Trend Analysis and Modeling of Health and Environmental Data: Joinpoint and Functional Approach

Ram C. Kafle
University of South Florida, rckafle@mail.usf.edu

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Part of the Environmental Sciences Commons, Epidemiology Commons, and the Statistics and Probability Commons

Scholar Commons Citation

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.
Trend Analysis and Modeling of Health and Environmental Data: Joinpoint and Functional Approach

by

Ram C. Kafle

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy Mathematics & Statistics College of Arts and Sciences University of South Florida

Major Professor: Chris P. Tsokos, Ph.D. Kandethody Ramachandran, Ph.D. Marcus McWaters, Ph.D. Rebecca Wooten, Ph.D.

Date of Approval: June 4, 2014

Keywords: Joinpoint Regression, Bayesian Statistics, Cancer Mortality, Functional Data Analysis, Global Warming

Copyright © 2014, Ram C. Kafle
Dedication

This doctoral dissertation is dedicated to my parents (Rishi Ram Kafle and Rohini Kumari Kafle), my wife (Arati Ghimeray Kafle), and my daughters (Arju Kafle and Arpita Kafle).
Acknowledgments

I would like to express my deepest gratitude to my advisor Professor Chris P. Tsokos for being a constant source of inspiration, motivation, encouragement, and invaluable advice during my graduate study. I am so grateful to him for all of his priceless efforts to grow me as a research scientist and a good person.

I would like to thank Dr. Kandethody Ramachandran, Dr. Marcus McWaters, and Dr. Rebecca Wooten for serving as the member of my Ph.D. committee, continuous support, and constructive advices during the preparation of this dissertation and during my study at USF.

Furthermore, many thanks to my friend Dr. Netra Khanal for his contribution in part of this thesis. In addition, my appreciation goes to my friends Dr. Keshav Pokhrel, Bhikhari Tharu, Hari Adhikari, Taysseer Sharaf, and Dr. Nana Bonsu for their support throughout this endeavor.

Finally, this work could not have been accomplished without the support and understanding of my wife, my daughters, my parents, my brothers and sisters, and my parents in-law.
# Table of Contents

List of Tables ................................................................. iii  
List of Figures ................................................................. iv  
Abstract ............................................................................... v  

Chapter 1  Introduction and Literature Review ......................... 1  
1.1 General Objectives ......................................................... 1  
1.2 Data Source ................................................................ 3  
  1.2.1 Crude and Age-adjusted Rates ................................. 4  
1.3 Joinpoint Regression Model ............................................ 5  
1.4 Annual Percentage Change (APC) .................................... 7  
1.5 Literature Review and Limitations of the Currently Applied Joinpoint Models ................................................................. 7  
1.6 Modeling Objectives and our Proposal ............................... 11  
1.7 Generalized Linear Models (GLM) ..................................... 12  
1.8 Bayesian Model Selection Criteria .................................... 14  
  1.8.1 Bayes Factor and Model Uncertainty ......................... 15  
  1.8.2 Deviance Information Criteria ................................. 16  
1.9 Markov Chain Monte Carlo Method (MCMC) ...................... 18  
  1.9.1 Gibbs Sampling Algorithm ................................. 20  
1.10 Prediction of Trends ...................................................... 21  
  1.10.1 Bayesian Model Averaging ..................................... 22  

Chapter 2  Joinpoint Regression Model .................................... 24  
2.1 Joinpoint Regression Program of National Cancer Institute ..... 25  
2.2 Bayesian Joinpoint Regression Model ................................ 28  
  2.2.1 Bayesian Inference and Specification of Priors ............... 31  
2.3 Age-Stratified Bayesian Joinpoint Regression Model ............ 32  
  2.3.1 Bayesian Inference and Specification of Priors ............... 37  
2.4 Conclusion ..................................................................... 41  

Chapter 3  Application of Bayesian Joinpoint Regression Model on Childhood Brain Cancer Mortality and its Comparison with NCI Approach ............... 43  
3.1 Statistical Analysis ......................................................... 45  
3.2 Model Validation .......................................................... 53  
3.3 Conclusion ...................................................................... 56  
  3.3.1 Contributions and future needs ............................... 57
List of Tables

Table 1  Parameter Estimates .................................................... 48
Table 2  DIC values for all five competing models for lung and bronchus cancer mortality .................................................. 62
Table 3  The posterior summaries of parameters for lung and bronchus cancer .................................................. 63
Table 4  DIC values for all five competing models for brain cancer mortality .................................................. 68
Table 5  The posterior summaries of parameters for brain cancer .................................................. 70
Table 6  Rank of the Variables by Xu and Tsokos (2013) .................................................. 80
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Posterior distribution of the number of joinpoints in child brain cancer mortality trend in the United States</td>
<td>46</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Box plot for parameters Beta of joinpoints</td>
<td>47</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Estimated time trend for the annual observed mortality rate per 100,000 children</td>
<td>49</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Mortality rates of child brain cancer obtained by using the NCI approach.</td>
<td>51</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Estimated Annual Percentage Change in child brain cancer rates over time per 100,000 children</td>
<td>52</td>
</tr>
<tr>
<td>Figure 6</td>
<td>95% Bayesian credible band for standardized residuals</td>
<td>53</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Trend Validation</td>
<td>54</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Difference in Chi-square statistics of observed and predicted mortality counts</td>
<td>55</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Comparison of actual and predictive frequencies</td>
<td>56</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Posterior probability of the number of joinpoints for lung and bronchus cancer mortality trend</td>
<td>61</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Fitted lung and bronchus mortality trends for male age groups</td>
<td>64</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Fitted lung and bronchus mortality trends for female age groups</td>
<td>65</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Estimated age-adjusted mortality trends of male and female lung and bronchus cancer</td>
<td>66</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Posterior probability of the number of joinpoints for brain cancer mortality trend</td>
<td>67</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Fitted brain cancer mortality trends for male age groups</td>
<td>71</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Fitted brain cancer mortality trends for female age groups</td>
<td>72</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Estimated age-adjusted mortality trends of male and female brain cancer</td>
<td>73</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Emission of Carbon Dioxide in the Atmosphere in U.S.A.</td>
<td>77</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Emission of Carbon Dioxide in the Atmosphere from Gas Fuels</td>
<td>81</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Emission of Carbon dioxide in the Atmosphere from Gas Fuels</td>
<td>86</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Estimated Model for the Emission of CO2 from Gas Fuels</td>
<td>89</td>
</tr>
</tbody>
</table>
Abstract

The present study is divided into two parts: the first is on developing the statistical analysis and modeling of mortality (or incidence) trends using Bayesian joinpoint regression and the second is on fitting differential equations from time series data to derive the rate of change of carbon dioxide in the atmosphere.

Joinpoint regression model identifies significant changes in the trends of the incidence, mortality, and survival of a specific disease in a given population. Bayesian approach of joinpoint regression is widely used in modeling statistical data to identify the points in the trend where the significant changes occur. The purpose of the present study is to develop an age-stratified Bayesian joinpoint regression model to describe mortality trends assuming that the observed counts are probabilistically characterized by the Poisson distribution. The proposed model is based on Bayesian model selection criteria with the smallest number of joinpoints that are sufficient to explain the Annual Percentage Change (APC). The prior probability distributions are chosen in such a way that they are automatically derived from the model index contained in the model space. The proposed model and methodology estimates the age-adjusted mortality rates in different epidemiological studies to compare the trends by accounting the confounding effects of age. The future mortality rates are predicted using the Bayesian Model Averaging (BMA) approach.

As an application of the Bayesian joinpoint regression, first we study the childhood brain cancer mortality rates (non age-adjusted rates) and their Annual Percentage Change (APC) per year using the existing Bayesian joinpoint regression models in the literature. We use annual observed mortality counts of children ages 0-19 from 1969-2009 obtained from Surveillance Epidemiology and End Results (SEER) database of the National Cancer
Institute (NCI). The predictive distributions are used to predict the future mortality rates. We also compare this result with the mortality trend obtained using joinpoint software of NCI, and to fit the age-stratified model, we use the cancer mortality counts of adult lung and bronchus cancer (25-85+ years), and brain and other Central Nervous System (CNS) cancer (25-85+ years) patients obtained from the Surveillance Epidemiology and End Results (SEER) data base of the National Cancer Institute (NCI).

The second part of this study is the statistical analysis and modeling of noisy data using functional data analysis approach. Carbon dioxide is one of the major contributors to Global Warming. In this study, we develop a system of differential equations using time series data of the major sources of the significant contributable variables of carbon dioxide in the atmosphere. We define the differential operator as data smoother and use the penalized least square fitting criteria to smooth the data. Finally, we optimize the profile error sum of squares to estimate the necessary differential operator. The proposed models will give us an estimate of the rate of change of carbon dioxide in the atmosphere at a particular time. We apply the model to fit emission of carbon dioxide data in the continental United States. The data set is obtained from the Carbon Dioxide Information Analysis Center (CDIAC), the primary climate-change data and information analysis center of the United States Department of Energy.

The first four chapters of this dissertation contribute to the development and application of joinpiont and the last chapter discusses the statistical modeling and application of differential equations through data using functional data analysis approach.
Chapter 1
Introduction and Literature Review

Cancer is a major public health problem in the United States and around the globe. Cancer accounts for nearly one quarter of the total deaths and ranks second after heart disease in the United States. The number of new cases and deaths in the United States is expected to be 1,660,290 and 580,350 in 2013 respectively [71]. Most of the cancers are in fact related to behavioral factors that can easily be modified. Some of the factors include genetic history, diet, tobacco use, physical inactivity, etc. Making progress against cancer is not a simple problem. It needs a commitment from all components that are associated with human factors. A good cancer research includes early detection, prevention, and reduction in mortality. Global and national policy to fight against the cancers are essential, and that requires a strong commitments from all sectors. The first part of this dissertation is the study of the mortality behavior of different cancers. Study of mortality trends in cancer are the most reliable study to measure the progress against cancer. This study reflects important insight in prevention, early detection, and treatment [32]. Study of mortality and incidence trends follow the same method, so throughout this dissertation we mention the incidence in parenthesis if it is applicable.

1.1 General Objectives

The general objective of the present study is to estimate the temporal trend for mortality (or incidence) of a particular disease in a large population setting. The statistical model which estimates and predicts the trend well is in essence a guideline for good management
practice to ensure the risk associated with the disease. These trends uncover the facts related to the cancer that helps to understand the risks and make health-related decisions in public policies to decrease the public’s risk of mortality or developing cancer. Having good estimates of the mortality (or incidence) rates will allow us to detect points in time where significant changes occur and provide the best possible predictions. Having good estimates and predictions of such mortality rates not only help us to monitor and evaluate the current status of the disease, but also to make an evidence based policy for resources allocation. More practically, it helps us to monitor the progress we are making in the particular disease, and evaluate the effectiveness of current treatment methods with respect to the mortality rate. The obtained numerical mortality (or incidence) rates, their Annual Percentage Change (APC) and predictions are in fact the measure of disease burden. These measures can be seen as tolerable measures in the national socio-economic context.

Given the estimated mortality (or incidence) rates at a particular time, we can measure the status of the disease. If the slope of estimated mortality (or incidence) rate decreases at a particular time, we can conclude that that we are making progress against the disease. If the slope of the estimated rate does not change (zero), then the mortality (or incidence) of the disease is constant indicating that we are not making any progress in the status of the disease. If the slope increases at a particular time, then this is the indication that the mortality (or incidence) rate is increasing recommending the policy makers to take an action against all the existing medical interventions.

Incidence (or mortality) due to cancer varies disproportionately among different population subgroups. These variations are due to tumor biology, genetics, hormonal status, lifestyle and behavior, screening policies, environmental exposure and risk, quality of interventions and response to therapy, and post-therapeutic surveillance. Understanding the actual behavior of the mortality trends due to cancer in society contributes to looking at the cancer interventions and helping to reduce the cancer burden in the United States. In fact, understanding the mortality (or incidence) behavior for different subgroups of the
population and over all in a population are an integral part to compare the trends between subgroups of patients that helps policy makers and scientists for planning public health programs and medical interventions. Our study helps to capture the variations among different age groups and other applicable covariates such as gender in the population while studying the mortality behavior.

1.2 Data Source

National Cancer Institute (NCI) routinely collects data on different types of cancers covering approximately 28 percent of the United States (U.S.) population through its Surveillance, Epidemiology, and End Results (SEER) program. This program is only an authoritative source of comprehensive source of population based information on cancer in U.S. and was funded by NCI in 1973. Currently SEER program collects the population based data from 20 registries covering around 28% of the U.S. population, and publishes an annual progress report to the nation on the status of the disease [53]. The data set for all application to develop and study the mortality trends are obtained from the Surveillance, Epidemiology, and End Results (SEER) program of National Caner Institute (NCI)[69].

This database is very popular among researchers around the world to study different kinds of cancers. It routinely collects data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. SEER works very closely with National Center for Health Statistics (NCHS) of Center of Disease Control (CDC)[18], North American Association of Central Cancer Registries (NAACCR) and the Census Bureau to obtain the information and to update its database annually. North American Association of Central Cancer Registries (NAACCR) has an interactive tool to access the data quickly for major cancer sites by age, sex, race etc. [52]. NCHS data source include birth and death certificates, patient medical records, standardized physical examination and lab tests, and facility information. SEER collaborates with NCHS to obtain these records. The SEER team has developed computer software to
disseminate, analyze and interpret the data. Since it covers almost one third of the U.S. population in its database, it takes time to collect and process its database, and it lags by three years.

According to SEER, the overall goal of this program is to collect complete and accurate data on all cancer patients from all registries, conduct a continual quality control and quality improvement program, periodically report the status of the disease to the nation, identify the unusual changes and differences in the patterns of cancers, describe the temporal changes in cancer incidence, mortality, extent of disease at diagnosis, therapy, and patient survival, monitor the occurrence of possible iatrogenic cancers, collaborate with other organizations on cancer surveillance activities, serve as a research source to researchers, and provide training and web-based training resources to the registries.

All researchers can access the SEER data after signing a contract with SEER program. The research data base is available in an ASCII test format or in the binary format using the SEER*Stat software (software developed by SEER to extract and analyze the data) [68]. There are three methods to access the data: 1) Using SEER*Stat through internet connection. 2) Downloading compressed files from the internet. 3) Obtaining the DVD containing the data via mailing. In this dissertation, we used the mortality data for different types of cancers accessed through the SEER*Stat software of NCI.

1.2.1 Crude and Age-adjusted Rates

We are interested in studying the mortality (or incidence) trends in a population and different sub groups of a population. SEER database provides the mortality (or incidence) counts and rates in the two different methods as described below.

A crude rate is obtained by dividing the total number of events at a specific time period by the total number of individuals in the population at risk. Generally these rates are provided in 100,000. To study the crude rates we pull out the record of the total number of events from all the age groups that we are interested in and divide it by the total number
of population of those age groups. When we are interested in the summary measure only, we study the crude mortality (or incidence) rates. This helps to study the overall burden of a particular disease in a population. Crude rates are the study of rates irrespective of other desirable factors such as age distribution.

Age-adjusted rates are the measure when we are comparing the rates of age-defined subgroups when rates are strongly age-dependent. We mostly use an age-adjusted rate to reduce the confounding effect of age while comparing rates across the different populations or within different subgroups of a population across time. Most of the public health studies demand the age-adjustment. Age-adjustment is done by multiplying the crude rate of different age groups by the standard population of that particular age group and sum the rates across all age groups. The age-adjusted rates at time \( t_i \) are given by

\[
r_i = \sum_{j=1}^{J} w_j \frac{d_{ij}}{n_{ij}}, \quad i = 1, 2, \ldots, n,
\]

where \( y_{ij} \) is the event for \( j^{th} \) year age group subjects at time \( t_i \), \( n_{ij} \) is the \( j^{th} \) year age group population at risk at time \( t_i \), \( w_j \) are the normalized proportion of mid-year population for the \( j^{th} \) age-group in the standard population such that \( \sum_{j=1}^{J} w_j = 1 \). NCI and SEER have divided the population in five year age groups to adjust the confounding effect of age in the population.

### 1.3 Joinpoint Regression Model

The study of the mortality (or incidence) trends of a data set comes with the detection of change points in the trends. It is important to determine whether the change has taken place or not while studying the trend behavior. Moreover, getting the smooth mortality (or incidence) curves including the capability of detecting change points is useful to actuaries and policy makers. The detection of those change points includes the location of the change points and their directions. One of the general objectives of the present study is to estimate
(detect the change points and estimate their slope) the temporal trend for mortality (or incidence) of a particular disease in a large population. In this process we wish to select the best model with significantly minimum number of change points that describes the trend. The process is carried out in such a way that if we add one more change point in the model, the model becomes statistically insignificant.

Although the concept of change point and joinpoint is same in studying the trend line, the estimation of change points using joinpoint regression and change point regression are different statistical procedures. As opposed to the change point regression analysis, which allows different sections of the data to follow different probability distributions and fitting the models based on each data sections, joinpoint theory works with a complete data set in time trend and searches for the peak points in the trend by estimating their locations and slopes. The common limitations of change point analysis include the determination of the number of change points in the time trend and sometimes likelihood of the data are not sufficient in between two change points to fit a separate model.

According to NCI, joinpoint regression model is a piecewise linear regression model that characterizes the trend behavior in the data by identifying the significant points where changes occur. This will be carried out by detecting the points and their locations within the data range. Although the jointpoint regression model can be used for different purposes, it is widely used in epidemiological studies such as incidence, mortality, or survival of a population to unveil the disease trend. The main objective of such a study is to give the reliable estimates of the incidence, mortality, or survival rates that provide up-to-date information and recent changes in its trend. The joinpoint regression model is preferable when analyzing the trend for several years as it enables the identification points in the trend where the significant changes occur (Kohler, et al., 2006). NCI uses joinpoint regression to study the trend of the disease as it is preferable to single linear regression when sufficient number of years are available [41].

We develop the Bayesian Joinpoint Regression Model and apply this model to study the
mortality (or incidence) trends in the population and different sub groups of population. Our developed model in this study can be fitted for both mortality and incidence data. However, we choose to study the mortality trends because of its importance in reflecting the real status of the disease in population as described in the beginning of the introduction.

1.4 Annual Percentage Change (APC)

Annual Percentage Change (APC) is used to characterize the behavior of the cancer trends. The estimated APC is the percentage change (increase or decrease) in the estimated cancer rates per year in the time trend. More specifically, it estimates the rate of change of mortality (or incidence) rate from \( t^{th} \) year to \((t + 1)^{th}\) year. This measure helps us to compare the different types of cancers among the different subpopulations across time. It is calculated by fitting a linear regression to the natural logarithm of the annual rates using the calendar year as the predictor variable as given by

\[
\ln(r_t) = b_0 + b \times t
\]

where \( \ln(r) \) is the natural log of the rate in year \( t \). Then the APC from year \( t \) to \( t+1 \) is given by

\[
\frac{r_{t+1} - r_t}{r_t} \times 100 = \frac{e^{(b_0+b(t+1))} - e^{b_0+b \times t}}{e^{b_0+b \times t}} \times 100 = (e^b - 1) \times 100.
\]

1.5 Literature Review and Limitations of the Currently Applied Joinpoint Models

In this section, we highlight some of the major contributions made so far in the development of joinpoint regression. Although joinpoint regression has been in practice under different
names such as change point regression, piecewise regression, segmented regression, spline regression from the early ’70s, it has received considerable attention among scholars when Kim et al. in [37, 38] proposed a nonparametric method of joinpoint regression. This model is widely used for analyzing and predicting the mortality and incidence data. NCI uses this methodology among others to find the trends in mortality, incidence, and survival of cancers in the United States. This method detects the joinpoints in the trends by using a numerical search method and fits the linear regression between two consecutive joinpoints using least square approach. The final number of joinpoints are selected by using a series of Permutation Tests Based (PTB) approach or the Bayesian Information Criteria (BIC). A short description of this approach is given in chapter 2. The method applied by NCI based on this approach may be useful to summarize the trend but it does not properly characterize the trend. The application to childhood brain cancer mortality rate is provided in chapter 3 as a comparison to Bayesian approach. Although this technique is in extensive use, its limitations are prominent [48].

In 1992, Charlin et al. developed hierarchical Bayesian analysis of changepoint problem in which they used an iterative Monte Carlo method [12]. This is one of the notable works to fit the Bayesian changepoint regression. In 2005, Tiwari et al. developed a Bayesian model selection approach for discrete joinpoint regression [72]. They obtained log of the age adjusted rates $y_i = \ln(r_i) = \ln\left(\sum_{j=1}^{J} w_j \frac{d_j}{n_{ij}}\right), i = 1, 2, ..., n$, and fitted the model considering the errors are independent and identical (IID) Normal distributions. They observed that this log-linear model is useful in modeling and interpreting the trends since the cancer rates arise from a Poisson distribution which is skewed. Later, they relaxed that assumption by assuming that the errors are normally distributed with mean zero and variance $\omega_i \sigma^2$ with known weights $\omega_i$. They assume that the spacing between two data points is constant. Later they relaxed that assumption by augmenting the data by inserting a certain number of equally spaced points. In their approach, they used two criteria to select the best model: one with the smallest BIC and the other with the Bayes Factor. Their result
performed better with BIC criteria, a significant improvement over the permutation test as discussed in [37, 38]. The mortality, incidence, and survival data in a given population are widely analyzed through the joinpoint software of NCI that is based on Permutation Test and Bayesian Information Criteria. The common impediment of both of these approaches is that the joinpoint occurs at the observed discrete time. Although the age-adjusted model fitted by Tiwari et al. provides a measure of uncertainty related to the number of joinpoints in trends, the assumptions made on the errors are IID normal similar to that of Kim et al (2000).

All of the previous studies assumed that the errors are IID normal for non adjusted mortality (or incidence) rates or they assume normal error for the logarithm of the age-adjusted rates. This is not relevant with real world applications, such as mortality and incidence due to a specific disease in a population. This normality assumption for error in modeling the joinpoint regression is relaxed by Ghosh et al. (2009) proposing a Bayesian approach on parametric and semi-parametric joinpoint regression model [24]. This was the first semi-parametric approach to fit the Bayesian joinpoint regression model. They introduced a continuous prior for the joinpoints induced by the Dirichlet distribution that allows the user to specify the minimum gaps in between two consecutive joinpoints. They relaxed the parametric assumptions using the Dirichlet Process Mixtures (DPM). They developed two semi parametric generalizations of the parametric model by modeling the error Dirichlet Prior and the slope Dirichlet Priors by relaxing the parametric assumption on the random slope and error. They applied Deviance Information Criteria (DIC) and Cross-validated Predictive Criteria to access the best model. Their error-DPM model provides robust prediction. However, they assumed fixed number of change points in the model and estimated the trends based on those fixed number of change points.

In 2009, Ghosh et al. [25] applied semiparametric Bayesian approaches to study the population based survival data using joinpoint regression. They use Bayesian joinpoint regression to study the survival trend. Their model is based on Poisson distribution to
study the relative survival in population. They used Dirichlet process mixture in studying
the regression slopes. In 2010, Ghosh et al.[26] fitted a semi-parametric Bayesian age-
stratified Poisson regression model to summarize the trends in cancer rates. All of the
previous works fitted the model by taking the logarithm of the age-adjusted rates as a linear
function of time. In their work, they considered the Poisson probability distribution for
the occurrence of death due to a particular disease in a population. They applied semi-
parametric Bayesian modeling in estimating the parameters by estimating the age-specific
intercepts parameters non-parametrically. Also, they assumed a mixture distribution with
point mass at zero for the slope that changes at the joinpoint. However, their method
assumes that the maximum number of joinpoints is known.

The generalized linear model with log link function in joinpoint regression model that
evaluates and incorporates the uncertainty in both model selection and model parameters
has been recently introduced and implemented by Martinez-Beneito et al. (2011)[48]. They
proposed a joinpoint regression model based on the Poisson assumption in which they find
a suitable reparametrization method to handle the joinpoints. They claimed that the de-
developed model is sophisticated enough to handle the uncertainty related to the model and
its parameters. In their application, they only used annually observed mortality counts
(non age-adjusted counts or crude counts) to fit the data without taking into account age-
standardized rates. Also, they did not consider the possible covariates that explain the
mortality (or incidence) in the model. Lack of both of these points is due to the computa-
tional burden in the model. However, despite the fact of the computational burden, the
possibility of incorporating the applicable covariates that explain the variation among the
different population sub-groups can not be undermined. Their developed model can be fit-
ted for an infinite number of joinpoints in the time trends. However, the uncertainty issues
related to the detection of those change points needs to be studied.
1.6 Modeling Objectives and our Proposal

The study of mortality (or incidence) trends is done in two different ways: the age-specific or age-groups mortality (or incidence) rates and age-adjusted rates. Both methods are equally important to study the behavior of trends. The age specific groups help us to study the overall trend for that group only. However, this information is important to know for more accurate future estimates for that particular age group, but in an epidemiological study, the potential confounding effect of age is another important factor if we are interested in comparing the mortality (or incidence) trends in different population subgroups. This effect is reduced by computing the age-adjusted incidence or mortality rates using the same standard population (NCI). These rates are indeed an important measure as they compare cancer trends in different population subgroups, areas, etc. Also, other major factors such as gender and race that influence the mean of the disease outcomes should be taken into consideration, especially when comparing trends. In practice, the covariates in the model are considered only for linear joinpoint regression models with the assumption of normality [24, 37, 72]. The models developed so far to analyze such trends lack at least the age standardization, or the incorporation of the covariates in the model or the Poisson model assumptions. Moreover, the potential effect of uncertainty related to the model and its parameters is always an important issue in model selection problem using Bayesian approach [14].

In the present study, we propose an age-stratified Bayesian joinpoint regression model with the adjustment of other applicable covariates in the model that can be fitted for both mortality and incidence data and the age-standardized mortality and incidence rates, and their Annual Percentage Change (APC) values can be investigated thereafter. Our work in this study extends the previous works in different dimensions. Being rare events, we assumed that the observed mortality counts are assumed to follow the Poisson probability distributions. The actual model is solely based on Bayesian method of model selection by
considering joinpoints as continuous random variables, often referred as a variable selection uncertainty problem [14]. We assume common slope on fitting the age-stratified models to reduce the computational burden. Here, our proposal is on the posterior quantification of post data uncertainty related to the detection of joinpoints, and since we can have an infinite number of joinpoints in the model, the manual elicitation of priors are not feasible [7]. In the proposed model, the belief propagation for performing model inferences and predictions is done with the help of parameter inferences (posterior search), model inferences and model averaging [58]. The inferences for parameters and uncertainty related to them are handled in such a way that the chosen priors are automatically derived from the model index contained in the model space [7]. Model inferences choose the best model based on the Bayes Factor with highest posterior probability and Deviance Information Criterion (DIC), and model averaging approach is applied to obtain the best possible predictions. In the following sections we define all the necessary terms and the literature review that we use in our study.

1.7 Generalized Linear Models (GLM)

Most of the continuous outcome data with $Y_i$ independent are fitted using the Linear Statistical Models of the form $E(Y_i) = \mu_i = X_i^T \beta$ where $X_i$ is vector of explanatory variables, and $\beta$ is vector of parameters, and $Y_i \approx N(\mu_i, \sigma^2)$. This describes the linear relationship between the response and the explanatory variables. However, in real life, the relationship between the variables may not be linear as explained above. Moreover, in nature, the response variables sometimes have distributions other than Normal, for example, categorical, or count data. Since we are interested in mortality (or incidence) of a disease in population, the response (count) does not follow the normal distribution. Nelder and Wedderburn (1972) developed generalized linear model as a natural advancement over the existing normal model and are based on the exponential family of distribution[55]. The major advances of this work are the recognition of exponential family of distribution, the family of distri-
butions which share the many properties of Normal distributions, and the estimation of the parameters vectors $\beta$ for nonlinear function.

A member of the exponential family has a probability density function that can be written in the following form;

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i \theta_i - \varphi(\theta_i)}{\phi} + c(y_i, \phi)\right),$$

where $y_i$ are a set of independent random variables, $\theta_i$ are unknown parameters associated with $y_i$, $\phi$ is scale parameter, and $\varphi(\theta_i)$ is a function that gives the conditional mean and variance of $y_i$. The distributions of each of $y_i$ has a canonical form and depends on a single parameter $\theta_i$.

Let $E(Y_i) = \mu_i$, where $\mu_i$ is some function of $\theta_i$, then the generalized linear model is given by

$$g(\mu_i) = X_i^T \beta,$$

where the function $g$ is a monotone, differentiable function called the link function, $X_i$ is vector of explanatory variables, and $\beta$ is vector of parameters.

A generalized linear model consists of the following three components.

1. A random component that specifies the conditional distribution of the response variable given the values of the explanatory variables. The response variables are assumed to have the same distribution that is coming from the exponential family of distributions.

2. A linear function of parameters vector and explanatory variables.

3. A smooth and invertible mathematical function, called the link function, which transforms the expectation of the response variable to the linear predictor.
1.8 Bayesian Model Selection Criteria

We assume that the jointpoints behave as continuous random variables and since the derivatives of the log likelihood with respect to joinpoints does not exist, the Bayesian approach is the reasonable choice. Moreover, the previous studies already focus on the advantages of using the Bayesian approach over the frequentist approach. The final model to estimate the trend is based on the Bayesian method of model selection by considering joinpoints as continuous random variables.

In statistical theory, to obtain an optimal statistical model from a set of competing models is always an important problem, the optimal model in the sense that it should be parsimonious, provide the best fit, and estimate best possible prediction with a certain level of confidence. In Bayesian model selection criteria, the solution is obtained in the form of parameter estimation by finding the posterior probability of all competing models.

In the Bayesian literature, there are different approaches to select the best model and each of these methods uses the rule based on the probability theory under different hypothesis. Some commonly used methods are described in a review paper by O’Hara and Sillanpaa [23]. To choose the best joinpoint regression model that best describes the mortality (or incidence) of trends with incorporation of uncertainty in the model and its parameters, we use the Bayesian method of model selection criteria based on Bayes Factor and Deviance Information Criteria (DIC). Bayes Factor is more robust, avoids model selection bias, evaluates evidence in favor or the null hypothesis, incorporates model uncertainty, and are suitable to test for non-nested models [36]. We also used Deviance Information Criteria (DIC) which is a Bayesian version equivalent to classical deviance for model assessment. DIC are suitable for comparing less than dozens number of candidate models [23]. In our work, as we know that mortality (or incidence) in a population due to a particular disease does not change significantly from year to year, we assume our change points (random predictor variables) along with other applicable covariates are not in big numbers. Moreover, DIC is
an efficient and straightforward way in defining an effective number of parameters in the
model and identifying the optimal model [2].

1.8.1 Bayes Factor and Model Uncertainty

Bayes Factor is a Bayesian method to test the hypothesis. The Bayesian method to test the
hypothesis was first developed by Jeffreys in 1935 [30]. According to him, the purpose of
hypothesis testing is to evaluate the evidence in favor of a scientific theory. This method
evaluates the evidence in favor of null hypothesis by incorporating the external information.

In statistical theory, model building process requires a lot of work. We usually have a
set of predictor variables, and we usually start our statistical analysis and modeling process
with determining whether those variables have any outliers or not. In the next step, we
check whether we need to transform those variables or not. Finally we would like to know
how many of the predictor variables explain the response statistically, or what are the pos-
sible combinations among the variables that best describe the response. To find the optimal
model, we compare different competing models with different set of parameters based on
a series of significance tests. If we are using complex models, we rely on approximate
asymptotic distributions to test the hypothesis. As explained by Kass and Raftery [36],
there are several problems associated with this process. According to Freedman (1983),
Miller (1984,1990), the sampling properties of individual and the overall test strategies are
not well understood [22, 50, 51]. The statistical model being tested are not nested are the
other problems associated with the tests. So, the selected statistical model and the infer-
ences based on that model are subject to questions related to the model uncertainty. All
of these uncertainties and the problems related to the selection of the best model can be
avoided by using the Bayesian method of model selctions based on Bayes Factor [44, 68].

The Bayesian comparison of two competing models \( m_1 \) and \( m_2 \) is done in the following
way using the Bayes Factor. Let \( p(D|m_1) \) and \( p(D|m_2) \) be the probability densities of
the data with respect to the models \( m_1 \) and \( m_2 \), where model \( m_1 \) or \( m_2 \) are the models under
the hypothesis $H_1$ and $H_2$. Then the Bayesian comparison of two competing models $m_1$ and $m_2$ is done by obtaining the ratio of their posterior probabilities as given below

$$PMO(\text{Posterior Model Odds}) = \frac{p(m_1|D)}{p(m_2|D)} = \frac{p(D|m_1)}{p(D|m_2)} \times \frac{p(m_1)}{p(m_2)} = B_{12} \times \frac{p(m_1)}{p(m_2)}$$

where $B_{12}$ is called the Bayes Factor of model $m_1$ versus model $m_2$, $p(m_1)$ and $p(m_2)$ are the prior model probabilities, and the marginal likelihood $p(D|m)$ for $m \in \{m_1, m_2\}$ is given by

$$p(D|m) = \int p(D|\theta_m, m)p(\theta_m|m)d\theta_m$$

where $p(D|m)$ is the likelihood of model $m$ with parameters $\theta_m$, and $p(\theta_m|m)$ is the prior of $\theta_m$ under model $m$.

In summary, Posterior Model Odds = Bayes Factor*Prior model odds. That is, Bayes Factor is the ratio of the posterior odds of model to its prior odds.

If no information is available regarding the model, then equal prior probabilities are considered for each of the competing models. If this is the case, model comparison and evaluation are based on Bayes Factor only. Also, the posterior odds ratio and its corresponding Bayes Factor actually evaluate the evidence in favor of null hypothesis. These are the added advantages of the Bayesian model testing using the Bayes Factor compare to the classical hypothesis test.

### 1.8.2 Deviance Information Criteria

Deviance Information Criteria (DIC) is a Bayesian method of model comparison and adequacy. This is the generalization of Akaike Information Criterion (AIC) for Bayesian models fitted using MCMC methods to choose the most parsimonious model, with wider applicability and can be applicable to any class of models [67]. In the frequentist approach,
the adequacy of the fitted model is checked by comparing it with a more general model with the maximum number of parameters in the model, called the saturated model \[21\]. Dampster (1974) suggested an approach for Bayesian model selection, analogous to frequentist approach by examining the posterior distribution. This approach is based on comparing the plots and the summary of the posterior means \[19\]. Spiegelhalter et al. (2002) developed the Deviance Information Criteria (DIC) as a Bayesian model choice criteria based on Dampster’s suggestion \[67\]. DIC consists of two components; the first measures the goodness of fit and the second is a penalty term for the model based on the number of parameters in the model. As the complexity of the models increases the penalty term also increases. DIC is mathematically represented by

\[
DIC = \overline{D} + P_D
\]

1. The Bayesian method of model fit is defined as the posterior expectation of the deviance as given by

\[
\overline{D} = E_{\theta|data}(D(\theta)) = E_{\theta|data}[ -2 \ln f(data|\theta)]
\]

where \(f(data|\theta)\) is the likelihood function. The model that fits the data well is called the better model. In this case the likelihood values are larger. Hence, in the above posterior expectation being negative value, the smaller value of \(\overline{D} \) is the better model.

2. The second component associated with the penalty term measures the complexity of the model that is based on an effective number of parameters, related with the term parsimonious model. The effective number of parameters \(P_D\) is defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean \(\overline{\theta}\) of the parameters.
\[ P_D = \bar{D} - D(\bar{\theta}) = E_{\theta|\text{data}}(D(\theta)) - D(E_{\theta|\text{data}}[\theta]) = E_{\theta|\text{data}}[-2 \ln f(data|\theta)] + 2 \ln f(data|\theta) \]

On rearranging the terms given above, the Deviance Information Criteria is given by

\[ DIC = \bar{D} + P_D = \bar{D}(\bar{\theta}) + 2P_D. \]

We can define \(-2 \ln f(data|\theta)\) as the residual in the data conditioned on the model parameters. This can be interpreted as a measure of uncertainty. Then the above expression can be regarded as the expected increase in the true residual over the estimated residual, indicating that \(P_D\) can be interpreted as the expected reduction in uncertainty due to estimation [2].

In this approach of model comparison, we find the DIC values of each of the competing models with different possible and applicable parameters and a model with smaller DIC value is selected as better-fitting model. The DIC should not be used in the case where the posterior distributions are not symmetric or unimodel. Because of the central assumption of DIC for posterior summary as good summary, it should be used with caution [56].

1.9 Markov Chain Monte Carlo Method (MCMC)

The integral involved in computing posterior probabilities under each models is not analytically tractable. The random number generating methods are very popular in Bayesian statistical inference. In Bayesian approach for every function of the parameter of interest not being analytically tractable, we generate samples from the posterior distribution and calculate its sample mean. This method is easy to use but the problem associated with it is to generate the samples from the posterior density. If the posterior distribution of the parameters are not analytically tractable, there are several methods derived in the litera-
ture, such as inverse cumulative distribution function, rejection sampling algorithm, and importance sampling. This is a direct method of simulating the posterior samples of the parameter of interest and is suitable only for one dimensional distributions. Some of these methods are good to use for the computation of specific integrals instead of obtaining the samples from the posterior distributions of parameter of interest.

The simulation techniques based on Markov Chain (MC) are called Markov Chain Monte Carlo (MCMC) methods which overcome the problems mentioned above. They are very flexible and general with great computing efficiency that can be used to estimate the posterior distributions of the parameters of interest with high accuracy. MCMC methods are based on the Markov Chain (MC) that converge to the posterior distribution of the parameter of interest. The samples obtained using the MCMC are iterative and the values produced in every step depend on previous steps as it is generated from Markov Chain. The algorithm is described as follows;

A Markov Chain is a stochastic process \( \{ \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots, \theta^{(t)} \} \) such that
\[
f(\theta^{(t+1)}|\theta^{(t)}, \theta^{(t-1)}, \ldots, \theta^{(1)}) = f(\theta^{(t+1)}|\theta^{(t)})
\]
This means that the distribution of \( \theta \) at step \( t + 1 \) given all of its previous steps depends only on its previous step. Also \( f(\theta^{(t+1)}|\theta^{(t)}) \) is not dependent on time and as \( t \to \infty \) the distribution of \( \theta^{(t)} \) converges to its equilibrium distribution and that is independent of the initial values of the chain \( \theta^{(0)} \).

Here we need to generate samples from \( f(\theta) \) and that is done by constructing a Markov chain in which \( f(\theta^{(t+1)}|\theta^{(t)}) \) is easy to generate and it should be the posterior distribution of the parameter of interest. Once we construct the Markov chain with the above properties, we follow the following steps;

1. Select an initial value for \( \theta^{(0)} \)

2. Generate samples until the target distribution is reached.

3. Monitor the convergence of the algorithm. This can be done by checking the con-
vergence diagnostics. Generate more samples from the target distribution until the algorithm reaches its equilibrium condition.

4. Disregard some initial observations as burn in period.

5. Consider the remaining samples after the burn in period as the sample for the posterior distribution.

6. Plot and obtain the summaries of the posterior distribution.

Metropolis Hasting and Gibbs Sampling are the two most popular MCMC methods. We apply the Gibbs sampling method in our study which will be discussed below;

1.9.1 Gibbs Sampling Algorithm

Gibbs sampling is a MCMC algorithm used to obtained a sequence of samples from the posterior distribution of the parameters when the direct sampling methods are very difficult. It was introduced by Geman and Geman in 1984. This is a special case of Metropolis-Hasting algorithm. We describe the Gibbs sampling algorithm in the following steps.

1. Set initial value for $\theta^{(0)}$

2. For $t = 1, 2, 3, \ldots, T$ repeat the following three steps

a) Set $\theta = \theta^{(t-1)}$

b) For $j = 1, 2, 3, \ldots, J$, update $\theta_j \sim f(\theta_j | \theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_J, y)$

c) Set $\theta^{(t)} = \theta$ and save it.

On applying this, for a particular value of $\theta^{(t)}$, we generate the parameters values as given by

$$
\theta_1^t \sim f(\theta_1 | \theta_2^{t-1}, \theta_3^{t-1}, \ldots, \theta_J^{t-1}, y)
$$

$$
\theta_2^t \sim f(\theta_2 | \theta_1^{t-1}, \theta_3^{t-1}, \ldots, \theta_J^{t-1}, y)
$$
\[ \theta_j^t \sim f(\theta_j | \theta_{1,j}^{t-1}, \theta_{3,j}^{t-1}, \ldots, \theta_{j-1,j}^{t-1}, \theta_{j+1,j}^{t-1}, \ldots, \theta_J^{t-1}, y) \]

Here, generating values from \( f(\theta_j | \theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_J, y) \) is relatively easy as it is a univariate distribution for \( \theta_j \) keeping the rest of variables as constant.

1.10 Prediction of Trends

One of the goals of statistical analysis is to make a forecast. By now, the literature is very rich to describe the predictive model for cancer mortality in both Frequentist and Bayesian approach. Extrapolation assumes that a future trend is the continuation of the past which is the basis for the most mortality forecasting methods [6]. Some examples of extrapolative forecasting methods are Box et al.[8], White[77], and Denton et al.[20]. Another well known forecasting method is Lee Carter model [43], which has shown to represent a large proportion of the variability in mortality rates in developing countries [73]. However, it assumes that the ratio of the rates of mortality change at different ages remains constant over time. It also lacks across-age smoothness and becomes increasingly spikey over time [27]. Czado, Delwarde and Denuit [17] used the Poisson log-bilinear model together with Bayesian approach to impose smoothness. Girosi and King[27] used the Bayesian method in forecasting mortality as an extension of the structural model. In this paper, we applied the Bayesian Model Average (BMA) approach to predict the future mortality rates. It was first proposed by Leamer in 1978 and applied it in linear regression model [42]. This approach
is coherent and effective to account for model uncertainty in which the predictions and inferences are based on a set of models that contribute proportionally based on the support it receives from the data [13]. More specifically, BMA averages all competing models by incorporating the model uncertainty into conclusions about the parameters which classical statistical analysis fails to do. Madigan and Raftery [47] mentioned that averaging over all the models provides better average predictive performance.

1.10.1 Bayesian Model Averaging

Model uncertainty is an important issue in the statistical analysis and modeling of data which is ignored in standard statistical procedure. In a certain statistical procedure, we are interested only in one model among the set of competing models. We believe that the selected model was generated by the given data ignoring the uncertainty in the model selection approach [28]. Instead of giving rise to a single model, the Bayesian Model Averaging (BMA) averages across a large set of models and make an inference based on a weighted averages on these models over the model space. The estimates of the parameters and models are robust as it calculates the posterior predictive distributions over parameters and models. In the model selection process using BMA each competing model with a set of variables receives some weight. The final model is the estimates of those weighted averages we get from each model. In this way, BMA incorporate all the applicable variables in the analysis by providing certain weight to the models including those variables. If the variables are not contributing a lot then it provides less weight for that model containing those variables and vice versa.

Let \( M = (M_1, M_2, \ldots, M_k) \) be the set of competing models where each model is comprised of certain attributable variables. Let \( \Delta \) be the quantity of interest which may be a model parameter or future observation, then the posterior predictive distribution of \( \Delta \) given the data \( D \) is given by
\[ p(\Delta|D) = \sum_{i=1}^{k} p(\Delta|D, M_i)p(M_i|D). \]

This is known as the average of the posterior predictive distribution of \( \Delta \) under each of the competing models weighted by the corresponding posterior model probability given data, i.e. \( p(M_i|D) \) where,

\[ p(M_i|D) = \frac{p(D|M_i)p(M_i)}{\sum_{i=1}^{k} p(D|M_i)p(M_i)}, \]

where

\[ p(D|M_i) = \int \ldots \int p(D|\theta_i, M_i)p(\theta_i|M_i)d\theta_i \]

is the integral likelihood of the model, \( \theta_i \) is the parameter vector, and \( p(\theta_i|M_i) \) is the prior distribution of the parameters, \( p(D|\theta_i, M_i) \) is the likelihood, and \( p(M_i) \) is the prior probability of the true model.

The BMA estimate of the parameter \( \theta \) is obtained by

\[ \hat{\theta}_{BMA} = \sum_{i=1}^{k} \hat{\theta}_k p(M_i|D) \]

The BMA point estimators and predictions both minimize the mean square error [63]. Also, The BMA estimation and prediction confidence intervals are better calibrated than the chosen one single best models confidence intervals [28].

In this chapter we discussed the epidemiological and modeling objectives of our proposal, Bayesian joinpoint regression model, literature review, the necessary terminologies, and methods that we applied to develop the Bayesian joinpoint regression model to study the mortality (or incidence) trends. The next chapter describes the commonly used joinpoint regression model by NCI, Bayesian joinpoint regression model by Beneito et al. and our proposed age-stratified joinpoint regression model and its Bayesian inference.
Chapter 2
Joinpoint Regression Model

One of the objectives of this dissertation is to develop a Bayesian joinpoint regression model that correctly estimates the mortality (or incidence) trends in a population and provides best possible future predictions. This chapter starts with a short description of the widely used statistical models for the joinpoint regression along with their limitations, and we develop an age-stratified joinpoint regression model to estimate and predict the mortality (or incidence) rates due to certain diseases in the population. We discuss the age-adjusted cancer mortality (or incidence) rates and their APC in population that incorporate the confounding effect of age in an epidemiological study and compare the rates in different population subgroups. We show that our statistical model and methodology helps to reduce the computational burden while adjusting the confounding effect of age estimating and predicting the future rates.

This chapter is divided into four main sections. In the first section, we discuss the Joinpoint Regression Method used by the National Cancer Institute in its Joinpoint Regression Program and its limitations. In the second section, we discuss the Bayesian Joinpoint Regression model to study mortality (or incidence) rates in the general population developed so far in the literature and their limitations. In the third section, we discuss our proposal on age-stratified joinpoint regression model and its Bayesian approach of estimation. We end this chapter with the discussion on the contributions we made in studying the mortality (or incidence) trends. In this chapter we aimed to answer the following questions.

1. What are the problems and limitations of commonly used joinpoint regression models
developed by NCI and other newly developed Bayesian joinpoint regression models in the statistical literature? and,

2. How to resolve the existence problems of joinpoint regression and find a statistical model that correctly estimates and predicts the mortality (or incidence) trends?

2.1 Joinpoint Regression Program of National Cancer Institute

The joinpoint Regression model used by National Cancer Institute (NCI) is a set of different linear statistical models connected together at the joinpoints that is used to describe the mortality, incidence, and survival trends in the data. NCI has its own software called Joinpoint Regression Program to analyze and estimate the trends of a particular disease in the population and uses those trends in its publications to report the status of the disease to the nation [54].

Joinpoint Regression Program takes the trend data produced by SEER*Stat software and fits the simplest statistical model with minimum number of joinpoints that fits the data. This program is very easy to use where the user determines the minimum and maximum number of joinpoints by looking at the observed data. The program starts with a simple linear statistical regression model (no joinpoints) and test it against the model with one joinpoint and so on using a sequence of permutation tests as developed by Kim et al. [37, 38].

Let \( y_i, i = 1, 2, 3, \ldots, n \) denote the mortality or incidence outcome process that describes the behavior as a function of time \( t_i, i = 1, 2, 3, \ldots, n \). Here \( t_i \) can be any covariates rather than time. Let, there be k change points in the data, then the joinpoint regression model with k joinpoints is given by

\[
y_i = \beta_0 + \beta_1 * t_i + \sum_{k=1}^{K} \delta_k * s_k(t_i) + \epsilon_i,
\]

where \( s_k(t_i) = (t - \tau_k)^+ \), and \( a^+ = a \text{ if } a > 0 \), and \( a^+ = 0 \), otherwise, \( \beta_k^\prime = (\beta_0, \beta_1, \delta_1, \ldots, \delta_k) \) are the regression parameters, and \( \tau_k^\prime = (\tau_1, \tau_2, \ldots, \tau_k) \) are the join-
points, and $\epsilon'_i$s are random errors with mean $=0$.

The response $y_i$ can either be count, crude, or age-adjusted rates. We can choose either linear or log-linear (log transformation of rates) model based upon how linear the observed rates or the logarithm of the observed rates are within the data range. The model can be tested for the normality of the residual obtained under both the linear or non-linear fit. The one main reason for using the log transformation for cancer mortality or incidence rates is based on the assumption that those arise from a Poisson distribution which is skewed especially when the cancer is rare. This is the standard way to approximate the skewed distributions to a Normal distribution. Another motivation of using log-linear model is for making the interpretation easy. It gives the constant rate of change per year in between two joinpoints.

The least square fit of this regression model is obtained by using either the grid search method as proposed by Lerman [44] or by using the continuous fitting algorithm proposed by Hudson [29]. The joinpoint software of NCI uses a series of permutation tests based on the grid search method to select the optimal number of joinpoints that best fits the observed data. This method detects the joinpoints in the trends by using a numerical search method and fits the linear regression between two consecutive joinpoints using least square approach. In this approach the permutation test is repeatedly used for testing between two models with a different number of joinpoints. For example, the test procedure sequentially conducts the tests of the null hypothesis of no joinpoint against the alternative of one joinpoint. This test is applied for all possible number of joinpoints that could possibly exists in the data and selects a final model with a certain number of joinpoints selected by using a series of Permutation Tests Based (PTB) approach or the Bayesian Information Criteria (BIC). This means the program tests whether more joinpoints are statistically significant to describe the nature of the observed rates or not for all possible number of joinpoint. The software chooses the minimum number of joinpoints that is sufficient to explain the trends in the data. If we add one more joinpoint in that model, then the model becomes statisti-
cally insignificant. Here, at each level of testing, the models with two different levels of joinpoints are fitted for each of the N permuted data where N is large to generate the permutation distribution of the test statistic [76]. In this approach, the age-adjusted rates are weighted averages (weights are the standard population weights from the census data) of age specific group rates.

The model is flexible enough to incorporate estimated variation for each point (age adjusted rates) and Poisson model of variation. The latest version of software also estimates the trend using the Bayesian Information Criteria (BIC) as developed by Tiwari et al. [72]. The BIC approach selects the model with the optimal number of joinpoints that best fit the data by penalizing the cost of extra parameters (join points). Since the applications have shown that the models selected by the BIC approach tend to fit the data well but they are less parsimonious, the permutation test approach is more favorable to BIC approach. The method proposed by NCI has the following limitations;

1. The model is based on the assumption of normal errors.

2. The model is used for descriptive purpose only. It cannot predict the future mortality (or incidence) rates.

3. The APC measured by the NCI method gives the single APC in between two joinpoins as the trend in between two joinpoins is described by a linear line. This is very unusual to assume that the cancer mortality (or incidence) trends increase or decrease over time at the same rate.

4. The joinpoint software of the NCI search the joinpoints at the observed data points only. However, the method proposed by Lerman [44] can be modified to observe the joinpoints at any point in the time trend, but the computation time for this type of grid search method increases dramatically [76].

5. If the mortality or incidence count is zero then the model handles it in different ways;
(a) In the linear model option, the data is analyzed normally.

(b) In the log-linear model option, it is dropped from the analysis.

(c) In the Poisson model, 0.5 is added to each of the counts. These approximations or dropping a particular year observations from the analysis may shift or affect the detection or locations of joinpoints affecting the analysis.

2.2 Bayesian Joinpoint Regression Model

The Bayesian Joinpoint Regression Model is considered very competitive to the Permutation Test Based approach discussed in earlier section. The method based on series of permutation tests to determine the unknown number of joinpoints tends to be conservative based on hypothesis testing and has computational limitations [39]. Moreover, if the data is not very informative the permutation test criteria is biased to the simple model with less number of joinpoints [48], and the quantification of the selected model compared to other competing models is very hard to determine [48]. Contrary to PTB, the main advantage of the Bayesian method is the posterior distribution of the number and location of the joinpoints. This information provides an additional insight in the plausibility of the other joinpoints models which could have been selected [72]. After the development of Joinpoint software and its extensive use in the study of cancer mortality, incidence, and survival rates, researchers around the world are interested in developing the statistical model that best describe the cancer trends. The theoretical research is mostly attractive to access the existence and the location of the joinpoints based on correct model assumption. More specifically, if we assume that the joinpoints are random variables that can occur at any locations within the data range, the log likelihood is not differentiable with respect to break points suggesting that the Bayesian method is a more realistic approach. While doing so, we are interested in using the correct statistical approach. Mostly, we are interested in a statistical model that is based on the probabilistic framework based on the real assumptions.
We want to detect the change points at any time (not only on the observed) in the trend, and the uncertainty related to the detection of joinpoints and the selected model are also an issue in the Bayesian model selection problem which needs to be addressed.

The main objective of this section is to provide a brief description of the Bayesian Joinpoint Regression model to study the mortality (or incidence) rates of the crude data and its estimation procedure currently exists in the literature. We focus on the method developed by Martinez-Beneito et al. (2011)[48] and we close this section with a discussion of some of the limitations of this method.

Let $Y_i, i = 1, 2, .., n$ be the number of mortality (or incidence) counts during a period of time $t_i$ in a population. Let there be $k$ change points that describe the behavior of the data, then the mean of the above outcome process can be expressed as the following generalized linear model

$$g[E(Y_i|t_i)] = \alpha + \beta_0(t_i - \bar{t}) + \sum_{j=1}^{k} \beta_j(t_i - \tau_j)^+, \quad (2.1)$$

where $\bar{t}$ is the mean of $t_i$, and $\tau_j$ is the change point in the model, and $g$ is monotonic and differentiable function, called the link function. The value of $(t_i - \tau_j)^+$ is $(t_i - \tau_j)$ if $(t_i - \tau_j)^+ > 0$ and 0 otherwise.

If there is no breakpoint in the model then

$$g[E(Y_i|t_i)] = \alpha + \beta_0(t_i - \bar{t});$$

and if we have one break point, the model becomes

$$g[E(Y_i|t_i)] = \alpha + \beta_0(t_i - \bar{t}) + \beta_1(t_i - \tau_1)^+.$$

The model with no breakpoint is named as $M_0$, one breakpoint as $M_1$ and so on. There will be $M_{k+1}$ nested models in total depending upon the number of breakpoints.

Since the model can choose an infinite number of breakpoints, we wish to impose some
restrictions on the position of the change points in the model. There are different ways of implementing these restrictions (see [48],[24]). To avoid such identifiability problem, the easiest way to impose such restriction is by choosing the joinpoints in such a way that $t_1 + 2 < \tau_1, t_2 + 2 < \tau_2, \cdots, t_k + 2 < \tau_k$.

The main goal of this modeling approach is to find the trend that describes the behavior of the data. This will be carried out by detecting the points and their locations where the significant changes occur within the data range. Finding such locations in this model selection problem is carried out by using Bayes Factor, in which data updates the prior odds to yield posterior odds. Bayes Factor summarizes the relative support for one model versus another for all competing models by selecting a model with highest posterior probability. Therefore, the posterior probability of each model will be calculated and the one with highest posterior probability will be selected as the best model.

In the proposed model given in (2.1), $\alpha, \beta_0$ represents the common parameters where as $\beta_i's$ are non-common parameters that are model-specific. $\beta_0$ together with $\beta_i's$ gives the slope for the different models with at least one change point. For all common parameters to give the same meaning across models, Martinez-Beneito et al.(2011) proposed an alternative parametrization imposing different conditions. On applying such reparametrization the model in (2.1) becomes

$$g[E(Y_i|t_i)] = \alpha + \beta_0(t_i - \bar{t}) + \gamma z(t_i) + \sum_{k=1}^{K} \delta_k \beta_k B_{rk}(t_i).$$

(2.2)

where $\delta_j, j = 1, 2, \cdots, k$ are binary indicators of the break point in the model. This means that

$$\delta_j = \begin{cases} 
1 & \text{for each break point} \\
0 & \text{otherwise.}
\end{cases}$$

Since the behavior of the mortality (or incidence) count data in the population is a rare event, characterized by Poisson distribution $(Y_i, Pois(\lambda_i, i = 1, 2, \cdots, n))$, it is modeled using natural log link function. Hence, the model in the equation (2.1) becomes
\[
\log(\lambda_i) = \log(n_i) + \alpha + \beta_0(t_i - \bar{t}) + \sum_{k=1}^{K} \delta_k \beta_k B_{\tau_k}(t_i) \tag{2.3}
\]

where \(n_i\) is the total number of population at time \(t_i\).

The estimated rates are obtained by using the following model,

\[
E(r_i) = \alpha + \beta_0(t_i - \bar{t}) + \sum_{k=1}^{K} \delta_k \beta_k B_{\tau_k}(t_i) \tag{2.4}
\]

### 2.2.1 Bayesian Inference and Specification of Priors

The introduction of prior distribution into the model has drawn a lot of interest recently and different criteria have been proposed by many researchers so far. In an objective Bayes solution to the model selection problem, the nature of the posterior distributions depends upon the selection of priors and is very sensitive if there are non-common parameters in the models as explained in Berger and Pericchi (2001) and Bayarri and García-Donato (2008) [1, 3]. For the common parameters \(\alpha\), and \(\beta_0\), we choose flat priors i.e. \(\pi(\alpha, \beta_0) \propto 1\). For non-common parameters, the generalization of Jeffreys divergence-based (DB) priors introduced in [11] and implemented in [48] is considered. As the parameter space is bounded, we can have \(\pi(\tau) \propto 1\). Based on the nature of \(\delta\), it is reasonable to choose independent Bernoulli priors with a probability of success \(p\) with hyper priors for \(p\) being \(Beta\left(\frac{1}{2}, \frac{k-1}{2}\right)\) where \(k\) is the number of joinpoints chosen as given in [48].

In Bayesian paradigm, finding a good candidate model from a set of nested models can be computationally intensive. The distribution of the posterior probability is not analytically tractable, so we used Gibbs sampler approach using WinBUGS software to obtain samples from the posterior distributions. The posterior distribution of the number of joinpoints in the mortality trend will be observed with the model having different number of joinpoints and we choose the one with highest posterior probability.
The model described above is in use in literature [5, 33]. We have also applied this methodology in studying the childhood brain cancer mortality and compare this result with the result obtained by joinpoint software of NCI. This will be discussed in the next chapter of this dissertation. It is based on the correct model assumption; however, it raises couple of questions regarding the estimation of the mortality (or incidence) of a particular disease in a population. The model developed cannot be used for the comparison purpose which is the basic need of epidemiological study while considering the population. The probability of mortality (or incidence) of an individual due to a particular disease in a population among various age groups is different and the model proposed in the above section failed to incorporate this issue. The mean of the outcome of the disease mortality (or incidence) may have significant differences among the different covariate factors, such as gender, race etc. The model proposed is computationally intensive; however, the statistical literature requires a novel approach that can address this issue despite the fact of computational burden. Moreover, the parameter and model uncertainty while applying the Bayesian approach is an important issue that need to be addressed. In the next section, we propose an Age-Stratified Bayesian Joinpoint Regression Model with the incorporation of applicable covariates in the model based on the parallel slope assumption. Our proposal model and its estimation procedure will address these problems we encounter in the statistical analysis and modeling of Bayesian joinpoint regression.

2.3 Age-Stratified Bayesian Joinpoint Regression Model

The Bayesian Joinpoint Regression Model developed so far in the literature as discussed above has several advantages over the joinpoint software of NCI. However, the models fail to address a couple of problems. The potential confounding effect of age is reduced by computing the age adjusted incidence or mortality rates. Also, other major factors such as gender and race that influence the mean of the disease outcome should be taken into consideration, especially when comparing trends. Reduction of computational burden while
adjusting for the confounding effect of age is another major problem that needs to be addressed. In practice, the covariates in the model are considered only for linear joinpoint regression model with the assumption of normality [24, 37, 72]. The models developed so far to analyze such trends lack at least the age standardization, or the incorporation of the covariates in the model or the Poisson model assumptions. Moreover, the uncertainty to detect the joinpoints that arise due to the parameterization by Beneito et al. is an important issue to be addressed.

We assume that there are \( nm \) observed independent responses \( y_{ij}, i = 1, 2, ..., n; j = 1, 2, ..., m \), each coming from an exponential family with probability density function of the form:

\[
f(y_{ij} | \theta_{ij}, \phi) = \exp \left( \frac{y_{ij}\theta_{ij} - \varphi(\theta_{ij})}{\phi} + c(y_{ij}, \phi) \right),
\]

where \( \theta_{ij} \) are unknown parameters associated with \( y_{ij} \), \( \phi \) is scale parameter, and \( \varphi(\theta_{ij}) \) is a function that gives the conditional mean and variance of \( y_{ij} \).

Let there be \( K \) change points that describe the behavior of \( y_{ij} \) as a function of time \((t_i, i = 1, 2, ..., n)\) and other covariates associated with such outcome process. Since each parameter \( \theta_{ij} \) associated with \( y_{ij} \) is not of our interest, we want to detect such \( K \) change points in \( m \) models based on the assumption of common slopes at a particular time for each \( j \)-group for the smaller set of parameters by using a generalized linear model of the form [49], that is,

\[
g[E(y_{ij}|t_i, z(t_i))] = \alpha_j + \beta_0(t_i - \bar{t}) + \gamma z(t_i) + \sum_{k=1}^{K} \beta_k (t_i - \tau_k)^+, \quad (2.5)
\]

where \( g \) is a monotonic, and differentiable function, called the link function; \( \alpha_j \) is the intercept for each group \( j \), \( \beta_0 \) and \( \gamma \) are common slopes, \( \tau_k \) is the location of \( k^{th} \) change point, \( \beta_k \) gives the change of slope at the \( k^{th} \) joinpoint and \((t_i - \tau_k)^+ = t_i - \tau_k \) if \( t_i - \tau_k > 0 \) and zero otherwise; \( z(t_i) \) is the univariate or multivariate covariate process associated with
outcome $y_{ij}$.

Although the model given in equation (2.5) can be used for different purposes with suitable link function $g$, our main goal in the present study is to estimate the temporal trend for mortality or incidence of a particular disease in a large population setting. The probability of a randomly chosen individual in a large population for incidence or mortality due to a particular disease at a given time is very small, then the counts $y_{ij}$ at time $t_i, i = 1, 2, ..., n$ and age-group $j, j = 1, 2, ..., m$ can be modeled by using the Poisson probability distribution, i.e. $y_{ij} \sim \text{Poi}(\mu_{ij}n_{ij})$. And, as exhibited in [35], the mean of the observed outcome process depends on the population size, period of observation and various characteristics of the population such as gender, races, etc. and is given by

$$\ln(\mu_{ij}) = \ln(n_{ij}) + \alpha_j + \beta_0(t_i - \bar{t}) + \gamma z(t_i) + \sum_{k=1}^{K} \beta_k(t_i - \tau_k)^+, \quad \forall i = 1, 2, 3, ..., n, \quad j = 1, 2, ..., m$$

(2.6)

where $n_{ij}$ is the population size at risk in $i^{th}$ year at $j^{th}$ age-group, and $\alpha_j$ is the intercept for the $j^{th}$ age-group. The above equation leads to the following expression that is used to estimate the rate:

$$E(r_{ij}) = \exp(\alpha_j + \beta_0(t_i - \bar{t}) + \gamma z(t_i) + \sum_{k=1}^{K} \beta_k(t_i - \tau_k)^+).$$

(2.7)

Here, inclusion of interaction term(s) between the covariate factors and time is an easy extension but may deviate from the model assumption of common slopes. For example, if $z(t_i)$ is a categorical variable then the magnitude of its interaction with time deviates the assumption of the model giving the changes in the effect of time on outcome across the different group. Even though the interaction terms are not considered in the model in (2.5), we have relaxed this assumption in our application to capture the steepness in the trend.

Since the crude mortality rate at time $t_i$ does not account for the distribution of the
population across various age-groups and is not of interest for epidemiological readership, we usually take the age-adjusted rates at time $t_i$ given by

$$r_i = \sum_{j=1}^{J} w_j \frac{y_{ij}}{n_{ij}}, i = 1, 2, ..., n,$$

where $w_j$ are the normalized proportion of mid-year population for the $j^{th}$ age-group in the standard population such that

$$\sum_{j=1}^{J} w_j = 1.$$

The annual age-adjusted mortality or incidence rate is estimated by

$$E(r_i) = E \left( \sum_{j=1}^{J} w_j \frac{y_{ij}}{n_{ij}} \right) = \sum_{j=1}^{J} w_j E(r_{ij}),$$

where $E(r_{ij})$ is the estimated rate at time $t_i$ for age-group $j$.

The proposed model (2.6) is equivalent to a single model for each group. If we consider different slope models for each group at different times, then we have the following models

$$\ln(\mu_{ij}) = \ln(n_{ij}) + \alpha_j + \beta_{0j}(t_i - \bar{t}) + \gamma_j z(t_i) + \sum_{k=1}^{K} \beta_{kj}(t_i - \tau_k)^+, i = 1, 2, ..., n, j = 1, 2, ..., m$$

with the estimated rate as given below;

$$E(r_{ij}^1) = \exp(\alpha_j + \beta_{0j}(t_i - \bar{t}) + \gamma_j z(t_i) + \sum_{k=1}^{K} \beta_{kj}(t_i - \tau_k)^+).$$

When we apply this to the annual age-adjusted mortality (or incidence) rate;

$$E(r_i) = \sum_{j=1}^{J} w_j E(r_{ij}^1),$$

$$= \sum_{j=1}^{J} w_j * \exp(\alpha_j + \beta_{0j}(t_i - \bar{t}) + \gamma_j z(t_i) + \sum_{k=1}^{K} \beta_{kj}(t_i - \tau_k)^+).$$
\[ \approx \sum_{j=1}^{J} w_j * (1 + \alpha_j + \beta_0 j (t_i - \bar{t}) + \gamma_j z(t_i) + \sum_{k=1}^{K} \beta_{kj} (t_i - \tau_k)^+) \]

\[ = \sum_{j=1}^{J} w_j * 1 + \sum_{j=1}^{J} w_j * \alpha_j + \sum_{j=1}^{J} w_j * \beta_0 j (t_i - \bar{t}) + \sum_{j=1}^{J} w_j * \gamma_j z(t_i) + \sum_{j=1}^{J} w_j * \sum_{k=1}^{K} \beta_{kj} (t_i - \tau_k)^+) \]

\[ = 1 + \alpha + \beta_0 * (t_i - \bar{t}) + \gamma_1 * z(t_i) + \sum_{k=1}^{K} \beta_k (t_i - \tau_k)^+ \]

\[ \approx \exp(\alpha + \beta_0 * (t_i - \bar{t}) + \gamma_1 * z(t_i) + \sum_{k=1}^{K} \beta_k (t_i - \tau_k)^+) \]

\[ = \exp(\alpha) * \exp(\beta_0 * (t_i - \bar{t}) + \gamma_1 * z(t_i) + \sum_{k=1}^{K} \beta_k (t_i - \tau_k)^+) \]

\[ \approx (\sum_{j=1}^{J} w_j * \exp(\alpha_j)) * \exp(\beta_0 * (t_i - \bar{t}) + \gamma_1 * z(t_i) + \sum_{k=1}^{K} \beta_k (t_i - \tau_k)^+) \]

\[ = \sum_{j=1}^{J} w_j * \exp(\alpha_j + \beta_0 * (t_i - \bar{t}) + \gamma_1 * z(t_i) + \sum_{k=1}^{K} \beta_k (t_i - \tau_k)^+) \]

\[ = \sum_{j=1}^{J} w_j E(r_{ij}) \]

This proves that fitting a parallel slope model is equivalent to fitting separate models for each age group when our ultimate goal is to find the age-adjusted mortality (or incidence) rates. Because of this equivalency, instead of fitting a separate model for each age-group, we can fit a parallel slope model for each age-group. The number of parameters to be estimated will be reduced greatly in this method hence reducing the computational burden. Moreover, as explained earlier the inference of the covariate factors and time will capture the trends by relaxing the parallel model assumptions.

The estimated annual percentage change (APC) is used to characterize the trends or the
change in rates over time. Estimated APC from $i^{th}$ year to $(i + 1)^{th}$ year is given as

$$\hat{APC} = \frac{E(r_{i+1}) - E(r_i)}{E(r_i)} \times 100.$$ 

### 2.3.1 Bayesian Inference and Specification of Priors

The assumption of the breakpoints in our proposed model is also random as used by the previous users and applying the Bayesian approach to detect them is a reasonable choice (see [24, 26, 48, 72]). For $k = K$ fixed, we develop a Bayesian model selection procedure to select the best model among $K+1$ nested models in the model space $\{M_0, M_1, \ldots, M_K\} = \Gamma$.

In our proposed model, for a particular age group, say $j = j^*$, the model with no joinpoints (global trend) is observed by estimating the parameters $\alpha_{j^*}$, $\beta_0$, and $\gamma$ where $\alpha_{j^*}$ is the intercept, and $\beta_0 + \gamma$ is the slope of the model i.e.

$$\ln(\mu_{ij^*}) = \ln(n_{ij^*}) + \alpha_{j^*} + \beta_0(t_i - \bar{t}) + \gamma(z(t_i))$$

and the model with one joinpoint is given by

$$\ln(\mu_{ij^*}) = \ln(n_{ij^*}) + \alpha_{j^*} + \beta_0(t_i - \bar{t}) + \gamma(z(t_i)) + \beta_1(t_i - \tau_1),$$

where $\beta_0 + \gamma + \beta_1$ represents the slope. We can assign same priors to all common parameters in all competing models only if they have the same meaning [3]. Beneito et al. [48] proposed an alternative parametrization arguing that the hypothesis of common parameters has the same meaning across the models. We have adopted their reparametrization method. Then the model in (1) becomes

$$\ln(\mu_{ij}) = \ln(n_{ij}) + \alpha_j + \beta_0(t_i - \bar{t}) + \gamma(z(t_i)) + \sum_{k=1}^{K} \delta_k \beta_k B_{\tau_k}(t_i),$$

with $\delta$ being binary indicators for the $k$ breakpoints in the model. This means that
\[ \delta_k = \begin{cases} 1 & \text{for each break point} \\ 0 & \text{otherwise.} \end{cases} \]

The assumption regarding the locations of \( \tau_k' \)s is not fixed in the model space and our goal is to find the minimum number of joinpoints that is sufficient to explain and predict the trend in the data. In such scenario, our problem becomes a variable selection problem.

Here, \( \delta \in \{0, 1\}^k \) is binary inclusion indicators for all non common \( \tau_k' \)s in the model known as latent vector where \( p(\delta|y) \) is the posterior distribution of \( \delta \), an encompassing model under which every other model is nested ([9–11, 45]). Here \( p(\delta|y) \) encapsulates the information about the effectiveness of joinpoints in the model and its inference will be carried out by using Bayes Factor ([31, 36]). The model inference over the model space \( \{M_0, M_1, \ldots, M_K\} = \Gamma \) is given by

\[ p(\delta|y) \propto p(y|\delta) \cdot p(\delta) \]

where the marginal likelihood is obtained as,

\[ p(y|\delta) = \int_{\mathbb{R}^{p+3}} p(y|\alpha, \beta_0, \gamma, \beta, \tau, \delta) \cdot p(\alpha, \beta_0, \gamma, \beta, \tau|\delta) d\alpha d\beta_0 d\gamma d\beta d\tau, \]

for \( \beta = (\beta_1, \beta_2, \ldots, \beta_k) \). The distribution of the posterior probability is not analytically tractable so we used Gibbs sampler approach to obtain samples from the posterior distributions. The posterior probabilities of the model with all the variables (joinpoints) enclosed is given by

\[ p(M_k|y) = \sum_{i=0}^{k} p\left( \sum_{i=0}^{k} \delta = i | y \right), \]

which is used to compare the different models and the one with the highest posterior probability will be chosen as the best model [48].
Since the model can choose an infinite number of breakpoints, we wish to impose some restrictions on the position of the change points in the model. There are different ways of implementing these restrictions (for example [48],[24]). To avoid such identifiability problems, a restriction is imposed in such a way that the model will only select the joinpoints more than two years apart leaving the first two and last two years in the time trend.

Bayesian paradigm discusses assigning prior distributions to all unknown parameters in the competing models and those priors get transformed to the posterior through the data. The posterior distribution is highly influential and sensitive to the choice of priors and the problem deepens if the models have both common and non-common parameters [3]. Furthermore, the choice of improper or vague priors would lead to arbitrary Bayes factor and make the result computationally challenging (see [48],[3]). Also, uncertainty issues with respect to the model and its parameters ([14]) is complicated when the nested models have common parameters that appear in all models and non-common parameters that are model specific [3]. Here, our concern is on the uncertainty related to the covariate vectors $B_{\tau_k}(t_i)$ coming from reparametrization. In the model selection process, these covariate vectors are considered as non-common variables in the models and $\delta$ indicates the existence of these variables.

The introduction of prior distribution into the model has drawn much interest recently and different criteria have been proposed by many researchers so far. In an objective Bayes solution to the model selection problem, the nature of the posterior distributions depends upon the selection of priors and is very sensitive if there are non-common parameters in the models as explained in [1, 3]. The specifications of prior distributions based on two types of parameters associated in the model are common and non-common parameters. The common parameters that parametrize the average linear predictors in all competing models for each age group model are $\alpha_j$, $\beta_0$, and $\gamma$ for which improper flat priors are assigned [1, 3].
Because of the uncertainty related to the reparametrization of joinpoints and the possible existence of an infinite number of breakpoints in the model, the manual eduction of all these priors is not possible and the priors which automatically derive from \( \delta \) that governs the breakpoints are attractive ([7]). Also, the model with the assumption of discrete prior on the location of change points provides poor convergence compared to continuous prior [24, 66]. In this context, we assign generalized hyper-\( g \) prior, an extension of the classical \( g \)-prior to generalized linear model proposed by Bové and Held ([7]) for non-common parameters \( \beta' \)s for which the proposed distribution is given by

\[
\beta_{\delta}\mid g, \delta, \alpha, \beta_0, \tau, \beta \sim N_{P_\delta}(0_{P_\delta}, g\phi c\Sigma)
\]

where \( g\phi \) is the scale dispersion with \( \phi = 1 \), being one parameter exponential family, and \( c \) has been proved to be equal to 1 for Poisson distribution with log link function [7]. Our decision in applying the generalized hyper-\( g \) prior for beta’s has an important advantage as the hyper prior on the hyper parameter \( g \) can be handled in such a way that any continuous proper hyper prior can be used giving rise to a large class of hyper-\( g \) priors [7]. The chosen prior has an important extension as it further allows us to implement a large class of hyper priors. In our study we use

\[
f(g) = IG(g\mid 1/2, n/2),
\]

corresponding to the Zellner and Siow approach [78].

It can be shown that the mode of this distribution is at \( \beta_{\delta} = 0_{P_\delta} \) (see [48], [7]) and in our model, the Fisher information matrix at \( \beta_{\delta} = 0_{P_\delta} \) is

\[
I = \Delta B^TWB\Delta
\]
where \( \Delta = \text{diag}(\delta) \), \( B = \{ B_{x_k} \}, W = \text{diag}(w_i) \) with

\[
w_i = \sum_{j=1}^{J} P_{ij} \exp(\alpha_j + \beta(t_i - \bar{t}) + \gamma z(t)).
\]

Since \( I \) is not a positive definite matrix for every choice of \( \delta \), we side step this problem by adding some quantity in diagonal element of the matrix as is done in [48], that is,

\[
\Sigma = n(\Delta B^T W B \Delta + \text{diag}(B^T W B - \Delta B^T W B \Delta))^{-1}.
\]

Similar to their argument in [48] for particular \( \delta^* \) with \( \sum \delta^*_i = K \), sub vectors of \( \beta \) corresponding to non-null \( \delta'_i \)s and null \( \delta'_i \)s are independent, and those null \( \delta'_i \)s behave as pseudo prior. This makes the estimation procedure an easy problem by assigning a single prior for \( \beta \). The prior for \( \tau \) is straight forward. As the parameter space is bounded, we can have \( \pi(\tau) \propto 1 \). Based on the nature of \( \delta \), it is reasonable to choose an independent Bernoulli prior with probability of success \( p \). Hyper priors of \( p \) are chosen as \( \text{Beta}(\frac{1}{2}, \frac{K-1}{2}) \) where \( K \) is the number of join points [48].

At every step of MCMC, we obtain a different estimation of the temporal trend based on a different number of joinpoints. The temporal trends are traced by averaging all the joinpoint curves at every step for each gender. As we know the analytical expression of the curve, we extend that curve beyond 2009 to obtain the 5-year prediction. The main advantage of this trend is that it does not depend on the unique value of the number of joinpoints. It averages the curves for different values of joinpoints as a function of the probability given by the value of delta.

### 2.4 Conclusion

We developed an Age-Stratified Bayesian Joinpoint Regression Model that has several theoretical and applied advantages over the existing Bayesian Joinpoint Regression Model.
The developed model can be applied to obtain better estimates of the mortality (or incidence) rates and public health personnals, government officials, and policy makers can use this to get the real status of the disease in the population. The model can also be used to compare the trends in the different subpopulations. Several advantages of the developed model are discussed below:

1. We proposed an Age-Stratified Bayesian Joinpoint Regression Model that can be used to study the age specific mortality (or incidence) rates which is suitable to incorporate the confounding effect of age in the population.

2. Our proposed parallel slope model reduces the computational burden which is equivalent to fitting separate models for each group under Poisson Model assumption. In the mean time the model can capture the trends for each age group by incorporating the interaction terms in the model.

3. Reliable and accurate age- adjusted rates and its Annual Percentage Change (APC) will be obtained to study and compare the mortality (or incidence) rates in the different population subgroups.

4. Since the developed model can have infinite number of joinpoints in the model and there is uncertainty related to the parametrization approach followed by Beneito et al., our choice of prior for beta (associated with joinpoints) is based on theoretical justification that helps to reduce the uncertainty related to the detection of joinpoints. Moreover, the chosen priors have an important property as they allow us to choose a large class of hyper g-priors.
Chapter 3
Application of Bayesian Joinpoint Regression Model on Childhood Brain Cancer
Mortality and its Comparison with NCI Approach

The social and economic burden due to cancer is rapidly growing in the United States and around the world. The study and the evaluation of the mortality trends due to cancer is an important factor in the current economic growth of any country and in measuring the potential future economic effect. Brain cancer (brain tumor and other central nervous system (CNS) cancers) is one of the leading cancers, ranking the second largest cause of childhood death due to cancers. Based on 1975-2007 incidence data reported by Kohler, et al. (2011), 65.2 percent of the children with brain tumors are diagnosed with malignant tumors whereas the percentage in adults is only 33.7 [41]. According to the National Cancer Institute (NCI), leukemias and the cancers of the brain and nervous system in children account for more than half of the new cases. Brain tumors are the most common solid tumors and are the second most common type of pediatric cancer. The central brain tumor registry of the United States reports that approximately 4300 children younger than age 20 are expected to be diagnosed with primary malignant and non-malignant brain cancer in 2013. According to Kleihues, et al. (1993), the histological appearances of childhood brain tumors differ significantly from that of adults and are classified into several large groups [40]. The overall distribution of these tumors also differ significantly [59–61]. Ullrich and Pomeroy (2003) reported in their paper that the Pilocytic astrocytoma is the main histologic types in children CNS tumors with relatively high frequency of occurrence [75]. According to Ries et al. (2007), the overall incidence for childhood brain cancer rose from 1975 to 2004 with the greatest increase occurring from 1983 through 1986 [64]. But, it is found
that the mortality rates are continuously decreasing, with relatively higher rate from 1969 to 1980 and slower rate from 1980 onwards. None of these works provided the better estimate of the rate of change of mortality in an early basis. All these previous works motivate us to study the mortality trend in childhood brain cancer using a statistical model that is based on realistic assumptions.

The main objective of this chapter is to study the crude (non age-adjusted) childhood brain cancer mortality trend using joinpoint model described in chapter 2. The main objective of this study is to give the reliable estimates of the measure of cancer mortality trend that provide up-to-date information and recent changes in childhood brain cancer which is also exhibited in [34]. Studied here is the mortality trend of childhood brain cancer data obtained from SEER database of NCI [69]. The model is fitted using softwares WinBUGS and R [46, 62]. We also fitted the trend line using the joinpoint software of NCI and compare the trend lines. We observe several advantages of Bayesian approach to the NCI approach. Here we divided this chapter into four sections: data description, statistical analysis, model validation, and contributions.

Brain tumor and other CNS cancer mortality data for children are considered for this study. We obtained the total annual observed mortality counts of children below 20 years of age from 1969-2009. The data set are extracted from the SEER data base of NCI using SEER*Stat software [70]. Being rare events, we assume the mortality counts are probabilistically characterized by the Poisson probability distribution and model them using log link function. We apply the Bayesian joinpoint regression model discussed in section 2.2 to obtain the mortality trend assuming that the break points are continuous over time. The joinpoint regression model using the joinpoint software of NCI is also fitted for the same data and compared these two results to see the theoretical difference in model fitting between. We observe that the model using Bayesian approach describes the data very well giving best possible short term predictions and performs a better improvement over the existing methods.
In this chapter, we study the childhood brain cancer mortality to address the following questions.

1. What is the annual estimated mortality rates for childhood brain cancer mortality using Bayesian joinpoint regression?

2. What is the future mortality rates for the childhood mortality in the population?

3. What is APC at each year for childhood brain cancer mortality trends and what is the difference in APC using NCI approach?

4. What are the advantages of Bayesian approach over the NCI method to study the mortality (or incidence) of trends in population?

3.1 Statistical Analysis

The model is described by four unknown joinpoints \((k = 4)\) to identify the years where a change over time in the slope of child brain cancer trend occurs. Since the posterior distributions are not analytically tractable and the high dimensionality of the integrals makes the model selection procedure even more complex, the Gibbs variable selection approach as discussed in Chapter 1 is used to select the best model with significantly minimum number of joinpoints that describes the trend. The process is carried out in such a way that if we add even one more joinpoint in the model, the model becomes insignificant.

We implemented two parallel chains in WinBUGS using different initial values. Each chain was run for 150,000 iterations giving 50,000 iterations as burn-in period. The posterior inferences is based on 100,000 iterations for each chain combining total of 200,000 iterations for each of the parameters. The posterior summaries for the parameters are given in Table 1. Out of competing five nested models, the model selection procedure using Bayes Factor selected the model with one joinpoint as given in Figure 1. For the selected model with one joinpoint, the posterior distribution of each of the parameters was observed.
by monitoring the trace, iterations, Monte Carlo errors, standard deviations, and density curves. The trace for each of the parameters satisfy the convergence criteria. Also, the Monte Carlo errors are within 0.1% of the posterior standard deviations.

As depicted in the graph given in Figure 1, the probability of the posterior distribution for one joinpoint is about 80%. The probability of the posterior distribution for no joinpoint is very low indicating that the linear trend is not a choice. Similarly, the probability of posterior distribution does not support two, three, and four joinpoints as well. This means that the childhood brain cancer mortality trends is best characterized by one joinpoint 80 percent of the time. The probability of the existence of the other competing models (other number of joinpoints) are significantly low compare to a model with one joinpoint supporting that model with one joinpoint is the best model.

The boxplot for the parameters $\beta_j, j = 1, 2, 3, 4$ associated with change points is plotted in Figure 2 are produced using the WinBUGS software. These plots are different compared to the boxplots obtained by using the frequentist approach. The middle bar of each box represents the posterior means and the two limits are the posterior quartiles. The two ends of the whiskers are represented by 2.5% and 97.5% posteriors percentiles. These percentiles give the Bayesian confidence interval called the credible interval. These intervals give the
interval estimation of the posterior probability of the parameters. Posterior means and 95% credible intervals of $\beta_j$’s suggest that their posterior distributions are not discriminable. This indicates that no more than one joinpoint is required and if more joinpoints are added, the model is not statistically significant.

We applied four joinpoints in the proposed model given in chapter 2. The analytical structure of the proposed model for the subject data is given by

$$\log(r_i) = \alpha + \beta_0(t_i - \bar{t}) + \beta_1(t_i - \tau_1) + \beta_2(t_i - \tau_2) + \beta_3(t_i - \tau_3) + \beta_4(t_i - \tau_4)$$

where $\bar{t}$ is the mean of $t_i$, and $\tau_j, j = 1, 2, 3, 4$ is the change point in the model $\alpha$ is the intercept parameter, and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are the regression parameters for the joinpoints.

The estimated rates for each year from 1969-2009 are obtained by averaging the estimates of joinpoint and other parameters in the model at every step of MCMC by using the WinBUGS software. The table below (Table 1) gives the estimates of the parameters in the model.

On applying the parameter estimates in the model, the estimates of the rate at any time $t_i$ is given by

$$\log(\hat{r}_i) = -11.76 - 0.01176 \ast (t_i - \bar{t}) - 0.0176 \ast (t_i - 8.366) - 0.01679 \ast (t_i - 15.13)$$

$$- 0.00151 \ast (t_i - 23.33) - 7.90E - 04 \ast (t_i - 31.98)),$$
Table 1: Parameter Estimates

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.50%</th>
<th>median</th>
<th>97.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>-11.76</td>
<td>0.006448</td>
<td>3.35E-05</td>
<td>-11.77</td>
<td>-11.76</td>
<td>-11.75</td>
</tr>
<tr>
<td>beta0</td>
<td>-0.01176</td>
<td>5.33E-04</td>
<td>2.79E-06</td>
<td>-0.01281</td>
<td>-0.01176</td>
<td>-0.01071</td>
</tr>
<tr>
<td>beta[1]</td>
<td>-0.0176</td>
<td>0.05287</td>
<td>7.68E-04</td>
<td>-0.09726</td>
<td>-0.02668</td>
<td>0.09301</td>
</tr>
<tr>
<td>beta[2]</td>
<td>-0.01679</td>
<td>0.09534</td>
<td>0.001723</td>
<td>-0.1736</td>
<td>-0.02925</td>
<td>0.1602</td>
</tr>
<tr>
<td>beta[3]</td>
<td>-0.00151</td>
<td>0.1265</td>
<td>0.001355</td>
<td>-0.218</td>
<td>-0.00167</td>
<td>0.2119</td>
</tr>
<tr>
<td>beta[4]</td>
<td>-7.90E-04</td>
<td>0.1114</td>
<td>0.001049</td>
<td>-0.1963</td>
<td>-1.52E-04</td>
<td>0.1938</td>
</tr>
<tr>
<td>delta[1]</td>
<td>0.5254</td>
<td>0.4994</td>
<td>0.01384</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>delta[2]</td>
<td>0.4684</td>
<td>0.499</td>
<td>0.01359</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>delta[3]</td>
<td>0.1156</td>
<td>0.3197</td>
<td>0.005156</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>delta[4]</td>
<td>0.05771</td>
<td>0.2332</td>
<td>0.001234</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tau[1]</td>
<td>8.366</td>
<td>2.891</td>
<td>0.1512</td>
<td>3.299</td>
<td>8.62</td>
<td>13.73</td>
</tr>
<tr>
<td>tau[2]</td>
<td>15.13</td>
<td>5.264</td>
<td>0.2663</td>
<td>7.273</td>
<td>14.13</td>
<td>27.41</td>
</tr>
<tr>
<td>tau[3]</td>
<td>23.33</td>
<td>6.327</td>
<td>0.2753</td>
<td>11.51</td>
<td>23.42</td>
<td>34.56</td>
</tr>
<tr>
<td>tau[4]</td>
<td>31.98</td>
<td>5.634</td>
<td>0.2222</td>
<td>18.19</td>
<td>33.36</td>
<td>38.78</td>
</tr>
</tbody>
</table>

and, the final model to estimate the rate curve is given by

\[ \hat{r}_i = \exp(-11.05127 - 0.04845 \times t_i) \]

The rates from 2010-2012 are predicted by applying the Bayesian Model Averaging approach as discussed in chapter 1. As we know that the Bayesian Model Averaging averages all sets of competing models and make an inference based on a weighted average on these models over the model space. We used 4 joinpoints to explain the model, and we used an encompassing model approach as given below (described in chapter 2) to choose the best competing model. In the mean time, this is the weighted average on these models over the model space. As we knew the analytical structure of the model, we extended that to get the future predictions. We recall from chapter 2 that the posterior probabilities of the model with all the variables (joinpoints) enclosed is given by

\[ p(M_k|y) = \sum_{i=0}^{k} p \left( \sum_{\delta = i}^{y|y} \right), \]
where $M_k$ is the set of competing models, $\delta \in \{0, 1\}^k$ is binary inclusion indicators for all $\tau'_k$s in the model known as latent vector, and $p(\delta|y)$ is the posterior distribution of $\delta$. Here, the posterior probability of sum of $\delta = i$ for all four joinpoints is used to estimate the model and it is extended to provide the future prediction making the encompassing model and Bayesian Model Averaging approach same. We used last year’s population information to predict the future rates. The future rates are predicted for short period of times as we are using log linear model which are not suitable for long term prediction and the population information is usually unavailable for the future.

The graph for the estimated rate and its prediction is given in Figure 3. The solid curve represents the estimated trend line for annual mortality rate whereas the dashed lines represent its 95% credible interval. The observed death rates are represented by unfilled circles. The extended graph beyond dashed vertical line represents the prediction of rate from 2009 to 2012.

![Child Brain and Other CNS Cancer Mortality Trend](image)

Figure 3.: Estimated time trend for the annual observed mortality rate per 100,000 children

49
The graph shows that the childhood cancer mortality rates declined faster from 1969 to 1978 compared to the rest of the time interval in a decreasing fashion. The overall mortality rate decreased from 1.056 to 0.63 per 100,000 by 2009 and is predicted to decrease continuously.

For the same data, the joinpoint regression model is fitted using the joinpoint software of NCI [54]. We assume the obtained mortality data are heteroscedastic in nature and use weighted least squares method. Since we assume the heteroscedasticity, we have to input the variance of each rate, for which we assume the Poisson variance with an autocorrelated errors based on the data. As only one independent variable is allowed, we have used calendar year as that variable. Grid search method is used to select the joinpoint model with grid size of 2 years leaving two years at the two ends of the data values to exactly match our condition we imposed for identifiability problem. The model selection method is performed using permutation test for four joinpoints with altogether five competing models. The overall significance level for the permutation test is considered 0.05. The number of permuted data sets for the permutation test is set as a default number of 4499. Usually, the large number of permutations give the more consistent p-values. We used the Bonferroni correction to adjust the significance level doing the multiple model comparisons. The joinpoint software also has Bayesian Information Criterion (BIC) approach as an alternative method to Peremuation Test Based (PTB) method to fit the best model. Many studies claims that PTB approaches performs better compare to BIC, we also choose PTB method to select the best model. The output is as shown in Figure 5. The solid line is the fit from the joinpoint software from NCI with a gap of minimum of two observations between two joinpoints.

As we know from section 2, the analytical expression for this method is given by

\[ y_i = \beta_0 + \beta_1 \ast t_i + \sum_{k=1}^{K} \delta_k \ast s_k(t_i) + \epsilon_i, \]

where \( y_i, i = 1, 2, 3, \ldots, n \) denote the observed mortality rates, \( k \) be the change points.
in the data, \( s_k(t_i) = (t - \tau_k)^+ \), and \( a^+ = a \) if \( a > 0 \), and \( a^+ = 0 \), otherwise, \( \beta_k^t = (\beta_0, \beta_1, \delta_1, \ldots, \delta_k) \) are the regression parameters, and \( \tau_k^t = (\tau_1, \tau_2, \ldots, \tau_k) \) are the join-points, and \( \epsilon_i's \) are random errors with mean \( = 0 \).

We observed one joinpoint by using the joinpoint software of NCI. The joinpoint exists at 1978 as shown in Figure 4. The model will be represented by two linear regression lines before 1978 and after 1978 with respective slopes -0.02 and -0.01.

![Figure 4. Mortality rates of child brain cancer obtained by using the NCI approach.](image)

The graph shows that the slopes of the rate curve before and after joinpoint are constant. It is not the case for the Bayesian joinpoint model as it gives the slope of the rate curve at any point. Also, the location of change point is discrete and occurs exactly at the whole number year in case of the regression trend given by joinpoint software, whereas the location of the change point is continuous in our case and can occur in between the years. The third difference is that the trend obtained from joinpoint software is descriptive but the regression trend we obtained can give insights for the mortality trend in the future with credible bands.
The estimated annual percentage change (APC) is used to characterize the trends or the change in rates over time. Estimated APC from \(i^{th}\) year to \((i + 1)^{th}\) year is calculated by

\[
\hat{APC} = \frac{\hat{E}(r_{i+1}) - \hat{E}(r_i)}{\hat{E}(r_i)} \times 100.
\]

where \(\hat{E}(r_i)\) and \(\hat{E}(r_{i+1})\) are the estimated rates at \(i^{th}\) and \((i + 1)^{th}\) year.

Figure 5.: Estimated Annual Percentage Change in child brain cancer rates over time per 100,000 children

The graph in Figure 5 gives the average rate of change in mortality rate per year from 1969 to 2009 and its estimates to 2011. APC was exactly -2.31 for the first three years and increased from -2.29 in 1973 to -1.12 in 1980. After 1980, APC looks almost constant with a fluctuation of 0.01 to 0.02 over the entire range. It means that the average rate of change per year in the childhood brain cancer mortality rate has not changed in recent years and is predicted to remain almost the same in the consequent years.
3.2 Model Validation

It is very important to evaluate how well the model fits the data in addition to its inference. To check the validity, goodness of fit, and assumptions of the proposed model, we perform different model validation techniques discussed in the literature.

The residual analysis is performed to check the robustness and fit of our developed model. We use the posterior simulation to examine the standard errors with their 95% credible intervals for checking the fit of each observation and the identification of outliers. Standardized residuals are obtained by taking the deviations of the data to their expectations for all measurements based on posterior simulations and dividing it by their standard deviations. Error bars with 95% credible bands are given in Figure 6. Most of the standardized residuals with their bands are randomly distributed within the range of -2 to 2. Also, the mean and standard deviation of the standardized residual are 0.000527 and 0.927 respectively. This indicates that the developed model fits the observed data very well.

We also validate the trend by fitting the model from 1969-2005 and tested the trend from
2006-2009. To validate the trend, we applied four joinpoints in the proposed model as explained above. The estimated mortality rate curve is produced to obtain the trend from 2006 to 2009 using the Bayesian Model Averaging (BMA) approach as discussed earlier in this chapter and in chapter 1. As shown in Figure 7 below, we observe that the observed mortality counts of childhood brain cancer from 2006-2009 falls within 95% credible interval of the projected mortality trend line.

![Brain Cancer Mortality Trend in Children](image)

**Figure 7.** Trend Validation

The goodness of fit for the obtained model is evaluated using Chi-square statistics. The posterior predictive distributions are used for checking the model assumptions and the goodness of fit in the model. We generally obtained the replicated values from the posterior predictive distributions that are evaluated at estimated parameter values. The replicated data values are the expected observations after replicating our experiment in the future considering the estimated model is true. If the model is true, then the observed data and replicated data should be very close. The comparison of actual and predicted values
gives the information regarding the model fit and the indication of possible outliers. The posterior predictive p-values are obtained by using the posterior distribution as follows

\[
Posterior\text{P\text{-}value} = P(D(y^{rep}, \theta) > D(y, \theta)|y)
\]

where \(D(y, \theta)\) is the deviance summary function that plays the role of a test statistic. The chi-square difference \(\chi^2(y^{rep}, \theta) - \chi^2(y, \theta)\) is also monitored. The difference of these two statistics is given in Figure 8. Also, their corresponding posterior \(p\)-value is obtained. The \(p\)-value based on the difference of Chi-squares obtained as a posterior mean using WinBUGS is 0.5513. The large \(p\)-value shows that the observed statistics is close from what is expected under the assumed model.

![Figure 8.: Difference in Chi-square statistics of observed and predicted mortality counts](image)

We calculate the Chi-square statistics for the observed mortality data and for the predicted data as well in each iteration of MCMC algorithm. The graph given in Figure 8 also proves that there is no significant difference between observed and expected frequencies supporting our Poisson model assumption.

Also, the distribution of future or replicated data is regenerated by using the predictive distribution and is compared with the observed data to satisfy the model assumptions. The
posterior predictive plots for frequencies with 95% credible intervals of replicated data are plotted with vertical segments for each year and compared with the observed data. From the graph in Figure 9, we find that the observed mortality counts not only fall inside the 95% posterior intervals of replicated data but also close to their mean values indicating that the assumption of Poisson distribution is valid.

![Posterior Predictive Plots of Replicated and Actual Death](image)

Figure 9.: Comparison of actual and predictive frequencies

### 3.3 Conclusion

In this study, we apply the Bayesian joinpoint regression model to uncover the patterns of childhood brain cancer mortality that provides important information pertaining to further study in the cases and control of the disease. Although different studies have shown that the childhood cancer mortality rates continue to decline dramatically by more than 50% in the past two decades ([64],[41]) in the United States, only few studies have considered the probability distribution of the observed counts as Poisson and the location of the change points continuous in time. The application discussed here is based on these probabilistic
assumptions. We obtained the trend that describes the behavior of the observed data very well and gives us the best possible short term predictions. The obtained temporal trend provides the different slopes of the rate curve at each point of time. In contrast, the joinpoint software of NCI gives the same slope at each year between two change points. Also, we are able to obtain the more accurate annual percentage change (APC) and we observed that the APC is almost constant from 1981 and is predicted to remain constant. SEER routinely collects the data covering 28% of the US population and there is a three year lag in time to collect and process the data. In this scenario, predictions in the temporal trend and APC are very helpful to evaluate the effectiveness of the current status of the disease. This improvement over the existing methods allows us to observe the real progress we are making in childhood brain cancer.

3.3.1 Contributions and future needs

In this chapter, we study childhood brain cancer mortality using the Bayesian approach as developed by Martinez-Beneito et al. (2011) and compare the result with the trends obtained by the joinpoint software of NCI. We observe several advantages of using the Bayesian approach over the NCI approach as discussed above. Here we would like to summarize a couple of points that we observed from this chapter.

1. Applied Bayesian joinpoint regression model is based on correct model assumptions to estimate and predict the mortality of childhood brain cancer.

2. We compared estimated mortality obtained by using the Bayesian approach of joinpoint regression with NCI approach and observed several advantages using Bayesian approach compared to NCI approach.

3. Bayesian model provides the slope of the rate curve at any point but the NCI approach has only two slopes: before and after the joinpoint.

4. Location of change points exists only in observed data points and is discrete with NCI
approach but it is continuous in Bayesian approach.

5. The trend obtained from joinpoint software is descriptive but the regression trend we obtained can give insights for the mortality trend in the future with credible bands.

The model applied in this section to study the trends has made several advantages over the NCI method. However, as we discussed in chapter 2, there are some limitations of using the Bayesian approach by Martinez-Beneito et al. (2011) to study the mortality (or incidence) rates in the population and different population subgroups to compare the trends by founding the model that should adjust the confounding effect of age. Moreover, there is need to extend this work to study the influence in the mean of the outcome by incorporating applicable covariates in the model, but the addition of covariates increases the complexity of the model increasing the computational time. Also, the Bayes Factors are sensitive to the prior specifications, and therefore further study is needed in selecting the objective priors by exploring different objective model selection criteria for priors that can deal with model uncertainty. Moreover, age standardized rates in this methodology was a further extension as discussed in the theoretical chapter. Also, we proposed an Age-stratified Bayesian joinpoint regression model that can overcome these issues. The following chapter is the application of our proposed Bayesian Age-stratified joinpoint regression model that overcomes the existence deficiencies in modeling using Bayesian approach.
Chapter 4

Application of Age-Stratified Bayesian Joinpoint Regression Model to Lung and Brain Cancer Mortality Data

In the previous chapters, we have discussed the necessity of developing a model that can estimate and predict the trend data well and proposed an age-stratified Bayesian joinpoint regression model. We also discussed several advantages of our model over the existing models. In this chapter, we apply our proposed model on the annually observed adult mortality counts of two cancer data drawn from the Surveillance, Epidemiology, and End Results (SEER) database of the National Cancer Institute (NCI) [69]. We study the annual age-stratified (5 years age group) mortality rates of lung and bronchus, and brain and other CNS cancers patients and we further apply these results to study the age-adjusted rates [35]. The data sets are obtained by using SEER*STAT software from 1969 to 2009 [70]. The age adjustment requires the weights for different age groups to adjust for the confounding effect of age and we use the weights of the US standard population of the year 2000 for age adjustment. Our study considers the yearly mortality counts of male and female for five years age group from 25 to 85+ years for lung and bronchus cancers, and 20 to 85+ years for Brain tumor and other CNS cancers. The justifications for considering the adult age group are: SEER*Stat does not give the counts for number of observations less than 10 and we have many of such cases specifically below the age of 25 for mortality counts. Although the proposed model can deal with zero counts, there are missing values or unknown values. Moreover, cigarette smoking is the most common cause of lung cancer [15] and this habit develops in adult ages. In brain tumor, the histological appearance and the distributions of the adult brain tumors are significantly different compared to children’s tumors and the
mode of the treatment and survival also differ significantly [41, 61].

The data is analyzed using the freely available softwares WinBUGS and R [46, 62]. We fitted the model for each age group first based on parallel model assumption. However, we estimated one more parameter for the interaction of time and gender. As we discussed in chapter 2, this is done to capture to trends for different groups and save the computational time instead of estimating the each age groups trend separately. Since the mortality of cancer does not depict too many changes from year to year, the model is fitted with maximum number of joinpoints equal to four for both cancer data. On fitting the model, we adjusted the interaction terms between gender and time to capture the trend across genders. For each cancer data, the model is run 150K iterations giving 50K iterations as burn in periods for a wide range of initial values for different parameters. The posterior inferences for the parameters are based on 100K iterations. For each of the selected models, the posterior distribution of parameters is observed by monitoring the trace, iterations, Monte Carlo errors, standard deviations, and the density curves. The trace of each of the parameters satisfy the convergence criteria. Also, the Monte Carlo errors are within 3% of the posterior standard deviations.

4.1 Lung and Bronchus Cancer Mortality Trends

Lung and bronchus cancer accounts for more deaths than any other cancers in the United States [65]. It causes even more deaths than the combined deaths due to colon, breast, and prostate cancers, which are the next three highly ranked cancer deaths after lung cancer in the United States. Incidence and mortality both due to lung cancer is three times higher in males than in females in the world [57]. According to NCI, the estimated new lung and bronchus cancer in 2014 are 116,000 for males and 108,210 for females, and the estimated death due to lung and bronchus cancers are 86,930 for males and 72,330 for females.

We fitted the model using four joinpoints. It is observed that the model selection process has selected a model with all four joinpoints based on its posterior probabilistic framework
and is found to occur almost 100% of the time (See Figure 10). This means that all four joinpoints have expressed the Annual Percentage Change in the trend of mortality rates. Also, the boxplot for the parameters $\beta_j, j = 1, 2, 3, 4$ associated with the change points and their posterior means and credible intervals were investigated and they suggest that their posterior distributions are discriminable.

![Figure 10: Posterior probability of the number of joinpoints for lung and bronchus cancer mortality trend](image)

The Deviance Information Criteria (DIC) is also used as a measure of model comparison and adequacy. DIC criterion and its application in our joinpoint model selection approach has been described in chapter 1. DIC values for all five competing models are given in Table 2 below. In the table below, $D_{\text{bar}}$ represents the posterior mean of deviance evaluated by an MCMC sample, $D_{\text{hat}}$ is a point estimate of the deviance obtained by substituting in the posterior means and theta, and $pD$ is the effective number of parameters given by the difference of posterior mean of the deviance and the point estimate of the deviance. Then, the Deviance Information Criteria (DIC) is given by

$$DIC = D_{\text{bar}} + P_D = D_{\text{hat}} + 2P_D$$

As the lowest value of DIC indicates the better fit, it also facilitates the requirement of
Table 2: DIC values for all five competing models for lung and bronchus cancer mortality

<table>
<thead>
<tr>
<th>Number of joinpoints</th>
<th>Dbar</th>
<th>Dhat</th>
<th>pD</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>No joinpoints</td>
<td>245960</td>
<td>245788</td>
<td>172.796</td>
<td>246133</td>
</tr>
<tr>
<td>One joinpoints</td>
<td>206236</td>
<td>206074</td>
<td>162.565</td>
<td>206399</td>
</tr>
<tr>
<td>Two joinpoints</td>
<td>180849</td>
<td>180689</td>
<td>159.784</td>
<td>181009</td>
</tr>
<tr>
<td>Three joinpoints</td>
<td>178993</td>
<td>178833</td>
<td>159.004</td>
<td>197152</td>
</tr>
<tr>
<td>Four joinpoints</td>
<td>178217</td>
<td>178058</td>
<td>159.086</td>
<td><strong>178376</strong></td>
</tr>
<tr>
<td>Total</td>
<td>990255</td>
<td>989442</td>
<td>813.235</td>
<td>991069</td>
</tr>
</tbody>
</table>

We applied four joinpoints in the proposed age-stratified joinpoints regression model given in chapter 2. The analytical structure of the proposed model for the subject data are given by

\[ \ln(r_{ij}) = \alpha_j + \beta_0 (t_i - \bar{t}) + \gamma z(t_i) + \gamma_1 (t_i - \bar{t}) + \sum_{k=1}^{K} \delta_k \beta_k B_{r_k}(t_i), \]

where \( \bar{t} \) is the mean of \( t_i \), and \( \tau_j, j = 1, 2, 3, 4 \) is the change point in the model \( \alpha \) is the intercept parameter, \( \delta \) is the indicator variable, and \( \beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \) are the regression parameters for the joinpoints.

The estimated rates for each year from 1969-2009 are obtained by averaging the estimates of joinpoint and other parameters in the model at every step of MCMC by using the WinBUGS software. The posterior summaries for parameters including the estimates of change points (tau) are given in Table 3. The table shows that the change points occur at \( t = 11.91, 21.81, 26.69, \) and 36.27 respectively.

The graphs shown in Figures 11 and 12 are the estimated crude mortality fits for each
Table 3: The posterior summaries of parameters for lung and bronchus cancer

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.50%</th>
<th>median</th>
<th>97.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>α1</td>
<td>-13.55</td>
<td>0.0226</td>
<td>1.63E-04</td>
<td>-13.6</td>
<td>-13.55</td>
<td>-13.51</td>
</tr>
<tr>
<td>α2</td>
<td>-12.23</td>
<td>0.01183</td>
<td>8.58E-05</td>
<td>-12.25</td>
<td>-12.23</td>
<td>-12.2</td>
</tr>
<tr>
<td>α3</td>
<td>-10.91</td>
<td>0.006255</td>
<td>4.49E-05</td>
<td>-10.92</td>
<td>-10.91</td>
<td>-10.9</td>
</tr>
<tr>
<td>α4</td>
<td>-9.784</td>
<td>0.003771</td>
<td>3.08E-05</td>
<td>-9.792</td>
<td>-9.784</td>
<td>-9.777</td>
</tr>
<tr>
<td>α5</td>
<td>-8.865</td>
<td>0.002554</td>
<td>2.46E-05</td>
<td>-8.87</td>
<td>-8.865</td>
<td>-8.86</td>
</tr>
<tr>
<td>α6</td>
<td>-8.135</td>
<td>0.001943</td>
<td>2.02E-05</td>
<td>-8.139</td>
<td>-8.135</td>
<td>-8.131</td>
</tr>
<tr>
<td>α7</td>
<td>-7.534</td>
<td>0.001623</td>
<td>1.99E-05</td>
<td>-7.537</td>
<td>-7.534</td>
<td>-7.53</td>
</tr>
<tr>
<td>α8</td>
<td>-7.044</td>
<td>0.001415</td>
<td>1.91E-05</td>
<td>-7.047</td>
<td>-7.044</td>
<td>-7.041</td>
</tr>
<tr>
<td>α9</td>
<td>-6.687</td>
<td>0.001317</td>
<td>1.88E-05</td>
<td>-6.689</td>
<td>-6.687</td>
<td>-6.684</td>
</tr>
<tr>
<td>α10</td>
<td>-6.446</td>
<td>0.001289</td>
<td>1.73E-05</td>
<td>-6.448</td>
<td>-6.446</td>
<td>-6.443</td>
</tr>
<tr>
<td>α11</td>
<td>-6.324</td>
<td>0.001364</td>
<td>1.74E-05</td>
<td>-6.326</td>
<td>-6.324</td>
<td>-6.321</td>
</tr>
<tr>
<td>α12</td>
<td>-6.302</td>
<td>0.00157</td>
<td>1.79E-05</td>
<td>-6.305</td>
<td>-6.302</td>
<td>-6.299</td>
</tr>
<tr>
<td>α13</td>
<td>-6.46</td>
<td>0.001888</td>
<td>2.00E-05</td>
<td>-6.464</td>
<td>-6.464</td>
<td>-6.456</td>
</tr>
<tr>
<td>β0</td>
<td>0.02854</td>
<td>7.37E-05</td>
<td>1.35E-06</td>
<td>0.02839</td>
<td>0.02854</td>
<td>0.02868</td>
</tr>
<tr>
<td>γ</td>
<td>0.9866</td>
<td>0.001032</td>
<td>1.91E-05</td>
<td>0.9846</td>
<td>0.9866</td>
<td>0.9886</td>
</tr>
<tr>
<td>γ * β0</td>
<td>-0.03255</td>
<td>8.74E-05</td>
<td>1.57E-06</td>
<td>-0.03273</td>
<td>-0.03255</td>
<td>-0.03238</td>
</tr>
<tr>
<td>β1</td>
<td>0.0474</td>
<td>0.002213</td>
<td>5.27E-05</td>
<td>0.04331</td>
<td>0.04728</td>
<td>0.05198</td>
</tr>
<tr>
<td>β2</td>
<td>0.1049</td>
<td>0.00738</td>
<td>3.71E-04</td>
<td>0.08578</td>
<td>0.1057</td>
<td>0.117</td>
</tr>
<tr>
<td>β3</td>
<td>0.04759</td>
<td>0.007184</td>
<td>3.61E-04</td>
<td>0.03454</td>
<td>0.04717</td>
<td>0.06508</td>
</tr>
<tr>
<td>β4</td>
<td>0.0132</td>
<td>0.001501</td>
<td>4.48E-05</td>
<td>0.01076</td>
<td>0.01295</td>
<td>0.01667</td>
</tr>
<tr>
<td>δ1</td>
<td>1</td>
<td>0</td>
<td>3.16E-13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>δ2</td>
<td>1</td>
<td>0</td>
<td>3.16E-13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>δ3</td>
<td>1</td>
<td>0</td>
<td>3.16E-13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>δ4</td>
<td>1</td>
<td>0</td>
<td>3.16E-13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>τ1</td>
<td>11.91</td>
<td>0.2627</td>
<td>0.004452</td>
<td>11.44</td>
<td>11.88</td>
<td>12.42</td>
</tr>
<tr>
<td>τ2</td>
<td>21.81</td>
<td>0.2405</td>
<td>0.009738</td>
<td>21.37</td>
<td>21.79</td>
<td>22.3</td>
</tr>
<tr>
<td>τ3</td>
<td>26.69</td>
<td>0.5413</td>
<td>0.02395</td>
<td>25.42</td>
<td>26.69</td>
<td>27.66</td>
</tr>
<tr>
<td>τ4</td>
<td>36.27</td>
<td>0.4354</td>
<td>0.01133</td>
<td>35.33</td>
<td>36.39</td>
<td>36.88</td>
</tr>
</tbody>
</table>
age groups for males and females. The crude lung and bronchus mortality trends for males on the higher age groups (50-85+ years) increase continuously until the period of 1990 and then start decreasing. Also, the mortality trends make clusters of different age groups. 75-79 and 80-84 age groups have the highest mortality rate all the time. The rate is almost parallel throughout the time and is expected to decrease in the same fashion for the next couple of years. 70-74 and 85+ years age groups make another cluster of parallel trends being the second highest mortality rate of clusters. The lower age groups (25-50 years) mortality rates are stable for the entire range of time. In the future, most of the higher age groups are expected to exhibit decline in mortality rates whereas the pattern remains the same for the lower age groups. For females, the mortality trends for the higher age groups increase continuously right from the beginning and become stable after 2005, but there is almost a linear trend for the lower age groups (25-50 years). The next five year projections also follow the similar pattern.

Figure 11.: Fitted lung and bronchus mortality trends for male age groups

The estimated age-adjusted curve in Figure 13 is obtained by using the equation
Figure 12.: Fitted lung and bronchus mortality trends for female age groups

\[ E(r_i) = E \left( \sum_{j=1}^{J} w_j \frac{y_{ij}}{n_{ij}} \right) = \sum_{j=1}^{J} w_j E(r_{ij}), \]

where \( w_j \) is the standard 2000 year population weight for each age group and \( E(r_{ij}) \) is the estimated rate at time \( t_i \) for age-group \( j \).

The observed data points are also changed into the crude annual death rate by using the expression

\[ r_i = \sum_{j=1}^{J} w_j \frac{y_{ij}}{n_{ij}}, i = 1, 2, \ldots, n. \]

The age-adjusted trend given in Figure 13 shows that the mortality in males increased steadily from 68.54 per 100,000 in 1968 to 91.97 per 100,000 in 1990. It started to decrease from 1991 to 2009 and is expected to exhibit continuous decline in the future. The decreasing temporal trend for male lung cancer patients shows that we are making progress against lung cancer in males, whereas the female mortality trend increases from 1969 to 2003 and
seems to be stabilized thereafter. NCI reports [71] that the overall lung cancer death rates began to decline in women from 2005. Our study does not show a significant symptom of decline. Based on our model, it can be argued that the mortality is predicted to remain the same for the next couple of years. We believe these changes in the trends are due to the advancement in treatment, and change in the smoking behavior among males and females.

Figure 13.: Estimated age-adjusted mortality trends of male and female lung and bronchus cancer

Here, we fitted each model $E(r_{ij})$ separately based on the common slope assumptions to reduce the computational burden, and used the US 2000 standard population weights for age standardization. As we see each of the age groups mortality trends are not consistent (lower age groups are almost linear and higher age groups are with three or four joinpoints) and male age groups have decreasing mortality trends whereas the female age groups have increasing mortality trends, the estimated age-adjusted rates can over- or under-estimate [16]. The 95% pointwise credible intervals are the intervals of the mean function and looks narrower in the graph shown in Figure 4. This may be due to two factors: first, the scale of model is substantially broader, and second, the gender explains a great amount of variability.
in the model. Despite the rapid advancement in treatment for adult women, the estimated model does not show any significant decrease in mortality in recent years. For men, the mortality is decreasing rapidly.

4.2 Brain and CNS Cancer

As a second application to the proposed model, we studied a comprehensive assessment of the crude age-specific groups mortality and age-adjusted mortality due to brain cancer patients by gender. The estimated new cases and deaths from brain and other nervous system cancers in the United States in 2014 are 23,380 and 14,320 respectively. There are no currently known specific causes of brain tumors. Cancers of the lung, breast, and melanoma are the most common cancers to metastasize to the brain.

![Figure 14: Posterior probability of the number of joinpoints for brain cancer mortality trend](image)

As depicted in the graph given in Figure 14, the probability of the posterior distribution for four joinpoint is about 80 percent indicating that the other models are not preferable choices. Similarly, the boxplot for the parameters $\beta_j, j = 1, 2, 3, 4$ associated with the change points is observed. Their posterior means and credible intervals suggested that their posterior distributions are discriminable.

The Deviance Information Criteria (DIC) is also used as a measure of model comparison
DIC criterion and its application in our joinpoint model selection approach has been described in chapter 1. The Deviance Information Criteria (DIC) values for all five competing models for brain cancer are also given in Table 10. In the table below, $D_{\text{bar}}$ represents the posterior mean of deviance evaluated by an MCMC sample, $D_{\text{hat}}$ is a point estimate of the deviance obtained by substituting in the posterior means and theta, and $pD$ is the effective number of parameters given by the difference of posterior mean of the deviance and the point estimate of the deviance. Then, the Deviance Information Criteria (DIC) is given by

$$DIC = D + P_D = D(\theta) + 2P_D$$

<table>
<thead>
<tr>
<th>Number of joinpoints</th>
<th>$D_{\text{bar}}$</th>
<th>$D_{\text{hat}}$</th>
<th>$pD$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>No joinpoints</td>
<td>20882</td>
<td>20878.1</td>
<td>3.849</td>
<td>20885.8</td>
</tr>
<tr>
<td>One joinpoints</td>
<td>20607.9</td>
<td>20603.5</td>
<td>4.41</td>
<td>20612.4</td>
</tr>
<tr>
<td>Two joinpoints</td>
<td>20161.6</td>
<td>20156.9</td>
<td>4.694</td>
<td>20166.3</td>
</tr>
<tr>
<td>Three joinpoints</td>
<td>19862.2</td>
<td>19857.6</td>
<td>4.644</td>
<td>19866.9</td>
</tr>
<tr>
<td>Four joinpoints</td>
<td>19845.8</td>
<td>19840.3</td>
<td>5.506</td>
<td><strong>19851.3</strong></td>
</tr>
<tr>
<td>Total</td>
<td>101360</td>
<td>101336</td>
<td>23.104</td>
<td>101383</td>
</tr>
</tbody>
</table>

The DIC values for four joinpoints and three joinpoints are 19851.3 and 19866.9 respectively. This clearly supports the conclusion of the posterior probability for four joinpoints (80% occurrence) and the fit statistics (3.531).

We applied four joinpoints in the proposed age-stratified joinpoints regression model given in chapter 2. The analytical structure of the proposed model for the subject data are given by

$$\ln(r_{ij}) = \alpha_j + \beta_0 \ast (t_i - \bar{t}) + \gamma \ast z(t_i) + \gamma_1 \ast z(t_i) \ast (t_i - \bar{t}) + \sum_{k=1}^{K} \delta_k \beta_k B_{\tau_k}(t_i),$$

where $\bar{t}$ is the mean of $t_i$, and $\tau_j, j = 1, 2, 3, 4$ is the change point in the model $\alpha$ is the
intercept parameter, $\delta$ is the indicator variable, and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$) are the regression parameters for the joinpoints.

The estimated rates for each year from 1969-2009 are obtained by averaging the estimates of joinpoint and other parameters in the model at every step of MCMC by using the WinBUGS software. The posterior summaries for parameters including the estimates of change points (tau) are given in Table 5. The table clearly shows that the change points are observed at $t = 9.57, 14.33, 23.76, \text{ and } 38.57$ respectively.

As shown in Figures 15 and 16, the crude mortality rate for both male and female brain cancer trends follows the similar patterns for similar age-groups. The death rates in lower age groups (20 to 44 years) are almost constant from 1969 to 2009 and are predicted to remain the same in the future. The overall mortality trends for higher age groups increase from 1969 to 1992 and decrease until 2006 but the trends seem to increase from 2006 to 2009. The different age groups are clustered together and show a similar pattern of mortality trends. For both male and female age groups, the 70-79 age groups shows a similar pattern of mortality trends being, with the 75-79 year age group being the highest all the time.

The rates from 2010-2014 are predicted by applying the Bayesian Model Averaging approach as discussed in chapter 1. As we know, the Bayesian Model Averaging averages all sets of competing models and make an inference based on a weighted average on these models over the model space. We used 4 joinpoints to explain the model, and we used an encompassing model approach as given below (described in chapter 2) to choose the best competing model. In the mean time, this is the weighted average on these models over the model space. As we knew the analytical structure of the model, we extended that to get the future predictions. We recall from chapter 2 that the posterior probabilities of the model with all the variables (joinpoints) enclosed are given by

$$p(M_k|y) = \sum_{i=0}^{k} p\left(\sum \delta = i | y\right),$$
Table 5: The posterior summaries of parameters for brain cancer

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.50%</th>
<th>median</th>
<th>97.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3.531</td>
<td>0.838</td>
<td>0.04671</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>α1</td>
<td>-12.18</td>
<td>0.01391</td>
<td>1.06E-04</td>
<td>-12.21</td>
<td>-12.18</td>
<td>-12.15</td>
</tr>
<tr>
<td>α2</td>
<td>-11.84</td>
<td>0.01195</td>
<td>9.35E-05</td>
<td>-11.87</td>
<td>-11.84</td>
<td>-11.82</td>
</tr>
<tr>
<td>α3</td>
<td>-11.42</td>
<td>0.009828</td>
<td>7.73E-05</td>
<td>-11.44</td>
<td>-11.42</td>
<td>-11.4</td>
</tr>
<tr>
<td>α4</td>
<td>-11.06</td>
<td>0.008484</td>
<td>7.15E-05</td>
<td>-11.08</td>
<td>-11.06</td>
<td>-11.05</td>
</tr>
<tr>
<td>α5</td>
<td>-10.67</td>
<td>0.007289</td>
<td>5.70E-05</td>
<td>-10.69</td>
<td>-10.67</td>
<td>-10.66</td>
</tr>
<tr>
<td>α6</td>
<td>-10.27</td>
<td>0.006315</td>
<td>5.56E-05</td>
<td>-10.28</td>
<td>-10.27</td>
<td>-10.26</td>
</tr>
<tr>
<td>α7</td>
<td>-9.867</td>
<td>0.005495</td>
<td>5.11E-05</td>
<td>-9.878</td>
<td>-9.867</td>
<td>-9.857</td>
</tr>
<tr>
<td>α11</td>
<td>-8.891</td>
<td>0.004684</td>
<td>4.28E-05</td>
<td>-8.891</td>
<td>-8.891</td>
<td>-8.881</td>
</tr>
<tr>
<td>α12</td>
<td>-8.84</td>
<td>0.005139</td>
<td>4.72E-05</td>
<td>-8.84</td>
<td>-8.84</td>
<td>-8.83</td>
</tr>
<tr>
<td>α13</td>
<td>-8.876</td>
<td>0.006306</td>
<td>5.15E-05</td>
<td>-8.889</td>
<td>-8.876</td>
<td>-8.864</td>
</tr>
<tr>
<td>β0</td>
<td>0.002805</td>
<td>2.06E-04</td>
<td>2.92E-06</td>
<td>0.002402</td>
<td>0.002806</td>
<td>0.003209</td>
</tr>
<tr>
<td>γ</td>
<td>0.4177</td>
<td>0.003207</td>
<td>4.77E-05</td>
<td>0.4115</td>
<td>0.4177</td>
<td>0.424</td>
</tr>
<tr>
<td>γ * β0</td>
<td>-0.002232</td>
<td>2.71E-04</td>
<td>3.84E-06</td>
<td>-0.002761</td>
<td>-0.002233</td>
<td>-0.001691</td>
</tr>
<tr>
<td>β1</td>
<td>0.03475</td>
<td>0.04732</td>
<td>0.001354</td>
<td>-0.07243</td>
<td>0.03966</td>
<td>0.08647</td>
</tr>
<tr>
<td>β2</td>
<td>-0.04826</td>
<td>0.05574</td>
<td>0.001746</td>
<td>-0.1079</td>
<td>-0.05797</td>
<td>0.08543</td>
</tr>
<tr>
<td>β3</td>
<td>0.1094</td>
<td>0.01024</td>
<td>5.15E-04</td>
<td>0.09158</td>
<td>0.1105</td>
<td>0.1284</td>
</tr>
<tr>
<td>β4</td>
<td>-0.007965</td>
<td>0.001874</td>
<td>3.53E-05</td>
<td>-0.0111</td>
<td>-0.007912</td>
<td>-0.005</td>
</tr>
<tr>
<td>δ1</td>
<td>0.7622</td>
<td>0.4258</td>
<td>0.02352</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>δ2</td>
<td>0.7711</td>
<td>0.4201</td>
<td>0.02326</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>δ3</td>
<td>0</td>
<td>0</td>
<td>3.16E-13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>δ4</td>
<td>0.9981</td>
<td>0.04378</td>
<td>0.001369</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>τ1</td>
<td>23.76</td>
<td>0.5338</td>
<td>0.2219</td>
<td>22.88</td>
<td>23.67</td>
<td>24.86</td>
</tr>
<tr>
<td>τ2</td>
<td>14.33</td>
<td>2.566</td>
<td>0.086</td>
<td>10.71</td>
<td>13.68</td>
<td>21.41</td>
</tr>
<tr>
<td>τ3</td>
<td>9.572</td>
<td>2.086</td>
<td>0.02865</td>
<td>3.946</td>
<td>9.842</td>
<td>14.75</td>
</tr>
<tr>
<td>τ4</td>
<td>38.57</td>
<td>0.4535</td>
<td>0.008553</td>
<td>37.82</td>
<td>38.65</td>
<td>38.98</td>
</tr>
</tbody>
</table>
where $M_k$ be the set of competing models, $\delta \in \{0, 1\}^k$ is binary inclusion indicators for all $\tau_k$‘s in the model known as latent vector, and $p(\delta|y)$ is the posterior distribution of $\delta$. Here, the posterior probability of the sum of $\delta = i$ for all four joinpoints is used to estimate the model. Since this sum is also the weighted average of all competing model, it is extended to provide the future predictions. We used the last years population information to predict the future rates. The future rates are predicted for short period of times as we are using log linear model which are not suitable for long term prediction and the population information is usually unavailable for future.

The 5-year predicted trends also continue to follow the increasing trend for every age group above 40 years of age for both males and females. However, mortality trends below 39 years of age groups are expected to remain constant keeping the same rates.

![Crude brain cancer mortality trends for male age groups](image)

**Figure 15.:** Fitted brain cancer mortality trends for male age groups

The estimated age-adjusted curve in Figure 17 is obtained by using the equation

$$E(r_i) = E\left(\sum_{j=1}^J w_j \frac{y_{ij}}{n_{ij}}\right) = \sum_{j=1}^J w_j E(r_{ij}),$$
where \( w_j \) is the standard 2000 year population weight for each age group and \( E(r_{ij}) \) is the estimated rate at time \( t_i \) for age-group \( j \).

The observed data points are also changed into the crude annual death rate by using the expression

\[
r_i = \sum_{j=1}^{J} w_j \frac{y_{ij}}{n_{ij}}, \quad i = 1, 2, \ldots, n.
\]

For the age-adjusted trend shown in Figure 17, there is a good qualitative agreement between male and female mortality rates. The age-adjusted mortality rates between men and women show similar patterns throughout the entire data range. The temporal mortality trends in both groups decrease significantly from 1990 to 2006 whereas the results are quite discouraging after 2006 as the rates are increasing and the model predicts the trend to increase in the future for both genders. Most interestingly, the gap on the mortality between genders has not changed in the last 41 years. Also, the narrow 95% pointwise
Figure 17.: Estimated age-adjusted mortality trends of male and female brain cancer

credible intervals, the interval of the mean function, suggests that the gender explains a greater amount of variability in the model.

4.3 Conclusion

We applied the proposed Bayesian Age-stratified joinpoint model in two different cancer mortality data. The model is used to estimate the age-groups specific rates as well as age-adjusted estimates for lung and brain cancer mortality data of the United States. The mortality of the two important cancers are explored and the changepoints for the mortality trends were identified. From this analysis, we observed several important information pertaining the lung and brain cancer mortality and future prediction with a certain level of confidence. The estimation of the joinpoints and the future prediction help in decision making regarding both cancers.

We observed two different results from male and female lung cancer mortality data.
The male age-adjusted lung cancer mortality rate is decreasing whereas the female rate is increasing and are expected to meet after certain years. This is a clear indication that policy makers should work to find the reason behind the increasing rate of female lung cancer mortality and act on it to reduce the mortality rate. Smoking is one of the primary risk factors for lung cancer. 85 to 90% of the lung cancer is estimated due to cigarette smoking [74]. The incorporation of smoking as one of the covariates in the model will explain the variation in mortality rates and can compare the mortality trends of lung cancer with or without smoking.

The age-adjusted brain cancer mortality estimates suggest that mortality for this cancer is increasing for both genders forcing public health officials to focus on medical interventions or early detection and to find any other causes that are responsible for an increase in mortality rates. These is information for the nation to act on these cancers as the overall mortality rate due to all cancers in the nation is decreasing.

It is estimated that approximately 12.1 and 4.5 billion dollars are spent in the United States each year on lung cancer and brain cancer treatments, respectively. We applied the developed model to study the age- adjusted trend of the mortality of lung and bronchus, and brain and other CNS cancers from the SEER database of the NCI. This information helps us to manage the on going research in lung and brain cancer as the produced estimates are good and give the short term predictions. We made several contributions from these studies in the modeling aspect which will be discussed below.

4.3.1 Contributions

We have shown that our proposed age-stratified model unveils the patterns in two different purposes in studying mortality (or incidence) of a disease in a population. It gives information for the age-specific age group trend and compares the trend in different population sub-groups. We summarize our contributions in the following points.

1. The proposed parametric Bayesian Joinpoint model can be used to identify the changes
in the age-adjusted mortality (or incidence) rates and their APC in the trend of different cancers.

2. Our modeling approach focuses posterior quantification of post data uncertainty in the estimation and detection of joinpoints giving more accurate results.

3. The proposed model uses the counts of each age-group and incorporates the changes in the effect of time on the outcome across the different population subgroups.

4. The proposed model can be extended easily to compare the trends among the different regions and can statistically compare the Annual Percentage Changes in the trends.
Chapter 5

Functional Data Analysis Approach to Study of the Rate of Change of Carbon Dioxide from Gas Fuel in the Atmosphere

The second part of this dissertation is the statistical analysis and modeling of carbon dioxide emission data. In this chapter, we develop a system of differential equations to study the rate of change of carbon dioxide in the atmosphere using functional data analysis approach. Global Warming is a growing concern as we experience an increase in the surface temperature of the earth with increase in carbon dioxide. Carbon dioxide including other air pollutants is the major causes of Global Warming. Atmospheric temperature and carbon dioxide are considered as the two main factors of Global Warming. The United States is one of the largest source of global warming pollution and currently ranks in number two in carbon dioxide emissions followed by China. China ranks first in carbon dioxide emission for more than two years. The United States is contributing 4 percent in the world pollution. It produces 25 percent of carbon dioxide by burning fossil fuel. Because of the rapid increment in global greenhouse gas emissions, all countries around the world are facing extreme pressure to reduce carbon dioxide emission.

The study of carbon dioxide emission trends estimates the rate of change of carbon dioxide in the atmosphere at any time. This type of study is an important entity to understand the behavior of carbon dioxide and global warming. This is the reflection of production of carbon dioxide and the estimation of the rate of change of production of carbon dioxide as a function of time. There are different variables that are significantly contributing to the emission of carbon dioxide in the atmosphere. The schematic diagram given in Figure 18 below shows the relationship among different attributable variables that contributes carbon
dioxide emission in the atmosphere [87].

![Figure 18.: Emission of Carbon Dioxide in the Atmosphere in U.S.A.](image)

### 5.1 Objective

The present and future objective of this study is to develop a system of differential equations using time series data on the major sources of the significant contributable variables of carbon dioxide in the atmosphere. We are interested in obtaining the good estimates of the rate of change of carbon dioxide in the atmosphere at a particular time in the trend.

Bringing the emission of carbon dioxide to an acceptable level is an important issue. It is very important to study the emission behaviour related to the different contributing factors. This type of knowledge helps the policy makers to determine which variables are significantly increasing or decreasing in terms of carbon dioxide emission at a particular time. Based on this rates, they can determine the necessary factors that have lead to this change and develop the appropriate policies. Moreover, they can develop a monitoring
system to avoid the uncontrollable emission of carbon dioxide. Recently, more research is being done to understand and control the emission behavior of carbon dioxide. This study estimates the rate of change of carbon dioxide in the atmosphere and helps the policymakers to prioritize and develop realistic strategies plans to address the problem.

Differential equation with respect to fitting the carbon dioxide emission data gives a representation of carbon dioxide in the atmosphere at any time. We have historical time series data on carbon dioxide emission for each of the major attributable variables. Having such a data for all covariates, we derive the system of differential equations that estimate the trend behavior of the carbon dioxide in the atmosphere. If we differentiate the function at any time in the time trajectory, we obtain the status of the carbon dioxide in the atmosphere at that time. Having this characterization, we can determine what happens to carbon dioxide emission rate at a particular time. We can use this information in planning purposes. If the rate of emission is above certain target, we need to take precautionary measures. If it stays below the target, we are making progress. If it stays at the same level, we are being able to control it, but are not being able to make any progress etc. If we continue our projection, it will give us a forecast of the carbon dioxide in the atmosphere due to particular covariate at a time in future.

5.2 Carbon Dioxide Emission Data

In this study, the data set is obtained from the Carbon Dioxide Information Analysis Center (CDIAC). CDIAC is the primary climate-change data and information analysis center of the United States Department of Energy. It collects the air samples for the U.S. data at Mauna Loa Observatory, Hawaii. It is located at the Oak Ridge National Laboratory (ORNL) and includes the World Data Center for Atmospheric Trace Gases. The World Data Centers (WDCs) provide archives for the data gathered during the International Geophysical Year (IGY) since 1957. WDCs operate under the the International Council of Scientific Unions (ICSU) and its main goal is to benefit the international scientific community by providing a
mechanism for international exchange of data related to the Earth, its environment, and the Sun. They collect data from scientists, projects, institutions, local and national data centers. CDIAC’s data provide estimates of carbon dioxide emissions from fossil-fuel consumption and land-use changes. It provides the records of concentrations of carbon dioxide and other gases in the atmosphere. It also provides the data on carbon cycle and terrestrial carbon management.

We obtained the yearly emission data from 1950 to 2010 in our analysis. All the carbon dioxide emission attributable variables are majored in thousands metric tons of carbon. Carbon Dioxide Information Analysis Center (CDIAC) and other studies defines Gas, Liquid, Solid, Cement, Flaring, Bunker are the major sources of the significant contributable variables of carbon dioxide in the atmosphere in the continental United States.

5.3 Literature Review

The literature is very rich with respect to carbon dioxide emission data. Some previous studies have ranked the attributable variables using statistical model approach that constitute the emission of carbon dioxide in the atmosphere [87, 88]. Xu and Tsokos (2013) did the parametric statistical analysis for the emission of carbon dioxide in the atmosphere. They rank the variables that contributes the emission of carbon dioxide in the atmosphere based on the continental United States data [87]. Their model ranks the variables based on individual contributions and their interactions. They ranked the variables and their interactions based on the contribution and is given the in following table. They found liquid, bunker, cement, gas flares, and gas fuels significantly contributing to the emission of carbon dioxide in the atmosphere. Moreover, they observe the five interactions also contribute to the emission.

The individual contributions and interactions along with their percentage of contribution is given in Figure 19 below

Here our goal is to develop a statistical model to study the emission trend of carbon diox-
<table>
<thead>
<tr>
<th>Rank</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Liquid</td>
</tr>
<tr>
<td>2</td>
<td>Liquid interact with Cement</td>
</tr>
<tr>
<td>3</td>
<td>Cement interact with Bunker</td>
</tr>
<tr>
<td>4</td>
<td>Bunker</td>
</tr>
<tr>
<td>5</td>
<td>Cement</td>
</tr>
<tr>
<td>6</td>
<td>Gas Flares</td>
</tr>
<tr>
<td>7</td>
<td>Gas Fuels</td>
</tr>
<tr>
<td>8</td>
<td>Gas Fuels interact with Gas Flares</td>
</tr>
<tr>
<td>9</td>
<td>Liquid interact with Gas Flares</td>
</tr>
<tr>
<td>10</td>
<td>Gas Flares interact with Bunker</td>
</tr>
</tbody>
</table>

The use of statistical models to understand the emission trend of carbon dioxide are not very promising in literature. The concept of study of rate of change of carbon dioxide with respect to time was started by Goreau in 1990 [80]. Tsokos and Xu (2009) modeled the carbon dioxide emission data from the Continental United States with a system of differential equations [86]. They fitted the differential equation of each of the attributable variables of yearly emission of carbon dioxide and the sum of all of these variables. They provided the analytical structure of the estimated differential equation for each of the variables. To develop their model, they used $R^2$ ($Adjusted R^2$), PRESS Statistic, and residual analysis to evaluate the quality of their proposed differential equations. They used these models to predict the emission of carbon dioxide for long term. To best of our knowledge, this is the first approach to represent the carbon dioxide emission data using the differential equation based on statistical modeling approach. Their fitted differential seems to represent the trend well but lacks of actual fit. Also the use of the normal assumption in fitting the differential equation is questionable in carbon dioxide emission data.

Tian and Jin used the dynamic system method to study the evolutionary rule of carbon dioxide emissions and dynamic evolutionary scenarios [85]. Their model can predict the carbon dioxide in future in China. We need different control function, carbon reduction
coefficient, and evolutionary coefficient in every other regions to apply their model, which may not be suitable to apply in general. As the modeling and understanding of the trends of emission using good statistical approach is indeed a need for the carbon dioxide emission data. In the next section, we focus on modeling objectives with respect to the study of rate of change of carbon dioxide in the atmosphere.

5.4 Modelling Objectives

Our aim is to develop differential equation for each of the components using functional data analysis approach that estimates the rate of change of carbon dioxide at a particular time in the continental United States. But, in this chapter, we will be studying the rate of change of carbon dioxide in the atmosphere due to gas fuels only. Emission of carbon dioxide due to Gas fuels include gas consisting primarily of methane. They include natural gas and other gases that provide energy through combustion. The study of the rate of change for the other contributable variables will be our future study and follow the same modeling approach.

In this study, we plan to develop a system of differential equations that best describe
the rate of change of Carbon dioxide in the atmosphere. The developed model expresses at least a substantial amount of variation in the carbon dioxide emission data and provides the best prediction of carbon dioxide emissions rate in the atmosphere in future. In this study, as developed by Ramsay and Silverman (2005), we are working on developing the differential equation using the carbon dioxide emission data. Differential Equations are useful to provide feedback to control the behaviour of the system and are getting very popular to model noisy data. Mostly, we are interested in the rate of change of carbon dioxide, so the behavior of a derivative is of more interest than the function itself. For short and medium time periods, we are mostly interested to know how it is changing with respect to time. Differential equations are appealing as they can imply function characteristics for different data that are difficult to model in other ways [82, 83]. We define the differential operator as data smoother and use the penalized least square fitting criteria to smooth the data. Finally, we optimize the profile error sum of squares to estimate the necessary differential operator. In the following section, we describe this approach in detail.

Although we are interested to develop a system of differential equations for the variables that significantly contribute to carbon dioxide emission in the atmosphere, in this chapter we focus on one variable, gas fuel, which ranks seventh in contributing carbon dioxide in the atmosphere. We will use the same statistical approach to develop the differential equation for other contributing variables in future.

5.5 Statistical Modeling Approach

We use the differential equation to study the rate of change of emission of carbon dioxide due to gas fuel in the atmosphere. Statistical modeling includes modeling of random variation in a data set obtained through a certain process. The modeling process is described by capturing and explaining the variation in the outcome process due to various input factors. The differential equations are important of being a dynamic aspect of the observed process based on which the rate of change are modeled. Ramsay and Silverman (2002, 2005) de-
developed new methods for fitting differential equations from noisy data that appears to be more appealing than the existing techniques as these methods are based on the development of functional data analysis approach [82, 83]. We apply this methodology to create and estimate the differential equation that best represent the trend behaviour of carbon dioxide in the atmosphere.

5.5.1 Functional Data Analysis

Functional data analysis is a statistical method to analyze the data based on information about the curves. This method analyzes the data obtained as a sample of functional variable [79, 81]. If the variance in the data set is very high (noisy), then we need a special type of analysis to capture those variances. In the functional data analysis, observations are transformed into the curve first using the repeated measurements and these curves are estimated. If we have discrete data across time with the assumption of observational error, we use smoothing to convert these data from discrete to continuous functions. In present method of statistical analysis, we look at the functional data (curves) as a whole instead of observations. The functional data is represented through a series of basis functions. This means functional data objects are constructed by specifying a set of basis functions. The basis function are any type of mathematical function that is suitable to represent the observed data. Some examples are fourier basis, spline basis etc. We are interested in the parameters of the basis function rather than the data itself. Here we would like to get the information regarding the slopes and curvature of the functional curves. This means we are estimating the slope and curvature of those basis functions.

As discussed by Ramsay and Silverman (2005), the common goals of the functional data analysis are to represent the data in different ways that help to produce further statistical analysis, to display the data so as to highlight its different characteristics and patterns it possesses, and to explain the variation in an outcome with the help of attributable variables [82, 83].
5.5.2 Linear Differential Operator

The application of a derivative is important if we are interested in the study of the rate of change. Of equal importance is the functional rate beyond the data range in time trend. The differential equation provides us real information in both the functional form itself and its derivative at the same time.

We define the linear differential operator based on the nature of the data.

\[ L_x(t) = \beta_0 x(t) + \beta_1 D x(t) + \beta_2 D^2 x(t) + \beta_3 D^3 x(t) + \beta_4 D^4 x(t) + \ldots \]

where operator L is the re-arrangement of the proposed differential equation, \( x(t) \) is the basis functions, \( \beta' s \) are the parameters to be estimated, and \( D^n x(t) \) is the \( n^{th} \) derivative of the basis function.

5.5.3 Fitting Differential Equation

Fitting a differential equation from noisy data means to fit the unknown parameters that are the coefficient functions that define the differential equation. we use the profile least square approach to estimate those unknown parameters. If we know the differential equation, then the operator L can be defined as a data smoother. The penalized least squares fitting criterion is given by:

\[ PENNSE = \sum_{i=1}^{N} \left[ (y_i - x(t_i))^2 + \lambda \int [L_x(t)]^2 dt \right] \]

where \( \lambda \) is a smoothing parameter, \( y \) is the vector of noisy observation to be smoothed, the second term in the right measures the penalty matrix with \( L_x(t) \) as a function of weight coefficients, and \( x(t_i) \) is the basis function.

Here, the penalized linear least square criterion given above is minimized for the minimum value of \( \lambda \) and we obtained the smoothing parameter \( \lambda \). Here we select the \( \lambda \) by
minimizing the generalized cross validation criteria. Once we obtain the minimum value of $\lambda$, we minimize the un-penalized profiled error sum of squares to estimate the linear differential operator by

$$PROFSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

with respect to the parameter vectors.

### 5.5.4 Two Levels of Fitting

This approach of fitting the statistical model using differential equations can always be explained by two levels of fitting. The low-dimensional fitting is defined by the solution of a differential equation. High-dimensional fitting is obtained by keeping smoothing parameter $\lambda$ possibly low so that the roughness penalty did not over or under fit the data. As explained by Ramsay and Silverman (2005), this is the partition of functional variance into two parts. The first one is the low dimensional part that is captured by the proposed differential equation and the second one is the balance between low and high dimensional fits.

### 5.6 Trend Analysis of Carbon Dioxide Emission from Gas Fuels

We apply the above outlined method to study the trend behaviour of carbon dioxide emission data from gas fuels. We obtained carbon dioxide yearly emission data from 1950 to 2010 due to gas fuel from the CDIAC center. The data set is measured in thousands of metric tons. The scatter diagram of the data is shown below. We use the statistical software R for functional data analysis to analyze the data [62, 84].

Here, our goal is to capture the trend of this emission data. First, we create the functional data objects constructed by specifying a set of basis functions. And we look at the pattern of the data. The nature of the trend is linear and periodic. As we know straight line solves the differential equation $D^2x(t) = 0$, and $\sin(\omega \ast t)$ and $\cos(\omega \ast t)$ solves $D^2x(t) = -\omega^2x(t)$
for the period $2\Pi/\omega$. Putting these all together gives

$$D^4x(t) = -\omega^2 D^2x(t)$$

We incorporate the damping effect in the following way

$$D^4x(t) = -\beta_1 D^2x(t) - \beta_2 D^3x(t)$$

Where $\beta_1 = -\omega^2$ and $-\beta_2 D^3x(t)$ allows for the exponential decay. Then the linear differential operator is defined by

$$L_{x(t)} = \beta_1 D^2x(t) + \beta_2 D^3x(t) + D^4x(t)$$

where operator $L$ is the re-arrangement of the proposed differential equation.

The solution of this differential equation represent the function that best describe the trend. Here the proposed equation is;
\[ D^4x(t) + \beta_2 D^3x(t) + \beta_1 D^2x(t) = 0 \]

Let \( D^2x(t) = y(t) \) then,
\[ y''(t) + \beta_2 y'(t) + \beta_1 y(t) = 0 \]

The characteristic equation is given by
\[ r^2 + \beta_2 r + \beta_1 = 0 \]
\[ \Rightarrow r = \frac{-\beta_2 \pm \sqrt{\beta_2^2 - 4\beta_1}}{2} \]

Therefore, \( y(t) = C_1 \exp\left(\frac{-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) + C_2 \exp\left(\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) \]

\[ x''(t) = C_1 \exp\left(\frac{-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) + C_2 \exp\left(\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) \]

\[ x'(t) = \frac{2C_1}{-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1}} \exp\left(\frac{-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) + \frac{2C_2}{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1}} \exp\left(\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) + C_3 \]

\[ x(t) = \frac{2C_1}{(-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1})^2} \exp\left(\frac{-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) + \frac{2C_2}{(-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1})^2} \exp\left(\frac{-\beta_2 - \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) + C_3 t + C_4 \]

when \( \beta_2^2 - 4\beta_1 < 0 \),
\[ \exp\left(\frac{-\beta_2 \pm \sqrt{\beta_2^2 - 4\beta_1}}{2}\right) = \exp\left(\frac{-\beta_2 t}{2}\right)[\cos((\sqrt{4\beta_1 - \beta_2^2})t) \pm i\sin((\sqrt{4\beta_1 - \beta_2^2})t)] \]

\[ x(t) = c_1 \exp\left(\frac{-\beta_2 t}{2}\right)\cos\left(\sqrt{4\beta_1 - \beta_2^2}\right)t + c_2 \exp\left(\frac{-\beta_2 t}{2}\right)\sin\left(\sqrt{4\beta_1 - \beta_2^2}\right)t + c_3 t + c_4 \]

Here, choosing a differential operator is first task and estimating the value of \( \beta' \)'s is another. We can use the functional regression to estimate the parameter estimates of the differential operator. We approach this problem in a different way. As we understand the nature of the data, we can provide the coefficient of the linear differential operator. This way we can simplifies the problem if we know the pattern of the trend. For the possible values of beta we can always obtained the residual mean square error and check the
effect of coefficient of operators. The linear differential operator with known values of coefficient that provides the minimum residual mean square error is chosen as the best differential operator. Since we have pre guess regarding the pattern of the data, we choose the coefficient of operator as $L_{x(t)} = (0, 0, (w)^2, 1, 1)$. This means, for this problem the values of $\beta_1 = (w)^2$ and $\beta_2 = 1$ and coefficient of $D^1x(t)$ is 1 by default gave us the minimum error giving us the best fit with minimum value of $\lambda$. In the mean time we apply the Generalized Cross Validation approach and obtained the minimum value of lambda. This fitting approach results in obtaining the minimum value of $\lambda$ 15.07. We fitted the data again using the same differential operator and the minimum value of $\lambda$ and the fitted trend is given in Figure 20. This fit is obtained with the minimum residual mean square error of 5403.066697. This error looks high but our data set is measured in thousands of metric tons. Moreover, this error is the minimum error we obtained on changing the coefficient of differential operator ($\beta'$s).

The obtained differential equation fits the data very well with by characterizing the emission behaviour of carbon dioxide due to gas fuels. The solution to this differential equation estimates the rate of change of carbon dioxide due to gas fuel in the United States at any time. The fitted trend line is significant improvement over the existing models to characterize the emission trend [86]
Figure 21.: Estimated Model for the Emission of CO2 from Gas Fuels

5.7 Conclusion

We develop the differential equation model based on functional data approach to study the rate of change of emission of carbon dioxide due to gas fuel in the atmosphere. The presented methods and methodology describe the emission behaviour trend of carbon dioxide due to gas fuels in the continental United States. The obtained differential equation estimates the rate of change at any time in the trend and provides good future predictions in the emission rate of carbon dioxide due to gas fuels. This helps the policy makers to establish new laws that requires the industries to cut the emission by certain percentages that helps to keep the global warming low. Climate stabilization is an important issue and needs special attention from all sectors that help us properly formulate the global policies. Moreover, they can fully utilize the resources to study and control the particular aspect that cause and increase the emission of greenhouse gases.

A statistical model that can help to provide the most reliable estimate of rate of emission due to gas fuels at a particular time is crucial to understand its contributions to carbon
dioxide. As global warming is an important issue, that needs to be addressed based on
the information from the data that helps the environmentalist to understand the emission
behaviour of carbon dioxide. A similar procedure can be used to develop differential equa-
tion for each of the contributing factors of carbon dioxide emission data and for the total
emission. The developed model provides the rate of change of carbon dioxide due to gas
fuel at any time in the atmosphere.

5.7.1 Contribution

In this chapter we made the following contributions.

1. We develop the system of linear differential equation, that can be helpful to estimate
   the rate of change of carbon dioxide in the atmosphere at a particular time.

2. The developed system of differential equations is based on historical time data that can
   be used to predict the amount of carbon dioxide due to gas fuel in the future in the
   atmosphere.

3. The scientific community can utilize this information in the rate of change of carbon
dioxide in their long term planning to control the emission of greenhouse gases.

4. This study helps the policy makers to identify the highest rate of change related to cer-
tain covariate factors and allocate the research fund to stabilize and better understanding
the emission.
Chapter 6

Future Work

6.1 Future Research in Bayesian Joinpoint Regression

In the first part of the dissertation, we propose a parametric Bayesian joinpoint model that identify the changes in the age-standardized mortality or incidence rates and their APC in the trend. Our proposal model is suitable to study the age-stratified rates to study the summary measure for each age group rates and the age-adjusted rates are studied to compare the mortality (or incidence) in a population. While doing so the assumption of parallel models help us to reduce the computational burden and takes care of the confounding effect of age in studying the trends. Moreover, our study also focuses on the posterior quantification of post data uncertainty related to the detection of possibly large number of joinpoints. The proposed model uses the counts of each age-groups and incorporates the changes in the effect of time on the outcome across the different population sub groups. The external factors such as socio economic status (education, wealth, and income), environment, nutrition, and lifestyles have an effect in mortality. The inclusion of these information in the study of cancer trends will be an added advantages.

A careful and full utilization of resources in cancer research is important. We need to understand the cancer, evaluate cancer control interventions, and estimate the future burden. As the burden of cancer is growing, having good estimates and predictions of such mortality (or incidence) rates not only help us to monitor and evaluate the current status of the disease, but also to make an evidence based policy for resources allocation. In fact, these measures are an integral part to compare the trends in mortality between subgroups
of patients that helps policy makers and scientists for planning public health programs and medical interventions. More importantly, there is a three year lag in time to collect and process the data. In this scenario, the proposed model is not only able to describe the data but produce the predictions based on Bayesian model averaging approach and is the most reliable way to incorporate the uncertainty in the model and its prediction [28, 47].

The model can be extended to account for the overdispersion. We know that incidence of lung cancer is highly correlated with smoking behavior. Smoking behavior have huge impact in the incidence and mortality of lung cancer among others. This analysis can be used to develop a model in the longitudinal data on the smoking rate and age-adjusted incidence rate jointly to explore the relationship between the two. This type of analysis will be an interesting continuation of the current study. Also, study of incidence and mortality rates at the same time will actually depict the clear picture of real improvements we are making in cancers. Moreover, we can clearly see the effect of smoking in the incidence and mortality of lung cancer in a population and its subgroups.

In addition to that, we can develop a parametric Bayesian joinpoint regression model for the population based survival data using the same methodology outlined above. We also plan to extend this method to develop the semiparametric Bayesian joinpoint regression model for relative survival data where the parametric assumptions in the model will be relaxed by modeling the distribution of regression slopes using the Dirichlet process mixtures.

6.2 Future Research in Differential Equation in Global Warming

The second part of this dissertation is also on the study of trend behaviour of the emission of carbon dioxide in the atmosphere. We develop a differential equation based on statistical approach to study the rate of change of carbon dioxide in the continental United States due to one attributable variable. Our obtained differential equation characterize the emission trend very well and can estimate the emission rate at any time.
To keep the carbon dioxide in control or below certain level is an important issue. A lot of factors are responsible for the carbon dioxide emission. However, the developed model to study the rate of change of carbon dioxide due to gas fuels and similar differential equations of other variables give the clear information to the policy makers to focus on different sectors to control the carbon dioxide emission. The investment should be rational to control the increasing behaviour of carbon dioxide in the atmosphere. This type of study will help to find the real rate of change of carbon dioxide in the atmosphere due to different attributable variables. This information can be utilized for further research to estimate the cost to keep or balance the carbon dioxide in the atmosphere.

In this approach we fitted the model by fixing the coefficient of the differential operator. The method can usually be extended to develop a system that fits the parameter of the differential operator as well. We are working on this approach to fit the model. Moreover, previous studies have already notified that a world wide monitoring system is required to keep the level of carbon dioxide emission below certain level. Information on per capita income and the emission of carbon dioxide give important information in the behavioural study of emission of carbon dioxide. The incorporation of per capita income helps to explain substantial amount of information in the trend study of carbon dioxide emission. We will consider this as an important advancement of our study. After the adjustment of per capita income in the model, we can compare the rate of carbon dioxide in the different regions around the world with respect to their development process. It helps us to understand the real behaviour of the rate of change of carbon dioxide compare the emission of CO2 in the continental United State models with other similarly developed model in the world and help to develop the global policy in atmospheric change. We believe that this will be an important information to facilitate the policy makers to introduce policies in reducing the carbon dioxide.
References


