of age, period and cohort effects. In order to overcome the nonidentifiability issues and random variations in APC model, Bayesian approaches have been used in fitting breast cancer incidence [Breslow and Clayton, 1993] and lung cancer mortality rates [Berzuini and Clayton, 1994] where *a priori* beliefs to smooth the temporal effects are incorporated and model constraints are added as well.

In this dissertation, I apply Bayesian approaches to the Armitage-Doll multistage carcinogenesis model to improve the precision in predicting cancer incidence and mortality rates. The Armitage-Doll multistage model has been successfully employed in carcinogenesis studies due to its simplicity in predicting cancer mortality rate. It estimates different numbers of stages for various types of cancer. This research is the first effort to use an alternative Bayesian approach in the Armitage-Doll multistage model. This carcinogenesis model describes a typical underlying disease process in which normal cells are transformed into cancer cells and age is a deterministic factor in the model. Therefore, I introduce the carcinogenesis model into an APC model to better explain the relationship between cancer mortality and temporal effects from the underlying disease process. Spatial-temporal pattern of cancer mortality rates are also studied in the area-age-period-cohort (AAPC) model.

In Chapter 2, I introduce the fundamental biological basis of cancer development and the concept of multistage carcinogenesis, focusing on the Armitage-Doll multistage model and the TSCE model through the literature review. The APC model and the Area-APC (AAPC) model are discussed in Chapter 2 as well. In Chapter 3, I develop the Bayesian Armitage-Doll multistage carcinogenesis model to derive posterior estimates with greater precision from the combined information of prior and likelihood. Different likelihoods and prior settings are discussed in this chapter. In addition, sensitivity analysis and model assessment conducted here show that the Bayesian Armitage-Doll model fits the cancer mortality data well. In Chapter 4, I develop the Bayesian extended APC model in which non-specific age effects are replaced by hazard functions derived from multi-stage carcinogenesis models. Autoregressive Gaussian priors are assigned to period and cohort effects while priors for carcinogenesis parameters are specified too. The proposed model is believed to help address the issue of lack of sound biological explanation and identifiability problems. In Chapter 5, Bayesian extended AAPC is described and applied to the study of spatio-temporal mappings of disease rates. Lastly, I present summary discussion and the direction for future works in chapter 6.

# Chapter 2

# Literature Review

## 2.1 Mechanisms of Carcinogenesis

Cancer is a term used for disease in which abnormal cells divide without control and are able to invade other tissues. In most cases, these cancer cells form a tumor. A benign tumor can proliferate but cannot grow into other tissues. In contrast, a malignant tumor can spread to other parts of body through the blood and lymph systems which is defined as metastasis. No matter where a cancer may spread, it is always named for the place where it started. For example, breast cancer that has spread to the liver is still called breast cancer, not liver cancer. Likewise, prostate cancer that has spread to the bone is metastatic prostate cancer, not bone cancer.

The mechanism of cancer development is still unclear. However, most of cancer researchers strongly believe there are multiple events involved in carcinogenesis, the process by which normal cells are transformed into cancer cells [Armitage and Doll, 1954, Moolgavkar and Venzon, 1979, Moolgavkar and Knudson, 1981, Moolgavkar and Luebeck, 1990]. Research shows that several genomic mutations occurred in the cells, such as the activation of oncogenes and inactivation of tumor suppressor genes, which demonstrated the multistage nature of carcinogenesis [Bishop, 1991, Fearon and Vogelstein, 1990]. Inheritable genetic alterations for neoplastic transformation also accounts for the carcinogenic process [Bishop, 1991, Fearon and Vogelstein, 1990].

Three consecutive phases are assumed in the carcinogenesis: initiation, promotion and progression. The first step is cell initiation, where normal stem cells are initiated to start to divide and expand slowly so that they can form the tumor. The substances that triggers the cell initiation is called initiator, for example, radiation, chemical agents or a virus. In the promotion phase, with the help of promoter, an initiated cell can expand clonally and reproduce a population of initiated cells. A promoter is a substance that stimulates the growth of initiated cells. During tumor progression, initiated cells can convert to malignant cancer cells via additional genetic alterations. The fast growth of malignant cancer cells can outpace cell apoptosis and enable cancer cells to invade other organs (metastasis) and build up their own vessel system for nutrition (angiogenesis). Aggressive metastatic tumors can kill their hosts quickly.

Stochastic models for carcinogenesis have been developed in the last 50 years to predict the risk of cancer. The process of carcinogenesis is widely views as normal cells deterioration in a number of stages to malignancy through initiation, promotion and progression [Moolgavkar and Venzon, 1979]. The Armitage-Doll multistage carcinogenesis model [Armitage and Doll, 1954] and the Moolgavkar-Venzon-Knudson Two-stage Clonal Expansion (TSCE) Carcinogenesis Model [Moolgavkar and Venzon, 1979, Moolgavkar and Knudson, 1981, Moolgavkar and Luebeck, 1990] will be introduced in the next section.

## 2.2 Carcinogenesis Models

### 2.2.1 Armitage-Doll Multistage Model

In the early 1950s, Armitage and Doll first proposed the multistage model of carcinogenesis which described the quantitative relationship between cancer mortality and age in industrialized nations [Armitage and Doll, 1954]. A stochastic model was used to describe the development of a malignant cancer from a normal cell as a finite number of stages on transitions. They found that the mortality rate ($r$) is proportional to age at death ($t$) raised to a power that is one less than the number of stages ($s$) between normal health and death, i.e., $r = \alpha \times t^{s-1}$, where $\alpha$ is a proportionality constant. The parameter $\alpha$ is affected by various factors such as gender, race, diet, genetic, environmental factors. This relationship is a direct consequence of the properties of a time-homogeneous birth process, the mathematical theory underpinning the multistage model. By taking the logarithm of this relationship we can write $\log(r) = \theta + (s-1) \times \log(t)$ where $\theta$ is some unknown constant. Thus the logarithm of the rate increases linearly with the log of age at death, and the slope of this line is $s-1$, that is, the number of stages less 1. Specifically Armitage and Doll noted that mortality increased with the sixth power of age, an observation that is consistent with the occurrence of seven successive cellular changes (i.e., stages) leading up to the development of cancer.

To derive this mathematical model, Armitage and Doll assumed a constant probability of occurrence of mutation throughout the lifetime and each mutation is a relatively rare event [1954]. Suppose during the time interval $[0, t]$ the probability that mutation $i$ happened is $p_i t$, then the probability that $n-1$ mutations happened is $\prod_{i=1}^{n} p_i t^{n-1}$. If order of $n$ mutations is not considered, there will be $(n-1)!$ possible orderings of the mutations. Therefore, the probability of one right-ordered mutation sequences is $\frac{1}{(n-1)!} \prod_{i=1}^{n} p_i t^{n-1}$. A detailed mathematical proof can be found in Armitage's paper [Armitage and Doll, 1954]. Moreover, Armitage and Doll were able to incorporate features into this model, such as varying hormonal levels with age, to allow for non-constant probabilities over time.

Due to the mathematical simplicity of Armitage-Doll model, many cancer researchers have been willing to apply this model to analyze experimental and observational data. Frank [2005] stated a distinct and predicted pattern at the population level. Biologists prefer simple models rather than complicated ones, that is why the Armitage-Doll model still remains popular among biologists 50 years after its publication. The Armitage-Doll model has contributed to the understanding of the underlying mechanism of carcinogenesis to cancer researchers and scientists. It demonstrated the application of mathematical models in the explanation of the biological events and built a bridge to connect those two fields: mathematics and biology.

### 2.2.2 Moolgavkar-Venzon-Knudson Two-stage Clonal Expansion (TSCE) Carcinogenesis Model

It is widely recognized that the clonal expansion of genetically altered cells (by cell division and cell death/differentiation) is fundamental in carcinogenesis. However, the Armitage-Doll multistage carcinogenesis model did not take into account such key features in cancer development. Moolgavkar,Venzon, and Knudson [1979, 1981, 1990] proposed the two-stage clonal expansion (TSCE) model which incorporates two steps of birth processes for the development of cancer. Two classes of the TSCE models were proposed with the first one being an entirely stochastic model and second one consisting of both deterministic and stochastic elements. Due to mathematical difficulties, Moolgavkar only derived the approximate hazard function for the second class of models. Three assumptions are established for the second class of models where normal cells assumedly grow deterministically and intermediate cells proliferate stochastically. In a small time interval $\triangle t$,

1. The transformation of normal cells to intermediate cells is a non-homogeneous Poisson process with intensity $\mu_1(t)\, X(t)$, where $\mu_1(t)$ is the rate of the first mutation per cell per unit time, and $X(t)$ indicates the total number of normal cells at time $t$. In a small time interval $\triangle t$, the probability of intermediate cells is $\mu_1(t)\, X(t) \triangle t + o(\triangle t)$.

2. The proliferation of intermediate cells is a Poisson process as well. An intermediate cell divides into two intermediate stem cells with probability $\alpha(t) \triangle t + o(\triangle t)$; dies or differentiates with probability $\beta(t) \triangle t + o(\triangle t)$; divides into one intermediate cell and one malignant cell with probability $\mu_2(t) \triangle t + o(\triangle t)$; the probability of more than one event occurring is $o(\triangle t)$.

3. The probability of a malignant cell developing into a malignant tumor is 1.

The hazard function $h(t)$ at time $t$ is given by

$$h(t) = \mu_2(t)\, E(Y(t)|Z(t) = 0)$$

where $Y(t)$, $Z(t)$ are the number of intermediate and malignant cells at time t respectively. Given a very small chance of cancer malignance, the conditional expectation can be simplified to the unconditional expectation as

$$h(t) = \mu_2(t)\, E(Y(t)).$$

Moolgavkar and his colleagues has obtained the approximate form for $h(t)$ through solving the differential equation for E(Y(t)) [Moolgavkar and Venzon, 1988]. The hazard function derived in TSCE model is approximately

$$h(t) \approx \mu_2(t) \int_0^t \mu_1(s) X(s) \exp[\int_s^t (\alpha(u) - \beta(u)) du]\, ds.$$

By assuming constant parameters $\mu_1$, $\alpha$, $\beta$, and $\mu_2$, Moolgavkar [1979] has presented a closed form solution for the exact hazard function:

$$h(t) = \frac{\mu_1}{\alpha} pq \frac{e^{-qt} - e^{-pt}}{qe^{-pt} - pe^{-qt}},$$

where

$$p = \frac{1}{2}(-\alpha + \beta + \mu_2 - \sqrt{(\alpha + \beta + \mu_2)^2 - 4\alpha\beta}),$$

$$q = \frac{1}{2}(-\alpha + \beta + \mu_2 + \sqrt{(\alpha + \beta + \mu_2)^2 - 4\alpha\beta}).$$

In contrast to the simple form of the hazard function in the Armitage-Doll model, the hazard function for the TSCE model is much more difficult. Four derived parameters will be summarized in the TSCE model, the rate of initation, $\mu_1$, the rate of division, $\alpha$, and death, $\beta$, of initial cells, and the rate of malignant conversion, $\mu_2$. $p$ and $q$ are the roots of a quadratic equation, with $p + q = (\alpha - \beta - \mu_2)$ and $pq = \alpha\mu_2$. Three estimated parameters $p$, $q$, and $r \equiv \mu_1/\alpha$ will be treated in the model. As we can see, the TSCE model requires one parameter more than the Armitage-Doll model.

## 2.3 Age-Period-Cohort (APC) Model

### 2.3.1 Background

Cancer mortality and incidence data are commonly presented in two-way tables with mortality and incidence rates distinguished by age-group and period. Recently, a three-factor multiplicative model, age-period-cohort (APC) model, has gained much attention among statisticians and epidemiologists studying the separate effects and trends due to age, period and cohort for cancer incidence and mortality rates [Holford, 1991, Berzuini and Clayton, 1994, Bray, 2002, Congdon, 2006b]. For example, children born during the years when diethylstilbestrol was a common prescription to pregnant women may have higher probabilities to get certain types of cancer as compared to children born at another time [Holford, 1991]. Cohort effects include both factors occurring at the year of birth and those that affect disease rate that is related to year of birth. The period effect is also a nonegligible factor. For instance, medical advances, such as the clinical application of cancer screening technology, have significantly reduced the cancer incidence and mortality in the past decades. It is necessary to include period effect in the model by taking those medical milestones into account. In the APC model, cancer incidence and mortality rates can be estimated as follows:

$$N_{ij} \sim Poisson(\lambda_{ij}),$$

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k,$$

$$i = 1, ..., I,\ j = 1, ..., J,\ k = j - i + I,$$

where $i, j$, and $k$ denote age, period and cohort respectively. $N_{ij}$ is the number of cancer cases at age-group $i$ and period $j$. $\lambda_{ij}$ refers to the unknown true mortality and incidence rate at age-group $i$ and period $j$, $\mu$ is the intercept term, and the $\alpha_i$, $\beta_j$, and $\gamma_k$ refer to effects due to age, period and cohort.

### 2.3.2 Identifiability Issue

Despite the popular application of APC model, equivalently log-linear model, it is well known that there is a non-identifiability problem associated with all three factors because of the exact linear relationship among them [Holford, 1983, 1991]. For any $c \in IR$, the linear predictor for $\log(\lambda_{ij})$ is not changed between the parameter set ($\mu$, $\alpha_i$, $\beta_j$, $\gamma_k$) and ($\mu^*$, $\alpha_i^*$, $\beta_j^*$, $\gamma_k^*$), where

$$\mu^* = \mu - c \cdot I\,; \quad \alpha_i^* = \alpha_i + c \cdot i\,; \quad \beta_j^* = \beta_j - c \cdot j\,; \quad \gamma_k^* = \gamma_k + c \cdot k.$$

Since $k = I + j - i$, we can have

$$\begin{aligned} \log(\lambda_{ij}^*) &= \mu^* + \alpha_i^* + \beta_j^* + \gamma_k^* \\ &= \mu - c \cdot I \,+\, \alpha_i + c \cdot i \,+\, \beta_j - c \cdot j \,+\, \gamma_k + c \cdot k \\ &= \mu \,+\, \alpha_i \,+\, \beta_j \,+\, \gamma_k \,-\, (I \,+\, j \,-\, i) \,+\, k \\ &= \mu \,+\, \alpha_i \,+\, \beta_j \,+\, \gamma_k \\ &= \log(\lambda_{ij}). \end{aligned}$$

Therefore, the age, period, and cohort effects cannot be identified and interpreted uniquely. Figure 2.1 demonstrates the problem of nonidentifibility in APC models. Here we choose the arbitrarily $c$ value as 1 and make the transformation on the age-period-cohort estimates as shown above. As we can see from Figure 2.1, the period effect is increasing in the top row but the new period effects in the bottom row has a decreasing trend over period j. Significant changes are also observed on age and cohort effects. In the top row, we can see the trends of age effects changing at age group of 5 but in the bottom row, a straight linear pattern is shown in the age effects. The cohort effects are orginially decreasing over the cohorts k, however, after transformation, the decreasing trend diminishes around cohort 6.

To overcome this non-identifiability problem, several parameter constraints or assumptions were proposed for the APC model [Osmond and Gardner, 1982, Clayton and Schifflers, 1987]. However, those constraints lack a sound biological explanation. To explain the

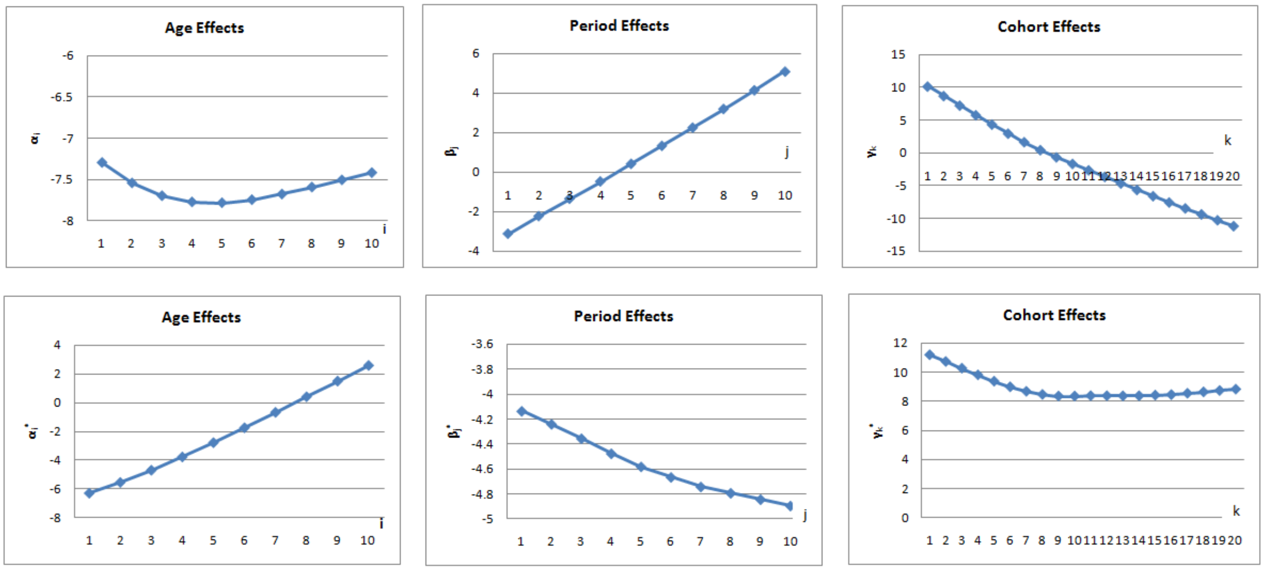

Figure 2.1: Two sets of posterior estimates for APC effects in colon cancer mortality; top row: ($\mu$, $\alpha_i$, $\beta_j$, $\gamma_k$); bottom row: ($\mu^*$, $\alpha_i^*$, $\beta_j^*$, $\gamma_k^*$) and $c = 1$.

biological meaning of age effects, the multistage carcinogenesis model was introduced to the APC model. Specially, the non-specific age effect of traditional APC models was replaced by the hazard function derived from multistage carcinogenesis models [Moolgavkar and Meza, 2009]. It is based on the assumption of fundamental role of age in determining the cancer incidence rates and subsidiary roles of period and cohort in modulating the age effect [Jeon and Moolgavkar, 2006]. However, Moolgavkar admitted that this model did not completely eliminate the non-identifiability problem even though the performance was better than the classical APC models based on AIC values [Moolgavkar and Meza, 2009].

### 2.3.3 Area-APC Model

To study the space and time variation for the risk of disease, many general or more heavily parameterized Bayesian approaches in Area-APC (AAPC) models have been proposed to study spatio-temporal mappings of disease rates [Waller et al., 1997, Carlin and Louis, 2009]. AAPC models have been studied in some recent work in considering spatial correlation in time or important cohort effects [Congdon, 2006a]. In an analysis of lung cancer rates in Tuscany, Lagazio et al. [2003] introduced a full area-age-period-cohort (AAPC) model to study the spatio-temporal pattern of disease risk. The model incorporates the main effect of area, age, period and cohort, and interaction terms such as the area-cohort and area-period interactions. The model is as follows:

$$\log(\lambda_{iap}) = \nu_i + \mu_i + \theta_a + \gamma_p + \delta_c + \varphi_{ip} + \varphi_{ic},$$

where $\lambda_{iap}$ is the relative risk for the $\alpha^{th}$ age group and the $p^{th}$ calendar period in the $i^{th}$ area, $\nu_i$ and $\mu_i$ are the spatial terms, $\theta_a$, $\gamma_p$, and $\delta_c$ are the age, period, and cohort main effects, $\varphi_{ip}$ is the space-period interaction and $\varphi_{ic}$ is the space-cohort interaction. Two different spatial effects considered in this model can be viewed as random effects where unstructured spatial effects $\nu_i$ represents the spatial heterogeneity and structured spatial effects $\mu_i$ considers the spatial clustering [Lagazio et al., 2003]. A Kronecker product of the structure matrix for the relevant dimensions [Congdon, 2006a, Lagazio et al., 2003, Schmid and Held, 2004] is used to derive the prior distribution for the interaction terms.

The joint distribution for spatio-temporal interactions is modeled as a multinormial distribution. For example, the joint spatio-period interactions $\varphi = (\varphi_{ip}, i = 1, ..., N, p = 1, ..., P)$ are taken as $\varphi \sim N(0, \tau_\varphi K_{\mu p})$. The structure matrix $K_{\mu p}$ is the Kronecker product of $K_\mu$ for the spatial effect and $K_p$ for the period effect, such as

$$K_{\mu p} = K_\mu \otimes K_p.$$

Since both $K_\mu$ and $K_p$ are symmetric and singular matrices, their Kronecker product $K_{\mu p}$ is symmetric and singular as well. Therefore, the joint density for spatio-temporal effects is improper [Waller et al., 1997]. Carlin and Louis [2009] pointed out that the proper posterior may not always result, thus extra care must be taken when using the improper priors. As an alternative solution, Congdon [2006b] introduced the parsimonious product interactions schemes with generic form

$$\alpha_i \beta_p, i = 1, ..., N, p = 1, ..., P,$$

where $\alpha_i$ are the structured spatial effects, subject to $\sum_i \alpha_i = 0$, while

$$\beta_p = \exp(\eta_p)/[1 + \sum_{p=1}^{P-1} \exp(\eta_p)], \quad p = 1, ..., P-1,$$

$$\beta_P = 1/[1 + \sum_{p=1}^{P-1} \exp(\eta_p)],$$

and $\eta_p$ are the period effects.

Similar interaction priors for spatio-cohort and spatio-age can be defined as well.

# Chapter 3

# Bayesian Armitage-Doll Multistage Carcinogenesis Model

## 3.1 Background

Armitage and Doll applied their model to study the cancer incidence data in England and Wales in 1950 and 1951 for various types of cancers. Considering too much variation in cell development at early age and unreliable source of death at old age, Armitage and Doll only include patients with age distribution between 25 and 74 into the study [Armitage and Doll, 1954]. The Armitage-Doll model fit the data very well for esophageal, stomach, pancreatic, colon and rectal cancer for both men and women (see Figure 3.1).

The model fitting is not very satisfactory for the cancer of lung, bladder, prostate, breast, ovary, cervix uteri and corpus uteri. Figure 3.2 displays the model fitting for breast and cervix uteri cancer using the Armitage-Doll model. Among those sex organ cancers, such as breast, ovary and cervix uteri for women and prostate for men, the hormonal control of growth should be considered as an important factor in determining the development of cancer cells. As we know, hormonal secretions differ considerably over a human's lifespan. A constant mutation probability may not be valid in the situation like this. Therefore, Armitage and Doll recommended an alternative approach to allow the probability of stage

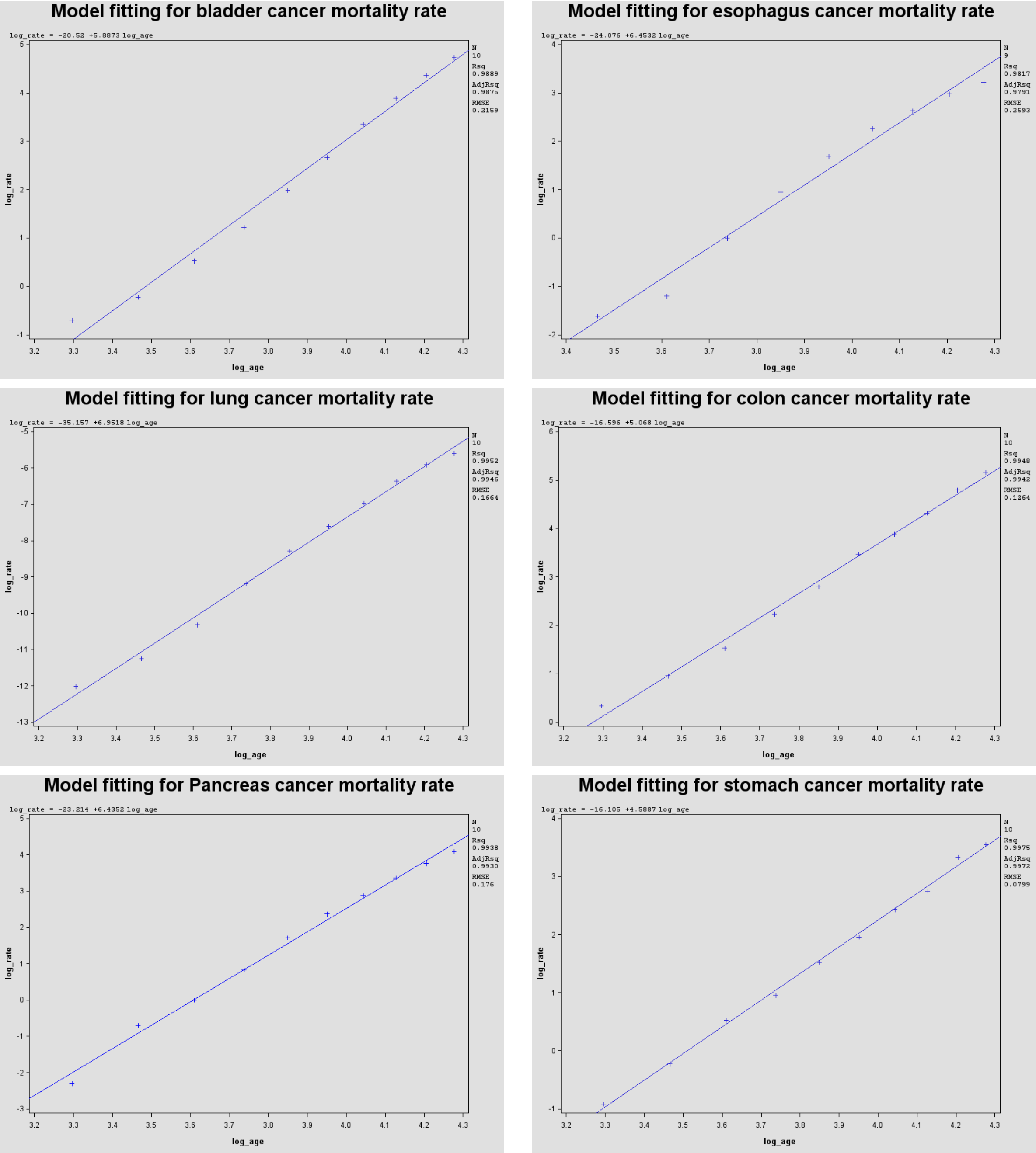

Figure 3.1: Fitting of Armitage-Doll model to non-hormone type of cancer

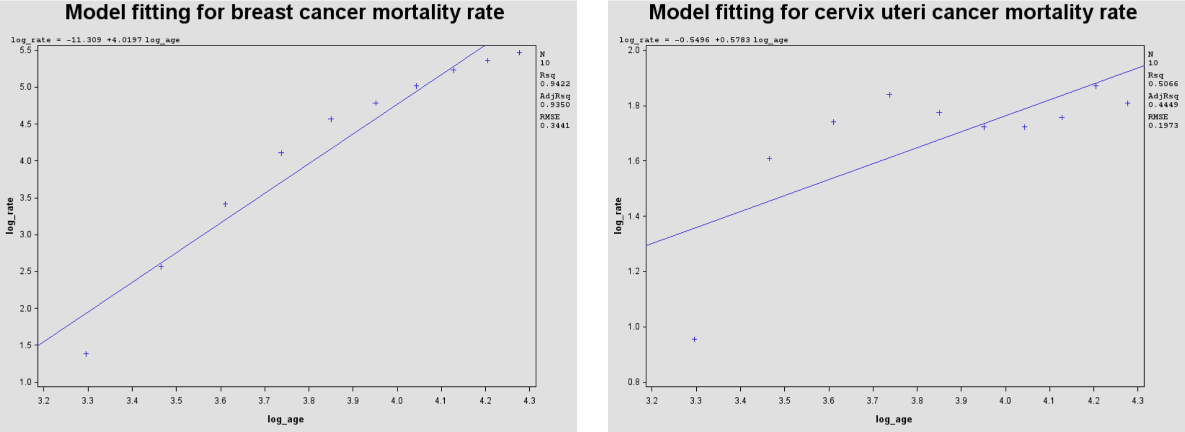

Figure 3.2: Fitting of Armitage-Doll model to hormone type of cancer

conversion varying over time, deriving the hazard rate in a complicated mathematical format [Armitage and Doll, 1954]. It was also widely recognized that considerable lung cancer cases are related to cigeratte smoking and a large proportion of bladder cancer cases is due to occupational hazards. Thus Armitage and Doll [1954] suggested a uniform relationship between death rates and any power of the age is impossible for cancer at these sites. Table 3.1 demonstrates the discrepancy in the estimation of parameter number of stages for cancinogenesis models at these sites (lung, bladder, prostate, breast, ovary, and cervix uteri).

Table 3.1: Fit of Armitage-Doll Model

| Cancer | Estimate | P-value | 95% CI | $R^2$ |
|---|---|---|---|---|
| Esophagus | 6.45 | <.0001 | (5.67, 7.24) | 0.9817 |
| Colon | 5.07 | <.0001 | (4.77, 5.37) | 0.9948 |
| Pancreas | 6.44 | <.0001 | (6.02, 6.85) | 0.9938 |
| Lung | 6.95 | <.0001 | (6.56, 7.34) | 0.9552 |
| Breast | 4.02 | <.0001 | (3.21, 4.83) | 0.9422 |
| Bladder | 5.89 | <.0001 | (5.38, 6.40) | 0.9889 |
| Stomach | 4.59 | <.0001 | (4.40, 4.78) | 0.9975 |
| Overy | 3.74 | <.0001 | (3.40, 4.07) | 0.9883 |
| Rectum | 4.46 | <.0001 | (4.17, 4.74) | 0.9938 |
| Cervix Uteri | 0.58 | .02 | (0.11, 1.04) | 0.5066 |

As we can see from Table 3.1, the application of the Armitage-Doll model leads to different estimates for the number of stages across various types of cancer. Biologically this variation by type reflects the different pathophysiological mechanisms leading to different cancer outcomes. In the standard statistical analysis of Armitage-Doll model, log transformations of cancer incidence rate and age have been conducted in order to fit a linear regression model. Normal distributions of residuals in the log-linear model are assumed. Maximum likelihood estimation (MLE) has been used to obtain the estimate for the number of stages. Considering Armitage-Doll model has not been applied with a Bayesian procedure, we first fit a Bayesain Armitage-Doll multistage model. The advantages of applying the Bayesian method includes the ability to formally incorporate prior information and easily interpretable results. Bayesian model doesn't need separate theories of estimation, testing and multiple comparisons [Carlin and Louis, 2009]. Bayesian method may also improve the precision of model estimates. Bayesian analysis provides a more intuitive interpretation of p-value and confidence intervals [Congdon, 2006a].

## 3.2 Significance and Innovation

Armitage-Doll multistage model has been successfully employed in the carcinogenesis studies due to its simplicity in predicting cancer mortality rate. It estimates different numbers of stages for various types of cancer. This research is the first effort to use an alternative Bayesian approach in the Armitage-Doll multistage model. One of the advantages of the Bayesian method over the classical method is the ability to formally incorporate prior information. Based on preliminary studies and literature reviews, knowledge of the number of stages could be translated into the prior distribution. Therefore, the posterior estimate derived from the combined information (prior and likelihood) could result in greater precision as compared to the classical estimators. In addition, it is much easier to interpret the confidence interval and probability values under the Bayesian frameworks than classical methods.

## 3.3 Build the Bayesian Armitage-Doll Multistage Carcinogenesis Model

To develop a Bayesian model, we need to make selections for the prior distribution and also obtain likelihood from the data we are using. In this project, I use two different prior settings in deriving the Bayesian Armitage-Doll multistage model. The first prior used is the conjugate prior where the posterior distribution is in the same family as the prior distribution. A conjugate prior gives a closed-form expression for the posterior and posterior modes could be conveniently derived from the expression. The normal-inverse-gamma(NIG) distribution will be used as a conjugate prior for the normal linear model. The second prior used is the noninformative prior which is the default choice in many situations when no reliable information concerning the parameter is available.

### 3.3.1 Normal Likelihood and Conjugate Prior

As explained in the background, after a logarithm transformation, the Armitage-Doll multistage carcinogenesis model can be written as the format of log-linear model:

$$\log(r) = \alpha + (s-1)\log(t).$$

To simplify the expression, the above linear model can be expressed as

$$y = X\beta + \epsilon,$$

where $y$ is an $n \times 1$ vector of observations, $X$ is an $n \times p$ design matrix and $\beta$ is a $p \times 1$ vector of parameters and $\epsilon$ an $n \times 1$ vector of random errors. In this example, $y = \log(r)$, $X = (1, \log(t))$, and $\beta = (\alpha, s-1)'$. In most scenarios, the error terms $\epsilon$ are assumed have a $N(0, \sigma^2)$ distribution and the model is described as the normal linear model where $y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$ and $I$ is an $n \times n$ identity matrix. The likelihood takes the format of

$$f(y|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-(y - X\beta)'(y - X\beta)/(2\sigma^2)\}.$$

The Normal-inverse-gamma (NIG) prior is then employed to define parameters $\beta$ and $\sigma^2$ in the form of

$$f(\beta, \sigma^2) \propto (\sigma^2)^{-(d+p+2)/2} \exp\{-[(\beta - m)'V^{-1}(\beta - m) + a]/(2\sigma^2)\},$$

where $a, d \in \mathbf{R}$. The prior distribution can be written as

$$(\beta, \sigma^2) \sim NIG(a, d, m, V)$$

and the normal-inverse gamma prior can be equivalently derived from

$$\sigma^2 \sim IG(a, d) \; and \; \beta|\sigma^2 \sim N(m, \sigma^2 V).$$

We can also obtain weak prior information about $(\beta, \sigma^2)$ within the conjugate family by letting prior variances go to infinity [O'Hagan, 1994]. If parameters $a$ and $d$ in the inverse gamma prior for variance $\sigma^2$ are taken to be 0, we can derive an improper prior for $\sigma^2$ as $f(\sigma^2) \propto \sigma^{-2}$ based on the probability density function for inverse gamma distribution. In addition, if we let $V^{-1} \rightarrow 0$, we can simplify the joint prior $f(\beta, \sigma^2) \propto \sigma^{-(p+2)}$.

The posterior can be derived after combining the likelihood and the prior. By the conjugate structure, the posterior is still with normal-inverse-gamma

$$f(\beta, \sigma^2|y) \propto f(y|\beta, \sigma^2)f(\beta, \sigma^2) \propto (\sigma^2)^{-(d+n+p+2)/2} \exp\{-Q/(2\sigma^2)\}$$

where

$$Q = (y - X\beta)'(y - X\beta) + (\beta - m)'V^{-1}(\beta - m) + a = (\beta - m^*)'(V^*)^{-1}(\beta - m^*) + a^*,$$

and the posterior distribution can be written as

$$(\beta, \sigma^2|y) \sim NIG(a^*, d^*, m^*, V^*)$$

where

$$V^* = (V^{-1} + X'X)^{-1},$$

$$m^* = (V^{-1} + X'X)^{-1}(V^{-1}m + X'y),$$

$$a^* = a + m'V^{-1}m + y'y - (m^*)'(V^*)^{-1}m^*,$$

$$d^* = d + n.$$

If $X'X$ is non-singular, the posterior estimate of $\beta$ which is $E(\beta|y) = m^*$ can be written as

$$m^* = (V^{-1} + X'X)^{-1}(V^{-1}m + X'X\hat{\beta}) = (I - A)m + A\hat{\beta},$$

where $A = (V^{-1} + X'X)^{-1}X'X$.

In the log-normal Armitage-Doll model, we have

$$(X'X)^{-1} = \begin{pmatrix} n & \sum logt_i \\ \sum logt_i & \sum (logt_i)^2 \end{pmatrix}^{-1} \quad and \quad X'y = \begin{pmatrix} \sum logr_i \\ \sum logt_i \, logr_i \end{pmatrix},$$

then we can calculate the classical estimate $\hat{\beta} = (X'X)^{-1}(X'y)$ and the posterior mean of $\beta$ is a weighted average of its prior mean $m$ and its classical estimate $\hat{\beta}$. The prior mean $m$ and variance $V$ of $\beta|\sigma^2$ will be approximated from the preliminary analysis results on the SEER data.

### 3.3.2 Poisson Likelihood and Noninformative Prior

Suppose $r = y/n$, where $r$ is the death rate, $y$ is the number of deaths, and $n$ is the size of the population at risk. A typical assumption for count data is that they take on a Poisson distribution. Thus, if we assume that $y \sim$ Poisson with parameter $\theta t^{s-1}n$ where $t$ is the age, $s$ is the number of stages, and $\theta$ is another unknown parameter, then the data distribution $y|s,\theta$ can be written as $f(y|s,\theta) \propto (\theta t^{s-1}n)^y \exp(-(\theta t^{s-1}n))$. The conjugate prior for Poisson likelihood is the Gamma distribution. However, the Poisson model parameter $\theta t^{s-1}n$ has two unknown parameters $\theta$ and $s$. It is difficult to derive specific distributions for $\theta$ and $s$ in order to obtain the Gamma distribution for $\theta t^{s-1}n$. Therefore, we consider the noninformative prior distributions for $\theta$ and $s$ then use Markov chain Monte Carlo (MCMC) sampler methods to obtain the posterior estimates for them.

*Prior 1:* $\theta \sim N(0, \sigma^2)$

A weak normal prior with large variance can be assigned to the unknown parameter $\theta$. Given little information we know about this parameter, the prior mean is set to zero. Gamma or Beta priors are not used for the unknown parameter $\theta$ since possible negative value is observed for $\theta$. The variance for the normal prior can be fixed at a large value (i.e., $10^6$), or defined as a hyperprior with Gamma distribution (i.e., $\sigma^2 \sim Gamma(0.1, 100)$).

*Prior 2:* $s \sim N(5, \sigma^2)$

Similarly to the unknown parameter $\theta$, a weak normal prior with large variance can be assigned to the parameter $s$. Based on preliminary literatures and classical approaches, we know the number of stages in the carcinogenesis model is about 6 to 7 for most cancers. Since one less the number of stages is assigned to the power of age in the Armitage-Doll model, we can simplify the prior mean for the $s$ as 5. The choice of $\sigma^2$ can be either a fixed large number or hyperprior with Gamma distribution.

*Prior 3:* $s \sim Gamma(\alpha, \beta)$

We can then assume a Gamma prior distribution for s with parameter $\alpha$ and $\beta$, i.e., $f(s|\,\alpha, \beta) \;\propto\; s^{\alpha-1}\, e^{-\theta/b}$ where $0 < \varphi < 1$. Since the mean of the Gamma distribution is $\alpha\beta$ and variance is $\alpha\beta^2$, we choose $\alpha = 1$ and $\beta = 10$ in order to design the Gamma distribution with mean at 10 and variance at 100.

Alternatively, we can define hyperpriors for the shape and scale parameters ($\alpha$ and $\beta$) in the Gamma distribution. By Bayes rule [Carlin and Louis, 2009] we have the posterior density

$$f(s, \varphi|y) \propto f(y|s)f(s|\varphi)f(\varphi),$$

then the marginal posterior distribution of s can be derived at

$$f(s|y) = \int f(s, \varphi|y)d\varphi \propto \int f(y|s)f(s|\varphi)f(\varphi)d\varphi$$

where $\varphi = (\alpha, \beta)$.

At this point there are several choices to be made on the hyperpriors with respect to the direction of inference, especially for the derivation of the marginal posterior, $f(s|y)$. Since the Gamma distribution is a two-parameter exponential family, I set the hyperprior for shape parameter $\alpha$ to be the Exponential distribution with mean at 5 and the hyperprior for the scale parameter $\beta$ as the Gamma distribution with the shape value 0.1 and scale value 10.

### 3.3.3 Binomial Likelihood and Noninformative Priors

The Poisson distribution was assumed previously for cancer death count.The Poisson distribution with parameter $\lambda = np$ can be used as an approximation to the Binomial distribution $B(n, p)$ if $n$ is sufficiently large and $p$ is sufficiently small. In the Bayesian framework for the Armitage-Doll model, we can also assume a Binomial observation model for cancer death count $y$ with an unknown cancer mortality rate $p$. In other words, $y \sim Bin(n, p)$ where $n$ is the total persons at risk and $p$ denotes the probability of persons at risk will die from cancer. We decompose the log-odds $\pi = \log(p/(1-p))$ of these probabilities into a linear combination of constant and a multiplication of number of stage (s) and log transformation of age $t$. More specifically, the logit model will be written as

$$\log \frac{p}{1-p} = \alpha + (s-1)logt$$

so

$$p = \frac{e^{\alpha+(s-1)logt}}{1+e^{\alpha+(s-1)logt}}.$$

### 3.3.4 Weibull Likelihood and Noninformative Priors

An alternative approximation for cancer mortality rates is the Weibull distribution [Berry, 2007]. It displays the distribution of failures, where the failure rate is proportional to a power of time. The probability density function of a Weibull random variable $x$ is

$$f(x; \lambda, k) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}e^{-(x/\lambda)^k}$$

where $x \geq 0$, the shape parameter $k > 0$, and the scale parameter $\lambda > 0$. The advantage of the Weibull distribution is that its failure rate (or hazard rate) is exactly in the format of the Armitage-Doll model:

$$h(x; k, \lambda) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}$$

where $k$ is equivalent to the number of stage $s$ in the Armitage-Doll model. To further demonstrate the application of Weibull distribution in modeling cancer mortality rate, we will treat the Bayesian Weibull model as an alternative model in this Bayesian framework for the Armitage-Doll model while taking noninformative priors for its shape parameter $k$ and scale parameter $\lambda$.

### 3.3.5 Results - NIG Priors

Normal-inverse-gamma priors are first employed in the Bayesian Armitage-Doll model to obtain the posterior estimate for the parameter of number of stages. The NIG priors are

$$\sigma^2 \sim IG(a, d) \ and \ \beta|\sigma^2 \sim N(m, \sigma^2 V),$$

where $a$ and $d$ are set as .001 and the prior for $m$ is defined as $(5, 5)'$. I let $V$ as $1000I$ and $I$ is 2 by 2 identity matrix.

As discussed previously, the posterior estimate of $\beta$ which is $E(\beta|y) = m^*$ can be written

Table 3.2: Fit of NIG priors in the Bayesian Armitage-Doll Model

| Cancer | Estimate | SD | MCSE | HPD | Geweke Diagnostics pvalue |
|---|---|---|---|---|---|
| Esophagus | 6.27 | 0.57 | .008 | (5.14, 7.42) | .89 |
| Colon | 4.99 | 0.28 | .004 | (4.44, 5.54) | .35 |
| Pancreas | 6.33 | 0.34 | .005 | (5.64, 7.08) | .28 |
| Lung | 6.84 | 0.36 | .005 | (6.13, 7.57) | .05 |
| Breast | 3.94 | 0.40 | .02 | (3.11, 4.73) | .80 |
| Bladder | 5.78 | 0.36 | .005 | (5.11, 6.53) | .85 |
| Stomach | 4.51 | .25 | .003 | (4.03, 5.00) | .11 |
| Overy | 3.67 | .25 | .003 | (3.17, 4.15) | .86 |
| Rectum | 4.37 | 0.26 | .004 | (3.87, 4.89) | .49 |
| Cervix Uteri | 0.56 | 0.22 | .01 | (0.15, 1.02) | .82 |

as

$$m^* = (V^{-1} + X'X)^{-1}(V^{-1}m + X'X\hat{\beta}) = (I - A)m + A\hat{\beta},$$

where $A = (V^{-1} + X'X)^{-1}X'X$. The assumption for the above formulation is that $X'X$ is non-singular matrix. In the log-normal Armitage-Doll model, I have

$$(X'X)^{-1} = \begin{pmatrix} n & \sum logt_i \\ \sum logt_i & \sum (logt_i)^2 \end{pmatrix}^{-1}$$

where age $t$ is chosen as the median age in every age group from 25 to 75 years old so that $t = (27, 32, 37, 42, 47, 52, 57, 62, 67, 72)$. Therefore, I can calculate the $X'X$ and $(X'X)^{-1}$ matrix as following

$$X'X = \begin{pmatrix} 10 & 38.56 \\ 38.56 & 149.67 \end{pmatrix}$$

and

$$(X'X)^{-1} = \begin{pmatrix} 15.65 & -4.03 \\ -4.03 & 1.05 \end{pmatrix}.$$

In order to compare the parameter estimates between classical approach (log-linear normal model) and Bayesian approach using NIG priors (Table 3.2), I generate Table 3.3 to demonstrate the relationship between the classical estimate and the posterior estimate for the parameter of number of stages. How the prior setting affects the posterior estimate is

Table 3.3: Different prior settings in affecting the posterior estimate

| | | Posterior $m^*$ | | |
|---|---|---|---|---|
| Cancer | Classical Estimate | $V = 1000I$ | $V = 100I$ | $V = 10I$ |
| Esophagus | 6.45 | 6.32 | 5.14 | 1.98 |
| Colon | 5.07 | 4.99 | 4.32 | 1.81 |
| Pancreas | 6.44 | 6.33 | 5.46 | 2.10 |
| Lung | 6.95 | 6.84 | 5.96 | 2.55 |
| Breast | 4.02 | 3.95 | 3.48 | 1.56 |
| Bladder | 5.89 | 5.79 | 5.01 | 2.01 |
| Stomach | 4.59 | 4.51 | 3.86 | 1.43 |
| Overy | 3.74 | 3.67 | 3.15 | 1.17 |
| Rectum | 4.46 | 4.37 | 3.79 | 1.42 |
| Cervix Uteri | 0.58 | 0.56 | 0.42 | -0.09 |

also displayed in Table 3.3. The classical estimates are obtained from the standard log-linear normal model which is shown in Table 3.1 and the prior mean $m$ is set as $(0, 0)$ as the starting point. Therefore, the prior variance $V$ plays an important role in determining the posterior mean of $\beta$. Strong prior gives more weight to the prior mean since $V$ is small, while weak prior gives more weight to the classical estimate. As we can see from Table 3.3, if we choose $V = 10I$ which means strong prior belief since the variance is small, the posterior estimate is close to the prior mean which is 0. However, if we increase $V$ to $1000I$ which indicates a weak prior due to large variance, the posterior estimates tend to shift to the classical estimate.

Inverse-gamma prior for the variance component in the normal model is a conditional conjugate prior. The posterior distribution for the variance in the normal model is still within inverse-gamma family. However, the inverse-gamma prior can become an ”improper” flat prior as the parameter $\epsilon$ in $IG(\epsilon, \epsilon)$ causes its variance approaches infinity. Therefore, the posterior inferences are sensitive to $\epsilon$. We chose different $\epsilon$ values in the inverse-gamma prior for variance component to see whether they affect the posterior inference on the model parameter s (number of stage -1). Uniform prior for the variance was also considered as well. Result shows the choice of inverse-gamma or uniform prior for the variance works well in the model to obtain proper posterior inference for model parameters.

### 3.3.6 Results - Noninformative Priors

The relative influence of the prior and data on the posterior belief depends on how much weight is given to the prior and the strength of the data. In general, if the sample was small with an informative prior, then the prior distribution would have a relatively greater influence on the posterior belief about the parameter of interest. However, a large data sample would tend to have a predominant impact on the posterior belief on the parameter of interest unless the prior was informative [Congdon, 2006a]. Since the sample size is small ($n = 10$) in this study, noninformative priors are used to avoid the possible negative influence on the posterior belief of the parameter of number of stages by wrongly defined informative priors. Likelihoods are computed based on different data assumptions such as Poisson, Binomial and Weibull distribution. Results from different noninformative priors are displayed in Table 3.4. Figure 3.3 shows the mean and 95% confidence intervals of cancer rates estimated by Bayesian Armitage-Doll model and compares with the observed colon cancer mortality rate. The model fits the data well except at older age.

Table 3.4: Different likelihood and noninformative priors in estimating the posterior for colon cancer

| | Prior $m$ | | | Posterior $m^*$ | | p value | |
|---|---|---|---|---|---|---|---|
| Likelihood | $\theta$ | $s$ | hyperprior | $\theta$ | $s$ | $\theta$ | $s$ |
| Poisson | $N(0, 10^6)$ | $N(5, 10^6)$ | NA | -31.57 | 5.80 | $< .001$ | $< .001$ |
| | $N(0, 10^6)$ | $Gamma(.1, 100)$ | NA | -31.60 | 5.80 | .51 | .52 |
| | $N(0, \sigma^2)$ | $N(5, \sigma^2)$ | $\sigma^2 \sim Gamma(.1, 100)$ | -31.61 | 5.78 | .26 | .28 |
| | $N(0, 10^6)$ | $Gamma(\alpha, \beta)$ | $\alpha \sim Exp(5)$;<br>$\beta \sim Gamma(.1, 100)$ | -31.57 | 5.80 | .85 | .88 |
| Binomial | $N(0, 10^6)$ | $N(0, 10^6)$ | NA | -31.58 | 5.80 | .67 | .67 |
| | $N(0, 10^6)$ | $Gamma(.1, 100)$ | NA | -31.58 | 5.80 | .20 | .19 |
| | $N(0, \sigma^2)$ | $N(0, \sigma^2)$ | $\sigma^2 \sim Gamma(.1, 100)$ | -31.58 | 5.80 | .83 | .77 |
| | $N(0, 10^6)$ | $Gamma(\alpha, \beta)$ | $\alpha \sim Exp(5)$;<br>$\beta \sim Gamma(.1, 100)$ | -31.58 | 5.80 | .83 | .82 |

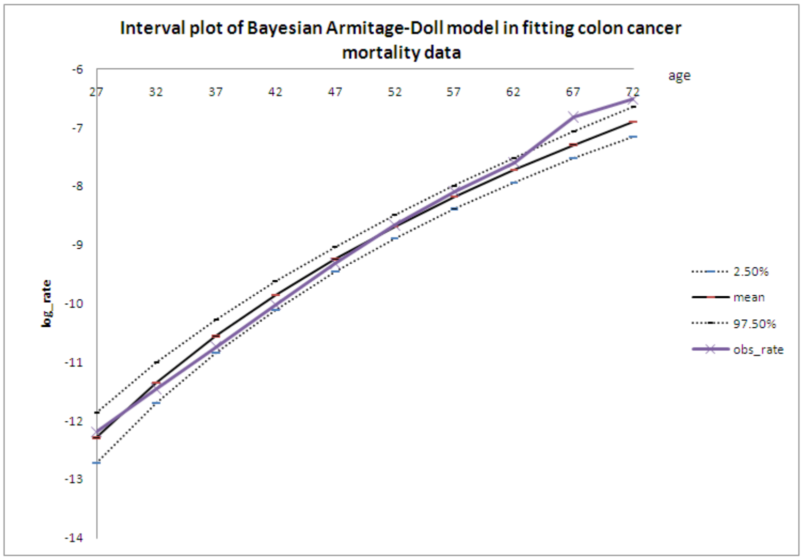

Figure 3.3: Fitting of Bayesian Armitage-Doll model

## 3.4 Assess the Bayesian Armitage-Doll Multistage Carcinogenesis Model

To prove the Bayesian model robustness, several possible concerns related to the prior distribution, the precise form of the likelihood, and the numbers of levels in the hierarchical model [Carlin and Louis, 2009] need to be addressed. In this section, I investigate the robustness of the conclusion from the Bayesian Armitage-Doll multistage model by checking whether the conclusion still holds subject to changes in the prior, likelihood, or some other aspect of the model.

### 3.4.1 Sensitivity Analysis

Consider the log-linear model after log transformation of the Armitage-Doll multistage model

$$\log(r_i) = \alpha + (s-1)\log(t_i) + \epsilon_i, i = 1, ..., n,$$

where $\log(r_i)$ is the cancer mortality rate for age group $i$, $\log(t_i)$ is the median age for age group $i$ and the $\epsilon_i$ are independent random errors having density $f$ with mean 0. Andrews and Mallows [1974] proved that defining the distribution of the error term to $\epsilon_i|\sigma^2, \lambda_i \sim N(0, \lambda_i\sigma^2), i = 1, ..., n$, and putting a prior on $\lambda_i$ can incorporate various familiar and more widely dispersed error densities. In other words, the scale mixture of normal densities is created as follows:

$$f(\epsilon_i|\sigma^2) = \int p(\epsilon_i|\sigma^2, \lambda_i)p(\lambda_i)d\lambda_i, \quad i = 1, ..., n.$$

The following list gives different formats for $p(\lambda_i)$ to obtain non-normal distributions.

- Student's t errors: $\lambda_i \sim IG(\nu/2, 2/\nu)$;
- Double exponential errors: $\lambda_i \sim expo(2)$;
- Logistic errors: if $1/\sqrt{\lambda_i}$ has the asymptotic Kolmogorov distance distribution, then $\epsilon_i|\sigma^2$ is logistic.

Due to the uncertainty about the error density and the impact of possible outliers, three different error densities are compared:

- $\epsilon_i \sim N(0, \sigma^2)$;
- $\epsilon_i \sim t(0, \sigma^2, \nu = 2)$;
- $\epsilon_i \sim double\ \ expo(0, \sigma))$.

Sensitivity analysis is implemented on the Bayesian Armitage-Doll model for colon, esophagus, breast and lung cancer, as shown in Table 3.5. In colon cancer, the parameter $s$ is more precisely estimated in the nonnormal errors (t and double exponential distribution) than in the normal errors, as indicated by the smaller standard deviations and narrower HPDs. The heavier tails in the t and double exponential errors can dissolve the negative effects caused by the possible large outliers more quickly than the normal errors. Therefore, the posterior estimates for the parameter $s$ can be achieved with higher accuracy and efficiency

[Carlin and Louis, 2009]. In esophagus, breast and lung cancer, we observe the disturbance in the posterior estimate for the number of stage $s$ as we change the prior assumptions for the model error terms. The posterior estimate in esophagus cancer is apparently inaccurate when t errors are defined. Furthermore, the standard deviations obtained in the t errors cases for breast and lung cancer are even higher than those in the normal errors cases, showing that it is not efficient in the posterior estimate as the error terms change from normal to t distribution. However, the performance of double exponential errors in esophagus, breast and lung cancer is satisfactory with improved efficiency and accuracy in obtaining the posterior estimate for the number of stage $s$.

Table 3.5: Sensitivity analysis results

| Cancer | Error $\epsilon_i$ | Estimate | SD | MCSE | HPD | p value |
|---|---|---|---|---|---|---|
| Colon | $N(0, \sigma^2)$ | 4.99 | 0.28 | .004 | (4.44, 5.54) | .35 |
| | $t(0, \sigma^2, \nu = 2)$ | 4.93 | 0.18 | .005 | (4.59, 5.25) | .74 |
| | $double\ \ expo(0, \sigma))$ | 4.93 | 0.17 | .003 | (4.59, 5.25) | .48 |
| Esophagus | $N(0, \sigma^2)$ | 6.27 | 0.57 | .008 | (5.14, 7.42) | .89 |
| | $t(0, \sigma^2, \nu = 2)$ | 0.40 | 0.29 | .013 | (-0.18, 0.98) | .95 |
| | $double\ \ expo(0, \sigma))$ | 5.64 | 0.52 | .013 | (4.55, 6.54) | .55 |
| Breast | $N(0, \sigma^2)$ | 3.94 | 0.40 | .02 | (3.11, 4.73) | .80 |
| | $t(0, \sigma^2, \nu = 2)$ | 1.48 | 0.55 | .108 | (0.57, 2.74) | $< .001$ |
| | $double\ \ expo(0, \sigma))$ | 3.63 | 0.29 | .005 | (3.03, 4.19) | .37 |
| Lung | $N(0, \sigma^2)$ | 6.84 | 0.36 | .005 | (6.13, 7.57) | .05 |
| | $t(0, \sigma^2, \nu = 2)$ | 6.63 | 0.48 | .067 | (5.75, 7.22) | .01 |
| | $double\ \ expo(0, \sigma))$ | 6.71 | 0.24 | .005 | (6.22, 7.13) | .02 |

### 3.4.2 Model Assessment

It is important to conduct diagnostic measurements to see whether the model is an adequate fit with justified assumptions. For example, standard linear regression requires the normality, independence, linearity, and homogeneity of variance [Carlin and Louis, 2009]. Similar to classical approach, we define a Bayesian residual as $r_i = y_i - E(Y_i|z),\ i = 1, ..., n$, where $z = (z_1, ..., z_m)$ is the fitting sample and $y = (y_1, ..., y_n)$ is the validation sample.

*Cross-validatory*

The limited number of observations in cancer mortality data per age group pushes us to consider the cross-validatory (or ”leave one out”) approach [Carlin and Louis, 2009], where the estimated value for $y_i$ is computed conditional on all the data except $y_i$. In other words, we calculate the Bayesian residual as $r_i = y_i - E(Y_i|y_{(i)}),\ i = 1, ..., n$, where $y_{(i)} = (y_1, ..., y_{(i-1)}, y_{(i+1)}, ..., y_n)'$. And the standardized residual is defined as

$$d_i' = \frac{y_i - E(Y_i|y_{(i)})}{\sqrt{Var(Y_i|y_{(i)})}}.$$

In the cross-validatory approach, the posterior mean and variance are computed based on the conditional predictive distribution

$$f(y_i|y_{(i)}) = \frac{f(y)}{f(y_{(i)})} = \int f(y_i|\theta, y_{(i)})\, p(\theta|y_{(i)})d\theta$$

which gives the likelihood of each point without taking itself into account.

Figure 3.4 shows the means of standard errors at each age group are approximately zero and the majority of standard errors (from the first quartile to the third quartile) are within the range of $\pm\, 1.5$. Therefore, the cross-validatory (leave-one-out) approach concludes that Armitage-Doll model is a good fit for the colon cancer mortality data.

*Posterior Predictive Distribution*

An alternative method for model assessment is using the posterior predictive distribution as a model checking tool. The posterior predictive distribution is defined as

$$p(y_{pred}|y) = \int p(y_{pred}|\theta)p(\theta|y)d\theta$$

and samples from $p(y_{pred}|y)$ are generated as $y^i_{pred}$ for $i = 1, ..., M$, where $M$ is the total number of replicates. Replicated samples are compared with the observed data to see whether there are any large and systematic differences. Bayesian p-value can be calculated as follows:

$$Pr(T(y_{pred}) > T(y)|\theta)$$

where $T(\cdot)$ denotes the test statistics, such as the mean, standard deviation, order statistics, and so on. $T(y)$ is based on the observed data while $T(y_{pred})$ are from the replicated data

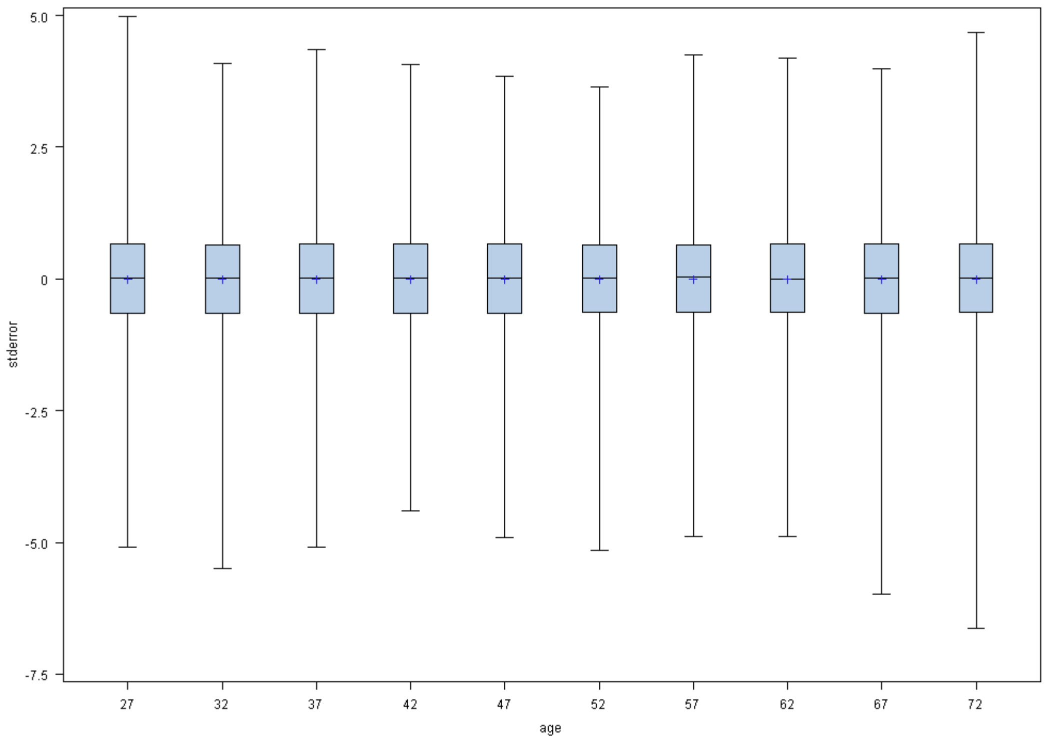

Figure 3.4: Box plot of standard error by age in fitting colon cancer mortality rate

sampled from the posterior predictive distribution.

In Figure 3.5, we observe that the Bayesian p-value for posterior means is 0.74 which indicates no overall lack of fit using the Bayesian normal model. Similarly, the Bayesian p-values for maximum and minimum values are 0.64 and 0.90 respectively, indicating no problem with fit in either tail of the predictive distribution. The small Bayesian p-value for standard deviation indicates the samples from the posterior predictive distribution have less variation than those from observed data. In summary, the model is a good fit for the data.

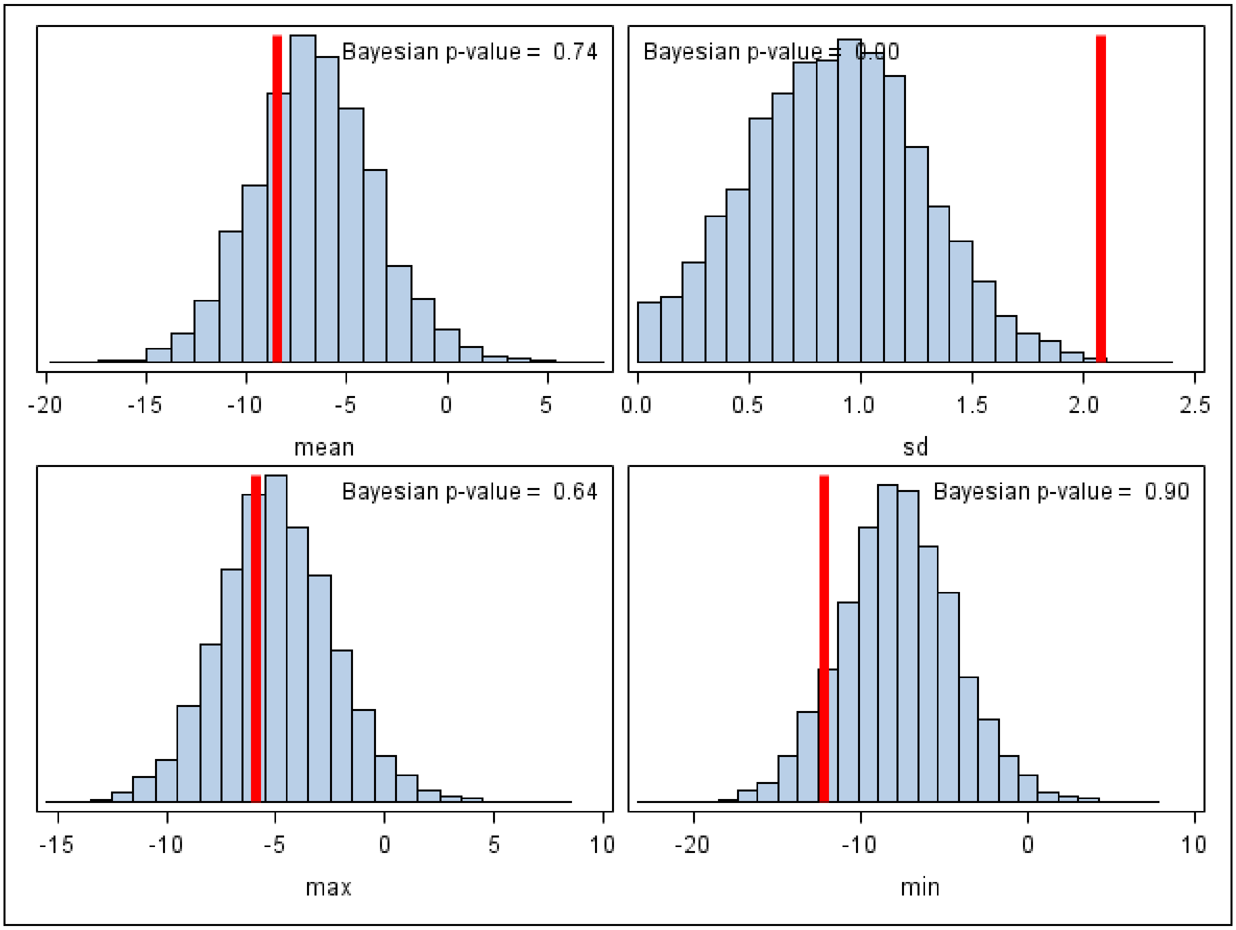

Figure 3.5: Model assessment through posterior prediction distribution

## 3.5 Simulation

For different types of cancer, the posterior estimates for the number of stages are rounded to the nearest integers. For example, Bayesian Armitage-Doll carcinogenesis model shows the posterior estimate for colon cancer is 4.99 (see Table 3.2). Therefore, I conclude there are six stages of transition from normal health cells to malignant cancer cells (the number of stages equals to the posterior estimate plus 1). Simulation studies are implemented to check the accuracy of such claims from the Armitage-Doll model.

Parametric Bootstrapping samples are generated using the Poisson distribution with the means equal to observed colon cancer death count for each age group, i.e.,

$$y_i \sim Poisson(d_i)$$

where $i = 1, ..., 10$ for ten different age groups, $d_i$ is the total colon cancer death count reported in SEER database at age group $i$ for the population with SEER coverage, and $y_i$ is the simulated colon cancer death count at age group $i$. The population size of each age group ($N_i$) is considered as constant and used as the denominator to compute the simulated colon cancer mortality rate $r_i$ as

$$r_i = \frac{y_i}{N_i}.$$

1000 datasets are generated and the Bayesian Armitage-Doll model is used to obtain the posterior estimate for the number of stages $\hat{\beta}$. Several numerical results are calculated as follows:

$$Bias = \hat{\beta} - 5,$$

$$MSE = Var(\hat{\beta}) + Bias^2,$$

where $\hat{\beta}$ is the sample mean of 1000 bootstrapped estimate for the number of stage and $Var(\hat{\beta})$ is their variance. The simulation result shows that the average of posterior estimates for $\beta$ is $\hat{\beta} = 4.9957$ and their variance is 0.00016. Then we can get

$$Bias = 4.9957 - 5 = -0.0043,$$

$$MSE = Var(\hat{\beta}) + Bias^2 = 0.00018$$

In conclusion, the simulation shows that the posterior estimate is a good approximation for the true value with negligible bias and MSE.

# Chapter 4

# Bayesian extended Age-Period-Cohort Model

## 4.1 Background

Trend analysis for disease incidence and mortality is very important to public health and those trends have been used by researchers in understanding disease etiology and making disease projections [Holford, 1991]. A display of age-specific rates is commonly used to present the age patterns in the distribution of disease incidence and mortality rates. Holford [1991] stated that age is an important factor in the etiology of most diseases and the risk for disease would vary as people aged from birth. Meanwhile, birth cohort effects are often observed when different birth cohorts exposed to different levels of risk factors and resulted in changes in disease incidence and mortality rates. For example, children born during the years when diethylstilbestrol was a common prescription to pregnant women may have higher probabilities to get certain types of cancer as compared to children born at another time [Holford, 1991]. In vital statistics, disease incidences are often reported by year of diagnosis and age. Similarly, disease death are also reported by year of death and age. One approach in trend analysis is to present the patterns by year of diagnosis for the incidence and by year of death for the mortality. The changes in the disease rates by calendar period are expected. For instance, medical advances, such as the clinical application of breast cancer screening technology in the late 1980s, have significantly reduced the cancer incidence and mortality since that period.

Therefore, a three-factor multiplicative model, age-period-cohort (APC) model, has gained much attention in statisticians and epidemiologists to study the separate effects and trends due to age, period and cohort for cancer incidence and mortality rates [Holford and McKay, 1994, Bray, 2002, Moolgavkar and Meza, 2009]. In the APC model, cancer incidence and mortality rates can be estimated as follows:

$$N_{ij} \sim \ Poisson(\lambda_{ij}),$$

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k,$$

$$i = 1, ..., I, \ j = 1, ..., J, \ k = j - i + I,$$

where $i, j$, and $k$ denote age, period and cohort respectively. $N_{ij}$ is the number of cancer cases for age-group $i$ and period $j$. $\lambda_{ij}$ refers to the unknown true mortality and incidence rate for age-group $i$ and period $j$, $\mu$ is the intercept term, and the $\alpha_i$, $\beta_j$, and $\gamma_k$ refer to effects due to age, period and cohort.

Despite the popular application of APC model, equivalently the log-linear model, it is well known that there is a non-identifiability problem associated with all three factors because of the exact linear relationship among them [Holford, 1983, 1991]. To overcome this non-identifiability problem, several parameters constraints or assumptions were proposed for the APC model [Osmond and Gardner, 1982, Clayton and Schifflers, 1987]. Since the age, period and cohort effects are linearly dependent, the conventional strategy is to transform at least one variable related to age, period, or cohort so that its relationship to the others is nonlinear. However, those constraints usually lack a sound biological explanation. In addition, statisticians and epidemiologists often find that it is difficult to explain the modifying effects of those temporal effects (age, period, and cohort) [Richardson, 2008]. Carcinogenesis models of a typical underlying disease process describe how the normal cells are transformed into cancer cells and age is a deterministic factor in the model. Typically conventional regression modeling techniques are used to smooth and summarize the cancer epidemiological data and goodness-of-fit of the model doesn't depends on the validity of any particular theoretical carcinogenesis model derived from those epidemiological data [Richardson, 2008]. However, theoretical carcinogenesis models can be used as a complimentary approach to the empirical models to explore the association between response and time from the underlying disease process.

To explain the biological meaning of age effects, multistage carcinogenesis models are introduced to the APC model. Specially, the non-specific age effect of traditional APC models is replaced by the hazard function derived from multistage carcinogenesis models [Moolgavkar and Meza, 2009]. It is based on the assumption of the fundamental role of age in determining cancer incidence rates and the subsidiary roles of period and cohort in modulating the age effect [Jeon and Moolgavkar, 2006]. However, Moolgavkar admitted that this model did not completely eliminate the non-identifiability problem even though the performance was better than the classical APC models based on AIC values [Moolgavkar and Meza, 2009].

Bayesian methods have been applied to the APC model to estimate the age, period and cohort effect in cancer incidence and mortality [Bray and Brennan, 2001]. A prior belief about the smoothness of the parameters was considered in the Bayesian model. Model constraints were implemented in the sampling procedures. However, the non-identifiability problem is still a challenge with Bayesian models.

## 4.2 Significance and Innovation

Bayesian APC models have been applied to study various cancer incidence and mortality rates in the UK and United States [Breslow and Clayton, 1993, Berzuini and Clayton, 1994, Bray, 2002]. Different from the classical approaches which make strong parametric assumptions, Bayesian models improve the precision in estimating the parameters by updating its posterior density from the combination of the prior belief and data [Carlin and Louis, 2009].

In this study, we first introduce multistage carcinogenesis models to the Bayesian APC model. Thus we can incorporate the biological meaning of age effects in the cancer development to predict cancer incidence and mortality rates while taking period and cohort effects into account as well. Prior settings for number of stages ($s$) and constant ($\theta$) we described in Chapter 3 are used in determining the prior for age effect. Noninformative priors are assigned to model parameters in the TSCE carcinogenesis model. Inspired by Berzuini and Clayton [1994] and Bray [2002], we implement a Gaussian autoregressive prior in the forward direction for cohort and period effects. The use of the autoregressive prior for cohort effects can avoid excessive variability problems in the data caused by few early and late cohorts. The autoregressive prior structure can be viewed as an exchangeable prior model for second differences of period and cohort effects which are all identifiable. An arbitrary linear constraint on the log-linear trend components of APC effects can be imposed to solve the identifiability problem and optimistically will have no effect on the prediction of the model.

The introduction of carcinogenesis model can make Bayesian APC model more biological sound in explaining cancer mortality and incidence. Furthermore, the entry of carcinogenesis model into APC model may help reduce the nonidentifiability concerns across the linear relationships in age, period and cohort effects. The Bayesian extended APC model can be used as a tool to estimate the cancer incidence and mortality rates with greater precision and to make more accurate projections for the near future. The estimation and prediction derived from the Bayesian extended APC model can be better used to inform public health policy makers in understanding the trend of cancer incidence and mortality.

## 4.3 Apply Armitage-Doll Multistage Carcinogenesis Model into APC Model

The hazard function derived from the Armitage-Doll multistage carcinogenesis model is introduced to the APC model. As I discussed before, Armitage-Doll multistage model estimates the hazard function as $h(t) = ct^{s-1}$ where $s$ is the number of stages, $t$ is the age and $c$ is constant. The log-linear model takes the format of $\log(h(t)) = \theta + (s-1)logt$ which will be plugged into the APC model to replace the age effect $\alpha_i$. The extended APC model will become

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k,$$

$$i = 1, ..., I,\ j = 1, ..., J,\ k = j - i + I,$$

where $\alpha_i = \theta + (s-1)\log t_i$ and $t_i$ is the median age in age group $i$.

***Model Constraints***

To improve the convergence rate in the MCMC samples, I add constraints to age, period and cohort effect to make them sum to zeros by substracting to their means [Berzuini and Clayton, 1994], i.e.,

$$\alpha_{ci} = \alpha_i - \frac{1}{I}\sum_{i=1}^{I}\alpha_i,$$

$$\beta_{cj} = \beta_j - \frac{1}{J}\sum_{j=1}^{J}\beta_j,$$

$$\gamma_{ck} = \gamma_k - \frac{1}{I+J-1}\sum_{k=1}^{I+J-1}\gamma_k.$$

Therefore we can get

$$\sum_i \alpha_{ci} = \sum_j \beta_{cj} = \sum_k \gamma_{ck} = 0.$$

Holford [1994] stated that further constraints are still needed in order to solve the non-identifiability issues in addition to the above constraints to center all three temporal effects. Additional constraints include the reparameterization of APC parameters, equating the first and second levels of APC effects, removing the linear trend in age effects, etc. In this APC model, I obtain the posterior estimate using a Bayesian setting. No additional constraints is imposed since in Bayesian modeling it is not crucial to ensure identifiability of latent parameters as long as the quantities in which we are interested (the prob of cancer risk in each subgroup) are identifiable (Knorr-Held and Rain, 2001).

***Priors***

I use prior settings introduced in chapter 3 for the number of stage $s$ and constant $\theta$ in the Armitage-Doll model. In addition, I use a Gaussian autoregressive prior model in the forward direction to smooth effects on period and cohort. Non-informative priors will be defined for the first two parameters for period and cohort. Suppose the projection is need for $N$ future periods, so for the $P+N$ period effects:

$$\beta_1 \sim N(0,\, 1000000\frac{1}{\tau_\beta}),$$

$$\beta_2|\beta_1 \sim N(0,\, 1000000\frac{1}{\tau_\beta}),$$

$$\beta_p|\beta_{1,\dots,p-1} \sim N(2\beta_{p-1} - \beta_{p-2}, \frac{1}{\tau_\beta})$$

$$3 \le p \le P+N,$$

$$\tau_\beta \sim gamma(0.0001, 0.0001).$$

For $C+N$ cohort effects:

$$\gamma_1 \sim N(0,\, 1000000\frac{1}{\tau_\gamma}),$$

$$\gamma_2|\gamma_1 \sim N(0,\, 1000000\frac{1}{\tau_\gamma}),$$

$$\gamma_c|\gamma_{1,\dots,c-1} \sim N(2\gamma_{c-1} - \gamma_{c-2}, \frac{1}{\tau_\gamma})$$

$$3 \le c \le C+N,$$

$$\tau_\gamma \sim gamma(0.0001, 0.0001).$$

***Likelihoods***

The Surveillance, Epidemiology and End Results (SEER) database at National Cancer Institute (NCI) provides cancer mortality data (i.e., cancer death count and population size) by their age group (5 years interval) and period (every year from 1969 to 2007). I then divide the cancer mortality data into 8 periods of 5 years each ($J = 8$) summarizing the death count and population size within each period. Since the Armitage-Doll model fit the cancer data well for the ages 25 to 74, I delete those observations with age less than 24 or greater than 74. The total number of age group is 10 (I=10). Therefore, the total number of cohort effect is 17 ($I + J - 1 = 17$) and the cohort ($k$) assignment for each observation is determined by its period ($j$) and age ($i$) where $k = j - i + I$. The index of age, period and cohort effects is listed in Table 4.1.

Table 4.1: Index of age, period and cohort effects in cancer mortality data

| Index (n) | Age (i) | Period (j) | Cohort (k) |
|---|---|---|---|
| 1 | 25-29 (i=1) | 1969-1973 (j=1) | 1942-1946 (k=10) |
| 2 | 25-29 (i=1) | 1974-1978 (j=2) | 1947-1951 (k=11) |
| 3 | 25-29 (i=1) | 1979-1983 (j=3) | 1952-1956 (k=12) |
| 4 | 25-29 (i=1) | 1984-1988 (j=4) | 1957-1961 (k=13) |
| 5 | 25-29 (i=1) | 1989-1993 (j=5) | 1962-1966 (k=14) |
| 6 | 25-29 (i=1) | 1994-1998 (j=6) | 1967-1971 (k=15) |
| 7 | 25-29 (i=1) | 1999-2003 (j=7) | 1972-1976 (k=16) |
| 8 | 25-29 (i=1) | 2004-2008 (j=8) | 1977-1981 (k=17) |
| ... | ... | ... | ... |
| ... | ... | ... | ... |
| ... | ... | ... | ... |
| 73 | 70-74 (i=10) | 1969-1973 (j=1) | 1897-1901 (k=1) |
| 74 | 70-74 (i=10) | 1974-1978 (j=2) | 1902-1906 (k=2) |
| 75 | 70-74 (i=10) | 1979-1983 (j=3) | 1907-1911 (k=3) |
| 76 | 70-74 (i=10) | 1984-1988 (j=4) | 1910-1916 (k=4) |
| 77 | 70-74 (i=10) | 1989-1993 (j=5) | 1917-1921 (k=5) |
| 78 | 70-74 (i=10) | 1994-1998 (j=6) | 1922-1926 (k=6) |
| 79 | 70-74 (i=10) | 1999-2003 (j=7) | 1927-1931 (k=7) |
| 80 | 70-74 (i=10) | 2004-2008 (j=8) | 1932-1936 (k=8) |

The Poisson distribution is used to model the cancer death count $Y_n$ at index $n$, i.e.

$$Y_n \sim \ Poisson(\mu_n)$$

and

$$\log(\mu_n) = \log(Popu_n) + \alpha_{ci} + \beta_{cj} + \gamma_{ck}$$

where $Popu_n$ is the population size at index $n$. The relationship between index $n$ and the values for $i$, $j$, $k$ is included in Table 4.1.

## 4.4 Apply TSCE Carcinogenesis Model into APC Model

The hazard function derived from the TSCE model is introduced to the APC model as well. In contrast to the simple format of the hazard function in the Armitage-Doll model, the hazard function from the TSCE model is more difficult. Four derived parameters will be summarized in the TSCE model, the rate of initiation, $\nu$, the rate of division, $\alpha$, and death, $\beta$, of initial cells, and the rate of malignant conversion, $\mu$. The hazard function in the TSCE model is given by Moolgavkar [1979, 1981, 1990, 2009] as

$$h(t) = \frac{\nu}{\alpha} pq \frac{e^{-qt} - e^{-pt}}{qe^{-pt} - pe^{-qt}},$$

where $p$ and $q$ are the roots of a quadratic equation, with $p + q = -(\alpha - \beta - \mu)$ and $pq = \alpha\mu$. Three estimated parameters $p$, $q$, and $r \equiv \nu/\alpha$ will be treated in the model. As we can see, the TSCE model requires one parameter more than the Armitage-Doll model. However, it adds additional complexity to the model, incorporates the stochastic feature of the carcinogenic process and characterizes the kinetics of clonal expansion.

***Priors***

In cancer development, the mutation rate $\mu$ is considered much smaller than the cell division rate $\alpha$ and the cell death rate $\beta$ [Moolgavkar and Meza, 2009]. Therefore,

$$p + q \approx -(\alpha - \beta).$$

The quantity $\alpha-\beta$ represents the net proliferation rate of intermediate cells [Moolgavkar and Luebeck, 1992]. For colon cancer, Moolgavkar and Luebeck [1992] estimated the quantity of $\alpha-\beta$ as approximately 0.107 per cell per year which indicates on the average level a stem cell experiences an effective clonal expansion approximately once every 10 years. In addition, Moolgavkar and Luebeck [1992] also estimated the quantity of $\beta/\alpha$ as close to 1 through fitting the two and three-stage clonal expansion carcinogenesis models. The parameter $q$ is much smaller than $p$ and can be estimated as

$$q \simeq \mu/(1 - \beta/\alpha).$$

Based on Moolgavkar and Luebeck's calculation [1992], the estimated mutation rates and initiation rates are on the magnitude of $10^{-6}$ to $10^{-8}$ . Therefore, I summarize the preliminary information and make our noninformative priors as follows:

$$r \sim Uniform(0, 10^{-5}),$$

$$p \sim Uniform(-0.2, 0),$$

$$q \sim Uniform(0, 10^{-5}).$$

## 4.5 Results

### 4.5.1 Convergence Diagnostics

In theory, Markov chain Monte Carlo (MCMC) samples after infinite runs will eventually converge to the stationary distribution, which is assumed to be the true posterior distribution [Carlin and Louis, 2009]. The way to know whether the samples have actually converged is to implement the convergence monitoring and diagnosis tests, both visual and statistical. The convergence diagnosic test is a way to detect the failure of the convergence, not a proof of convergence.

***Gelman and Rubin Diagnostic***

Among all the convergence diagnostic test, the Gelman and Rubin multiple sequence diagnostic test [Gelman and Rubin, 1992] is perhaps the most popular approach. Here I run $m \geq 2$ chains of length $2n$ from overdispersed starting points and discard the first $n$ samples in each chain. The within-chain ($W$) and between-chain ($B$) variances are calculated as follows:

$$W = \frac{1}{m}\sum_{j=1}^{n} s_j^2$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_j)^2.$$

Here $\theta_{ij}$ is the $i^{th}$ MCMC samples at $j^{th}$ chain and $\bar{\theta}_j$ is the mean of MCMC samples at$j^{th}$ chain. Basically, $s_j^2$ is the variance for the $j^{th}$ chain and $W$ is the mean of the variances of each chain.

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2$$

where

$$\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j.$$

The variance of the stationary distribution is measured as a weighted average of $W$ and $B$.

$$V\hat{ar}(\theta) = (1 - \frac{1}{n})W + \frac{1}{n}B.$$

Then the convergence is monitored by the estimated scale reduction factor at

$$\hat{R} = \sqrt{\frac{V\hat{ar}(\theta)}{W}},$$

where $\hat{R}$ close to 1 suggests good convergence [Carlin and Louis, 2009]. Brooks and Gelman[1998] extended the Gelman and Rubin approach to generalize a multivariate analysis for simultaneous convergence diagnosis for every parameter in a model [Carlin and Louis, 2009]. The Brooks-Gelman-Rubin approach is available in the latest WinBUGS software and is shown as "bgr diag" in its sample monitor tool. From Table 4.2, I conclude the convergency of the posterior estimates is met since the estimated scale reduction factors $\hat{R}$ are close to 1 for all parameters. In addition, I select $\tau_a$, $\tau_p$ and $\tau_c$ in Figure 4.1 and $\alpha_{c4}$, $\beta_{c4}$ and $\gamma_{c4}$ in Figure 4.2 as examples to demonstrate the convergency visually.

Table 4.2: Brooks-Gelman-Rubin Convergence Diagnostics

| $Parameter$ | $\hat{R}$ | $Parameter$ | $\hat{R}$ |
|---|---|---|---|
| $\alpha_{c1}$ | 1.000475 | $\gamma_{c1}$ | 1.003891 |
| $\alpha_{c2}$ | 1.003729 | $\gamma_{c2}$ | 1.005464 |
| $\alpha_{c3}$ | 1.001545 | $\gamma_{c3}$ | 1.007351 |
| $\alpha_{c4}$ | 1.000746 | $\gamma_{c4}$ | 1.007656 |
| $\alpha_{c5}$ | 1.000757 | $\gamma_{c5}$ | 1.007936 |
| $\alpha_{c6}$ | 1.001491 | $\gamma_{c6}$ | 1.005536 |
| $\alpha_{c7}$ | 1.003153 | $\gamma_{c7}$ | 1.004627 |
| $\alpha_{c8}$ | 1.005947 | $\gamma_{c8}$ | 1.003926 |
| $\alpha_{c9}$ | 1.008871 | $\gamma_{c9}$ | 1.001019 |
| $\alpha_{c10}$ | 1.008653 | $\gamma_{c10}$ | 1.00023 |
| $\beta_{c1}$ | 1.002581 | $\gamma_{c11}$ | 1.000781 |
| $\beta_{c2}$ | 1.002195 | $\gamma_{c12}$ | 1.000707 |
| $\beta_{c3}$ | 1.001924 | $\gamma_{c13}$ | 1.000048 |
| $\beta_{c4}$ | 1.000472 | $\gamma_{c14}$ | 1.000017 |
| $\beta_{c5}$ | 1.000168 | $\gamma_{c15}$ | 1.001583 |
| $\beta_{c6}$ | 1.002058 | $\gamma_{c16}$ | 1.006247 |
| $\beta_{c7}$ | 1.001891 | $\gamma_{c17}$ | 1.005061 |
| $\beta_{c8}$ | 1.003495 | $\gamma_{c18}$ | 1.003596 |
| $\beta_{c9}$ | 1.000556 | $\gamma_{c19}$ | 1.002865 |
| $\beta_{c10}$ | 1.000424 | $\gamma_{c20}$ | 1.002452 |
| $\beta_{c11}$ | 1.000383 | $\tau_a$ | 1.036245 |
| $s$ | 1.001584 | $\tau_c$ | 1.000367 |
|  |  | $\tau_p$ | 1.00019 |

### 4.5.2 Model Comparison

To compare different models in fitting cancer incidence and mortality data, I use the deviance information criterion (DIC) [Spiegelhalter et al., 2002] which is the posterior average of the deviance plus a measure of complexity. The DIC is an addition of two statistics $\bar{D}$ and $p_D$, where $\bar{D}$ is the posterior mean deviance which can be computed from the distribution of posterior deviance $D(\lambda_{iap})$ and $p_D$ is the effective number of parameters which is used to penalize increasing model complexity [Schmid and Held, 2004]. The posterior deviance is computed as

$$D(\lambda_{ij}) = -2\sum_{ij}(l(\lambda_{ij}) - l(\hat{\lambda}_{ij})),$$

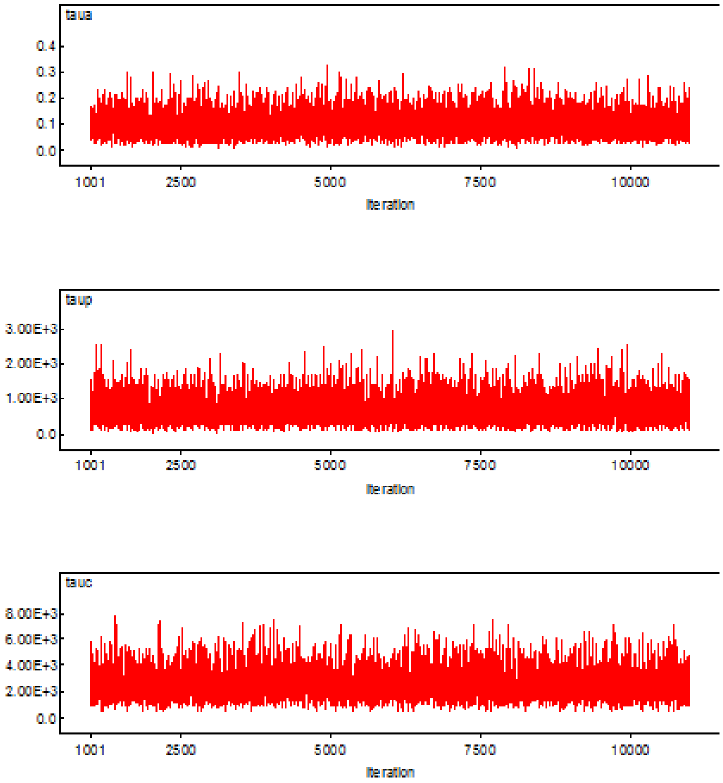

Figure 4.1: Convergency plots for posterior estimate $\tau$

where $l(\lambda_{ij})$ is the observed log likelihood and $\hat{\lambda}_{ij}$ is the estimated cancer mortality rate at age $i$ and period $j$. The effective number of parameters can be derived as

$$p_D = \bar{D} - D(\bar{\lambda}_{ij})$$

where $\bar{\lambda}_{ij}$ denotes the posterior mean of $\lambda_{ij}$. The DIC is calculated as

$$DIC = p_D + \bar{D}.$$

Table 4.3 displays the DIC and posterior estimates for model parameters. When the Armitage-Doll carcinogenesis model is considered in the Bayesian extended APC model, we get DIC equals to 1287 and the posterior estimate for number of stage $(s)$ at approx-

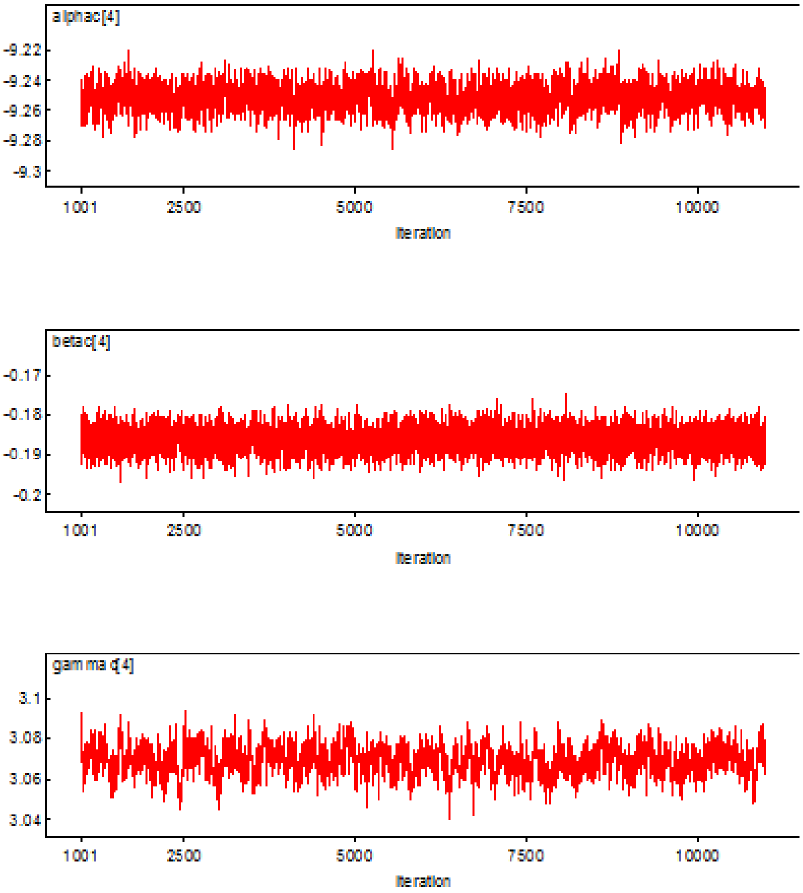

Figure 4.2: Convergency plots for posterior estimate apc effects

imately 5. The DIC value is almost the same when the TSCE carcinogenesis model is incorporated into Bayesian extended APC model. For both Bayesian extended APC models, autoregressive priors are chosen for period and cohort effects. Model constraints as illustrated before are added. Compared with the DIC value (DIC=1286.43) derived from conventional Bayesian APC model where all age, period and cohort effects are taking autoregressive priors, we don't see much difference on the improvement of DIC values between these two models. However, the age effect in the extended BAPC model has more sound biological meanings since we replace it with the hazard function from the carcinogenesis model.

To consider alternative prior settings for TSCE model parameters, three different uniform

Table 4.3: Model comparisons using DIC

| Extended BAPC Model | DIC | Parameter | Posterior Estimate | 95% HPD |
|---|---|---|---|---|
| Armitage-Doll | 1287.0 | $s$ | 4.867 | [4.237, 5.5] |
| TSCE | 1286.98 | $p$ | -0.1495 | [-0.1984, -0.05376] |
| | | $q$ | $4.9 \times 10^{-4}$ | $[2.7 \times 10^{-4}, 9.8 \times 10^{-4}]$ |
| | | $r$ | $7.5 \times 10^{-6}$ | $[2.6 \times 10^{-6}, 9.9 \times 10^{-6}]$ |

priors for model parameter p were chosen. Result shows that the posterior inference varied when we change the uniform intervals in the prior setting. From the biological perspective, we chose the first prior setting where p is uniformed distributed between -0.2 and 0.

### 4.5.3 Rate Estimation and Projection

In the Bayesian extended APC model, I also compute the estimated mortality rate at each iteration of MCMC samplings as

$$\lambda[i, j] = 100,000 \times \exp(\alpha_{ci} + \beta_{cj} + \gamma_{ck})$$

where

$$i = 1, ..., I, \ j = 1, ..., J + N, \ k = j - i + I,$$

$N$ denotes the future period where the projected rates are computed. Here, I choose $N = 3$ to represent the future periods at 2009-2013, 2014-2018, and 2019-2023. The posterior estimates for age, period and cohort effects are listed in Table 4.4.

Through visually checking the history of posterior estimates on the mortality rates, I found the estimated mortality rates converge. Figure 4.3 compares the observed cancer mortality rate with the estimated rates derived from the Bayesian extended APC model using Armitage-Doll model. In the younger age groups (25-29, 30-34 and 35-39), the estimated rates seem to deviate a little from the observed rates. A closer look at the cancer mortality rates at younger age groups, I found those rates at very small scale and the difference between the observed rates and estimated rates is magnified as compared with its absolute small value. For example, the observed colon cancer death count at age group 30-34 in

Table 4.4: Posterior estimates for age, period and cohort effects in the Bayesian extended APC models

| *Age* | $\hat{\alpha}_{ci}$ | *Period* | $\hat{\beta}_{cj}$ | *Cohort* | $\hat{\gamma}_{ck}$ |
|---|---|---|---|---|---|
| 25-29 (i=1) | -9.663 | 1969-1973 (j=1) | -2.004 | 1897-1901 (k=1) | 4.872 |
| 30-34 (i=2) | -9.492 | 1974-1978 (j=2) | -1.387 | 1902-1906 (k=2) | 4.271 |
| 35-39 (i=3) | -9.369 | 1979-1983 (j=3) | -0.776 | 1907-1911 (k=3) | 3.672 |
| 40-44 (i=4) | -9.252 | 1984-1988 (j=4) | -0.186 | 1912-1916 (k=4) | 3.069 |
| 45-49 (i=5) | -9.154 | 1989-1993 (j=5) | 0.332 | 1917-1921 (k=5) | 2.447 |
| 50-54 (i=6) | -9.116 | 1994-1998 (j=6) | 0.843 | 1922-1926 (k=6) | 1.835 |
| 55-59 (i=7) | -9.171 | 1999-2003 (j=7) | 1.361 | 1927-1931 (k=7) | 1.218 |
| 60-64 (i=8) | -9.302 | 2004-2008 (j=8) | 1.818 | 1932-1936 (k=8) | 0.591 |
| 65-69 (i=9) | -9.497 | 2009-2013 (j=9) | 2.275 | 1937-1941 (k=9) | -0.041 |
| 70-74 (i=10) | -9.728 | 2014-2018 (j=10) | 2.733 | 1942-1946 (k=10) | -0.685 |
| | | 2019-2023 (j=11) | 3.189 | 1947-1951 (k=11) | -1.3 |
| | | | | 1952-1956 (k=12) | -1.884 |
| | | | | 1957-1961 (k=13) | -2.475 |
| | | | | 1962-1966 (k=14) | -3.055 |
| | | | | 1967-1971 (k=15) | -3.622 |
| | | | | 1972-1976 (k=16) | -4.181 |
| | | | | 1977-1981 (k=17) | -4.73 |
| | | | | 1982-1986 (k=18) | -5.28 |
| | | | | 1987-1991 (k=19) | -5.829 |
| | | | | 1992-1996 (k=20) | -6.379 |

1969-1973 is about 1.2 per 100,000 while the estimated death count is 1 per 100,000 which makes their difference at only 2 per 1 million people. Considering more than 90% of colon cancer cases are in people age 50 and older, we are more interested in precisely estimating the mortality rates among older age groups. In Figure 4.3, I found high consistency between the estimated rates and observed rates for those older age groups ($\geq 45$). It shows the Bayesian extended APC can be used in determining the colon cancer mortality rates among age, period and cohort. In addition, I project the mortality rates heading downward in future periods (2009-2013, 2014-2018, and 2019-2023). Our projection on the decreasing trend of colon cancer mortality rate can be supported by the fact that colon cancer is often highly treatable and the promising development of cancer screening methods in the near future.

The results displayed here are computed from 10000 iterations after a burn-in of 1000 iterations in WinBUGS software. MCMC samples were drawn to derive the posterior distribution about model parameters and function of these parameters, such as rates. The estimated

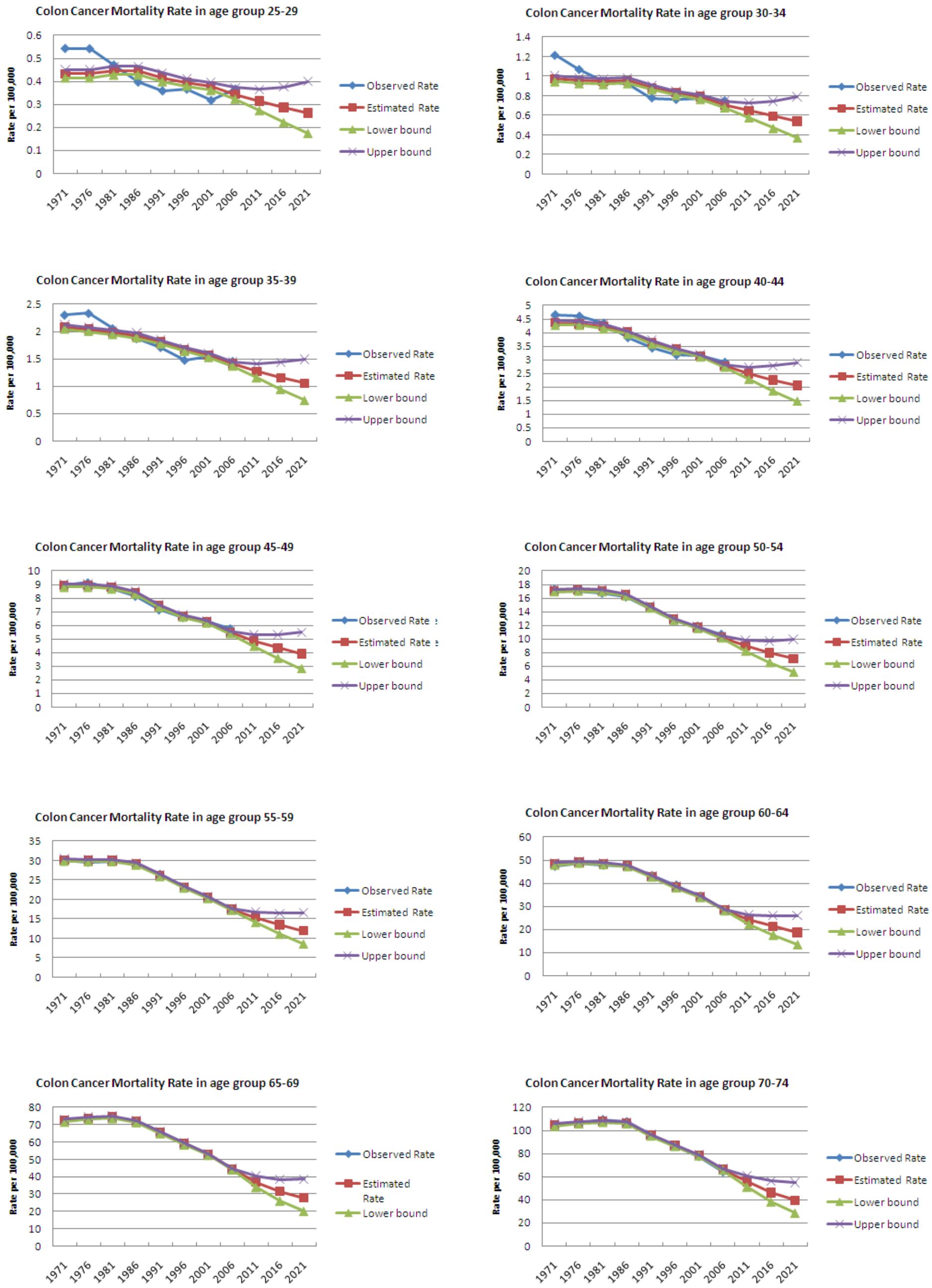

Figure 4.3: Posterior estimates for the colon cancer mortality rates

mortality rates are calculated by medians of the parameters (age, period and cohort) and their 95% credible intervals are based on the 95% highest posterior density (HPD) regions of these parameters. Figure 4.3 shows the posterior estimates and their 95% HPD for the estimated colon cancer mortality rates for each age group at different period. The observed rates are also displayed in Figure 4.3.

# Chapter 5

# Bayesian extended Area-Age-Period-Cohort (AAPC) Model

## 5.1 Background

Disease mapping is an important topic in epidemiology to study the space and time variation for the risk of disease. Many general or more heavily parameterized Bayesian models have been proposed to study the spatio-temporal mappings of disease rates [Waller et al., 1997, Carlin and Louis, 2009]. Disease incidence and mortality data may vary considerably among different geographical regions. Areas with a small population could result in an extreme observation of incidence and mortality due to the small population at risk. Therefore, to consider the high sampling variability in small areas when estimating disease incidence and mortality rates across each region, we usually add a weight matrix to the set of model parameters to smooth variation among neighboring areas and improve the estimation in the small regions by borrowing strength from their adjacent regions [Buenconsejo and Holford, 2008].

In an analysis of lung cancer rates in Tuscany, Lagazio, Dreassi, and Biggeri [2003] proposed a full spatio-temporal Bayesian model which include main effects of area, age, period and cohort as well as area-period and area-cohort interactions. Gaussian first-order and second-order random walk priors (RW1, RW2) were given to age, period and cohort effects in Lagazio's full model. To better explain the biological meaning of age effect in determining cancer mortality rate, multistage carcinogenesis models will be considered in AAPC models to replace the main age effects. It is based on the assumption of fundamental role of age in determining the cancer incidence rates and subsidiary roles of period and cohort in modulating the age effect [Jeon and Moolgavkar, 2006].

## 5.2 Significance and Innovation

The AAPC model provides a general framework to jointly study the evolution in time and the spatial pattern of the risk of disease [Lagazio et al., 2003]. The interaction terms over area can reduce the identifiability burden in the standard APC model [Clayton and Schifflers, 1987]. Gaussian RW1 and RW2 structures on the age, period and cohort effects can improve model estimation and prediction of future mortality rates [Schmid and Held, 2004, Knorr-Held and Rainer, 2001]. Model constraints can be further implemented in the Bayesian framework to handle the identifiability issues in APC models.

In this study, I develop a new Bayesian extended AAPC model where multistage carcinogenesis models are introduced into the AAPC model to incorporate more biological meaning of the age effects in studying the spatio-temporal pattern of cancer mortality rates. The prior means of age effects in the AAPC model are replaced by the log transformation of hazard functions derived from the Armitage-Doll multistage carcinogenesis model and the TSCE model. The proposed extended AAPC model is also compared with the conventional AAPC model where age effects are assigned as RW1 or RW2 priors in fitting cancer mortality data. Model selection procedures (DIC) are implemented to compare the performance of several alternative models.

## 5.3 Bayesian AAPC Model

Lagazio et al. [2003] introduced a full area-age-period-cohort (AAPC) model to study the spatio-temporal pattern of disease risk. The model incorporates the main effect of area, age, period and cohort, and interaction terms such as the area-cohort and area-period interactions. The model is as follows:

$$\log(\lambda_{iap}) = \nu_i + \mu_i + \theta_a + \gamma_p + \delta_c + \varphi_{ip} + \varphi_{ic},$$

where $\lambda_{iap}$ is the relative risk for the $\alpha^{th}$ age group and the $p^{th}$ calendar period in the $i^{th}$ area, $\nu_i$ is the unstructured spatial term for the spatial heterogeneity effects, $\mu_i$ is the structured spatial term to incorporate spatial clustering effect, $\theta_a$, $\gamma_p$ and $\delta_c$ are the age, period, and cohort main effects, $\varphi_{ip}$ is the space-period interaction and $\varphi_{ic}$ is the space-cohort interaction.

In the prior assumptions, $\nu_i$ is an unstructured area effect, and $\mu_i$ follows an intrinsic conditional Gaussian spatial autoregressive model (ICAR). For the unstructured spatial effect $\nu_i$, we usually assign a homoscedastic distribution to them such as $\nu_i \sim N(0, \sigma^2)$ where $\sigma$ can be further defined as a hyperprior with an inverse-gamma distribution, i.e., $\sigma \sim IG(\alpha_\nu, \beta_\nu)$. For the structured spatial effect $\mu_i$, Congdon [2006a] illustrated the joint distribution for spatial effects $\mu = (\mu_1, ..., \mu_n)$ derived from their pairwise differences and a variance term $\kappa$ as follows:

$$P(\mu_1, ..., \mu_n) \propto \exp[-0.5\kappa^{-1} \sum_i \sum_j c_{ij}(\mu_i - \mu_j)^2],$$

where the $c_{ij}$ are contiguity measures based on spatial adjacency between areas $i$ and $j$.

$$c_{ij} = \begin{cases} 1, & \text{if areas } i \text{ and } j \text{ are first-order neighbours;} \\ 0, & \text{otherwise.} \end{cases}$$

Congdon [2006a] further demonstrated that the conditional prior of $\mu_i$ given the remaining spatial effects $\mu_j$ where $j \neq i$ follows a normal distribution.

$$P(\mu_i|\mu_{[i]}) \sim N(\omega_i, \tau_i^{-1}),$$

where $\mu_{[i]}$ refers to remaining spatial effects $\mu_j$ where $j \neq i$. The weighted average $\omega_i$ is computed as

$$\omega_i = \frac{\sum_j c_{ij}\mu_j}{\sum_j c_{ij}} = \sum_j \omega_{ij}\mu_j$$

and

$$\tau_i^{-1} = \frac{\kappa}{\sum_j c_{ij}}$$

are conditional variances. This is recognized as the intrinsic conditional autoregressive (ICAR) prior since the conditional distribution involves row-standardised weights [Congdon, 2006a].

The effects $\gamma_p$ and $\delta_c$ are modeled as Gaussian RW1 and RW2 [Berzuini and Clayton, 1994]. The Kronecker product of these structure matrices $K_{\mu p} = K_\mu \otimes K_p (or\ K_{\mu c} = K_\mu \otimes K_c)$ defines the structure matrix for the joint prior and provides a prior for the interaction terms $\varphi_{ip}(or\ \varphi_{ic})$ [Congdon, 2006a]. For example, the joint spatio-period interactions $\varphi = (\varphi_{ip}, i = 1, ..., N, p = 1, ..., P)$ are taken as $\varphi \sim N(0, \tau_\varphi K_{\mu p})$. Usually, the hyperprior $\tau_\varphi$ is set as a noninformative gamma distribution [Lagazio et al., 2003], i.e., $\tau_\varphi \sim Gamma(\alpha_\tau, \beta_\tau)$. The structure matrix $K_{\mu p}$ is the Kronecker product of $K_\mu$ and $K_p$ where for the RW1 prior in the period effect

$$K_{p[cd]} = \begin{cases} -1, & \text{if periods c and d are adjacent;} \\ 0, & \text{if periods c and d are not adjacent;} \\ 1, & \text{if c=d=1 or c=d=P;} \\ 2, & \text{if c=d=k where k is not equal to 1 or P.} \end{cases}$$

and for the spatial effect

$$K_{\mu[ij]} = \begin{cases} -1, & \text{if areas i and j are adjacent;} \\ 0, & \text{for non-adjacent areas;} \\ L_i, & \text{when i=j.} \end{cases}$$

and $L_i$ is the cardinality of area $i$ which is the measure of its total number of neighbors. Lagazio [2003] derived the conditional distribution of the spatio-period interaction term $\varphi_{ip}$ at area $i$ and period $p$ given the remaining terms have Normal distribution with mean

$$\bar{\varphi}_{ip} = \begin{cases} \varphi_{i,p+1} + \frac{\sum_{j \in S_i} \varphi_{jp}}{n_i} - \frac{\sum_{j \in S_i} \varphi_{j,p+1}}{n_i}, & \text{if } p = 1 \\ \frac{\varphi_{i,p+1}+\varphi_{i,p-1}}{2} + \frac{\sum_{j \in S_i} \varphi_{jp}}{n_i} - \frac{\sum_{j \in S_i}(\varphi_{j,p+1}+\varphi_{j,p-1})}{2n_i}, & \text{if } p = 2, ..., P-1 \\ \varphi_{i,p-1} + \frac{\sum_{j \in S_i} \varphi_{jp}}{n_i} - \frac{\sum_{j \in S_i} \varphi_{j,p-1}}{n_i}, & \text{if } p = P \end{cases}$$

and precision

$$\tau_{ip} = \begin{cases} n_i \tau_{\varphi}, & \text{if } p = 1 \text{ or } p = P \\ 2n_i \tau_{\varphi}, & \text{if } p = 2, ..., P-1 \end{cases}$$

where $S_i$ is the set of areas adjacent to area $i$, $n_i$ is the number of areas adjacent to area $i$ and hyperprior $\tau_{\varphi}$ is taken as noninformative Gamma distribution.

## 5.4 Bayesian extended AAPC Model - Introduction of Carcinogenesis Model into AAPC Model

Inspired by recent works on using carcinogenesis models into APC model to study cancer trends [Moolgavkar and Meza, 2009, Jeon and Moolgavkar, 2006], I use various forms of multistage carcinogenesis models in this study to represent age effects in the extended AAPC model. Due to its flexibility in prior settings and advantages in handling identifiability problems, the Bayesian method is used to obtain posterior estimates of model parameters in the AAPC model. The likelihood and the choice of priors are discussed below.

***Likelihood***

Cancer death counts $y_{iap}$ at area $i$, age $a$ and period $p$ are modelled as Poisson distribution with parameters rate $\lambda_{iap}$ and population size $N_{iap}$, i.e.,

$$y_{iap} \sim \ Poisson(N_{iap}\lambda_{iap}),$$

where $\lambda_{iap}$ is the relative risk for the $\alpha^{th}$ age group and the $p^{th}$ calendar period in the $i^{th}$ area, and can be modeled as

$$\log(\lambda_{iap}) = \nu_i + \mu_i + \theta_a + \gamma_p + \delta_c + \varphi_{ip} + \varphi_{ic}.$$

Therefore, we can write the likelihood as

$$\begin{array}{lll} P(y|\lambda) & \propto & \prod_{i,a,p} e^{N_{iap}\lambda_{iap}}(N_{iap}\lambda_{iap})^{y_{iap}} \\ & \propto & \prod_{i,a,p} e^{\lambda_{iap}}(\lambda_{iap})^{y_{iap}} \end{array},$$

where $\lambda = (\lambda_{iap})$ includes all $i$, $a$, and $p$ and represents all the parameters which need to have prior distributions. $N_{iap}$ is a known quantity which can be ignored here.

***Priors***

The joint prior distribution can be written as

$$P(\lambda) = P(\nu, \mu, \theta, \gamma, \delta, \varphi_1, \varphi_2),$$

where $\nu$ is the joint spatial unstructured effect, $\mu$ is the joint spatial structured effect, $\theta$ is the age effect, $\gamma$ is the period effect, $\delta$ is the cohort effect, $\varphi_1$ and $\varphi_2$ are the joint spatial-period and spatial-cohort interactions, respectively.

*Age effects*

As we introduced before, we apply the carcinogenesis model into the AAPC model by replacing the age effects with hazard functions derived from the carcinogenesis model. For example, the age effect $\theta_a$ at age group $a$ is given a noninformative prior as below:

$$\theta_a \sim N(\bar{\theta}_a, \tau_a)$$

where

$$\bar{\theta}_a = \log(h(t_a)),$$

and

$$\tau_a \sim Gamma(0.001, 0.001).$$

Armitage and Doll [1954] assumed multiple transformations happened in stages for a cell to grow into a cancerous tumor. In the Armitage-Doll model, $h(t) = ct^{s-1}$ where $s$ is the number of stages, $t_a$ is the age and $c$ is constant. Therefore, we have the following age effects in a nonlinear function of age,

$$\bar{\theta}_a = c + (s-1) * \log(t_a).$$

Hyperpriors are given to $c$ and $s$ at

$$c \sim N(0, \tau_c)$$

$$s \sim N(5, \tau_s)$$

where Gamma hyperpriors are given to two precision terms as below:

$$\tau_c \sim Gamma(0.001, 0.001)$$

and

$$\tau_s \sim Gamma(0.001, 0.001).$$

The hazard function derived from the TSCE model is introduced to the AAPC model as well. In contrast to the simple format of the hazard function in the Armitage-Doll model, the hazard function from the TSCE model is more difficult. Four derived parameters will be summarized in the TSCE model, the rate of initiation, $\nu$, the rate of division, $\alpha$, and death, $\beta$, of initial cells, and the rate of malignant conversion, $\mu$. The hazard function in the TSCE model is given by Moolgavkar [1979, 1981, 1990, 2009] as

$$h(t) = \frac{\nu}{\alpha} pq \frac{e^{-qt} - e^{-pt}}{qe^{-pt} - pe^{-qt}},$$

where $p$ and $q$ are the roots of a quadratic equation, with $p+q = -(\alpha - \beta - \mu)$ and $pq = \alpha\mu$. Three estimated parameters $p$, $q$, and $r \equiv \nu/\alpha$ will be treated in the model. As we can see, the TSCE model requires one parameter more than the Armitage-Doll model. However, it adds additional complexity to the model by incorporating the stochastic feature of the carcinogenic process and characterizing the kinetics of clonal expansion. The prior mean for age effect is written as

$$\bar{\theta}_a = \log(rpq) + \log(e^{-qt_a} - e^{-pt_a}) - \log(qe^{-pt_a} - pe^{-qt_a}).$$

For the TSCE parameters, we can assign noninformative priors as follows:

$$r \sim Uniform(0, 10^{-5}),$$

$$p \sim Uniform(-0.2, 0),$$

$$q \sim Uniform(0, 10^{-5}).$$

*Spatial effects*

The prior for $\nu_i$ is taken as

$$\nu_i \sim N(0, \tau)$$

where $\tau$ is a hyper prior which is defined as gamma distribution

$$\tau \sim Gamma(0.001, 0.001).$$

Independence is assumed to all spatial unstructured effects.

The intrinsic Gaussian conditional autoregressive (ICAR) priors are considered for the spatial structured effects. Congdon [2006a] showed the joint prior for the structured spatial effect $\mu$ can be taken as

$$P(\mu_1, ..., \mu_n) \propto \exp[-0.5\kappa^{-1} \sum_{i \sim j} c_{ij}(\mu_i - \mu_j)^2].$$

In Winbugs, we use the distribution ***car.normal*** to assign ICAR priors to the joint spatial structured effects $\mu = (\mu_1, ..., \mu_C)$.

*Period and Cohort effects*

The Gaussian RW2 [Berzuini and Clayton, 1994] priors are assigned to the period and cohort effects, which result in a joint form as improper multivariate normals

$$\gamma \sim N(0, \tau_p K_p^-),$$

$$\delta \sim N(0, \tau_p K_c^-),$$

where $K_p$ is the structure matrix for period effects with generalized inverse $K_p^-$, and $K_c$ is the structure matrix for cohort effects with generalized inverse $K_c^-$. Suppose there are six period effects ($P = 6$), we can derive the structure matrix in the RW2 priors for period effects ($K_p$) at

$$K_p = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

and structure matrix for cohort effect ($K_c$) whose pattern is similar to those for period effects.

*Interactions*

The joint distribution for spatio-temporal interactions is modeled as a multinormal distribution. For example, the joint spatio-period interactions $\varphi = (\varphi_{ip}, i = 1, ..., N, p = 1, ..., P)$ are taken as $\varphi \sim N(0, \tau_\varphi K_{\mu p})$. The structure matrix $K_{\mu p}$ is the Kronecker product of $K_\mu$ for the spatial effect and $K_p$ for the period effect, such as

$$K_{\mu p} = K_\mu \otimes K_p.$$

Since both $K_\mu$ and $K_p$ are symmetric and singular matrices, their Kronecker product $K_{\mu p}$ is symmetric and singular as well. Therefore, the joint density for spatio-temporal effects is improper [Waller et al., 1997]. Carlin and Louis [2009] pointed out that the proper posterior may not always result, thus extra care must be taken when using improper priors. As an alternative solution [Congdon, 2006b], we consider the parsimonious product interactions schemes with generic form

$$\alpha_i \beta_p, i = 1, ..., N, p = 1, ..., P,$$

where $\alpha_i$ is the structured spatial effects, subject to $\sum_i \alpha_i = 0$, while

$$\beta_p = \exp(\eta_p)/[1 + \sum_{p=1}^{P-1} \exp(\eta_p)], \quad p = 1, ..., P-1,$$

$$\beta_P = 1/[1 + \sum_{p=1}^{P-1} \exp(\eta_p)],$$

and $\eta_p$ is the period effects.

***Posterior***

Given the likelihood and prior density, we can derive the posterior distribution for model parameters

$$\begin{array}{rcl} P(\lambda|y) & \propto & P(y|\lambda)P(\lambda) \\ & \propto & \prod_{i,a,p} e^{\lambda_{iap}} (\lambda_{iap})^{y_{iap}} P(\nu, \mu, c, s, \gamma, \delta, \varphi_1, \varphi_2) \end{array}.$$

Markov chain Monte Carlo (MCMC) methods have been used to sample from the posterior density. Due to the non-identifiability issues in the model parameters, certain constraints are added in the model to improve numerical stability and mixing [Bray, 2002]. Area, age, period and cohort effects are each adjusted by subtracting their respective means [Bray, 2002]. Winbugs software is used to derive the posterior distribution about model parameters from 10000 iterations after a burn-in of 1000 iterations. The convergence plots of model parameters are provided along with convergence diagnostics. The posterior estimates of main effects and interactions in the AAPC model are summarized by the mean and 90% highest posterior density (HPD) derived from posterior samples.

Statistical software R and WinBUGS are used in this study.

## 5.5 Example 1 - Lung Cancer Mortality in Iowa

The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. There are 17 SEER registries in the U.S. which cover about 26% of the US population. To study the performance of the AAPC model at the county level in the specified state, we chose states which have full coverage in the SEER surveillance network. Lung cancer mortality records in the state of Iowa are retrieved from SEER database for this study. The death counts are aggregated by county, age group, and period. The state of Iowa has 99 counties and the county adjacent matrix is obtained using the open-source Geographic Information Software - GeoDa. Since the Armitage-Doll model fit the cancer data well for the age between 25 and 74, I drop those observations with age less than 24 or greater than 74. The total number of age group is 10. To calculate the mortality rate, I visited the US Census website to download files containing population sizes per age group in each county of Iowa in the 1980s, 1990s and 2000s. I took five consecutive calendar years as one period group and group them into six period groups (1980-1984, 1985-1989, 1990-1994, 1995-1999, 2000-2004, 2005-2008). Table 5.1 describes age and period specific rates (×100,000) and number of cases. Figure 5.1 displays Iowa lung cancer mortality rate in 1980-2008 at each county level for all age groups from 25 to 74.

Table 5.1: Age-period specific mortality rates (×100,000) and number of deaths. Lung cancer, Iowa, 1980-2008

| | Period | | | | | |
|---|---|---|---|---|---|---|
| Age | 1980−1984 | 1985−1989 | 1990−1994 | 1995−1999 | 2000−2004 | 2005−2008 |
| 25-29 | 0.50 (6) | 0.70 (8) | 0.52 (5) | 0.55 (5) | 0.69 (6) | 0.67 (5) |
| 30-34 | 1.02 (11) | 1.28 (14) | 1.75 (19) | 1.78 (17) | 1.31 (12) | 1.73 (12) |
| 35-39 | 4.22 (36) | 3.91 (39) | 5.30 (58) | 5.51 (61) | 4.65 (47) | 2.27 (17) |
| 40-44 | 16.50 (120) | 12.73 (103) | 9.74 (98) | 12.60 (140) | 14.24 (161) | 8.83 (73) |
| 45-49 | 37.39 (249) | 35.25 (241) | 34.08 (271) | 26.87 (266) | 33.00 (362) | 22.81 (206) |
| 50-54 | 77.30 (533) | 83.27 (524) | 73.61 (497) | 62.65 (494) | 58.00 (562) | 45.03 (386) |
| 55-59 | 137.41 (976) | 156.62 (1008) | 137.11 (848) | 137.03 (911) | 116.97 (888) | 80.02 (594) |
| 60-64 | 203.54 (1363) | 227.81 (1482) | 232.45 (1446) | 223.90 (1348) | 212.21 (1310) | 150.32 (847) |
| 65-69 | 260.07 (1556) | 295.36 (1797) | 335.60 (2009) | 316.06 (1755) | 319.67 (1708) | 233.29 (1049) |
| 70-74 | 300.95 (1496) | 351.17 (1801) | 372.57 (1969) | 417.04 (2168) | 406.16 (2029) | 303.45 (1154) |

To compare different models in fitting lung cancer mortality data in Iowa, I use the deviance information criterion (DIC) [Spiegelhalter et al., 2002] which is the posterior average of the deviance plus a measure of complexity. Table 5.2 displays the DIC and posterior estimates for model parameters.

Table 5.2: Model comparisons using DIC

| | Age Effect | | |
|---|---|---|---|
| Model | ICAR prior | Armitage-Doll Model | TSCE model |
| AA | 20966.60 | 20973.90 | 20969.20 |
| AAP | 19499.80 | 19506.30 | 19501.30 |
| AAC | 19813.90 | 19886.80 | 19842.40 |
| AAP + AP | 19592.40 | 19606.60 | 19575.70 |
| AAC + AC | 19814.40 | 19838.20 | 19823.00 |
| AAPC | 19231.30 | 19233.00 | 19234.70 |
| AAPC + AP | 19021.30 | 19230.70 | 19240.70 |
| AAPC + AC | 18985.00 | 18996.30 | 19235.80 |
| AAPC + AP + AC | 19227.00 | 19228.30 | 19223.40 |

Compared with the DIC values derived from the conventional Bayesian APC model where all age, period and cohort effects are taking autoregressive priors, we don't see much difference on the improvement of DIC values when the multistage carcinogenesis models are incorporated into AAPC model. However, the age effect in the Bayesian extended AAPC model has a more sound biological meaning since we replace it with the hazard function from the carcinogenesis model. The temporal evolutions of age effects derived from the carcino-

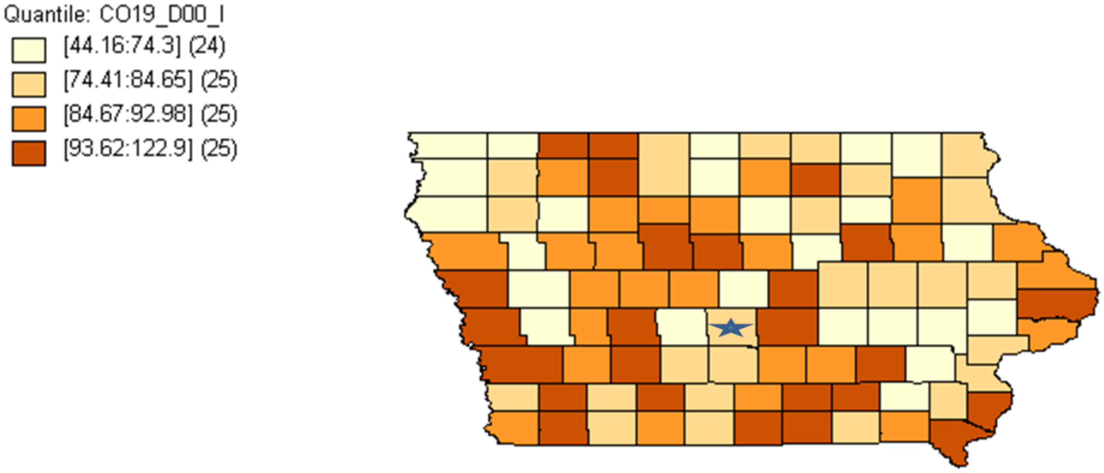

Figure 5.1: Lung cancer mortality rate in Iowa from 1980 to 2008 for all age groups (blue star represents Des Moines, Iowa capital)

genesis model clearly demonstrates the association between age and mortality rate. With the introduction of the carcinogenesis model into AAPC model, we reduce the complexity of any possible linear or nonlinear age effects in determining the mortality rate, except for those derived from the Armitage-Doll model and the TSCE model. With similar model fitting criterior (DIC values), the extended AAPC model outperforms the conventional AAPC model due to its strong biological meaning of age effects.

As an alternative, I also use predictive model selection procedure where the full posterior predictive distribution is utilized to sample replicated data $y_{new}$ and calculate the discrepancy function $d(y_{new}, y_{obs})$. Since Poisson likelihood is used for cancer death counts, Waller [1997], Carlin and Louis [2009] recommended to choose the discrepancy function as

$$d(y_{new}, y_{obs}) = 2\sum_{l} y_{l,obs} log(y_{l,obs}/y_{l,new}) - (y_{l,obs} - y_{l,new}),$$

where $l$ is the index in $y$. For every model $M_i$, I compute the expected predictive deviance (EPD) as $E[d(y_{new}, y_{obs})|y_{obs}, M_i]$ and select the model with the lowest EPD. The predictive model selection procedure also detects the similarity in EPD values between conventional model and extended AAPC models with carcinogenesis models incorporated. Similar to

DIC procedure, the predictive model selection also find the model AAPC+AC achieves lowest EPD values than other types of model do.

Figure 5.2 and 5.3 show the convergence plots for posterior estimates from Armitage-Doll and TSCE carcinogenesis model. Clearly the convergence is met for the Bayesian extended AAPC model.

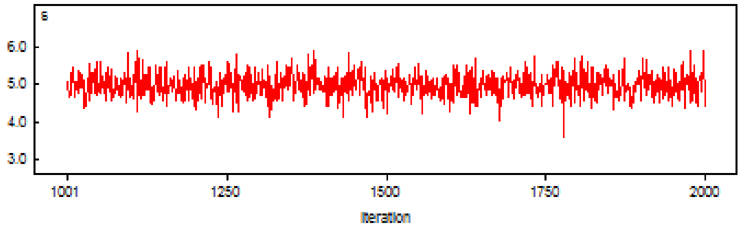

Figure 5.2: Convergency plots for posterior estimate of Armitage-Doll model parameter

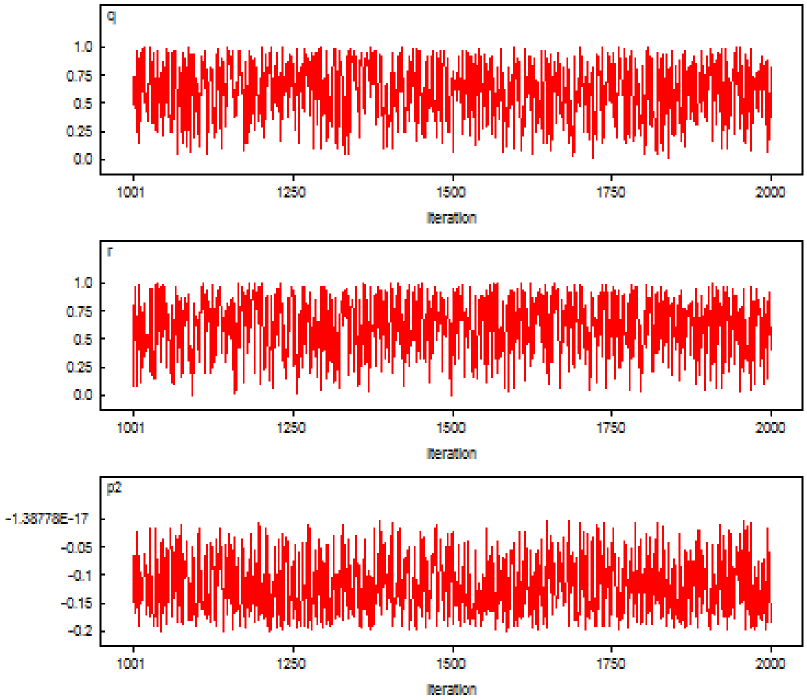

Figure 5.3: Convergency plots for posterior estimates of TSCE model parameters

The model we chose to first fit Iowa lung cancer mortality data is AAPC+AC using Armitage-Doll carcinogenesis model due to its low DIC value and good convergence. The posterior estimate for age, period and cohort effects are listed in Table 5.3.

Table 5.3: Posterior estimates for age, period and cohort effects in the Bayesian extended AAPC models

| *Age* | $\hat{\alpha}_{ci}$ | *Period* | $\hat{\beta}_{cj}$ | *Cohort* | $\hat{\gamma}_{ck}$ |
|---|---|---|---|---|---|
| 25-29 (i=1) | -3.021 | 1980-1984 (j=1) | -0.330 | 1908-1912 (k=1) | 0.969 |
| 30-34 (i=2) | -2.289 | 1985-1989 (j=2) | -0.125 | 1913-1917 (k=2) | 0.907 |
| 35-39 (i=3) | -1.375 | 1990-1994 (j=3) | 0.014 | 1918-1922 (k=3) | 0.840 |
| 40-44 (i=4) | -0.529 | 1995-1999 (j=4) | 0.157 | 1923-1927 (k=4) | 0.792 |
| 45-49 (i=5) | 0.211 | 2000-2004 (j=5) | 0.308 | 1928-1932 (k=5) | 0.641 |
| 50-54 (i=6) | 0.757 | 2005-2008 (j=6) | -0.024 | 1933-1937 (k=6) | 0.473 |
| 55-59 (i=7) | 1.209 | | | 1938-1942 (k=7) | 0.270 |
| 60-64 (i=8) | 1.522 | | | 1943-1947 (k=8) | -0.021 |
| 65-69 (i=9) | 1.714 | | | 1948-1952 (k=9) | -0.285 |
| 70-74 (i=10) | 1.801 | | | 1953-1957 (k=10) | -0.394 |
| | | | | 1958-1962 (k=11) | -0.518 |
| | | | | 1963-1967 (k=12) | -0.694 |
| | | | | 1968-1972 (k=13) | -0.865 |
| | | | | 1973-1977 (k=14) | -0.996 |
| | | | | 1978-1982 (k=15) | -1.12 |

Age, period and cohort main effects are displayed in Figure 5.6 where the age, period and cohort effects are centered by their means. The inclusion of those constraints is to improve numeric stability and mixing in the MCMC samplings. The age effects show an increasing pattern over time, which means older age leads to higher cancer mortality rate than younger age does as we controlled for other covariates. The age pattern can be easily explained by Armitage-Doll carcinogenesis model since we assume a log-linear relationship between age and hazard function. The change point for the period effects is in period 2000-2004. The period effects are increasing before the year 2000 but sharply decreasing after the year 2000. The anti-smoking campaign has been introduced in the U.S. in the 90s and since then people's behaviors on smoking have signficantly changed which explain the decreasing trend in period effects since 2000. The lung cancer mortality rate is continuously declining by birth cohorts. The main area effects are displayed in Figure 5.4 which shows a higher lung cancer mortality rate in the south of Iowa as compared to that in the north of Iowa. In southern Iowa, there are higher rates of radon gas and a higher rate of smoking, blue-collar workers, and manufacturing jobs where coal mining previously occurred. The scatter plot of main spatial effects versus Iowa county population from 2000 to 2004 is displayed in Figure 5.5. It supports the conclusion of higher spatial effects in largely populated areas, such as Polk county where Iowa capital city Des Moines is located. Compared to the main area and cohort effects in the AAPC model, the coefficients of area-cohort interactions are much smaller and can be ignored. In the extended AAPC model, there is no intercept to indicate the overall mean. The negative value of log transformation of mortality rate (at a scale of $10^{-4}$ to $10^{-6}$) falls into main spatial effects since temperate effects were small. To solve this problem, we can add an intercept term in the extended AAPC model or adjust the normal mean of noninformative prior for unstructured spatial effects from 0 to negative values, i.e., -5.

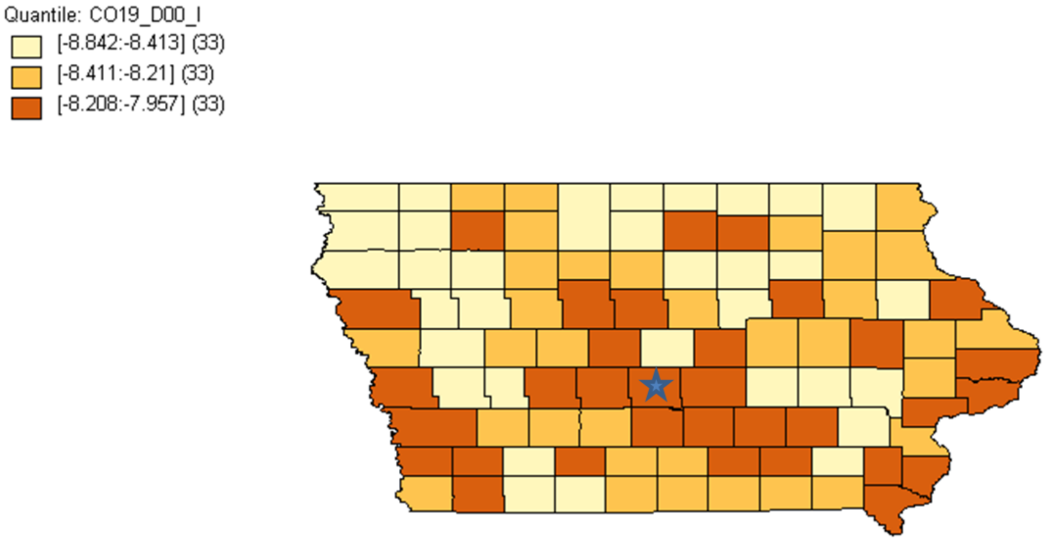

Figure 5.4: Area main effects in AAPC model (blue star represents Des Moines, Iowa capital)

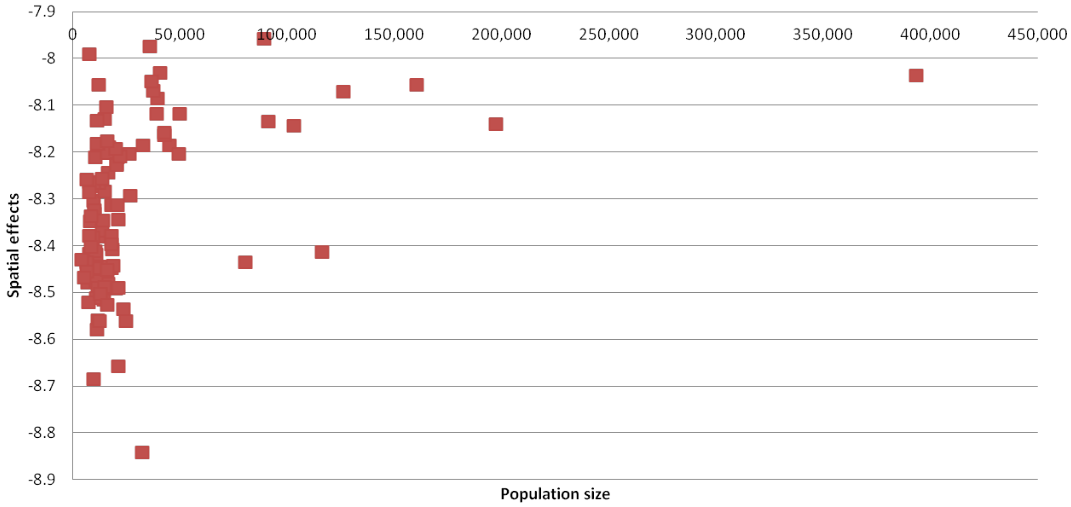

Figure 5.5: The scatter plot of main spatial effects versus Iowa county population from 2000 to 2004 in lung cancer mortality

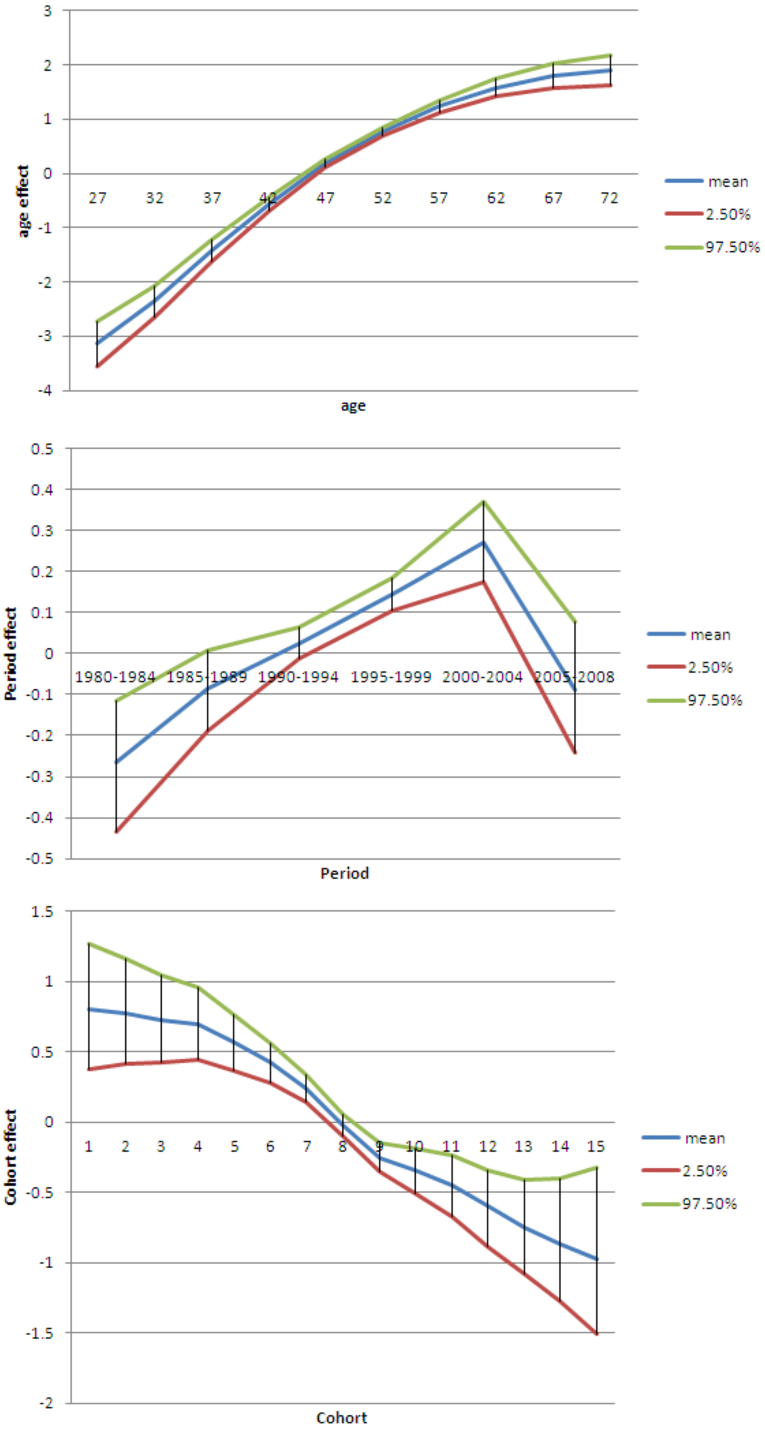

Figure 5.6: Age, period and cohort main effects in AAPC model

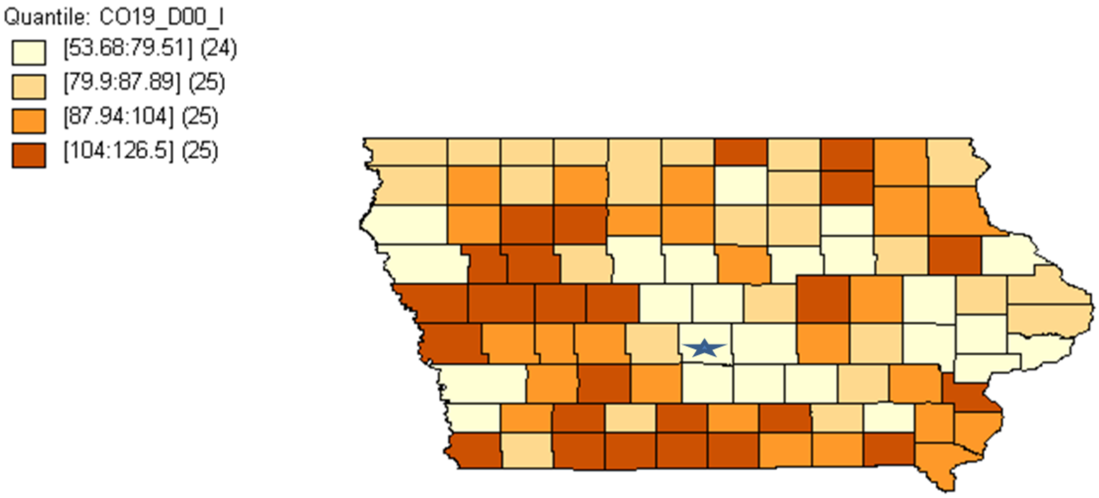

Figure 5.7: Colon cancer mortality rate in Iowa from 1981 to 2007 for age 50 and more (blue star represents Des Moines, Iowa capital)

## 5.6 Example 2 - Colon Cancer Mortality in Iowa

Colon cancer mortality data in Iowa is used to further study the performance of the Bayesian extended AAPC model in estimating cancer mortality rates. Since the most colon cancer death is associated with older population, I limit the study population with the age 50 or above. The age group use five years interval and eight age groups are reported (50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85+). The calendar period is chosen from 1981 to 2007. The 2010 SEER program provides cancer mortality data in three years intervals. Eight period groups are reported (1981-1983, 1984-1986, 1987-1989, 1990-1992, 1993-1995, 1996-1998, 1999-2002, 2003-2007). Table 5.4 describes age and period specific rates (x100,000) and number of cases. Figure 5.7 displays Iowa colon cancer mortality rate in 1981-2007 at each county level for population at age 50 or more.

Moolgavkar's TSCE carcinogenesis model is used to update the age effect in the AAPC model for colon cancer since Armitage-Doll model only consider the age group from 25 to 74. Figure 5.8 displays the distribution of unstructured spatial effects for colon cancer mortality county-wide in Iowa. At those remote counties especially in south and west, we

Table 5.4: Age-period specific mortality rates (×100,000) and number of deaths. Colon cancer, Iowa, 1981-2007

| | Period | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age | 1981−1983 | 1984−1986 | 1987−1989 | 1990−1992 | 1993−1995 | 1996−1998 | 1999−2002 | 2002−2007 |
| 50-54 | 24.38 (101) | 27.23 (104) | 25.18 (95) | 21.62 (85) | 13.78 (59) | 18.71 (88) | 15.27 (113) | 11.84 (123) |
| 55-59 | 38.52 (164) | 43.80 (178) | 38.01 (143) | 37.11 (136) | 37.10 (141) | 31.20 (124) | 25.10 (144) | 19.79 (173) |
| 60-64 | 62.33 (251) | 50.34 (200) | 56.88 (220) | 50.07 (190) | 57.36 (208) | 47.34 (170) | 35.13 (170) | 32.77 (219) |
| 65-69 | 89.65 (322) | 82.32 (299) | 72.65 (266) | 68.20 (247) | 73.51 (259) | 75.15 (251) | 65.72 (279) | 52.63 (288) |
| 70-74 | 120.95 (361) | 117.50 (358) | 110.17 (341) | 97.31 (308) | 91.47 (291) | 96.67 (301) | 80.27 (328) | 69.92 (334) |
| 75-79 | 160.88 (375) | 166.54 (402) | 148.64 (373) | 142.12 (366) | 135.30 (357) | 124.84 (336) | 115.87 (419) | 101.96 (442) |
| 80-84 | 212.24 (343) | 212.88 (357) | 205.07 (358) | 199.40 (362) | 194.78 (367) | 176.34 (344) | 152.81 (417) | 136.45 (477) |
| 85+ | 287.82 (416) | 265.32 (408) | 282.39 (455) | 276.29 (467) | 260.94 (465) | 270.67 (507) | 235.43 (625) | 215.35 (789) |

observe higher spatial effects in estimating the colon cancer mortality rates. Part of the reasons are due to the lack of colorectal cancer screening. Cancer statistics show that the survival rate is high if the colon cancer is diagnosed early [ACS, 2009]. However, the early stage diagnosis colorectal cancer ratio varies across counties in Iowa [Iowa Department of Public Health, 2006]. The rural counties have the lowest rate of early stage diagnosis than the urban counties or metropolitan counties have [Iowa Department of Public Health, 2006]. Such negative association between spatial effects and population size can be demonstrated in Figure 5.9, where we see highly populated areas have smaller spatial effects while less populated areas can have larger spatial effects.

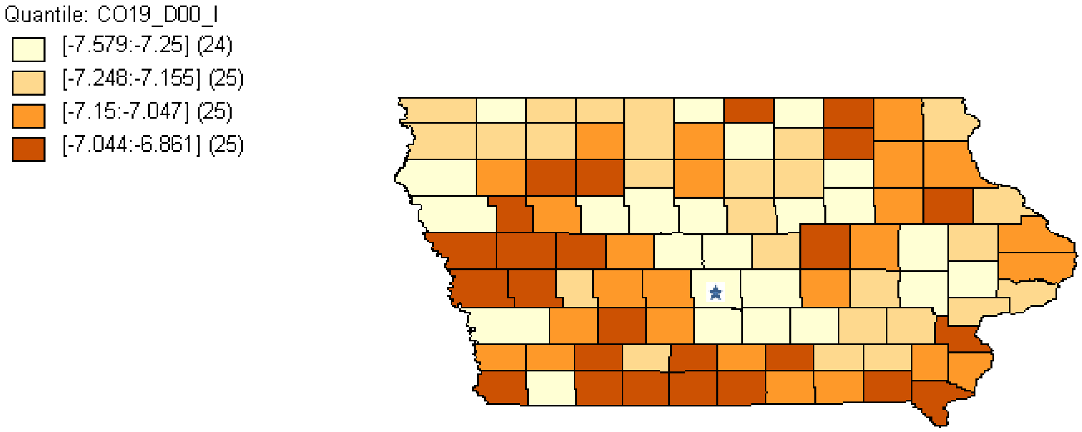

Figure 5.8: Unstructured spatial effects for colon cancer mortality in Iowa (blue star represents Des Moines, Iowa capital)

Figure 5.10 shows the three main temporal effects in the AAPC model. The overall age effect is in increasing mode. The upward trend is higher in 50s than that in later age groups. The period effect is increasing while the birth cohort effects is decreasing over time.

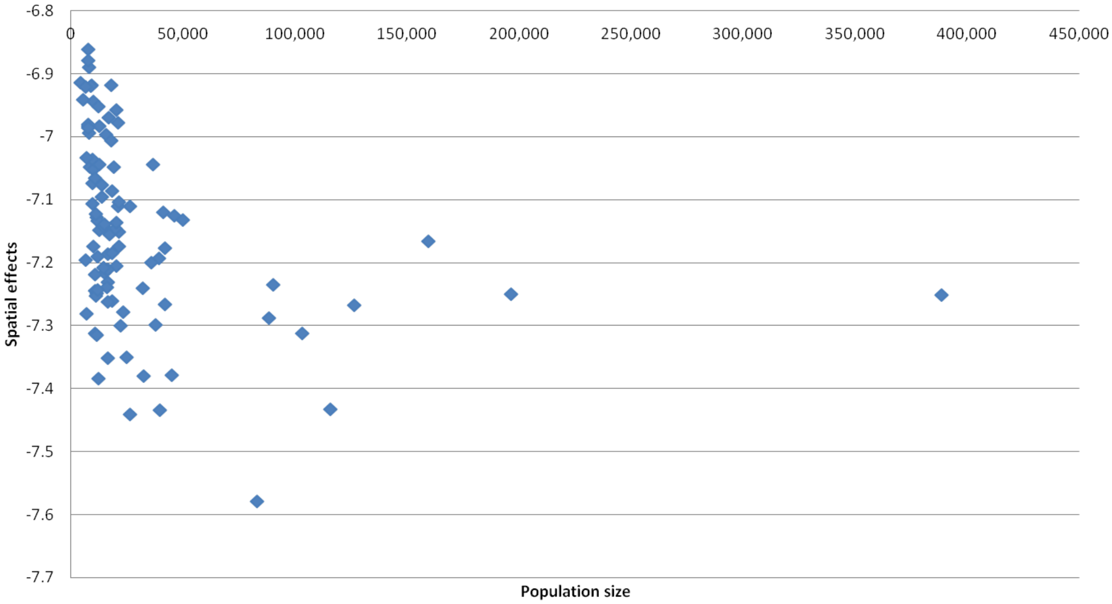

Figure 5.9: The scatter plot of main spatial effects versus Iowa county population from 1999 to 2002 in colon cancer mortality

## 5.7 Future Work

It is very important to determine the trends of disease risk both temporally and spatially [Lagazio et al., 2003]. However, it is difficult to explain some of those temporal effects (age, period, and cohort) [Richardson, 2008] due to lack of biological meanings. Carcinogenesis models of a typical underlying disease process describe how normal cells are transformed into cancer cells and age is a deterministic factor in the model. Therefore, we propose a new extended AAPC model by incorporating carcinogenesis model into our study to improve our prior knowledge of age effects in determining disease trends. Both Armitage-Doll and TSCE carcinogenesis model are considered in this study. The lung cancer mortality study shows the extended AAPC model with area-cohort interaction and Armitage-Doll age effects can be used to estimate lung cancer risk while we control the age effect from the underline disease process. The colon cancer mortality study also demonstrates the use of extended AAPC model in estimating colon cancer risk from the carcinogenesis process. The convergence of model parameters is guaranteed as well. The extended AAPC model can be used in studying spatial-temporal pattern of cancer mortality with strong biological prior beliefs in

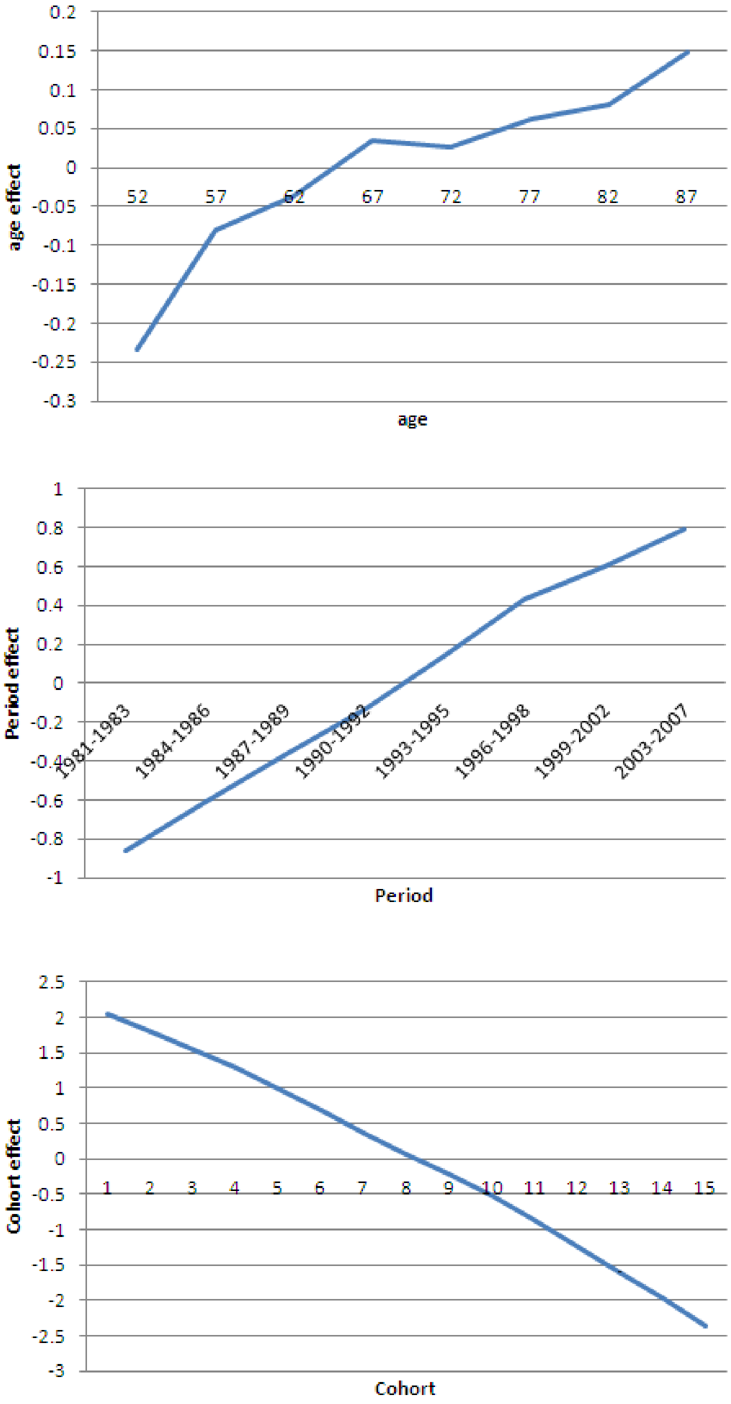

Figure 5.10: Temporal effects for colon cancer mortality in Iowa

the age effects.

Non-identifiability is the common challenge in fitting APC models. We have added model constraints in our extended AAPC models in considering the identifiability issues. However, further works are still needed in the extended AAPC models in this area. A different sampling technique which uses multivariate Metropolis steps [Lagazio et al., 2003, Rue, 2001] would be a better approach to handle efficiently the identification problems. Choosing different priors for temporal effects instead of autoregressive Gaussian distribution can also be considered in the Bayesian model. More complicated forms for spatial priors can be added in the future study. For example, Waller [1997] suggested to include the distance between county $i$ and $j$ in the formula of computing the weights $\omega_{ij}$. Furthermore, covariates such as smoking status, socialeconomic status of the counties might be included in the model.

# Chapter 6

# Conclusion

The development of carcinogenesis models has improved our understanding of the biological processes in the formation of cancer and provided scientific evidences in evaluating the risk of cancer. In this dissertation, I apply a new Bayesian approach to study the Armitage-Doll multi-stage carcinogenesis model in estimating cancer mortality rates. Mortality data for different types of cancer are retrieved from SEER database. The posterior estimates for the number of stages conclude there are six to seven stages of transitions involved in cancer formation. The sensitivity analysis and model checking show that the Bayesian Armitage-Doll model fits the data well.

To enhance the biological meanings of age effects in the APC model, I introduce the extended APC model where age effects are taking the format of multi-stage carcinogenesis models. Both Armitage-Doll and TSCE carcinogenesis models are considered in the extended APC models. Bayesian approaches are used in order to ease the concern of non-identifiability problems commonly presented in the APC models. The Bayesian extended APC model obtains the similar DIC value as conventional APC does (DIC=1286.98 for extended APC model using TSCE model, 1287.00 for extended APC model using Armitage-Doll model, 1286.43 for conventional APC model). However, the explanation of age effects in the Bayesian extended APC model has more sound biological meanings that that in APC model. The Bayesian extended APC model is applied to study colon cancer mortality rate in the US, achieving high consistency between estimated and observed rates for older age groups ($\geq 45$). Furthermore, I apply the Bayesian extended APC model to study the spatio-temporal variation in cancer mortality rates. The Bayesian extended Area-APC (AAPC) model is developed to study the county level lung and colon cancer mortality data in Iowa. The study shows the Bayesian extended AAPC model with area-cohort interaction and Armitage-Doll age effects can be used to estimate lung cancer risk while the age effects are controlled by the underline disease process. The colon cancer mortality study also demonstrates the use of extended AAPC model in estimating colon cancer risk from the carcinogenesis process. The convergence of model parameters is guaranteed as well. The Bayesian extended AAPC model can be used in studying spatial-temporal pattern of cancer mortality with strong biological prior beliefs in the age effects.

I apply the extended Bayesian APC model to project the mortality rates in the future periods (2009-2013, 2014-2018, and 2019-2023). My projection on the decreasing trend of colon cancer mortality rate can be supported by the fact that the colon cancer is often highly treatable and the promising development of cancer screening methods in the near future. In addition, the *a priori* beliefs in smoothing period and cohort effects in the Bayesian approach produce more precise posterior estimates of mortality rates than those derived from the maximum likelihood approach in the classical APC model. I use the autoregressive priors in our Bayesian model to avoid the considerable increase in the number of parameters in the classical APC model which could lead to large standard errors and decreased precision for making projections [Hakulinen and Dyba, 1994]. Using a Bayesian approach, I also avoid the strong parametric assumptions often made in the classical APC models. In the Bayesian version of this method the most appropriate degree of smoothing can be learned from the data. The Bayesian model is also more flexible than the classical linear models [Hakulinen and Dyba, 1994, Dyba and Hakulinen, 1997] since it copes with both increasing and decreasing trends. In the extended Bayesian AAPC model, I study the spatial-temporal pattern of cancer mortality rates by examining the main effects for area, age, period, and cohort and their interactions. I compare different models which take into account the interaction effects between period and space or cohort and space and substitute the main age effects with hazard functions from carcinogenesis model. With the introduction of carcinogenesis model into AAPC model, I reduce the complexity of any possible linear or nonlinear age effects in determining the mortality rate, except for those derived from the Armitage-Doll model and the TSCE model. With similar model fitting criterion (DIC values), the extended AAPC model outperforms the conventional AAPC model due to its strong biological meaning of age effects.

Due to a large number of a priori-dependent parameters in the Bayesian model, a block updating MCMC algorithms can be used in this study to avoid slow mixing of the Markov chain and to allow for a proper incorporation of identifiability constraints such as sum-to zero constraints [Schmid and Held, 2004]. However, additional development in jointly updateing hyperparameters and parameters is needed. Alternative prior settings beside the Gaussian prior distributions can be considered in our Bayesian model. The reparameterization of the age, period and cohort effects can be used as well to include a joint cumulative effect for period and cohort. Knorr-Held [2001] introduced four different prior distributions for the spatial-temporal interactions such as area-period or area-cohort interactions which could also be applied to our Bayesian extended AAPC model.

In the lung cancer study, I group period effects by 5-year intervals from cancer mortality data at individual calendar years. Age groups are predetermined in SEER program by 5-year interval. Therefore, the cohort effect for the same age and period groups could vary with a maximum of 10 years. This could reduce the precision of model parameter estimations. One possible approach is to model period effects by each year (not 5-year intervals) which could lead to more accurate birth cohorts. However, the unobserved age-specific cancer death in certain calendar years could reduce the power of parameter estimation in the classical APC models. To address this problem, we can apply the MCMC method in the Bayesian extended APC model to simulate data for the missing observations and derive the posterior estimates from the full samples. The Bayesian model suggests period could be modeled by year.

In addition to estimation in an individual calendar year, we can consider including some important factors that affect cancer incidence and mortality into the Bayesian models.

Smoking status has proven to be associated with lung cancer incidence and mortality [NCI, 2011]. Obtaining relevant information about smoking prevalence and cigarette consumption geographically would help to improve the precision of estimating lung cancer rates. However, it is difficult to get reliable data on the occurrence of lung cancer by smoking status because most cancer registries including the SEER program are population-based systems and generally do not gather information related to individual smoking habits. The SEER database provides information about tobacco use information at population level, but it can only be used to determine occurrence of lung cancer amongst never smokers within wide geographic areas. In two separate lung cancer mortality studies in Germany and Japan [Knorr-Held and Rainer, 2001, Kaneko and Sobue, 2003], the available smoking data has been applied to the APC models and used to explain the increasing or decreasing trend in period and cohort effects. To further extend the research in investigating the spatial-temporal pattern of lung cancer, we would strongly recommend obtaining available smoking data and modifying the AAPC model to incorporate significant covariates such as smoking prevalence.

Air pollution is another important risk factor for cancer. Thirty-six out of 1 million U.S. residents will develop cancer due to breathing toxic air pollution, according to estimates by the Environmental Protection Agency (EPA). Large cities appear to carry greater cancer risk because of a higher volume of cars, trucks, construction equipment, gas stations, and in some cases, dry cleaners. However, there are also many rural industrial areas where residents have an elevated risk, according to the EPA report. The air quality data at each county level could be a good resource to be considered to be included in the model to study the cancer mortality rates.

# Bibliography

Cancer facts and figures, 2009. URL http://www.cancer.org/Research/CancerFactsFigures/CancerFactsFigures/index.

Lung cancer prevention, 2011. URL http://www.cancer.gov/cancertopics/types/lung.

D.F. Andrews and C.L. Mallows. Scale mixtures of normality. *Journal of the Royal Statistical Society, Ser. B*, 36:99–102, 1974.

P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 8:1–12, 1954.

G. Berry. Relative risk and acceleration in lung cancer. *Statistics in Medicine*, 26:3511–3517, 2007.

C. Berzuini and D. Clayton. Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, 13(8):823–838, 1994.

J.M. Bishop. Molecular themes in oncogenesis. *Cell*, 64(2):235–248, 1991.

I. Bray. Application of markov chain monte carlo methods to projecting cancer incidence and mortality. *Applied Statistics*, 51(2):151–164, 2002.

I. Bray and et al. Brennan, P. Recent trends and future projections of lymphoid neoplasms-a bayesian age-period-cohort analysis. *Cancer Cause and Control*, 12(9):813–820, 2001.

N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

Fish D. Childs J. Buenconsejo, J. and T. Holford. A bayesian hierarchical model for the estimation of two incomplete surveillance data sets. *Statistics in Medicine*, 27:3269–3285, 2008.

B. Carlin and T. Louis. *Bayesian Methods for Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC, 2009.

D. Clayton and E. Schifflers. Models for temporal variation in cancer rates. i: Age-period and age-cohort models. *Statistics in Medicine*, 6(4):449–467, 1987.

P. Congdon. *Bayesian Statistical Modeling.* London: John Wiley & Sons Ltd, 2nd edition, 2006a.

P. Congdon. A model framework for mortality and health data classified by age, area, and time. *Biometrics*, 62:269–278, 2006b.

T. Dyba and et al. Hakulinen, T. A simple non-linear model in incidence prediction. *Statistics in Medicine*, 16(20):2297–2309, 1997.

E.R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5): 759–767, 1990.

S. Frank. Age-specific incidence of inherited versus sporadic cancers: A test of the multistage theory of carcinogenesis. *Proceedings of National Academy of Sciences*, 102(4):1071–1075, 2005.

A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457–511, 1992.

T. Hakulinen and T. Dyba. Precision of incidence predictions based on poisson distributed observations. *Statistics in Medicine*, 13(15):1513–1523, 1994.

T.R. Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39(2):311–324, 1983.

T.R. Holford. Understanding the effects of age, period and cohort on incidence and mortality rates. *Annu Rev Public Health*, 12:425–457, 1991.

Zhang Z. Holford, T. and L. McKay. Estimating age, period and cohort effects using the multistage model for cancer. *Statistics in Medicine*, 13:23–41, 1994.

The University of Iowa College of Public Health Iowa Department of Public Health. Increasing colorectal cancer screening in iowa: needs and strategies for improvement, 2006. URL `http://www.canceriowa.org/Files/News/crcmonograph.aspx`.

Luebeck E.G. Jeon, J. and S.H. Moolgavkar. Age effects and temporal trends in adenocarcinoma of the esophagus and gastric cardia (united states). *Cancer Cause and Control*, 17(7):971–981, 2006.

Ishikawa K. Yoshimi I. Marugame T. Hamashima C. Kamo K. Mizuno S. Kaneko, S. and T. Sobue. Projection of lung cancer mortality in japan. *Cancer Science*, 94(10):919–923, 2003.

L. Knorr-Held and E. Rainer. Projects of lung cancer mortality in west germany: a case study in bayesian prediction. *Biostatistics*, 2:109–129, 2001.

E. Levi F. Decarli A. La Vecchia, C. Negri and P. Boyle. Cancer mortality in europe: effects of age, cohort of birth and period of death. *European Journal of Cancer*, 34: 118–141, 1998.

C. Lagazio, A. Biggeri, and E. Dreassi. Age-period-cohort models for disease mapping. *Environmetrics*, 14:475–490, 2003.

Dewanji A. Moolgavkar, S.H. and D.J. Venzon. A stochastic two-stage model for cancer risk assessment. i. the hazard function and the probability of tumor. *Risk Analysis*, 8(3), 1988.

S.H. Moolgavkar and A.G. Knudson. Mutation and cancer: a model for human carcinogenesis. *Journal of National Cancer Institute*, 66:1037–1052, 1981.

S.H. Moolgavkar and E.G. Luebeck. Multistage carcinogenesis: population-based model for colon cancer. *Journal of National Cancer Institute*, 84 (8):610–618, 1992.

S.H. Moolgavkar and G. Luebeck. Two-event model for carcinogenesis: biological, mathematical and statistical considerations. *Risk Analysis*, 10:323–341, 1990.

S.H. Moolgavkar and et al. Meza, R. Pleural and peritoneal mesotheliomas in seer: age effects and temporal trends, 1973-2005. *Cancer Cause and Control*, 20(6):935–944, 2009.

S.H. Moolgavkar and D.J. Venzon. Two-event model for carcinogenesis: incidence curves for childhood and adult tumors. *Mathematical biosciences*, 47:55–77, 1979.

A. O'Hagan. Bayesian inference. *Kendall's Advanced Theory of Statistics*, 2B, 1994.

C. Osmond and M.J. Gardner. Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1(3):245–259, 1982.

David B. Richardson. Multistage modeling of leukemia in benzene workers: A simple approach to fitting the 2-stage clonal expansion model. *American Journal of Epidemiology*, 169(1):78–85, 2008.

H. Rue. Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society B*, 63:325–338, 2001.

V. Schmid and L. Held. Bayesian extrapolation of space-time trends in cancer registry data. *Biometrics*, 60:1034–1042, 2004.

D. Spiegelhalter, N. Best, B. Carlin, , and van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

L. Waller, B. Carlin, H. Xia, and A. Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617, 1997.

# Appendices

# Appendix A

# R2WINBUGS code

## A.1 R code

```
library("arm")

library(R2WinBUGS)

setwd("G:/Ph.D. research/Computations/AD")

aapc <- read.table ("iowa.txt", header=TRUE)

adj<-read.table("IA_counties_adj.txt")[,1]

num<-read.table("IA_counties_num.txt")[,1]

Nneigh<-588

A <- 10

P<-6

Co<-15

C<-99

county <- aapc[,1]

age <- aapc[,2]

period <- aapc[,3]

cohort <- A+period-age

pyr <- aapc[,4]
```

```
   cases <- aapc[,5]

   rate<-100000*cases/pyr


#create adjacent matrix adj_max<-mat.or.vec(C,C) begin<-1 end<-0 for

(i in 1:C){

    end<-end+num[i]

    for (j in begin:end){

        adj_max[i,adj[j]]<-1

    }

    begin<-begin+num[i]

}



vec10<-c(1:10)*0.01 vec6<-c(1:6)*0.01 vec99<-c(1:99)*0.01

vec15<-c(1:15)*0.01


# area and age

   data <- list ("Nneigh","adj","num", "C","county","age", "A","cases", "pyr")

   inits <- function() {list (tauc=1, taua=1, r=0.07, p2=-0.0885, q=0.000029, alpha=vec1

   parameters <- c("tauc", "u","alphac","taua","v", "r", "p2", "q", "cases_dev")

   aapc.sim <- bugs (data, inits, parameters, "aa.bug", n.chains=1, n.iter=10000, debug=

   ##AD model

   inits <- function() {list (tauc=1, taua=1, s=5,c=-10, alpha=vec10,u=vec99,v=vec99)}

   parameters <- c("tauc", "u","alphac","taua","c", "s", "cases_dev")

   aapc.sim <- bugs (data, inits, parameters, "aa_AD.bug", n.chains=1, n.iter=10000, deb

   ##Conventional Age model

   inits <- function() {list (tauc=1, taua=1, alpha=vec10,u=vec99,v=vec99)}

   parameters <- c("tauc", "u","alphac","taua","cases_dev")

   aapc.sim <- bugs (data, inits, parameters, "aa_conv.bug", n.chains=1, n.iter=10000, d
# area and age,period
   data <- list ("Nneigh","adj","num", "C","county","age", "A","period","P","cases", "py
   inits <- function() {list (tauc=1, taua=1,taua2=10, s=5,c=-10, u=vec99,alphac=vec10,v
   parameters <- c("tauc", "u","alphac","taua2", "taua", "c", "s","taup","betac")
   aapc.sim <- bugs (data, inits, parameters, "aap.bug", n.chains=1, n.iter=100, debug=T

# area, age, and cohort
   data <- list ( "Nneigh","adj","num","C","county","age", "A", "cohort","Co","cases", "
   inits <- function() {list ( taua=1, tauco=1,r=0.07, p2=-0.0885, q=0.000029,alpha=vec1
   parameters <- c( "u","v","alphac","r", "p2", "q","taua", "gammac","tauco")
   aapc.sim <- bugs (data, inits, parameters, "aac.bug", n.chains=1, n.iter=1000, debug=

# area and age,period and area-period interaction
   data <- list ("Nneigh","adj","num","C","county","age", "A", "period", "P", "cases", "
   inits <- function() {list ( taua=1, taup=1,r=0.07, p2=-0.0885, q=0.000029, alpha=vec1
            u=vec99,v=vec99)}
   parameters <- c("tauc", "u","v","alphac","r", "p2", "q","taua", "betac", "taup", "phi
   aapc.sim <- bugs (data, inits, parameters, "aap+ap.bug", n.chains=1, n.iter=10000, de

# area, age,period and cohort
   data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
   inits <- function() {list ( taua=1, taup=1,tauco=1,r=0.07, p2=-0.0885, q=0.000029, al
   parameters <- c( "u","v","alphac","r", "p2", "q","taua", "betac", "taup","gammac","ta
   #parameters <- c( "cases_dev")
   aapc.sim <- bugs (data, inits, parameters, "aapc.bug", n.chains=1, n.iter=10000, debu




# AAPC + AP
   data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
   inits <- function() {list ( taua=1, taup=1,tauco=1,r=0.07, p2=-0.0885, q=0.000029, al
   parameters <- c( "u","v","alphac","r", "p2", "q","taua", "betac", "taup","gammac","ta
   aapc.sim <- bugs (data, inits, parameters, "aapc+ap.bug", n.chains=1, n.iter=1000, de

# AAPC + AC
   data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
   inits <- function() {list ( taua=1, tau2=1,taup=1,tauco=1,cons=-10,s=5, alpha=vec10,
   parameters <- c( "u","v","alphac","s","cons","taua", "betac", "taup","gammac","tauco"
   aapc.sim <- bugs (data, inits, parameters, "aapc+ac.bug", n.chains=1, n.iter=10000, d
   #TSCE model
   inits <- function() {list ( taua=1, tau2=1,taup=1,tauco=1,r=0.07, p2=-0.0885, q=0.000
   parameters <- c( "u","v","alphac","r","p2","q","taua", "betac", "taup","gammac","tauc
   aapc.sim <- bugs (data, inits, parameters, "aapc+ac_TSCE.bug", n.chains=1, n.iter=10,



# AAPC + AC2
   data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
   inits <- function() {list ( taua=1, tau2=1,taup=1,tauco=1,alpha=vec10, beta=vec6,u=0,
   parameters <- c( "u","v","alphac","taua", "betac", "taup","gammac","tauco", "phi_area
   aapc.sim <- bugs (data, inits, parameters, "aapc+ac2.bug", n.chains=1, n.iter=1000, d


# AAPC + AP+ AC
   data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
   inits <- function() {list ( taua=1, taup=1,tauco=1, s=5, cons=0,alpha=vec10, beta=vec
   parameters <- c( "u","v","alpha","alphac","taua", "s", "cons","betac", "taup","gammac
   aapc.sim <- bugs (data, inits, parameters, "aapc+apac.bug", n.chains=1, n.iter=1000,

# AAPC (car.normal for period and cohort)
```

```
  data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
  inits <- function() {list ( taua=1, taup=1, tauc=1, tauco=1,s=5, cons=0,alpha=vec10,
  parameters <- c( "u","v","alpha","alphac", "s", "cons","betac","gammac")
  aapc.sim <- bugs (data, inits, parameters, "aapc2.bug", n.chains=1, n.iter=100, debug


# area and age,cohort and area-cohort interaction
  data <- list ("Nneigh","adj","num","C","county","age", "A", "cohort", "Co", "cases",
  inits <- function() {list ( taua=1, tauco=1,r=0.07, p2=-0.0885, q=0.000029, alpha=vec
      u=0,v=vec99)}
  parameters <- c("tauc", "u","v","alphac","r", "p2", "q","taua", "gammac", "tauco", "p
  aapc.sim <- bugs (data, inits, parameters, "aacac.bug", n.chains=1, n.iter=1000, debu

# area and age,cohort and area-cohort interaction
  data <- list ("Nneigh","adj","num","C","county","age", "A", "cohort", "Co", "cases",
  inits <- function() {list ( taua=1, tauco=1, alpha=vec10, gamma=vec15,
      u=0,v=vec99)}
  parameters <- c("tauc", "u","v","alphac","taua", "gammac", "tauco", "phi","phi_period
  aapc.sim <- bugs (data, inits, parameters, "aapc_aacACold.bug", n.chains=1, n.iter=10



# area, age,period and area-period interaction
  data <- list ( "adj_max","Nneigh","adj","num","C","county","age", "A", "period", "P",
  inits <- function() {list (tauc=1, taua=1, taup=1,tauphi=1,s=5, alpha=mat.or.vec(10,1
      u=0,v=mat.or.vec(99,1), phi=mat.or.vec(99,6)+0.001)}
  parameters <- c("tauc", "u","v","alphac","s","taua", "betac", "taup", "phi","tauphi")
  aapc.sim <- bugs (data, inits, parameters, "aapc_inter.bug", n.chains=1, n.iter=1000,
# area, age,period and area-period interaction-Condgon formula
  data <- list ("Nneigh","adj","num","C","county","age", "A", "period", "P", "cases", "
  inits <- function() {list (tauc=1, taua=1, taup=1,s=5, alpha=vec10, beta=vec6,
      u=0,v=vec99, phi=mat99by6)}
  parameters <- c("tauc", "u","v","alphac","s","taua", "betac", "taup", "phi","phi_peri
  aapc.sim <- bugs (data, inits, parameters, "aapc_condgon.bug", n.chains=1, n.iter=100

# Without area main effect APC + AC
  data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
  inits <- function() {list ( taua=1, taup=1,tauco=1,cons=-10,s=5, alpha=vec10, beta=ve
  parameters <- c( "alphac","s","cons","taua", "betac", "taup","gammac","tauco", "phi_a
  aapc.sim <- bugs (data, inits, parameters, "apc+ac.bug", n.chains=1, n.iter=100, debu

# Without cohort main effect AAP + AC
  data <- list ( "Nneigh","adj","num","C","county","age", "A", "period", "P","cohort","
  inits <- function() {list ( taua=1, taup=1,tauco=1,cons=-10,s=5, alpha=vec10, beta=ve
  parameters <- c( "alphac","s","cons","taua", "betac", "taup","phi_area","phi_cohort",
  aapc.sim <- bugs (data, inits, parameters, "aap+ac.bug", n.chains=1, n.iter=100, debu
```

## A.2 BUG code

```
model {

#likelihood for age effect for(i in 1:5940) {
        cases[i] ~ dpois(mu[i])
        log(mu[i])<-log(pyr[i]) +  u+ v[county[i]]+ alphac[age[i]]+  betac[period[i]]+ga
            +phi[county[i],cohort[i]]
        cases_exp[i] ~ dpois(mu[i])
```

```
    cases_diff[i]<-(cases[i]+0.5)*log((cases[i]+0.5)/(cases_exp[i]+0.5))-(cases[i]-c

    #log(mu[i])<-log(pyr[i]) + u[(county[i]+1)/2]  + alphac[age[i]]

} cons~dflat() #for (i in 1:C){ #   u[i]~dnorm(0,tauc) #} u ~

dnorm(0,tauc) tauc~dgamma(1, 1.0E-2)


# ICAR(1) Spatial Prior v[1:C] ~ car.normal(adj[], w[], num[],

kappa) kappa ~ dgamma(1,0.001) for (j in 1:Nneigh) {w[j] <- 1}



for (a in 1:A) {

  #alphamean[a]<-log((r/100000)*p2*q2*(exp(-q2*age[a])-exp(-p2*age[a]))/(q2*exp(-p2*ag

  alphamean[a]<-cons+s*log(27+5*(a-1))

  alphaprec[a]<-taua

} for (a in 1:A){

  alpha[a]~dnorm(alphamean[a],alphaprec[a])

} taua~dgamma(1.0E-1,1.0E-1) s~dnorm(5, tau2)

tau2~dgamma(0.001,0.001) #remove linear trend from the age effects

for (a in 1:A) {

  ivec[a]<-a-(A+1)/2

  aivec[a]<-ivec[a]*alpha[a]

  #alphac[a]<-alpha[a]-ivec[a]*sum(aivec[])/(A*(A+1)*(A-1)/12)

  alphac[a]<-alpha[a]-mean(alpha[1:A])

}



betamean[1]<-0.0 betaprec[1]<-taup*1.0E-6 betamean[2]<-0.0

betaprec[2]<-taup*1.0E-6 for (p in 3:P) {

  betamean[p]<-2*beta[p-1]-beta[p-2]

  betaprec[p]<-taup

} for (p in 1:P){
    #beta[p]~dnorm(betamean[p],betaprec[p])

    beta[p]~dnorm(betamean[p],betaprec[p])

    betac[p]<-beta[p]-mean(beta[1:P])

} taup~dgamma(1.0,1.0E-3)


gammamean[1]<-0.0 gammaprec[1]<-tauco*1.0E-6 gammamean[2]<-0.0

gammaprec[2]<-tauco*1.0E-6 for (c in 3:Co) {

    gammamean[c]<-2*gamma[c-1]-gamma[c-2]

    gammaprec[c]<-tauco

} for (c in 1:Co){

    #gamma[c]~dnorm(gammamean[c],gammaprec[c])

    gamma[c]~dnorm(gammamean[c],gammaprec[c])

    gammac[c]<-gamma[c]-mean(gamma[1:Co])

} tauco~dgamma(1.0,1.0E-3) #spatial-cohort interaction for (i in

1:C){

    for (j in 1:Co){

        phi[i,j]<-phi_area[i]*phi_cohort[j]

    }

} # ICAR(1) Spatial Prior phi_area[1:C] ~ car.normal(adj[], w2[],

num[], kappa2) kappa2 ~ dgamma(1,0.001) for (j in 1:Nneigh) {w2[j]

<- 1}


#unstructured cohort effect for (j in 1:Co){

    beta2[j]~dnorm(0,100)

} for (j in 1:Co-1){

    beta2_exp[j]<-exp(beta2[j])

    phi_cohort[j]<-beta2_exp[j]/(1+sum(beta2_exp[1:Co-1]))

} phi_cohort[Co]<-1/(1+sum(beta2_exp[1:Co-1]))

cases_dev<-sum(cases_diff[])*2 }
```