6-9-2016

Statistical Analysis and Modeling Health Data: A Longitudinal Study

Bhikhari Prasad Tharu
University of South Florida, bhikhari@mail.usf.edu

Statistical Analysis and Modeling Health Data: A Longitudinal Study

by

Bhikhari Prasad Tharu

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 M. Ramachandran, Ph.D.
Dan Shen, Ph.D.
Lu Lu, Ph.D.

Date of Approval:
June 7, 2016

Keywords: Histogram Smoothing, Bayesian Statistics, Lung Cancer Mortality, Functional Data Analysis, Total Serum Cholesterol Level, Longitudinal Data

Dedication

I dedicate my dissertation work to my parents Chongal and the late Basdaiya, wife Sitarani, son Srijan, and daughter Diya.

Acknowledgments

I would like to express my sincere gratitude to my advisor Professor Chris P. Tsokos for his continuous guidance, encouragement, and support during the development of this work. His enthusiasm, motivation, and passion for research and teaching are inspirational.

I am indebted to the members of my dissertation committee: Dr. Kandethody M. Ramachandran, Dr. Dan Shen, and Dr. Lu Lu for their insight, suggestions, and support which are valuable to me. I am thankful to Dr. Getachew A. Dagne for chairing the session.

I would like to thank my friend Dr. Ram C. Kafle for his contribution in this work. In addition, my heartfelt thanks to Dr. Netra Khanal, Dr. Keshav P. Pokhrel, and Dr. Krishna Khatri for their suggestions.

Finally, I am grateful to my parents and wife for their everlasting love, support, and encouragement.

Table of Contents

List of Tables

List of Figures

Abstract

Lung cancer has been considered one of the leading causes of deaths while cancer remains the second most common cause of deaths in the USA. Understanding the behavior of a disease over time could yield important information to make decisions about the disease. Statistical models could provide crucial clues and help to make a decision about the disease, budget allocation, evaluation, and implement prevention. Longitudinal trend analysis of the diseases helps to understand long term effects and nature. Cholesterol level is one of the most contributing risk factors for Coronary Heart Disease. Studying cholesterol statistically helps to know more about its nature and provides crucial information to mitigate its effectiveness in diagnosing its impact to public health.

In our study, we have analyzed lung cancer mortality in the USA based on age at death, period at death, and birth cohort to investigate its nature in longitudinal effects. The attempt has been made to estimate mortality rate based on age for different age groups and to find the relative risk of mortality due to period effect and relative risk due to birth cohort for lung cancer in the United States. Our statistical analysis and modeling are based on the data obtained from Surveillance Epidemiology and End Results (SEER) program of the United States.

We have also investigated the probabilistic behavior of average cholesterol level based on gender and ethnicity. The study reveals significant differences with respect to the distribution they follow and their basic inferences which could be beneficial to draw conclusions in various ways in addressing related issues. At the same time, the change of cholesterol level over time for an individual might be a good source to study the association of cholesterol level, coronary heart disease and their effects on age. The cholesterol data is obtained

from inter-university Consortium for Political and Social Research and National Health and Nutrition Examination Survey (NHANS) of the United States.

Understanding the average change in total serum cholesterol level over time as people get older could be vital to explore it. We have studied the longitudinal behavior of the association of sex and time with cholesterol level. It is observed that age, sex, and time have an individual effect and can impact differently upon collective considerations. Their adverse effect in increasing cholesterol level could promote to worsen the cholesterol related issues and hence heart related diseases. We believe our study pivots knowing more about target population of cholesterol level and helps to have the useful inference about cholesterol levels for public health.

Finally, we also analyzed the average cholesterol data using a functional data analysis approach to understand its nature and effect on age. Since functional data analysis approach presents more flexibility in modeling, it could provide more insight in studying cholesterol level.

Chapter 1

Introduction

In this chapter we briefly discuss some of the statistical methods, algorithms and estimation procedures that we used in the present studies.

Mortality due to lung cancer in the United States is really alarming with respect to public health safety. It is one of the major causes of cancer deaths, where cancer alone is approaching the number one cause of deaths in the USA. Analyzing cancer deaths over time might provide important clues to the epidemiologists, public health practitioners, and decision makers about the disease to better understand its nature and efficacy of past decisions.

Lung cancer has been considered the second most commonly diagnosed cancer among men and women in the United States [1, 2]. In our second chapter, we have estimated the lung cancer mortality rate based on different five year age groups from median age 22 years to 82 years from 1971 to 2010, the most recent data available for the USA. We have also estimated the relative risk due to period effects and due to birth cohort of lung cancer for the USA. Most of Chapter Two has already been published in the journal "Epidemiology Biostatistics and Public Health" [3].

Normality assumption may not always be the case to study the behavior of the phenomenon we are interested in. In the third chapter, we investigate total cholesterol level of an individual based on gender and race to better understand it probabilistically. Parametric analysis can perform well with skewed and non normal distributions when the data follow a particular distribution. Basic properties of distribution could reveal basic but crucial information about the cholesterol level we are interested in.

To characterize changes in cholesterol level over time and to investigate between and within subject variations, a longitudinal study appears to be useful. Linear mixed model approach is considered a popular procedure for analyzing repeated measures and clustered data. Its application is wide including biology, epidemiology, economics, etc. In the fourth chapter, we have investigated the total cholesterol level of individuals to better understand how it changes over time, as age, gender, and time are considered to be risk factors. Finally, we have also investigated average cholesterol level through a functional data analysis approach where data are transformed into a curve considered as basis functions. It provides more flexibility in modeling and smoothing the data. Most of this chapter has been accepted for publication as proceeding of Dynamic Systems and Application [4].

1.1 Generalized Linear Model (GLM)

The generalized linear model (GLM) is an extension of the linear model framework to variables that are not normally distributed. A generalization of linear models that connects the random distribution of the dependent variable (the distribution function) to the linear predictor-effect through a link function. It is a nice way of unifying various statistical models, like linear regression, logistic regression, and Poisson regression, etc, under a single framework. If the distributions of observations are or can be approximated with the normal distribution, the linear models are useful. Even if it is not the case, normal distribution is a safe assumption for large samples. However, a non-linear model should be used for many cases where linear models are not appropriate, such as when response is restricted to binary or count and variance of response depends on the mean. GLM extends the linear model framework to address both of these issues. There are three basic components of a generalized linear model. The components are:

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 systematic component (design matrix multiplied by parameter vector).

3. A smooth and invertible mathematical function, called the link function that links the systematic component to the random component.

1.2 Bayesian Model Selection Criteria

In this section we briefly introduce the deviance data criteria that we use in the model selection process.

1.2.1 Deviance Information Criteria

The deviance information criterion (DIC)[5] is a measure of model comparison and adequacy. It is given by

$$DIC(m) = 2\overline{D(\theta_m, m)} - D(\bar{\theta_m}, m) = D(\bar{\theta_m}, m) + 2p_m \tag{1.1}$$

where $D(\theta_m, m)$ is the usual deviance measure and is given by $D(\theta_m, m) = -2logf(y|\theta_m, m)$ and $\overline{D(\theta_m, m)}$ is its posterior mean. p_m can be taken as the number of effective parameters for the model m given by $p_m = \overline{D(\theta_m, m)} - D(\bar{\theta_m}, m)$. $\bar{\theta_m}$ is posterior mean of the parameters involved in model m. The smaller the DIC, the better fitting the model. The DIC is approximately equal to Akaike's information criterion (AIC)[6]. It is important to note that the DIC should not be used when the posterior distributions are highly skewed or bimodal since it assumes symmetry.

1.3 Markov Chain Monte Carlo Method (MCMC)

Evaluating the integral of the posterior probability distribution function is often not analytically tractable. Inverse cumulative distribution function, importance sampling, and

rejection sampling algorithm, etc, are some sample methods to address the issue. They refer mostly to unidimensional distributions and cannot be used to obtain samples from any posterior distribution of interest and the samples are independent. The techniques based on Markov chain overcome such problems because of flexibility and generality.

MCMC is basically Monte Carlo integration using Markov chains. It is required to integrate over possibly high-dimensional posterior probability distributions to make an inference about the parameters given the data. It samples from posterior distribution by Markov chain and computes the mean. MCMC methods are based on construction of Markov Chain (MC) that converges to the posterior distribution of the parameter of interest. The MCMC output is a dependent sample since in every step they produce values depending on the previous one as it is generated from Markov chain. The iterative procedure can be described in the following way:

A Markov Chain is a stochastic process $\{\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, ..., \alpha^{(T)}\}$ such that $f(\alpha^{(t+1)}|\alpha^{(t)}, \alpha^{(t-1)}, ..., \alpha^{(1)}) = f(\alpha^{(t+1)}|\alpha^{(t)})$. The distribution of α at sequence $t+1$ given all of its previous α values depends only on the value α^t of previous sequence t. Also, $f(\alpha^{(t+1)}|\alpha^{(t)})$ is independent on time t. As $t \to \infty$ the distribution of $\alpha^{(t)}$ converges to its equilibrium distribution and that is independent of the initial values of the chain $\alpha^{(0)}$. Construction of a Markov chain is essential to generate a sample from $f(\alpha|y)$ with two desired properties:

- $f(\alpha^{(t+1)}|\alpha^{(t)})$ should be easy to generate.
- Equilibrium distribution of the selected Markov chain must be the posterior distribution of interest $f(\alpha|y)$.

We then follow these steps after constructing Markov chain:

1. Select an initial value for $\alpha^{(0)}$
2. Generate samples until the equilibrium distribution is reached.

3. Monitor the convergence of the algorithm with convergence diagnostics. We generate more samples from the target distribution upon failure.

4. Cut off some initial "B" observations as burn in period.

5. Consider $\{\alpha^{(B+1)}, \alpha^{(B+2)}, ..., \alpha^{(T)}\}$ samples after the burn in period as the sample for the posterior distribution.

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

Gibbs Sampling and Metropolis Hasting are the two most popular MCMC methods used by researchers to estimate the corresponding posterior distributions with accuracy. We will discuss briefly the Gibbs sampling, the method we used in our study.

1.3.1 Gibbs Sampling Algorithm

The Gibbs sampler was introduced by Geman and Geman in 1984. It is a MCMC algorithm used to obtained a sequence of samples from the posterior distribution of the parameters when the direct sampling methods are difficult to perform. This is a special case of single-component Metropolis-Hasting algorithm using as proposal density, the full conditional posterior distribution. Gibbs sampling algorithm can be summarized in the following steps:

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

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

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

   (b) For $j = 1, 2, 3, ..., D$, update α_j from $\alpha_j \sim f(\alpha_j|\alpha_1, \alpha_2, ..., \alpha_{d-1}, \alpha_{d+1}, ...\alpha_D, y)$.

   (c) Set $\alpha^{(t)} = \alpha$ and save it as the generated set of values at $(t + 1)$ iteration of the algorithm.

   For a given particular state of chain $\alpha^{(t)}$, we generate the new parameters values by α_1^t from $f(\alpha_1|\alpha_2^{t-1}, \alpha_3^{t-1}, ..., \alpha_J^{t-1}, y)$

α_2^t from $f(\alpha_2|\alpha_1^{t-1}, \alpha_3^{t-1}, ..., \alpha_J^{t-1}, y)$

$\vdots$

α_j^t from $f(\alpha_j|\alpha_1^{t-1}, \alpha_3^{t-1}, ..., \alpha_{j-1}^{t-1}, \alpha_{j+1}^{t-1}, ..., \alpha_J^{t-1}, y)$

$\vdots$

α_J^t from $f(\alpha_J|\alpha_1^{t-1}, \alpha_2^{t-1}, ..., \alpha_{J-1}^{t-1}, y)$

Generating values from $f(\alpha_j|\alpha_1, \alpha_2, ..., \alpha_{j-1}, \alpha_{j+1}, ..., \alpha_J, y)$ is relatively easy as it is a univariate distribution and can be written as $f(\alpha_j|\alpha_1, \alpha_2, ..., \alpha_{d-1}, \alpha_{d+1}, ...\alpha_D, y) \propto f(\alpha|y)$, keeping the rest of variables constant except α_j.

1.4 Histogram Smoothing Prior

The statistical technique used to reduce or eliminate short-term irregularities and extract real trends and patterns from data results is smoothing. It reduces the variance and improves the accuracy of estimates. Different smoothing methods are applied depending on the type of data and purpose of the study, for instance, random walk, moving average, simple exponential, linear exponential, and seasonal exponential smoothing, etc. Many smoothing methods involve a series of observations $y_1, y_2, ..., y_n$ at equally spaced observations.

We have considered a histogram smoothing prior which is applicable to frequency plots, including those deriving from original data with equally spaced intervals. We assumed that continuous variable data are grouped into frequency counts with f_q being the observation between points t_{q-1} and t_q. It is thought that $g(x)$ possesses a continuous first derivative for all points on that interval to follow some similar property of smoothness. We treat the bins of histogram under the assumption that they are related in a certain manner. The limitation of histogram smoothing is that its estimate for $g(x)$ can be discontinuous at several points in an interval where it does not satisfy the smoothness property. Our concern is to overcome with independent observations whose common distribution is not limited to any particular family of distribution in Bayesian approach.

We have considered independent and identically distributed (i.i.d.) with unknown density which is concentrated on a finite interval of real line. Let π_q be the probability density that an observation is in the q^{th} interval and g(x) be an underlying density of smoothness:

$$\pi_q = \int_{t_q-1}^{t_q} g(x)dx$$

Since observations are mutually independent, the observed deaths (f_q) are assumed to be multinomial with parameters $\{\pi_1, \pi_2, ..., \pi_Q\}$ satisfying $\sum \pi_q = 1$, and index $n = \sum f_q$. We express π_q through a multiple logit model given by

$$\pi_q = \exp(\phi_q)/\sum_k \exp(\phi_k) \tag{1.2}$$

ϕ_k are considered to be multivariate normal distribution with means $\{m_1, m_2, ..., m_Q\}$ and $Q \times Q$ covariance matrix "V". We have assumed prior on $\pi_q's$: $\pi_q = p_q/\sum_k p_k$ or

$$m_q = \log(p_q/\sum_k p_k) \tag{1.3}$$

where p_q is population at risk at q^{th} interval. Covariance matrix (V) is assumed to be a first order dependence between neighboring frequencies on the histogram with common variance σ^2 and correlation ρ. Therefore,

$$V_{ij} = \rho^{|i-j|}\sigma^2 \tag{1.4}$$

where $(i = 1, 2, ..., Q; j = 1, 2, ..., Q)$, $\sigma > 0$, and $\rho > 0$. In detail, the elements of

covariance matrix V are given by:

$$V = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \cdots & \sigma^2\rho^{n-1} \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho^{n-2} \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n-1} & \sigma^2\rho^{n-2} & \sigma^2\rho^{n-3} & \cdots & \sigma^2 \end{bmatrix} \tag{1.5}$$

The elements of the matrix V take variance to be σ^2 and the correlation between adjacent bins of the histogram to be $\rho^{|i-j|}$. This correlation tends to zero as $|i-j|$ is large and to $|\rho|$ as $|i-j|$ tends to unity. We observe that $|i-j|$ a unity for consecutive bins for the histogram. Hence, for $i \neq$j and $\rho > 0$, bins are most closely related when $|i-j| = 1$, and become less closely related as $|i-j|$ gets larger than 1. These assumptions utilize prior beliefs about the smoothness of the histogram of the data. The elements of the matrix V^{-1} are given by

$$V^{-1} = \begin{bmatrix} \frac{1}{\sigma^2(1-\rho^2)} & -\frac{\rho}{\sigma^2(1-\rho^2)} & 0 & \cdots & 0 \\ -\frac{\rho}{\sigma^2(1-\rho^2)} & \frac{1+\rho^2}{\sigma^2(1-\rho^2)} & -\frac{\rho}{\sigma^2(1-\rho^2)} & \cdots & 0 \\ 0 & -\frac{\rho}{\sigma^2(1-\rho^2)} & \frac{1+\rho^2}{\sigma^2(1-\rho^2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sigma^2(1-\rho^2)} \end{bmatrix} \tag{1.6}$$

Chapter 2

Bayesian Age-Period-Cohort Model of Lung Cancer Mortality

2.1 Introduction

Lung cancer (LC) is the leading cause of cancer deaths in the US, where cancer remains the second most common cause of deaths [2]. In 2011, 14% of all cancer diagnoses and 27% of all cancer deaths were due to LC. More people in the United States die from LC than any other type of cancers which is true for both men and women. After increasing for decades, LC rates are decreasing nationally as fewer people smoke cigarettes [7]. However, it is still one of the biggest threats for public health. About 221,200 new cases and an estimated 158,040 deaths were reported in the American Cancer Society‘s estimates for LC in the United States for 2015. It is of strong public interest to study the trend, number of LC deaths, and the corresponding mortality rates for public health.

Age-period-cohort (APC) modeling has been a well-known issue in epidemiology [8–14]. The non-identifiability problem for parameters estimation has been drawing the attention of many researchers. It exists because of linear relationship of age, period, and cohort. To resolve this problem, several approaches have been suggested by researchers to analyze the trend in cancer epidemiology [12, 15–17]. This study applied Holford approach [12] to analyze the effects.

In this study, we analyze LC mortality of the USA based on age at death, period at death, and birth-cohort through Bayesian APC model with histogram smoothing priors. The Bayesian method extracts the necessary information from the data to describe the trend observed by exploring the uncertainty associated with functions of parameters. Var-

ious studies have been carried out through Bayesian APC analysis [18–20] with different smoothing priors. But, to the best of our knowledge, histogram smoothing prior has not been considered in Bayesian modeling. We have assumed that the densities of APC possess similar property of smoothness to adapt histogram smoothing prior. Our hypothesis was that the mortality rate decreases in 21st century. Hence, the reference period 1996-2000 was considered. The analysis has been executed with statistical packages R and WinBUGS.

2.2 Objective

The objective of our study is to estimate the time trend for the mortality or incidence of a particular disease for a large population. The analysis of trends in epidemiology is an important way of monitoring the behavior of diseases. It can be used to monitor etiology of disease and assess the effect of public health policies in the form of prevention, improve treatment, and cost assessment. The model can be used to forecast trends in health care needs and effect of public health proposals for disease control. Vital statistics provide epidemiologists with a useful first approach to understanding the significance of a disease, and they generally provide hints on etiology. The effect of age, period, and cohort can be analyzed through longitudinal trend. This trend provides crucial information about the status of a particular disease whether trend is increasing, decreasing, or stationary. These three things have different significant meaning to epidemiologists. They help to provide an insight into the dynamics of the spread of disease in the population.

In this study, we attempt to study mortality rate and incidence due to lung cancer in the USA based on the different age groups, period groups, and birth-cohort groups. It has been estimated the mortality rate due to lung cancer in the USA for different age groups. We have also estimated the relative risk due to period effect, and relative risk due to birth-cohort over time in longitudinal trend. Since it is rare events, observed mortality counts due to lung cancer have been assumed to follow Poisson probability distribution to apply to generalized linear model in Bayesian modeling. To address the uncertainty and infer-

ences with respect to parameters in Bayesian modeling, the priors are chosen from data itself. The best model has been chosen based on the Deviance Information Criterion (DIC). To overcome inability to address the independent observation whose common distribution is not limited to any particular distribution, we have attempted to implement histogram smoothing prior in Bayesian modeling and analysis.

2.3 Literature Review

Longitudinal trends of incidence and mortality rates of a particular disease generally provide important information for disease etiology to epidemiologists. Age (age at death), date at death (period), and date of birth (birth-cohort) are three time factors commonly considered for investigation to a disease. Different combinations of these three factors have been a great concern to investigators for the study. The effect of time in terms of longitudinal response could be an important implication from various perspectives. Age plays an important role in the risk of most of the diseases, so that it is included in almost any analysis of disease incidence and mortality. Date of birth or birth-cohort also can be another important temporal factor in disease incidence because a particular level of exposure may be clustered by generation. This factor can be identified by the date of enrollment in the study. All three temporal effects - age, period, and cohort - are thought to be useful by epidemiologists.

The literature is very rich with respect to estimation of parameter in age period cohort model. D. Clayton and E. Schifflers proposed the age cohort and age period model in 1987 while modeling for temporal variation in cancer rates I [8]. They analyzed Bladder cancer incidence in the region of Birmingham based on age cohort and age period model. While doing so they noticed that these factors together could provide epidemiologists more insight about any disease. Hence they proposed age period cohort model in 1987 [10] considering all three factors together to analyze breast cancer mortality in Japan with special reference to the phenomenon of Clemmesen's hook. D. Clayton in 1987 handled this situation through drift model approach. They faced un identifiability property for estimating

parameter for the model because of linear relationship among them. Theodore R. Holford, in 1983, suggested to address this issue through estimable function [21]. In 1992, he again came up with various approaches to address non identifiability property in parameter estimation. He has suggested estimable contrasts of parameter, design matrix for linear and curvature partition, interaction with non temporal variables, interaction with temporal variables, and alternative parameter constraints where he has suggested as zero slope, equate two or more effects, and restricted range for slopes [12]. Leonhard Knorr-Held in 2001 used a generalized Bayesian age-period- cohort (APC) model to a data set to obtain the trend [18]. Robert M. O'Brien in 2010 also has suggested different approaches with two fundamental problems associated with estimation of parameter [22]. Ramon Cleries (2010) applied the model through constraint parameter [20]. Irene O.L. et.al suggested through curvature approach [19].

We have adapted Holford approach to address un-identifiability problem by introducing a constraint while estimating parameters [12]. We have assumed the slope of period zero at first stage to allow independent estimation of parameters. Then, we introduced the fitted values to only period effect. The rationale to this approach is the following:

- An age effect has a strong biological association on cancer, is a well established result. Hence we can not ignore age effect.
- Cohort had a stronger association with incidence than period, has been shown in empirical results.
- Assuming slope of period is zero itself is less restrictive than ignoring the whole period effect.

2.4 Data Source

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is an authoritative source of information on cancer incidence and survival

in the United States [23]. It is dedicated to provide information on cancer statistics in order to minimize cancer burden among the U.S. population. SEER Program currently collects and publishes cancer incidence and survival data for approximately 30% of the US population. It includes 26 percent of African Americans, 38 percent of Hispanics, 44 percent of American Indians and Alaska Natives, 50 percent of Asians, and 67 percent of Hawaiian/Pacific Islanders. The SEER Program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data [24]. This program, which was established in 1973, collects data on cancer cases through 20 different registries from various locations and sources throughout the United States.

The SEER data are being used by many researchers, clinicians, public health officials, legislators, policymakers, community groups, and the public. It is updated annually and provided as a public service in print and electronic formats. The data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status are routinely collected by the SEER program. The SEER reports mortality data and it can be obtained from National Center of Health Statistics. The population data used in calculating the cancer rates is obtained periodically from Census Bureau [25]. The SEER based data has been improved through web-based access to the data and analytic tools and connected to other national data sources. A statistical software named 'SEERStat' is being used to extract the data from database.

2.5 Statistical Analysis

The data contains incidence and corresponding mortality due to lung cancer (LC) in the USA from 1971 to 2010, the most recent available data in the database. In our study, we have considered age at death due to lung cancer as "age", period at death due to lung cancer as "period", and date of birth of the person as "cohort". For statistical analysis, we have aggregated incidence and mortality into thirteen 5-year age groups (20-24 years to 80-84

years). Age groups are 22, 27, 32, ..., 72, 77, 82 years. There are eight 5-year periods (1971-1975 years to 2006-2010 years) and are 1973, 1978, ..., 2003, 2008.

We have considered these age groups (22-82 years) in our study because SEERStat does not give the counts for less than 10 numbers of observations and we have many such cases particularly below the age of 20 years and above 84 years for mortality counts. Also, cigarette smoking is the most common cause of LC and this habit is likely to develop in adult ages. Ages, periods, and cohorts are represented by their medians during our study.

These age groups and calendar periods involved 20 (13 age groups+8 periods -1) possibly overlapping 5-year cohorts [8, 10]. Total number of birth-cohort has been computed as follows: $K = I + J - 1$ (I =Age groups, J =period groups). A person died due to lung cancer in 1973 and he/she could have been 82 years old. So, he/she could be in birth-cohort 1891. Similarly, if a person died with lung cancer in 2008 and he/she was 22 years old, the possible birth-cohort could be 1986. Hence, if we divide time between 1891 to 2008 into 5-year group, we have ended up to 20 possible birth-cohorts given by 1891, 1896, 1901, ..., 1976, 1981, 1986.

Table 1: Number of Ages and Periods Used to Compute Possible Birth-Cohorts

	Periods							
Ages	1973	1978	1983	1988	1993	1998	2003	2008
22								
27								
.								
.								
.								
.								
82								

The following relation; $k = I + j - i$, is useful to calculate required ages, periods, and birth-cohorts. Where (k =Cohorts, I =Age groups, j =periods, i =ages). Ages, periods, and birth-cohorts calculated are used for statistical analysis. In this manner, we have computed and arranged the available data into five different columns named age, period,

Table 2: Calculated Ages, Periods, and Corresponding Cohorts

k	I	j	i
13	13	1	1
14	13	2	1
⋮	⋮	⋮	⋮
20	13	8	1
12	13	1	2
13	13	2	2
⋮	⋮	⋮	⋮
19	13	8	2
⋮	⋮	⋮	⋮
1	13	1	13
2	13	2	13
⋮	⋮	⋮	⋮
8	13	8	13

cohort, mortality and corresponding incidence due to lung cancer and we perform basic preliminary statistical analysis.

Age specific LC mortality rates seem stable for age groups 40 years and lower. Within every period, it can be noticed that mortality rate is increasing until age groups, 77 years and decreasing afterwards, Figure 1. Age specific mortality rates are lower for lower age in all birth cohorts. In similar fashion, it is significantly high for older birth cohorts. Older people of all birth-cohorts might be at high risk of ending up as lung cancer patients as shown by Figure 2. Period specific mortality rates for lung cancer are lower for lower age groups for every period. The greater the age, the more mortality rates seem to be for each period. The mortality rate for age group 82 years was lower than age groups 67, 72, and 77 in early periods. However, it seems to be greater in 1993 and recent periods, Figure 3.

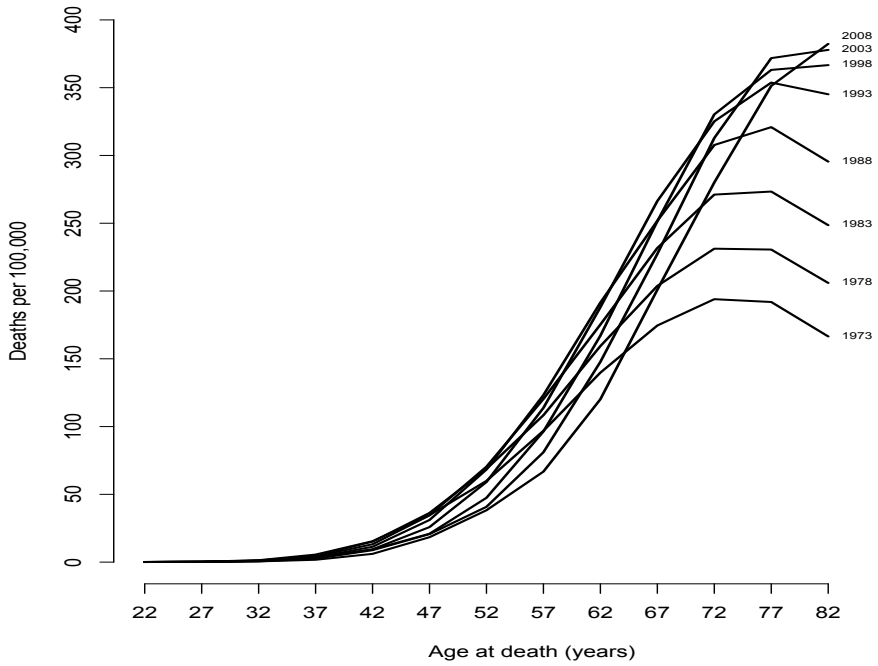


Figure 1.: Age-Specific Lung Cancer Mortality Rates per 100,000 in the USA by 5-Year Period

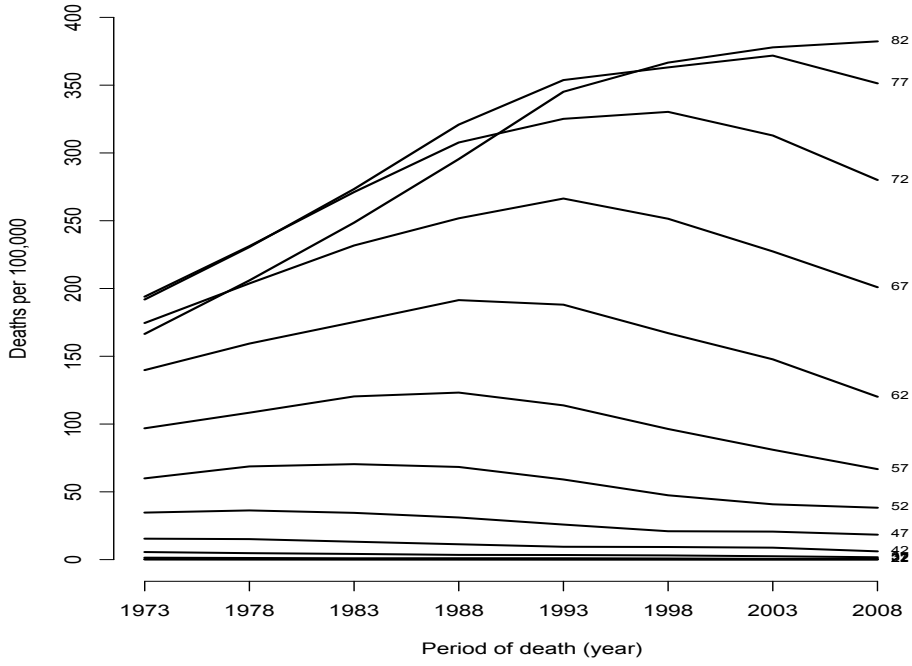


Figure 2.: Period-Specific Lung Cancer Mortality Rates per 100,000 in the USA by 5-Year Age Group

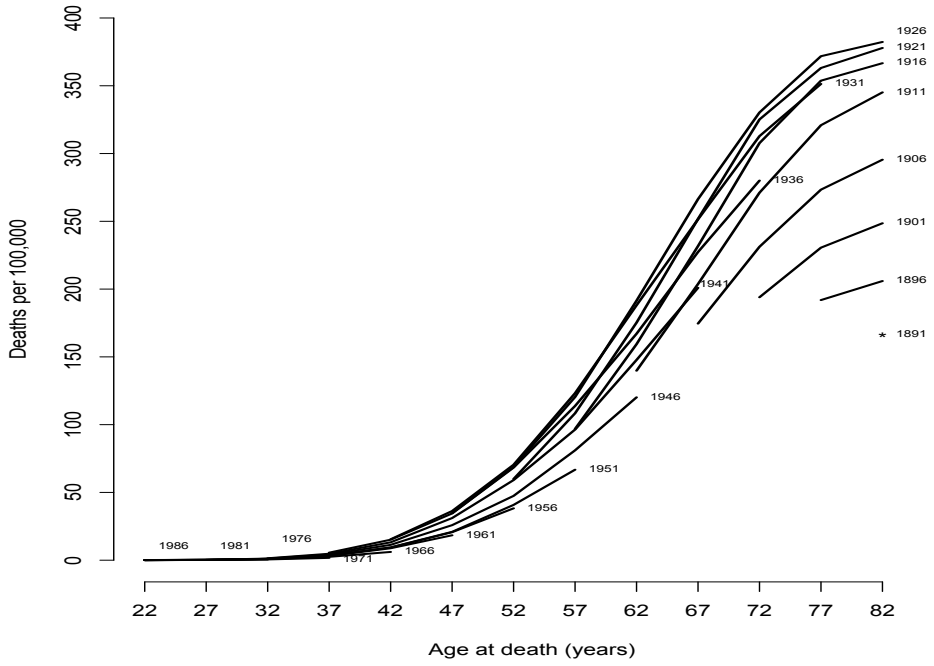


Figure 3.: Age-Specific Lung Cancer Mortality Rates per 100,000 in the USA by 5-Year Birth-Cohort

Cohort specific mortality rates due to lung cancer are higher for age groups in early birth-cohorts. However, it seems to be decreasing for lower age groups for recent birth-cohorts. We have noticed that older people are at more risk of dying due to lung cancer than younger people throughout the birth-cohorts. Within the same age group, the mortality rate of younger birth-cohort is relatively lower than older birth-cohort, as shown by Figure 4 [21].

We have plotted mortality rates as an initial exploration whether rates are proportional between periods or cohorts. These plots might help to have preliminary idea that model possibly includes as contributing factors. Log scale rate plots of Figure 1 and Figure 2 will exhibit almost parallel lines if age specific rates are proportional between periods which might indicate age-period model. Similarly, log scale rate plots of Figure 3 and Figure 4 will exhibit parallel lines if age specific rates are proportional between cohorts, which

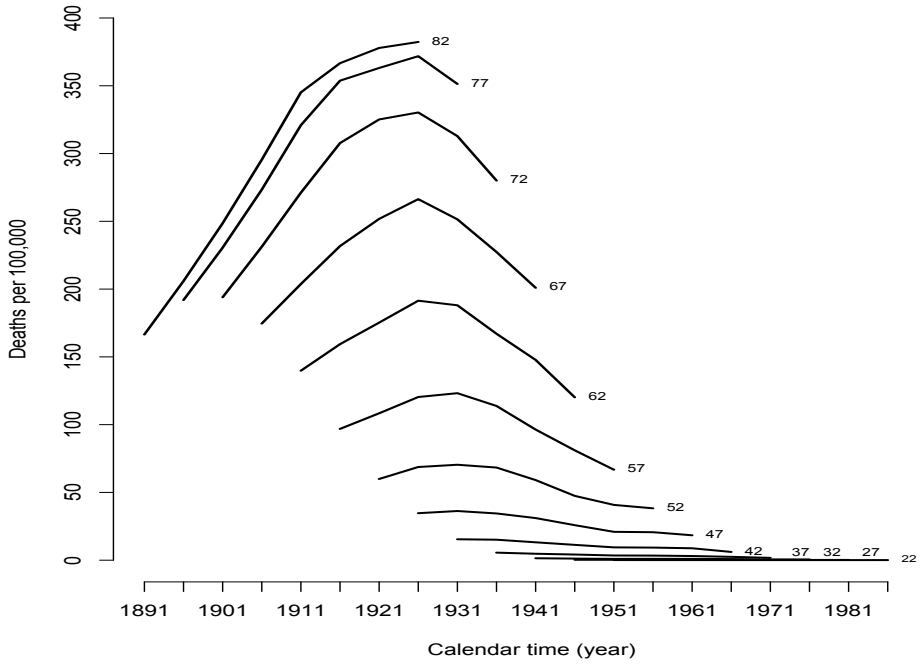


Figure 4.: Cohort-Specific Lung Cancer Mortality Rates per 100,000 in the USA by 5-Year Age Group

possibly indicates age-cohort model [26].

2.6 Modeling Lung Cancer Mortality

d_{ij} and n_{ij} are observed number of deaths for age group i and period j and total number of person-year at risk respectively. It has been assumed that $d_{ij} \sim Poisson(\lambda_{ij} n_{ij})$

$$log(\lambda_{ij}) = log(n_{ij}) + \alpha_i + \beta_j + \gamma_k \tag{2.1}$$

α_i is age effect $(i = 1, 2, \ldots, I)$, β_j is period effect $(j = 1, 2, \ldots, J)$, and γ_k is cohort effect $(k = 1, 2, \ldots, K)$ where $K = I + J - 1$ and $k = I + j - i$.

The choice of prior is always an important issue in Bayesian Statistical Analysis. Normality is a common assumption for time effect-specific (age, period, and cohort) smoothing prior formulations in most of the smoothing approaches including power link model [27].

Our assumptions regarding age, period, and cohort are not restricted to any particular family of distributions. Histogram smoothing is a technique for the analysis of independent and identically distributed (i.i.d.) observations with unknown density which is concentrated on a finite interval of the real line [28]. A histogram helps to visualize the data since it adapts replacing a large point set with a compact approximation of the underlying distribution. It eliminates the random fluctuation that usually occurs with estimate of parameters and prevents instability in the situation where there are very few counts of deaths as in the younger age groups in our study. The variance parameters (age, period, and cohort variances) provide information about degree of smoothness. The larger the values, the greater the degree of smoothing [29]. The trends corresponding to age, period, and cohort were smoothed using histogram smoothing prior. That being said, it is crucial to notice that the model did not appear to be sensitive to the prior of variance (roughness) parameters.

The relation, $cohort = period - age$, leads to a non-identifiability problem where a constraint should be introduced [21, 30]. We have adopted Holford approach to represent the effects [11, 12]. It is observed that age-cohort model with an unstructured error term is enough to describe the extra Poisson variation [31]. Hence, we estimated age and cohort effects assuming slope of period is zero considering 11th (1941) cohort as reference. Then, the fitted values are introduced to a model that includes only period effects considering 6th (1998) period as reference. In this way, we can obtain independent effects of age, period, and cohort. A similar approach with respect to choice of effect has been previously adopted [19, 20]. The parameter estimates for the model are obtained from posterior distribution. Median is the point estimate. The model goodness-of-fit was measured by the posterior mean deviance $\bar{D}$ [32]. The deviance information criterion (DIC) has been considered to compare the models which adjusts the posterior mean deviance for the number of parameters in the model [32]. A smaller DIC indicates the better fit. 95% credible intervals were obtained using 2.5th and 97.5th percentiles of the Monte Carlo Markov Chain run.

We fitted partial as well as full models and compared the different models based on DICs

obtained. We investigated age, period, and birth-cohort models with their different possible interactions. However, we observed no evidence of significance interactions as indicated by DICs. The full model, which contains age at death, period at death, and birth-cohort, shows best fit with lowest DIC in comparison of partial models Table 3.

Table 3: DIC Values for Different Combinations for Age, Period, and Cohort Models for Deaths Due to Lung Cancer in the USA

Model	DIC
age	165722
period	1,923,440,000
cohort	1,041,650,000
age, period	116,025
period, cohort	233,250,000
period, cohort, period*cohort	850,328,000
age, period, age*period	120,595
age, period, cohort, age*period	11924
age, cohort, age*cohort	8,021.3
age, cohort	7171.17
age, period, cohort	2119.65

2.7 Result

We observed that it is stable around 1 or 2 deaths per 100,000 for age groups 32 and below and is around 5 deaths for age group 37, Figure 5, on average annually. However, it had an upward inflection with age groups 42 and above. It is consistent with higher age which is also observed in preliminary analysis as shown in Figure 1. We have observed that it reaches the peak of 325 (95% CI: 323-326) deaths annually per 100,000 for the age group 82 years. The 95% point wise credible intervals are the intervals of the mean function which looks narrower in our model. This might be due to the scale of model which is substantially broader and has greater variability.

We have witnessed a continuous increase in the relative risk in period effects peaking in

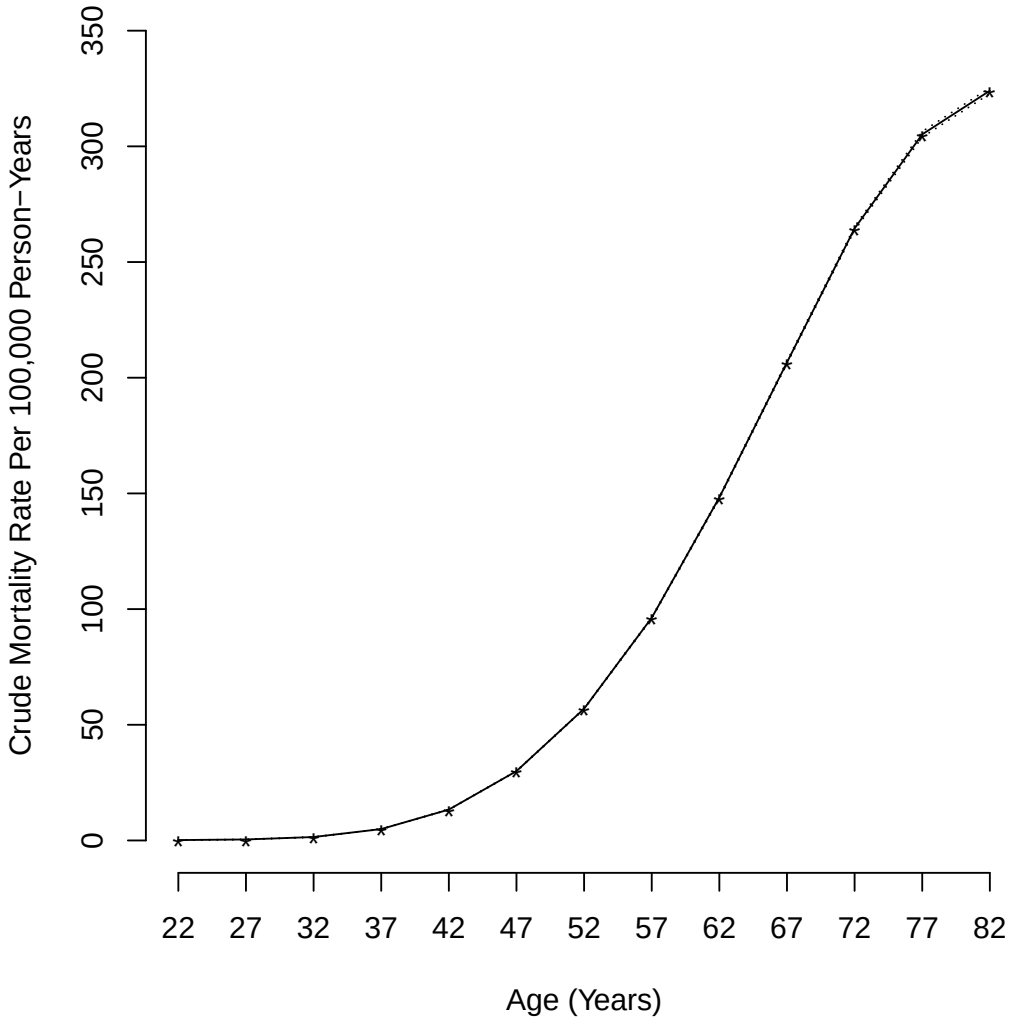


Figure 5.: Estimated Age-Specific Annual Mortality Rates (Age Effects) in 5-Year Age Group

period 1993 and then downward inflection afterwards, Figure 6, consistent with Figure 2.

It indicates that risk of LC mortality rate is improving after this calendar period. The result is well underlined by research [33]. We have noticed that risk of mortality rate decreased most rapidly in periods 2003-2008, which has been reinforced in similar research [1, 33]. It should be noticed that risk of birth-cohort increased sharply among cohorts from the late 1800s into the early 1900s before reaching a plateau and then declining, Figure 7.

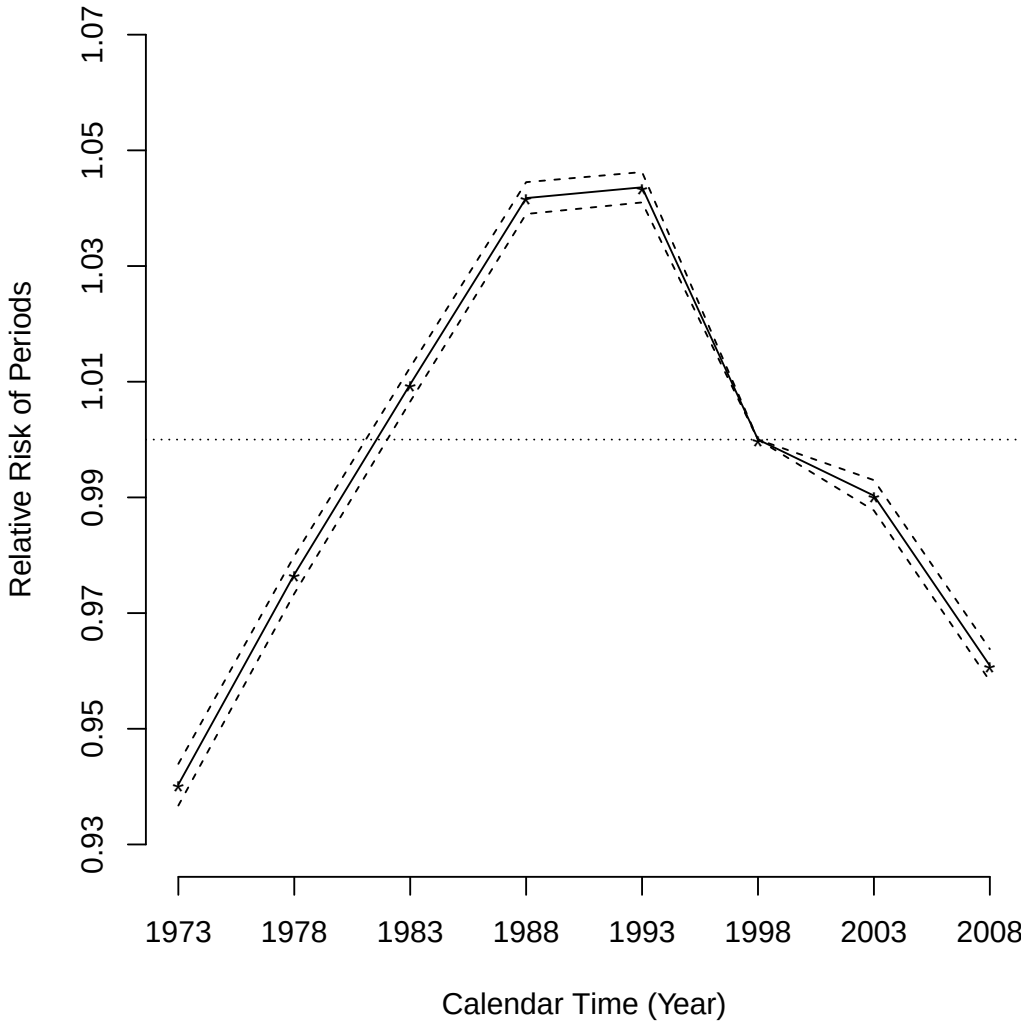


Figure 6.: Estimated Relative Risks for 5-Year Calendar Periods (Period Effects) with 95% Credible Interval with Respect to Reference Period 1998

The curvature of cohort effect depicts an increase in peak risk at birth cohort 1926 and declines continuously until 1950 as it was observed in Figure 4. The risk of LC mortality peaked in birth-cohort 1926-1931 that is supported by similar research [33–35]. We have observed an increase in birth-cohort slope in 1950, indicating a deteriorating of birth-cohort trend in LC mortality after 1950. The results have strongly agreed in similar research [33]. It can be seen in a slowing of decline in risk after cohort 1950 in almost all age intervals,

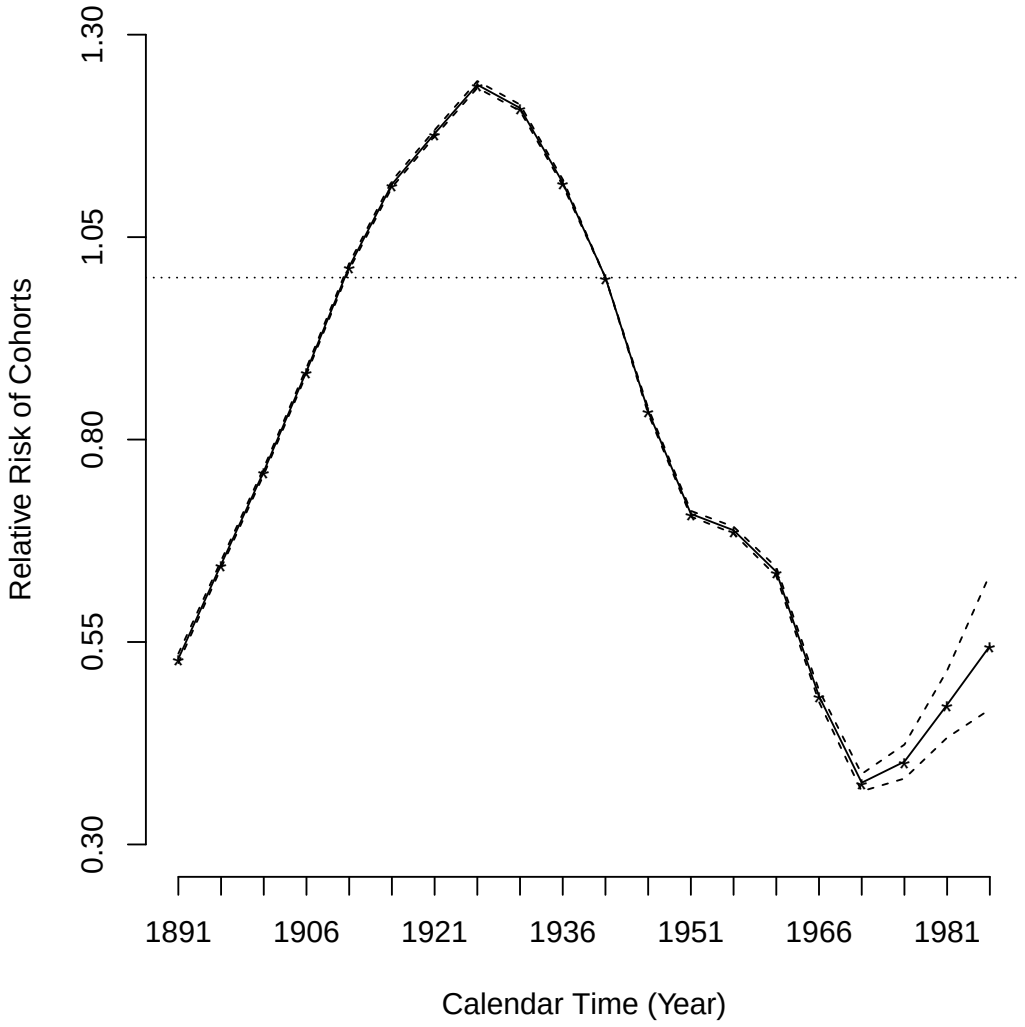


Figure 7.: Estimated Relative Risks for 5-Year Birth Cohorts (Cohort Effects) with 95% Credible Interval with Respect to Reference Cohort 1941

which indicates that the worsening birth-cohort risk is not an artifact of the model fitting, Figure 4. We have noticed declining birth-cohort risk after 1960. The relative risk due to LC mortality by birth-cohort reflects upward inflection for people born around 1975 which is consistent with mortality rates by cohort described in Figure 3. Wider 95% credible interval was observed in the latest cohorts because of fewer LC death points leading to greater uncertainty. The estimated parameter values of age, relative risks of period, and cohort components with their 95% credible intervals are described in Table 4.

Table 4: The Posterior Summaries of Parameters for Lung Cancer

node	mean	sd	MC error	2.50%	median	97.50%
α_1	-13.32	0.02952	2.53E-04	-13.38	-13.32	-13.27
α_2	-12.36	0.01746	1.59E-04	-12.4	-12.36	-12.33
α_3	-11.12	0.009049	9.68E-05	-11.14	-11.12	-11.11
α_4	-9.92	0.005063	8.18E-05	-9.93	-9.92	-9.911
α_5	-8.922	0.003286	7.88E-05	-8.928	-8.922	-8.915
α_6	-8.113	0.002531	8.22E-05	-8.118	-8.113	-8.108
α_7	-7.471	0.002214	8.72E-05	-7.475	-7.471	-7.468
α_8	-6.946	0.002105	9.17E-05	-6.95	-6.946	-6.943
α_9	-6.514	0.002113	9.72E-05	-6.518	-6.514	-6.511
α_{10}	-6.182	0.00214	1.02E-04	-6.185	-6.182	-6.179
α_{11}	-5.935	0.002463	1.20E-04	-5.938	-5.935	-5.932
α_{12}	-5.792	0.00257	1.25E-04	-5.796	-5.792	-5.788
α_{13}	-5.732	0.002731	1.29E-04	-5.736	-5.732	-5.728
β_1	-0.06153	0.001943	2.95E-05	-0.06532	-0.06151	-0.05774
β_2	-0.02363	0.001686	2.30E-05	-0.02695	-0.02362	-0.02035
β_3	0.009501	0.001491	1.58E-05	0.006575	0.009505	0.01241
β_4	0.04091	0.001349	9.89E-06	0.03826	0.04092	0.04355
β_5	0.04269	0.001285	8.66E-06	0.04018	0.04269	0.04521
β_6	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
β_7	-0.00972	0.001353	2.39E-05	-0.01232	-0.00973	-0.007082
β_8	-0.03982	0.001489	3.46E-05	-0.04259	-0.03985	-0.03696
γ_1	-0.6358	0.005803	1.50E-04	-0.6465	-0.6359	-0.6247
γ_2	-0.4373	0.003925	1.41E-04	-0.4439	-0.4374	-0.4303
γ_3	-0.2734	0.003309	1.38E-04	-0.2786	-0.2734	-0.268
γ_4	-0.1228	0.002905	1.29E-04	-0.1271	-0.1229	-0.1182
γ_5	0.01356	0.002659	1.22E-04	0.009697	0.01341	0.01745
γ_6	0.1092	0.002518	1.19E-04	0.1056	0.109	0.1129
γ_7	0.1639	0.002406	1.15E-04	0.1605	0.1638	0.1674
γ_8	0.2139	0.002306	1.10E-04	0.2106	0.2137	0.2172
γ_9	0.1904	0.002225	1.03E-04	0.1873	0.1903	0.1936
γ_{10}	0.111	0.002207	9.68E-05	0.1077	0.1109	0.1144
γ_{11}	0.0000	0.00000	0.00000	0.0000	0.0000	0.0000
γ_{12}	-0.1799	0.002305	8.37E-05	-0.1838	-0.1799	-0.1758
γ_{13}	-0.3449	0.002637	8.19E-05	-0.3496	-0.345	-0.340
γ_{14}	-0.3739	0.003178	7.87E-05	-0.3798	-0.374	-0.3678
γ_{15}	-0.4515	0.004365	7.68E-05	-0.4599	-0.4515	-0.4429
γ_{16}	-0.7276	0.007546	8.76E-05	-0.7425	-0.7276	-0.7127
γ_{17}	-0.9779	0.01449	1.36E-04	-1.006	-0.9778	-0.9495
γ_{18}	-0.911	0.02636	2.25E-04	-0.9635	-0.9107	-0.8602
γ_{19}	-0.7505	0.04461	3.55E-04	-0.8389	-0.7499	-0.6639
γ_{20}	-0.6082	0.07892	6.07E-04	-0.7648	-0.6073	-0.4561

Box plots 8, 9, and 10 have slightly different box plots from the traditional box plots. The middle bar represents the posterior mean of the parameter, while limits of each box represent posterior quartiles. The ending whisker lines represent the 2.5% and 97.5% posterior percentiles respectively. They provide Bayesian confidence interval called credible interval which give interval estimation of parameter of interest.

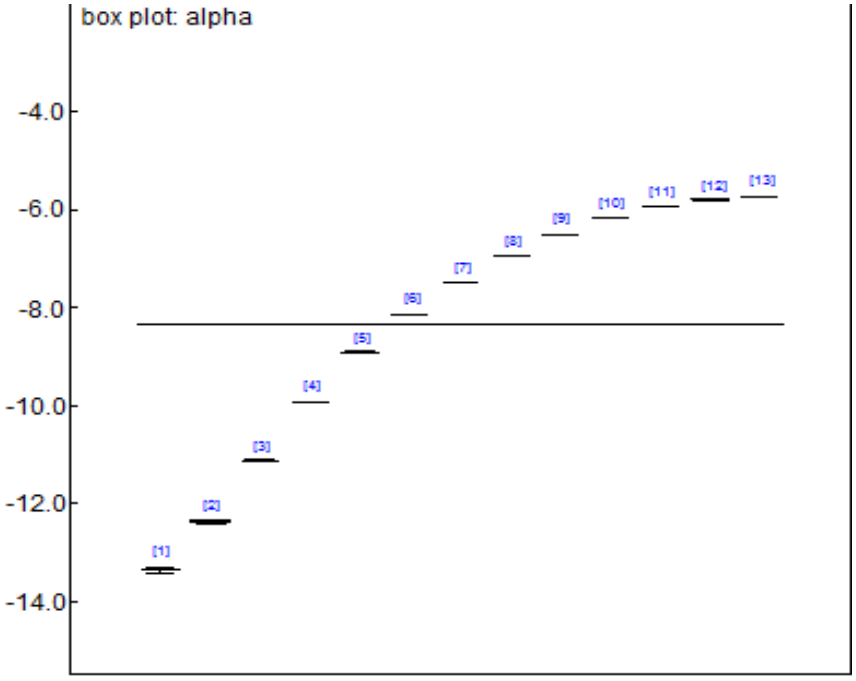


Figure 8.: Box Plot of Estimated Parameters for Age Group

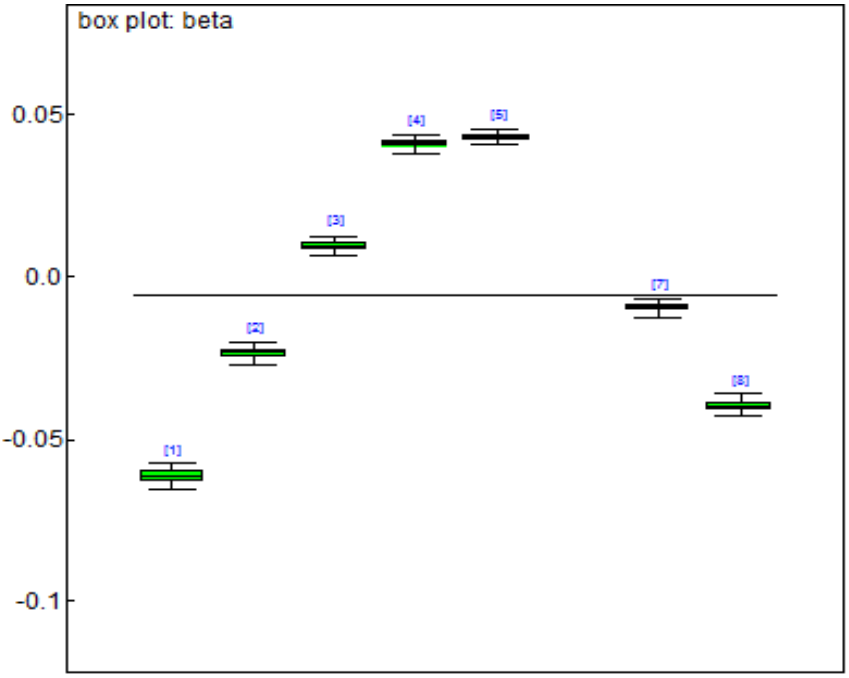


Figure 9.: Box Plot of Estimated Parameters for Relative Risk of Period

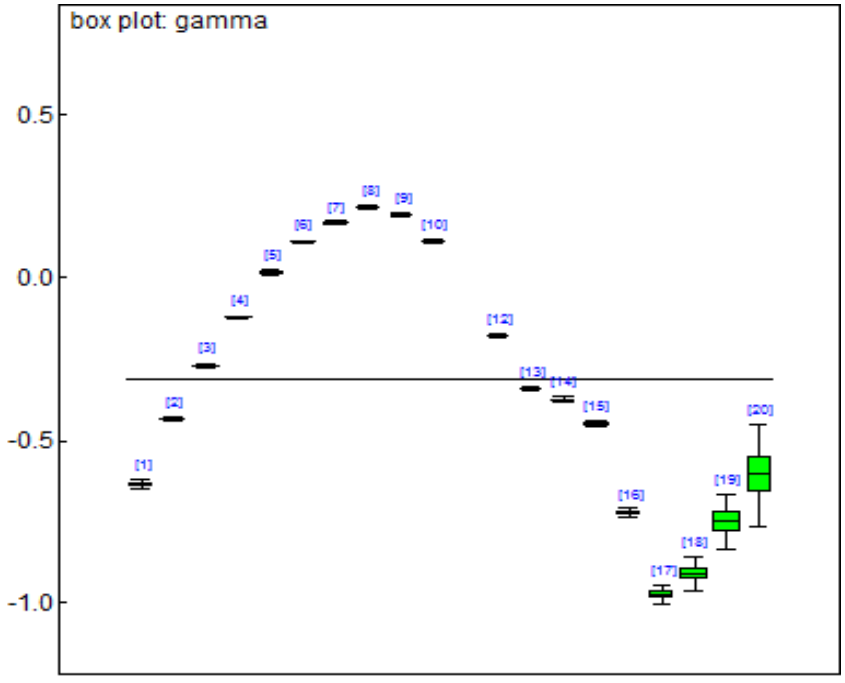


Figure 10.: Box Plot of Estimated Parameters for Relative Risk of Cohort

2.8 Model Validation

The inferences of a Bayesian analysis are conditional on the appropriateness of an assumed probability model. It is essential to be satisfied that our assumptions are a reasonable

approximation to reality, though it is not generally believed any model is actually true. Various aspects of an assumed model might be questioned: observations that do not fit, the distributional assumptions, link functions, which covariates to include, and so on.

We have analyzed the residual to check the robustness and fit of the model. To examine the standard errors with their 95% credible intervals for fitting of each observation and the identification of possible outliers, we have considered posterior simulation. The standardized residuals are obtained to measure the deviation between observations and estimated expected values, and should ideally be assessed on data that has not been used to fit the model. It is obtained by taking deviations of the data to their expectations for all observations based on posterior simulations and dividing by their standard deviations. We have used Pearson residual given by:

$$residual_i(\theta) = \frac{y_i - E(y_i|\theta)}{\sqrt{var(y_i|\theta)}} \tag{2.2}$$

It is a function of θ (parameter) and therefore has a posterior distribution. If it is considered as a function of random y_i for fixed θ, it has mean 0 and variance 1 and so we might broadly expect values between -2 and 2 for adequate model. Standardized errors with 95% credible intervals are given in Figure 11.

Almost all of the standardized residuals with their bands are randomly distributed around zero within the range of -2 to 2, a clear indication of a well fitted developed model to the data.

2.9 Conclusion

We observed that there is strong evidence of significant changes in risk from LC by birth cohorts. Initial increasing risk of mortality trend is observed in the early 19th century. There could be many possible etiologic factors like increased air pollution by gases and

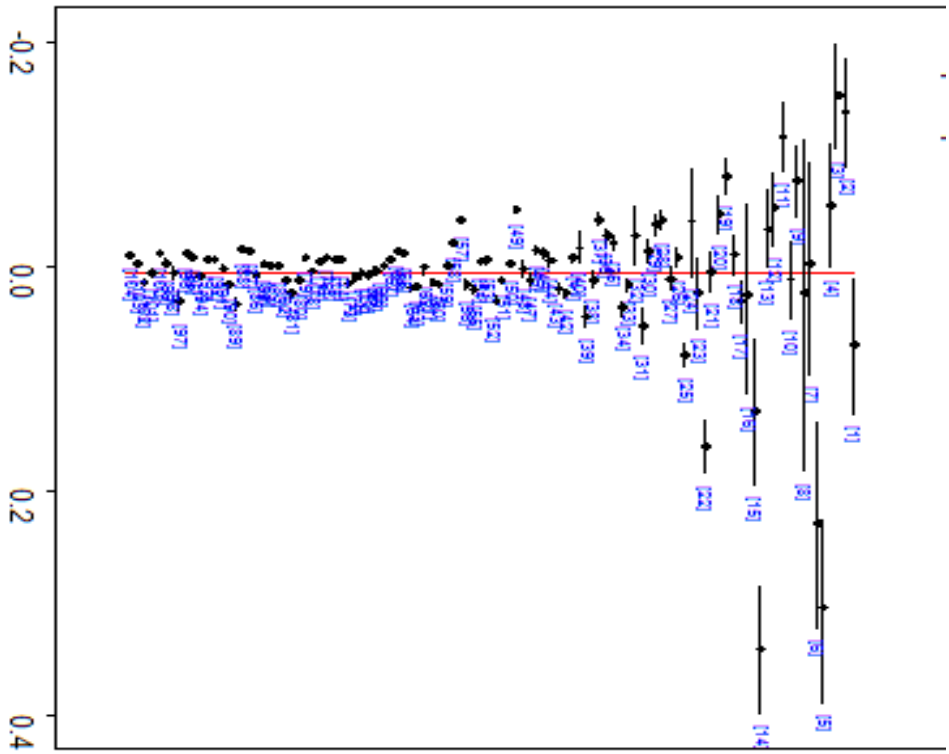


Figure 11.: Standardized Residual Plot

dusts caused by industries, the asphalting of roads, the increase in automobile traffic, exposure to gas in World War I, the influenza pandemic of 1918, working with benzene or gasoline [36, 37]. Although LC can be caused by environmental exposures as well, 80-90% of LCs are attributed to cigarette smoking and secondhand smoke [1]. Tobacco use has been identified as the most contributing risk factor for LC in developed countries and is approaching to be first in developing countries [38]. The peak of birth-cohort risk occurred in 1926, which may be caused by the earlier use of cigarettes [35, 39, 40]. These cohorts had the highest prevalence of cigarette smoking during World War II [40]. Over half of the young population used to smoke cigarettes. Mortality rate is observed to be higher for age groups 60 years and above. People born around 1926 had reached 60 years and above after 1980, where the higher risk has been observed in period effect. The risk of LC grew with increasing age which might be due to the lifetime number of cigarettes increasing with age as the risk can be interpreted as a reflection of past smoking [41–43].

There has been no significant breakthrough in lung cancer treatment that explains the decrease in mortality rates after calendar period 1993 [44]. However, we have witnessed the decrease in slopes of incidence rate curves after 1993. It might suggest that the decrease in mortality rates is because of a decrease in risk of LC instead of an improvement in survival. The impact on initiation stage of decrease in tobacco carcinogen exposure and increase in smoking cessation beginning around 1960 might have caused the substantial decline in calendar-period risk after 1993. Because of lower risk in mortality from period 1993, decreasing slope of risk to birth-cohort 1961-1971 might have been observed. A reduction in risk of death in the USA due to LC is observed through periods 1991-2010 from period-specific trend. Older people are relatively at more risk than younger ones. A similar conclusion has been discussed in the study [2].

A decreasing mortality trend was observed during cohorts 1926-1951 due to prevalence of smoking filter cigarettes and low-tar cigarettes [35]. The risk of death due to LC was slightly increasing in 1951-1961 possibly because of the promotion of deeper inhalation of smoke [37]. It may reflect a failure of widespread tobacco control efforts by private and public health agencies in the 1960s [38] to break through social and cultural aspects which influenced teenage-smoking [45, 46]. Marijuana contains the same carcinogen as is found in cigarettes [47]. It is possible that increased smoking of marijuana by teenagers and young adults in the 1960s and in the 1970s contributed to an increase in risk of birth-cohort around 1950.

Since 1964 when the first Surgeon General's report on the health consequences of smoking was published, cigarette smoking cessation rates increased and cigarette smoking initiation rates decreased more rapidly among men than women [37]. The increased use of cigarettes and marijuana since 1991 among teenagers might likely be reflected by an increase in birth-cohort risk for people born around 1975. Increased smoking from 1971 possibly increased the relative risk of death from LC in birth-cohort around 1980 and contemporary cohorts [37]. People born during 1880-90 had a relatively high prevalence of

cigarette smoking with mixed tobacco and tar that could have led to an increase in risk. However, insufficient data earlier than period 1971 prevented precise estimation of risk trend among older periods.

Our study is concentrated in well-defined population of the USA. The proposed model fits mortality data well in general. Hence, it is reasonable to argue that our approach extracts the necessary information from the data to explain the possible trend. The Bayesian approach can be adapted to incidence and mortality data. Of course, it can provide further information on potential benefits of analyzing LC mortality.

2.9.1 Contribution

In this chapter, we have studied lung cancer mortality rate in the USA through Bayesian analysis with histogram smoothing prior suggested by T. Leonard [28]. It is noted that we have discovered several advantages using Bayesian approach and with histogram smoothing prior. We would like to list some of the important contributions as a result of the current study.

1. We have implemented histogram smoothing prior in Bayesian modeling to model mortality data and can be implemented to identify mortality rate and corresponding incidence in different cancer and other related fields.

2. We have implemented Bayesian approach in addressing independent observations whose distribution is not limited to any particular family of probability distributions. We have assumed independently and identically distributed (i.i.d.) observations with unknown mortality density which is concentrated on a finite interval of real line.

3. The developed statistical model includes counts of death for each age group, period, and birth-cohort and estimates the effect of LC as function of time.

4. The developed statistical model can be easily extended for further improvement for estimation of parameters with proper approach of breaking the component of precision

matrix.

5. We have identified mortality rates based on different age groups, relative risks due to period effect and relative risk due to birth-cohort effect in lung cancer in the USA.

6. The proposed statistical model can be used to predict future age specific mortality rates due to lung cancer and relative risks due to period and birth-cohort.

Chapter 3

A Parametric Analysis of Serum Cholesterol Levels by Gender and Race

3.1 Introduction

Cigarette smoking, high blood pressure, and high blood cholesterol are the most clearly established risk factors that have been identified as being strongly associated with coronary heart disease (CHD) [48]. Total serum cholesterol level (SCL) is a major risk factor for CHD which is the leading cause of death in the United States [49–51]. CHD is responsible for more deaths than all forms of cancer combined [48]. A recommendation has been made that total SCL for adults should be below 200mg/dl and individuals with values between 200mg/dl to 239 mg/dl should be considered as borderline high risk; those with values more than 240 mg/dl should be regarded as high risk for CHD [48–52]. Hence, a detailed study of SCL is essential for public health.

A better understanding of lipoprotein production and removal, lipoprotein receptors, and apolipoprotein is needed because they are considered the most important factors in cholesterol. Various studies have been carefully executed to reduce SCL through different means through diets and drugs [53, 54]. Significant positive changes have been achieved through dietary means as well as through drugs in reducing cholesterol levels in test subjects. Relevant of effects of drugs on SCL can be carefully handled through drugs that have previously been recognized. Studies have been carried out in order to address these issues as well [55–57]. Attempts have been made to make prediction about the SCL based on age in order to better understand the relationship between age and cholesterol levels [58–60].

Precisely defined diets and pharmacologic interventions to reduce blood cholesterol and

other lipids are presently being studied in individuals under carefully controlled conditions to investigate effects of drugs on SCL [61]. The compounds that are more effective, economical, and safe for people in the reduction of blood cholesterol are under intensive research. Longevity of life in the elderly population who have a high-density lipoprotein cholesterol has also been investigated in order to better understand its effects on survival.

That being said, there are very few studies that have been carried out addressing SCL, based on gender and race. SCL is heavily dependent on two factors: it is strongly influenced by food intake of an individual and it varies by race. In addition, the resistance to disease capabilities varies by gender and race as well. It is therefore equally important for the study of SCL to take both factors into consideration.

In this chapter, we performed parametric analysis of the SCL and statistically discuss the behavior based on gender and race. We identified the probability distributions that best describe the cholesterol level for different genders and races. Such a characterization will be crucial in obtaining central tendency, dispersion, skewness, and kurtosis of distributions. Although it would be helpful to speculate on the effect of cholesterol level on race and gender, our study will be able to provide further insight through parametric analysis into its nature and will assist in exploring SCL's various aspects. Succinctly stated, the purpose of this study is to enable researchers to identify subgroups of the population who are at risk with respect to SCL and to identify distributional differences among the population subgroups of epidemiological interest.

3.2 Objective

The purpose is to study, parametrically, the cholesterol level of individuals based on gender and race and to determine the probability distributions that the cholesterol level resembles. The study concludes that the cholesterol level for different genders and races exhibits significant racial and gender differences in terms of probability distributions. Vital statistics obtained from identified probability distributions could be useful information. We intend

to obtain skewness and kurtosis based on gender and ethnicity of average cholesterol level. A recommendation of any kind of means for an individual, such as diets or drugs, might differ from individual to individual, even though the individuals have the same cholesterol level. Our study will further widen inside the behavior of cholesterol on the individuals. It should also be taken into consideration that high level risk and borderline risk levels of cholesterol might also be different for races and genders.

3.3 Data Source

The data utilized in this paper were made available by the inter-university Consortium for Political and Social Research and the data for National Health and Nutrition Examination Survey (NHANS) II, 1976-1980: Serum Cholesterol was originally collected by United States Department of Health and Human Services.

NHANS II was conducted on a nationwide probability sample with approximately 28,000 persons. The target population for the survey was the civilian non-institutionalized population of the United States (including Alaska and Hawaii). The NHANES II serum cholesterol data files contain two parts of the extensive data available. One part consists of the demographic information obtained from household interview and the other part is laboratory results. The survey started in February 1976 and was completed in February 1980. Samples were selected so that certain population groups thought to be at high risk of malnutrition (person with low incomes, preschool children, and the elderly) were oversampled. Adjusted sampling weights were then conducted for persons over the age of 76, sex, and race categories in order to inflate the sample in such a manner as to closely reflect the estimated civilian non-institutionalized U.S. population.

In addition to the general examination components, several more detailed examinations were performed on subsamples of the population. Our study included 11,864 persons for SCL cases with 9,602 males and 2,262 females. The information relating to SCL in NHANS II survey considered codes 355-357. Primary site codes were 1 and 2 for males

and females respectively. Similarly, codes for white, black, and other were 1, 2, and 3 respectively. Male data included 8,536 white, 881 black, and 185 other individuals. On the other hand, female data included 1,769 white, 456 black, and 37 other individuals. Descriptive information for the total number of individuals by sex and race is presented in schematic diagram, Figure (12) which explains how the study was systematically organized.

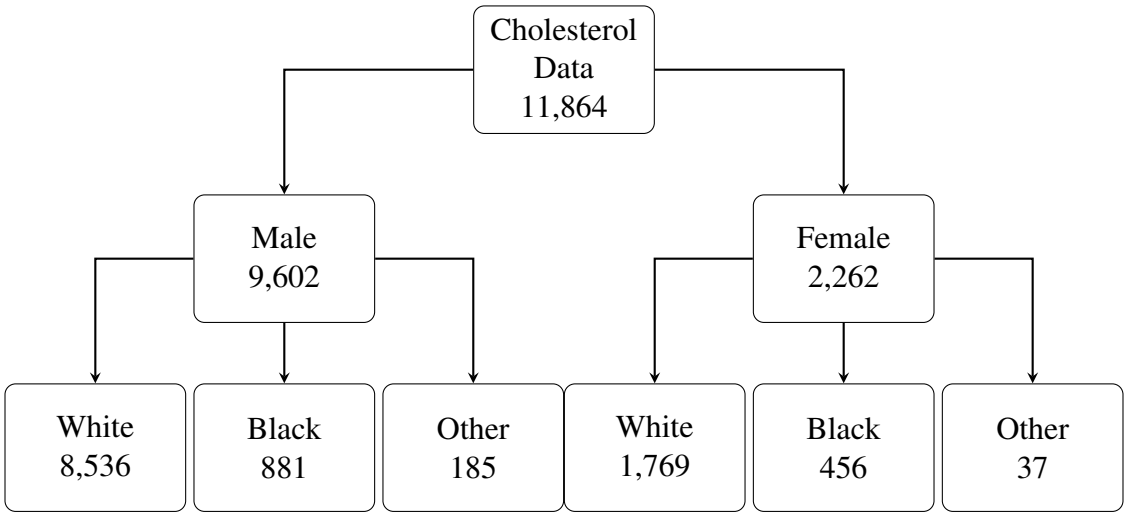


Figure 12.: Schematic Diagram of Total Number of Individuals by Sex and Race

3.4 Statistical Analysis

In order to perform parametric analysis of cholesterol levels, the probability distributions that best fit the cholesterol data are lognormal and gamma distribution. The probability distributions were confirmed, taking into consideration all available data. Kolmogorov Smirnov goodness of fit test was performed to identify the underlying distributions.

The probability density function of three-parameter lognormal distribution is given by:

$$h(x) = \begin{cases} \frac{1}{(x-\gamma)\sigma\sqrt{2\pi}} \exp(-\frac{[ln(x-\gamma)-\mu]^2}{2\sigma^2}), & \gamma < x < \infty, \sigma > 0, -\infty < \mu < \infty \\ 0, & \text{elsewhere} \end{cases} \tag{3.1}$$

The probability density function of two-parameter lognormal distribution is given by:

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{[ln(x)-\mu]^2}{2\sigma^2}), & x > 0, \sigma > 0, -\infty < \mu < \infty \\ 0, & \text{elsewhere} \end{cases} \tag{3.2}$$

The probability density of Gamma distribution is given by:

$$g(x) = \begin{cases} \frac{1}{\Gamma\gamma\beta^\gamma} x^{\gamma-1} \exp(-\frac{x}{\beta}), & x, \gamma, \beta > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{3.3}$$

$$\Gamma\gamma = \int_0^\infty x^{\gamma-1} \exp(-x) dx, \gamma > 0 \tag{3.4}$$

where $\Gamma\gamma$ is gamma function. With respect to the lognormal distributions, μ and σ respectively are mean (scale parameter), standard deviation (shape parameter), and γ a location parameter shown in equation (3.1) and (3.2), which characterize the cholesterol levels of most of the male cases. Gamma probability distribution represents cholesterol levels of mostly female cases, which is shown in equation (3.3).

3.5 Result

In this section, cholesterol level by sex and race was investigated. Distinctively different cholesterol levels may have important implications. Being aware of the complexity of the data, all pairwise comparisons were performed non-parametrically, not relying on any particular distribution. The mean cholesterol levels of males and females for overall behaviors were investigated. The study performed the Mann-Whitney-Wilcoxon non parametric test in order to identify whether or not the males and females were two identical population, with respect to cholesterol levels.

We were able to reject the null hypothesis that male and female cholesterol levels were coming from two identical populations (p-value=4.8e-09). It was therefore, crucial to investigate male and female, independently. We also examined the equality of mean cholesterol levels for white, black, and other in each of the cases for males and females, to understand if they were coming from identical populations. All comparisons were made using the non-parametric Kruskal-Wallis test. In each of the cases, we were able to reject the null hypothesis with a p-value=0.01289 for males and a p-value=0.006276 for females and concluded that the cholesterol levels were coming from independent populations with different means. We also tested, pairwise, a non- parametric test for white, black, and other. The null hypothesis was rejected in each of those cases.

Table 5: The Identified Probability Distributions with the Estimated Parameters that Best Fit the Cholesterol Level Data

	Males	Female
All races	Lognormal(2p):$\hat{\mu} = 5.35, \hat{\sigma} = 0.22$	Gamma: $\hat{\gamma} = 18.27, \hat{\beta} = 12.24$
White	Lognormal(3p):$\hat{\gamma} = -27.19, \hat{\mu} = 5.48, \hat{\sigma} = 0.19$	Gamma: $\hat{\gamma} = 12.11, \hat{\beta} = 18.60$
Black	Lognormal: $\hat{\mu} = 5.33, \hat{\sigma} = 0.24$	Lognormal: $\hat{\mu} = 5.35, \hat{\sigma} = 0.24$
Other	Gamma: $\hat{\gamma} = 26.22, \hat{\beta} = 8.02$	Burr: $\hat{\theta} = 173.06, \hat{\alpha} = 0.18, \hat{\gamma} = 23.05$

Once the study identified racially classified populations from both males and females that exhibit different subpopulations, the study performed the Kolmogorov goodness of fit test to determine the underlying distribution the data follows. The study identified the overall male cholesterol level that exhibited lognormal probability distribution (p-value=0.06) whereas the overall female population exhibited gamma probability distribution (p-value=0.051). Similarly, the study identified and concluded that white males followed three parameters lognormal probability distribution, black males followed as lognormal probability distribution, and other males resembled gamma probability distribution. In addition, the study determined that white females exhibited gamma distribution, black females exhibited lognormal distribution, and that the other females satisfied the Burr distribution. Since the sample size for Burr distribution was very small, the study makes a disclaimer for careful interpretation. Estimated parameters for fitted distributions are presented in Table 5.

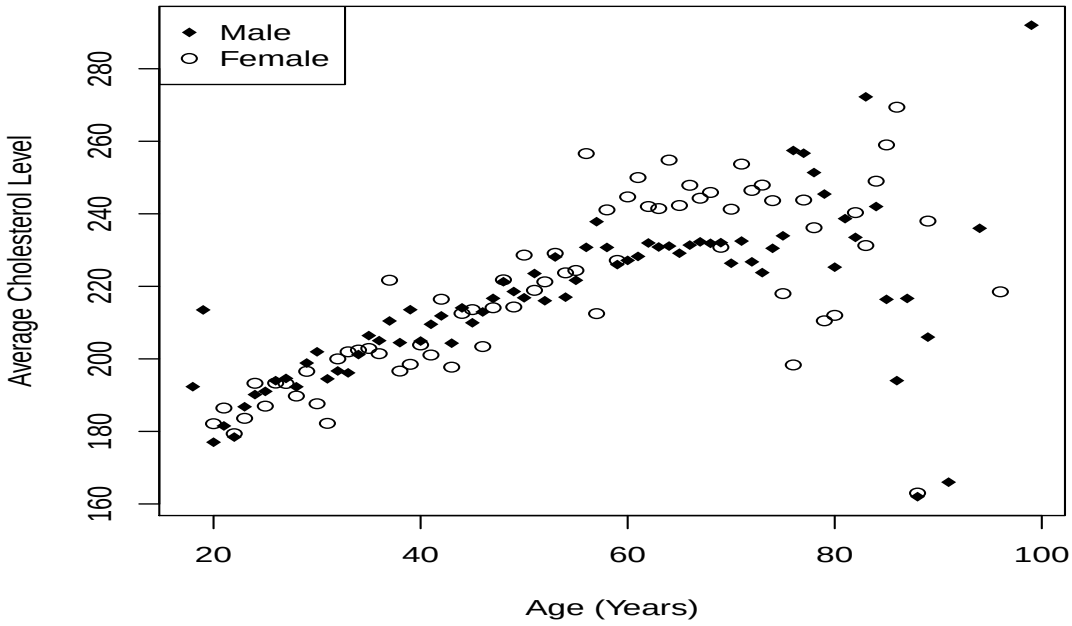


Figure 13.: Age Specific Average Cholesterol Levels by Gender

Age specific average cholesterol levels by sex were plotted to present the variability for various ages. Cholesterol levels varied greatly in both age and sex. Mean cholesterol levels were computed and compared statistically by gender specific averages, Figure 13. Clearly, the average cholesterol level of an individual over 60 years of age exhibited greater variability than those under the age of 60 and the average was consistent in both males and females. Mean cholesterol levels were progressively and consistently higher in each succeeding age, prior to the age of 60. This was true for both males and females. The SCL appeared to be less volatile for ages under 60 years old and displayed a clear linearly increasing pattern for both genders. There was no clear pattern of cholesterol levels after the age of 60 for both genders. However, cholesterol level for females appeared to be consistently higher than males. The study noted that above 80 years old there were relatively few data values, as compared to other ages.

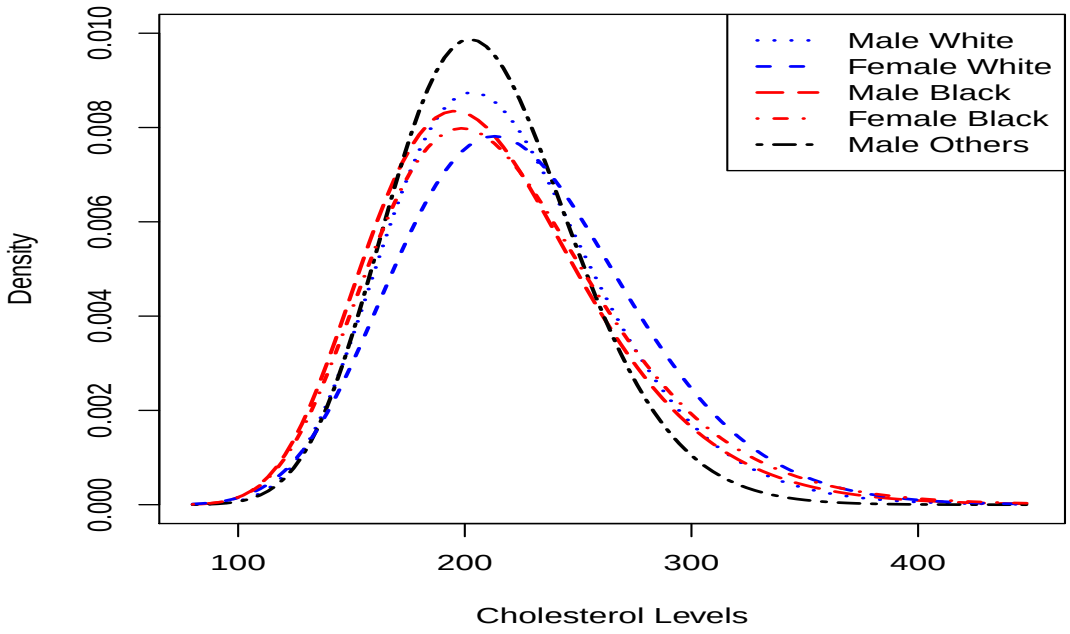


Figure 14.: The pdf of Identified Probability Distributions that Best Fit the Cholesterol Level Data

The study plotted probability density functions that best described the cholesterol level for races, for both males and females, Figure (14). There was a clear difference between cholesterol levels of males and females. The distributions that SCL exhibited were skewed to the right with different means and different skewness. Having known the underlying distributions, we might have a better understanding of the variability of cholesterol level and estimates of basic statistics from which we would have been able to draw proper inferences for different subpopulations.

Cumulative distribution function (CDF) can be used to obtain expected level of cholesterol for genders and races and it also provides the confidence interval, Figure 15.

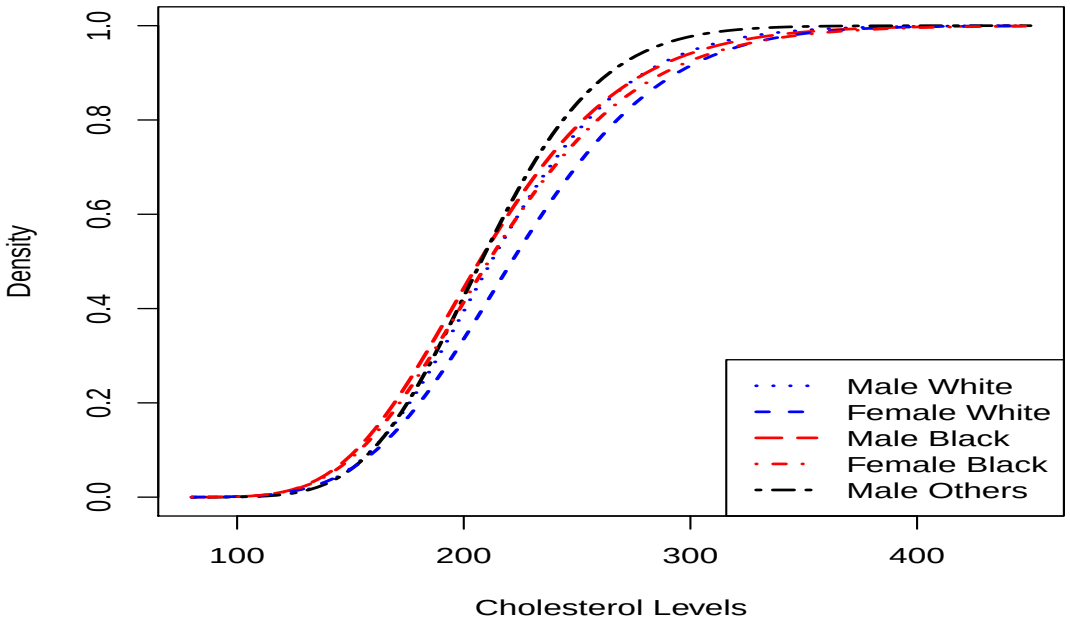


Figure 15.: The CDFs of Identified Probability Distributions that Best Fit the Cholesterol Level Data

Statistics based on distributions have been presented in Table 6. Such measurements and analysis are common but have great implications in evaluations and treatments. Females

Table 6: Estimated Central Tendency and Variability of Cholesterol Level Classified as Race and Gender Using Estimated Parameters of Fitted Distributions

	Mean		Variance		Skewness		Kurtosis	
	Male	Female	Male	Female	Male	Female	Male	Female
All races	216.26	223.6	2345.47	2736.34	0.65	1.24	4.25	11.46
Whites	216.71	232.25	2285.8	2728.39	0.68	-86.77	4.49	384.16
Black	212.83	217.29	2629.64	2897.67	7.2	-63.97	34.75	255.86
Other	210.26	223.8	1769.04	2567.45	0.11	0.80	2.28	2.61

appear to have a higher cholesterol level, on average, than males and the higher cholesterol level is consistent in all the races. This finding is also applicable for variance. Significantly, however, white individuals, both males and females, appear to have higher cholesterol levels than other subpopulations. White females have the highest average cholesterol level among all and have the highest kurtosis. Higher kurtosis implies more of a variance due to infrequent extreme deviation from the mean. It was determined that white females have the highest extreme variation from the mean cholesterol level.

3.6 Conclusion

The study identified the probability distributions for males and females that are respectively lognormal and gamma, which were included in the study. A probability model was applied to investigate its effect on SCL. The study will help researchers to have a greater insight into risk factors of cholesterol levels and their behaviors. The majority of the individuals in the data set were white and had a greater variability in both males and females. The differences in distributions might be used as benchmarks for racial and gender comparisons, and as possible indicators of changes in factors known to influence serum cholesterol, such as diets and drugs.

We have witnessed that SCL increases as age increased and it is possible that SCL becomes more volatile as an individual becomes older, most notably when above 60 years old, Figure 13. SCL is most likely to develop during the late teens to mid-40s [62, 63].

Different behaviors between the race-sex groups were observed in our study. We identified a resemblance to different probability distributions. Our findings are supported by research [62]. The higher the level of cholesterol that an individual has, the greater the risk of subsequently developing CHD would be. The findings of the study were proven with prospective studies such as the Framingham Study [64]. Since the likelihood of developing high SCL is dependent upon race and gender, borderline high risk and high risk level of SCL are likely to be different, according to race and gender. This observation may be subject to a further area of research. The deeper one delves into the issues, the more useful information will be obtained. Further information derived from future research will help researchers speculate more precisely about SCL.

Our study suggests that the average SCL for males and females is different indicating that to a greater extent, reliance upon the analytical precision and accuracy of laboratory measurements will be required before making a generalized assumption about the degree of risk of an individual. An informed decision which is based on male information cannot be generalized to the female and vise-versa [62]. It has been clinically proven that lowering elevated cholesterol levels will reduce the risk of CHD. The degree of risk is relative to gender and race. Efforts have been made to investigate the relationship of some demographic variables to SCL as they may relate to SCL. Our study will assist in developing guidelines which will better inform physicians and public health practitioners of best practices, when determining how to best treat an individual, dependent upon gender and race. The results of this study will inform individuals and be an aid in preventing premature deaths [50]. Further, the study provides data which will aid in better understanding the relationship of certain risk factors to the development of high SCL. The study pivots on aspects of gender and race.

The study identified differences in probability distributions with respect to gender and race to characterize SCL, which in turn was dependent upon diet intake. Therefore, more careful studies of diet, eating patterns, as well as the attitudes and life style of general pop-

ulation and specific subpopulations will help to better understand their relative relationship to SCL. A number of retrospective case-control and cohort studies have investigated the associations between intakes of dietary fats and cholesterol and their associated risk [65, 66]. Since these habits can vary the results, based upon on race and gender, inferences made, which do not take into account these factors, will be misleading which will most likely have serious consequences [67].

This study has resulted in identifying the underlying distributions which now provide a better understanding of the variability of cholesterol level and estimates of basic statistics which include gender and racial difference in terms of probability distributions. Based upon this information, one is able to draw proper inferences for different subpopulations.

3.6.1 Contribution

The study suggested various interesting findings that could be practically and statistically meaningful. Vital statistics can give crucial information about the status of cholesterol level of an individual which can lead to identifying whether an individual is at risk or not. Statistical analysis based on gender and ethnicity also can play a major role for identifying the associated risk due to cholesterol level. We have listed some useful notes associated with this study.

1. The study identified the probability distributions based on gender and ethnicity that the cholesterol level might follow. It shows that the probability distributions followed by males and females are different which is also true with respect to ethnicity.

2. To get more information about the behavior of the cholesterol level of an individual, this study provides more insight about its understanding. It is interesting to note that the risk for high cholesterol depends heavily on gender and ethnicity.

3. It has been noticed that people above 60 years of age are more at risk which is true for both males and females.

Chapter 4

A Longitudinal Study of Serum Cholesterol Levels

4.1 Longitudinal Data

The observations collected repeatedly from the same unit over time in a prospective cohort study is a longitudinal data. Multiple measurements are taken into consideration from each individual at different times. Our aim in the present investigation is to study the changes of the response variable or variables of interest over time and its' association with treatments and other covariates. That is, the objective is to describe how the response changes over time and how these changes depend on the given characteristics of the individuals. The goal in our study is using longitudinal data to characterize the changes with respect to time and the factors that affect the changes. We are working with clustered data that consists of repeated measurements obtained from units at different occasions or time points. We also want to use this data to distinguish among different sources of variation. A longitudinal data analysis (LDA), which includes baseline measurements that increases the efficiency and explains the baseline variability. An important aspect of this type of data is to characterize the changes of the response variable over time, the effect of the baseline covariates, and among subject variation to the response.

4.2 Linear Mixed Model for Longitudinal Data

Linear mixed models (LMM) are popular procedures for analyzing repeated measurements and clustered data. In the present study, the model includes the correlation through the inclusion of random effects that are routinely assumed to be normally distributed [68–71].

Several methods have been proposed in the literature to replace the normality assumption for random effects [31, 72, 73]. Normality of within individual deviations may be a reasonable model for many problems involving continuous measurements. However, in other situations the normality assumption on a transformed scale is better but it offers challenges in interpreting the data. We can study the change in the variable of interest over time either at the population level or the individual level through a longitudinal study. Such a study allows us to separately estimate the cross sectional effects and longitudinal effects. The longitudinal study is more powerful to detect the association of interest than cross sectional study and provides more insight for our interpretation.

In the present study there is only one response for each subject and responses are assumed to be independent, which is the key assumption for classical regression models. However, since the observations are collected from the same subject repeatedly and they tend to be similar and dependent in our longitudinal data analysis. Although observations collected from different subjects are still assumed to be independent. Ignoring dependency of observations taken from the same subject implies ignoring the correlation which could lead to invalid inference. To take care of the dependent property of longitudinal data, linear mixed models are used where the random part addresses the correlation structure.

The linear mixed-effects model is probably the most widely applied method for analyzing longitudinal data [74–82]. It is useful for hierarchical or clustered data that can be traced back to ANOVA paradigm which is useful to analyze longitudinal data [83]. The method includes allowing certain regression coefficients to vary randomly across individual to individual. This idea leads to a two-stage approach of analysis in longitudinal data. The two-stage approach of analyzing longitudinal data has been greatly researched by various article in literature [84, 85]. Since measurements are taken under different situations which are not under control in general, there is a likelihood of considerable variation among individuals. These variations are handled through the random effects approach. Hence, random effects are subject to specific explanation between subject variation and its correlation.

A linear mixed model (LMM) is an extension of a linear regression model to model correlated (longitudinal) data. It has two main components fixed effects and random components. The covariate effects that are fixed across subjects in the study are considered as the fixed effects. If we are to fit a simple linear regression model for each individual profile, the regression coefficients in the classical regression models are fixed effects. The fixed effects explain the expected values of the observations. That is, a simple linear regression model is given by

$$y = \beta_0 + x\beta + \epsilon \tag{4.1}$$

where β_0 is the population intercept, β is the population slope of the regression line, and ϵ is independent error term, $\epsilon \sim N(0, 1)$. β_o and β are estimated by the least square method or maximum likelihood estimators. The covariate effects that change among individuals is called the random effects. Since these values are subject specific, they are considered as random, that is, each subject is a random subject from the population. The random effects explain the variance and covariance of the observations, [85], such a model is written by.

$$y_{ij} = \beta_0 + \beta_1 x_{i,j} + \cdots + \beta_p x_{ij} + b_i + \epsilon_{ij}, \tag{4.2}$$

where β's are fixed effects of interest, $b_i \sim N(0, \sigma_b^2)$ are random effects, and $\epsilon \sim N(0, \sigma_\epsilon^2)$ are independent errors. The b_i represents the deviation of the intercept of individual "i" from the population intercept β_0. The σ_b^2 is the variance between the subjects and σ_ϵ^2 is the variance within each subject. Hence, the total variance of y becomes $Var(y_{ij}) = \sigma_b^2 + \sigma_\epsilon^2$. The expected (average) increase in y with an increase in x while assuming other factors are constant is given by:

$$E(y_{ij}) = \beta_0 + \beta_1 x_{i,j} + \cdots + \beta_p x_{ij}. \tag{4.3}$$

The probability distribution for repeated measurements has the same form for each individual under this type of formulation. However, the parameters of that probability distribution

vary from individual to individual. The marginal probability distribution of the repeated measurements is a multivariate normal probability distribution with a specific covariance structure.

4.3 Total Serum Cholesterol Level

The total serum cholesterol level is a useful screening means for consideration of lipids in individuals younger than 50 years old. There is a direct association of the cholesterol levels with mortality rates, [86]. There is a strong association between cholesterol levels and coronary heart disease as well. Lowering of the cholesterol level is associated with a reduction in coronary heart disease mortality rates, [49, 87]. Initial low serum cholesterol levels and obesity combined appear to indicate a four times greater risk for colon cancer than people with average values of both variables, [88]. The present study is to assess the cause of changes of total cholesterol levels and high density of lipoprotein cholesterol change in the adult population.

It has been discovered that coronary heart disease developed in every fifth man and every 17^{th} woman by the age of sixty that the total cholesterol level has been shown to be an excellent predictor of coronary heart disease in the ages less than 50 years, [89]. Studying the cholesterol level provides crucial information about potential coronary heart disease and other complexities in the human body. Note that various drug treatment are avoidable for those individuals with cholesterol level 300 or above by many physicians. Necessary therapy is suggested for individuals with cholesterol level above 250 and it is believed that the patients, in whom, coronary heart disease will eventually develop seems to have similar cholesterol levels, [89]. The higher the concentration of cholesterol level, the greater the risk of coronary heart disease, [90].

4.4 Objectives of Our Study

In the present study, we will fit a linear mixed model to analyze longitudinal total serum cholesterol levels and investigate how it is changing over time based on age and time. Very little research if any has been done to identify the behavior of cholesterol level over time with age. We believe the present study strengthens the understanding of the behavior of cholesterol level and helps to reduce the risk due to cholesterol related medical issues. The aims of the present study is to analyze the total serum cholesterol data through frequentist approach. We intend to analyze how the cholesterol levels change over time on the average as people get older. Also, we wish to answer the questions: Is the change of cholesterol levels associated with gender and age? How the cholesterol levels change either at population level or individual level? Is there a significance difference between changes in cholesterol level between males and females on the average?. We shall also explore the interaction effects of age, gender, and cholesterol level.

4.5 Statistical Analysis

We have categorized the statistical analysis of the study in to two parts as follows.

4.5.1 Statistical Discussion of the Data

The data used in the present study was obtained from the Inter-university Consortium for Political and Social Research and the data from the National Health and Nutrition Examination Survey (NHANS) II, 1976-1980: Serum Cholesterol data was originally collected by United States Department of Health and Human Services.

The target population for the survey was the civilian non-institutionalized population of the United States. The NHANES II serum cholesterol data files contain two parts of the extensive data available. One part consists of the demographic information obtained from household interviews and the other part is from laboratory results. The survey started in

February 1976 and was completed in February 1980. Samples were selected so that certain population groups thought to be at high risk of malnutrition (person with low incomes, preschool children, and the elderly) were oversampled. Adjusted sampling weights were then conducted for persons over the age of 76, sex, and race categories in order to inflate the sample in such a manner as to closely reflect the estimated civilian non-institutionalized U.S. population.

In addition to the general examination components, several more detailed examinations were performed on subsamples of the population. The information relating to cholesterol level in NHANS II survey considered codes 355-357. Primary site codes were 1 and 2 for males and females, respectively. They illustrated semi-parametric approach on repeated cholesterol data from randomly selected individuals. The data consists of participants' cholesterol levels measured at the beginning of the study and then every 2 years for 10 years, age at baseline, and gender. The following diagram shows the data that was used in our study.

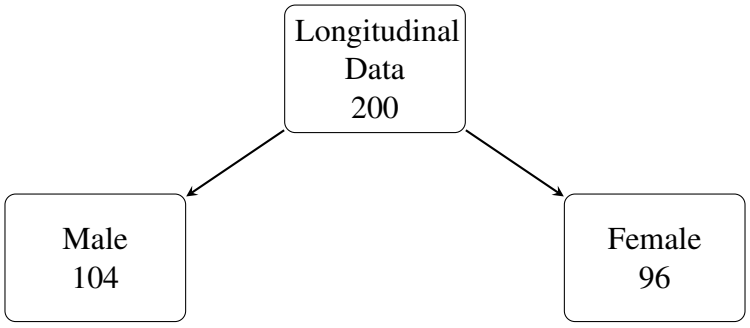


Figure 16.: Schematic Diagram of Total Number of Individuals by Gender

Summary of the basic statistics of cholesterol level measured for each time of an individual is presented, Table 7 below. It shows that the minimum cholesterol level is fluctuating over time. A careful review, we discovered that the quartiles as well as the average cholesterol level over time is increasing with age. However, there is no visual difference in mean and median cholesterol level of an individual.

Table 7: The Summary of Cholesterol Data

	Min.	1st Q.	Median	Mean	3rd Q.	Max.	NA
time1	133	191	213	219.5	247.5	340	
time2	150	195	217	224.2	246	360	
time3	129	200	227	231.7	257	380	
time4	144	209.2	232.5	238.9	264	403	8
time5	144	209	235	242.1	270	430	7
time6	153	220	243	249.3	277.5	378	
age	31	36	41.5	42.49	48	62	

We also notice that the maximum cholesterol level is also increasing upward with age which is true for both males and females. Since there are about seven and eight missing data points, we perform a t-test both (pooled and Satterthwaite) and fail to reject the null hypothesis that there is no difference in population means for missing and non-missing age groups at 5% level of significance.

Table 8 below shows the correlation among cholesterol levels for different times that the observations have been measured. As observations are collected from the same individual over time, they seem to be highly correlated. We have observed strong positive correlation among the responses.

Table 8: Correlation Table of Cholesterol Data

	time1	time2	time3	time4	time5	time6
time1	1					
time2	0.673931	1				
time3	0.784145	0.754992	1			
time4	0.711845	0.707773	0.801818	1		
time5	0.711208	0.731132	0.812008	0.803965	1	
time6	0.660881	0.64175	0.728224	0.815582	0.76791	1

We are considering the longitudinal data of cholesterol level of 200 individuals. The data consists of participants' cholesterol measurements every two years for 10 years. The purpose of such data is to characterize the change in cholesterol level over time, and among-

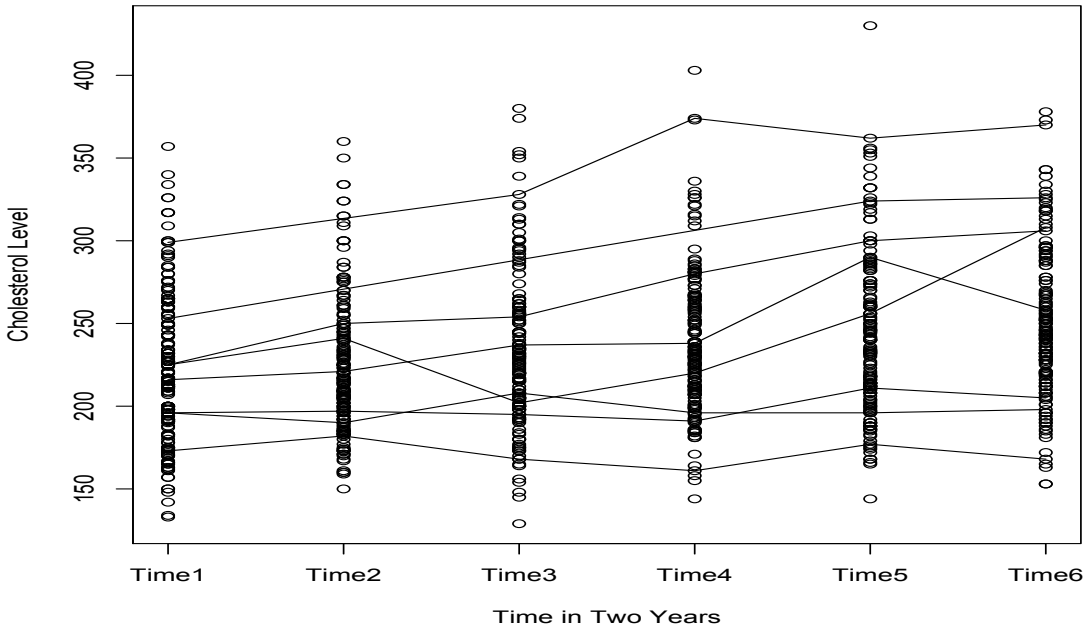


Figure 17.: Longitudinal Cholesterol Level with Trajectories for 8 Random Subjects

subject variation in cholesterol. Figure 17, Figure 18, and Figure 19 show that cholesterol level increases over time for most subjects but substantial inter-subject variation. The parameter estimates for fixed and random effects are given in Table 13. Note that the test of fixed effects on the model came to be significant at 5% level of significance.

A spaghetti plot of the cholesterol data explain that the cholesterol level of an individual over time is some what similar across the individuals regardless of age or gender. Some of the individuals have unusual trajectory line with erratic changes. Detail analysis of the data will give us more insight of the information. Sample plot of randomly selected eight individuals provides us more visual insight about the nature of the data. It explains more clearly that there is a random slope and intercept which suggests in adapting linear mixed effects model as shown in Figure 17. It depicts that there is a variation of cholesterol level over time for each individual and suggests that cholesterol level increases almost linearly over time for most of the subjects with different inter-subject variation. In other words, each

subject seems to have his/ her own trajectory line with a possibility of different intercepts and slopes suggesting possibly two types of variation simultaneously called within and between variations.

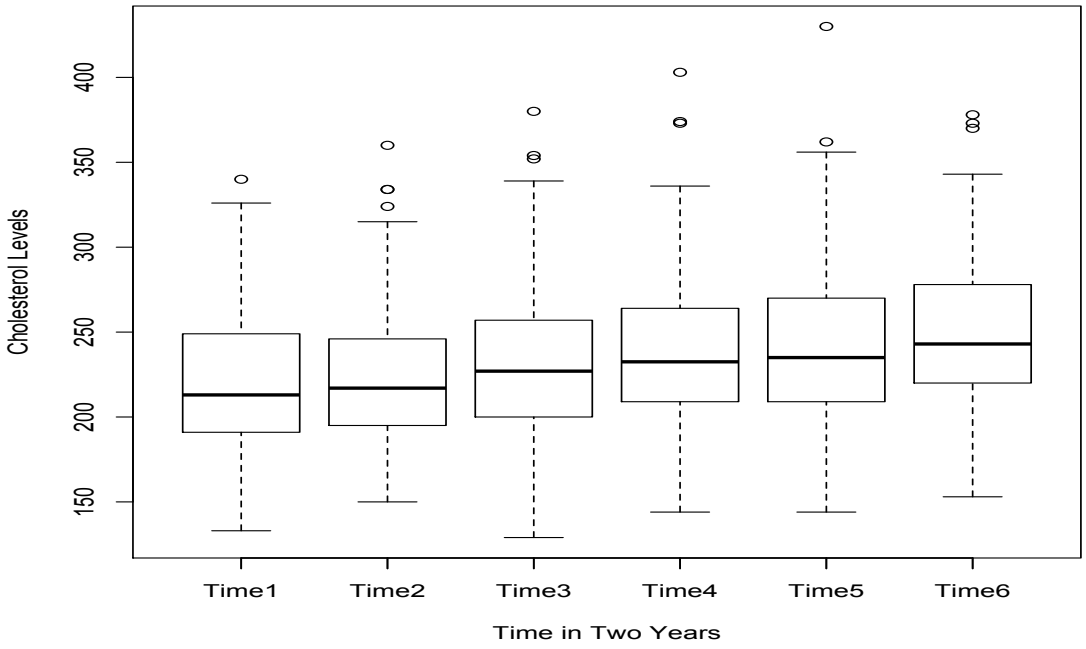


Figure 18.: Box Plot of Cholesterol Data

The box plot, Figure 18 and mean trend graph, Figure 19, of cholesterol level also indicate that there is a change in slope and intercept over time suggesting that the mixed effect model might explain its behavior. Over all, there is two way variation of cholesterol level. Each subject has on average of 5 observations and the data is not balanced because some of the data are missing.

To explore the nature of the data more closely, we fitted simple linear regression over time individually. We found that the pooled residual plot follows approximately normal probability distribution which agrees with the assumption of within individual normality. Figure 20, shows the cholesterol level for individuals over time. It has been shown in the literature that the inference on fixed effects appears to be robust even for non normality of

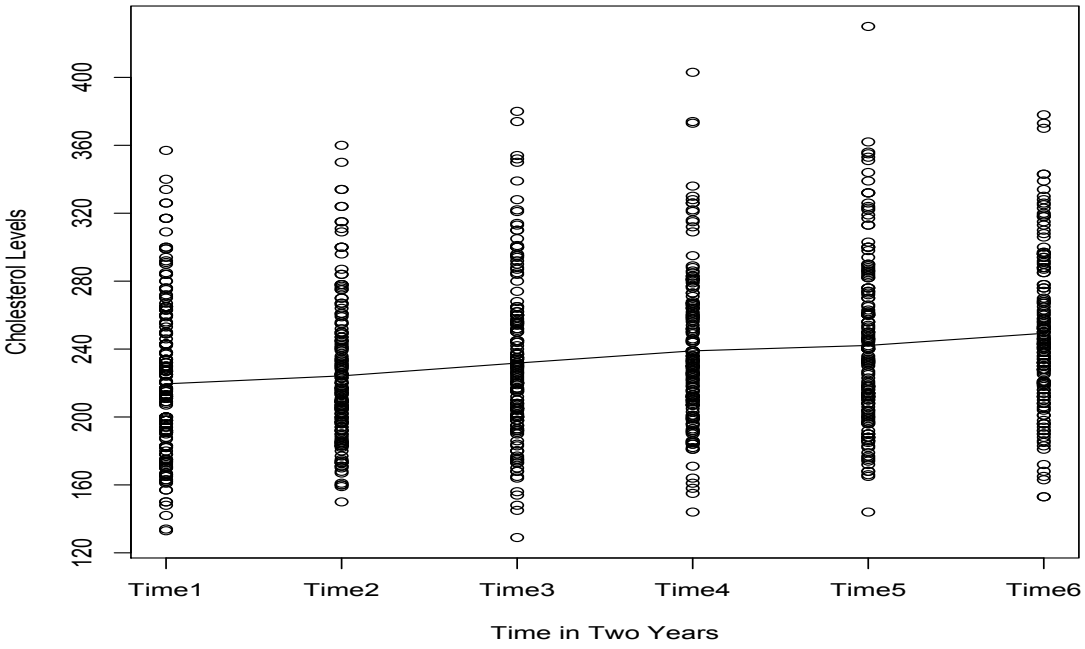


Figure 19.: Average Cholesterol Level Over Time

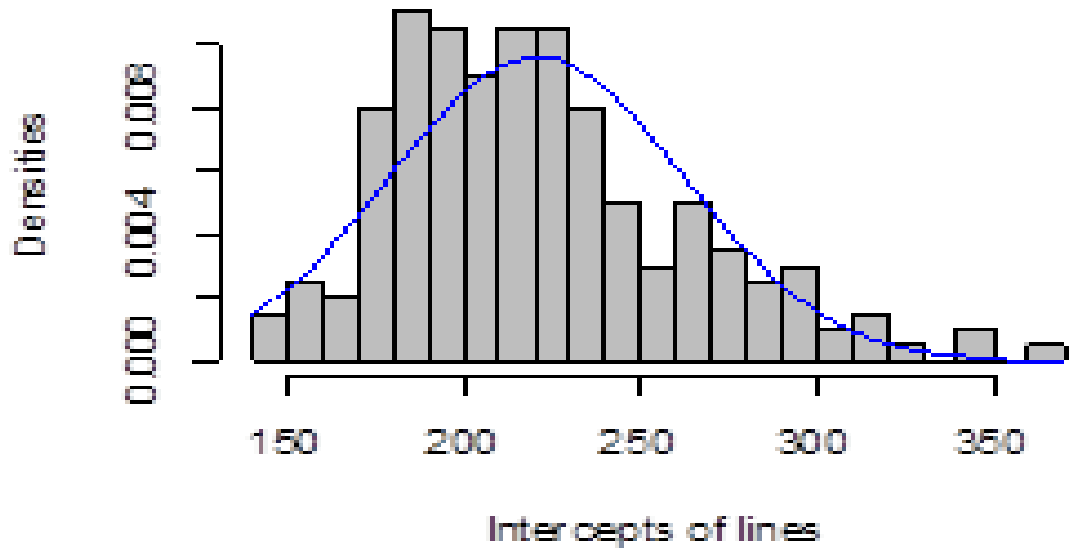


Figure 20.: Histogram of Subject Specific Intercepts in Simple Linear Reression

random effects, [73, 91].

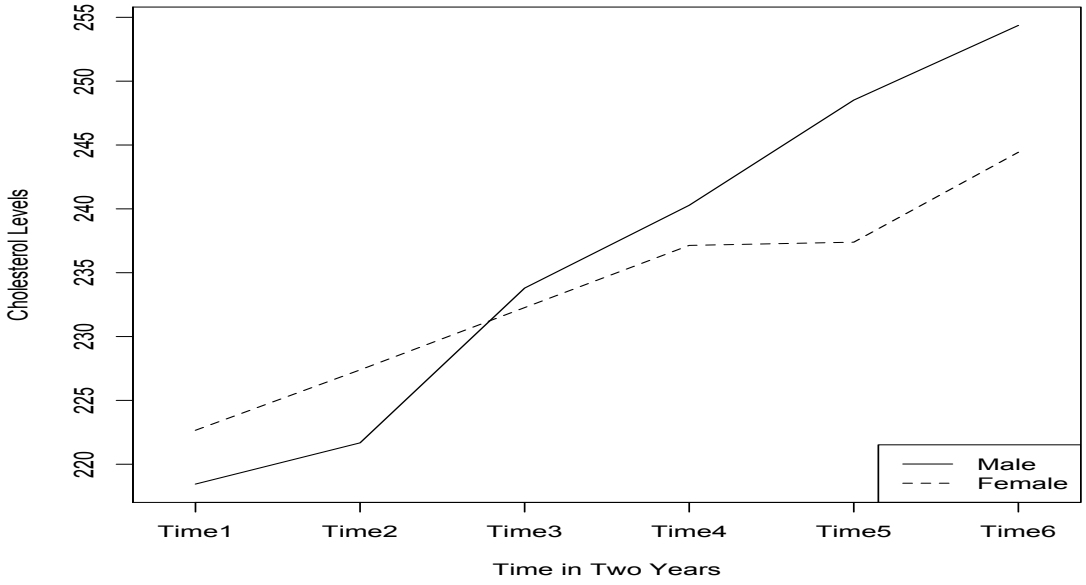


Figure 21.: Interaction Plot

We fitted a simple linear regression considering cholesterol as response and their corresponding time point as covariates. A summary of the statistics for intercept and slope for individual regression line shows that there is a difference in variances of intercept and slope. Intercepts for individual regression has mean cholesterol level of around 221 with high standard deviation of around 42 which explains that there is high variation among the individuals cholesterol levels. At the same time, there are slopes with mean 2.55 with standard deviation of 3.63 indicating there is a variation of slopes as a random effects, Table 9.

Table 9: Summary of Slopes and Intercepts

Variable	N	Mean	Std. Dev.	Sum	Minimum	Maximum
Intercept	200	220.6894	41.68917	44138	141.1429	360.1667
Slope	200	2.55025	3.62947	510.0506	-14	11.74286

4.5.2 Statistical Modeling of the Data

The general form of the linear mixed effects statistical model is given by

$$Y = X\beta + Zb + \epsilon, \tag{4.4}$$

$Y \sim N(\mu, V)$, where Y is the response vector, μ is the mean vector, and V is the variance covariance matrix. X and Z are fixed and random design matrices, respectively. Also, β is a vector of unknown fixed effects, $b \sim N(0, \sigma_b^2)$ is an unknown random effects, and $\epsilon \sim N(0, \sigma_\epsilon^2 I)$ is the unknown random independent errors. The expected value of Y is given by

$$E(Y) = X\beta = \mu,$$

and the variance by

$$Var(Y) = Var(Zb + \epsilon) = ZVar(b)Z' + Var(\epsilon) = ZGZ' + R = V,$$

where $R = \sigma_\epsilon^2 I$ and $G = \sigma_b^2$. Y is k - dimensional random response vector that may include repeated measures of the same variable or measurement of correlated variables. X and Z are known $k \times p$ and $k \times q$ matrices of covariates, β is p-dimensional vector of fixed effects, b is q-dimensional vector of random effects, and R is a $k \times k$ covariance matrix, and G is a $p \times q$ covariance matrix. Treating b_i as random variable so as to insure the inference for the whole population from which the sample is drawn. The correlation between $y_{i,j}$ and $y_{i,j'}$ is

given by:

$$Cor(y_{i,j}, y_{i,j'}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2} = \rho. \tag{4.5}$$

The analytical statistical model in this study is given by:

$$\begin{aligned} cholestlevel_{ij} = \beta_0 + \beta_1 * time_{i,j} + \beta_2 * gender_i + \beta_3 * age_i \\ + \beta_4 * time_{i,j} * age_i + \beta_5 * time_{i,j} * gender_i \\ + b_{0,i} + b_{1,i} * time_{i,j} + \epsilon_{i,j}. \end{aligned} \tag{4.6}$$

4.5.3 Covariance Structures for Repeated Measurements: Development of the Model

The covariance structure for the random behavior within subject effects which is also called repeated measurements, is challenging considering non independence and is defined by a series of $k \times k$ within subject covariance matrices. We model matrix "V" by setting up the random effects design matrix Z and by specifying covariance structure for G and R. We denote it by sigma (Σ) where k represents the number of times the subject is observed. The covariance structure incorporates the error structure of marginal residuals as they are defined at individual observation levels. Given below are some of key entities in the development of the proposed statistical model, that is, unstructured (UN), variance components (VC), compound symmetry (CS), first order autoregressive (AR (1)), and toeplitz (TOEP), etc. Given below is a brief summary of its' structure.

1. **Unstructured (UN)**

   The elements of the matrix are assumed to be different in this covariance structure. It requires maximum parameters to be estimated. The number of parameters to be estimated is given by the relation $\frac{k(k+1)}{2}$, where k is number of repeated measurements.

The elements in the covariance matrix are represented by $\sigma_{ij} = \sigma_{ji}$, that is,

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2(k-1)} & \cdots & \sigma_k^2 \end{bmatrix}. \tag{4.7}$$

2. **Variance Components (VC)**

   This structure assumes that the correlation of errors within a subject is zero and the covariance matrix is given by

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix}. \tag{4.8}$$

3. **Compound Symmetry (CS)**

   Compound symmetry assumes homogeneous variances and constant correlation regardless of how far apart the measurements are. It also requires that the two parameters to be estimated in the statistical modeling. The elements in matrix form are represented by

$$\sigma_{ij} = \begin{cases} \sigma_1^2 + \sigma^2, & i = j \\ \sigma_1^2, & \text{elsewhere} \end{cases} \tag{4.9}$$

$$\Sigma = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \cdots & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 & \cdots & \sigma_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2 & \sigma_1^2 & \cdots & \sigma^2 + \sigma_1^2 \end{bmatrix}. \tag{4.10}$$

4. **First Order Autoregressive AR(1)**

   This covariance structure assumes homogeneous variances and exponentially declining correlations as a function of distance. It implies that the two values that are close to each other are highly correlated and less correlated as the measurements are farther apart. It requires two parameters to be estimated and the elements in the matrix are represented by $\sigma_{ij} = \sigma^2\rho^{|i-j|}$, as follows,

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho^{k-1} \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho^{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{k-1} & \sigma^2\rho^{k-2} & \cdots & \sigma^2 \end{bmatrix}, \tag{4.11}$$

   where σ represents the variance of responses and ρ represents their correlation.

5. **Toeplitz (TOEP)**

   This structure of covariance is similar to AR(1) which assumes that all the measurements next to each other have the same correlation, the measurements that are two apart have the same correlation, different from the first, the measurements that are three apart have the same correlation different from first two and so on. It requires that "k" parameters to be estimated, the elements of the matrix can be represented by, $\sigma_{ij} = \sigma_{|i-j|+1}$, and the covariance matrix is given by

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_1 & \cdots & \sigma_k \\ \sigma_1 & \sigma^2 & \cdots & \sigma_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_k & \sigma_{k-1} & \cdots & \sigma^2 \end{bmatrix}. \tag{4.12}$$

4.5.4 Estimation in Linear Mixed Model (LMM)

Estimating the parameters is more difficult in the linear mixed model than the linear model because of not only we need to estimate β but also the unknown parameters in b, Z, and R. Generalized least square (GLS) method is used to estimate the parameters by minimizing $(Y - X\beta)'V^{-1}(Y - X\beta)$. There are mainly two approaches of estimating the necessary parameters. Let $Y_i \sim N(X_i\beta, V)$ and then the probability density function of Y_i is given by

$$f(Y_i) = (2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}}e^{-\frac{1}{2}(Y_i - X_i\beta)'V^{-1}(Y_i - X_i\beta)}$$

where $i = 1, 2, ..., m$. Given below are the two statistical methods of obtaining estimates of the parameters in the proposed model.

1. **Maximum likelihood estimator (MLE) method:**

   We have the likelihood function:

$$\begin{aligned} L(\beta, V; Y) &= \prod_{i=1}^{m} f(Y_i) \\ &= \prod_{i=1}^{m} (2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}}e^{-\frac{1}{2}(Y_i - X_i\beta)'V^{-1}(Y_i - X_i\beta)} \\ &= (2\pi)^{-\frac{nm}{2}}|V|^{-\frac{m}{2}}e^{-\frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)'V^{-1}(Y_i - X_i\beta)}. \end{aligned}$$

   Setting

$$D = \sum_{i=1}^{n}(Y_i - X_i\beta)'V^{-1}(Y_i - X_i\beta),$$

   the log-likelihood function is given by:

$$logL(\beta, V) = -\frac{nm}{2}\log(2\pi) - \frac{m}{2}\log(|V|) - \frac{1}{2}D.$$

Taking the partial derivative with respect to β and letting equal to zero, that is,

$$\frac{\partial log L(\beta, V)}{\partial \beta} = 0,$$

the maximum likelihood estimator of β and b are given by

$$\hat{\beta} = (X_i' \hat{V}^{-1} X_i)^{-1} X_i \hat{V}^{-1} Y$$

and

$$\hat{b} = \hat{G} Z' \hat{V}^{-1} (Y_i - X_i \hat{\beta}).$$

2. **Restricted maximum likelihood estimator (REML) method:**

Based on Bayesian approach, we proceed to assume a uniform prior for β. Thus, the likelihood function is given by

$$L(V; Y) = \int L(\beta, V) d\beta,$$

where

$$L(\beta, V) = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} e^{-\frac{1}{2}(Y_i - X_i\beta)' V^{-1} (Y_i - X_i\beta)}.$$

After integrating out over β, we obtain

$$L(V) = (2\pi)^{-\frac{1}{2}(n-r)} |V|^{-\frac{1}{2}} |X_i' V^{-1} X_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(Y_i - X_i\hat{\beta})' V^{-1} (Y_i - X_i\hat{\beta})}.$$

Now the log-likelihood function is given by

$$\log L(V; Y) = -\frac{n-r}{2} \log(2\pi) - \frac{1}{2} \log |V| - \frac{1}{2} \log |X_i' V^{-1} X_i| - \frac{1}{2} (Y_i - X_i\hat{\beta})' V^{-1} (Y_i - X_i\hat{\beta})$$

where "r" is the rank of design matrix X. Taking partial derivative of the log-likelihood

and equating it with zero, that is,

$$\frac{\partial \log L(V;Y)}{\partial V} = 0.$$

There is no closed form solution to the resulting equation and we proceed to obtain numerical approximations. We consider some reasonable estimate for V to estimate G and R. The approximate restricted maximum likelihood or residual maximum likelihood estimator of V by solving the equation given by

$$\hat{\beta} = (X_i'\hat{V}^{-1}X_i)^{-}X_i\hat{V}^{-1}Y.$$

4.5.5 Model Selection

We have considered graphical representation to visually identify possible random effects, Figure17. **Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)** have been used to select the best model. The AIC criteria is given by

$$AIC = -2\left\{l(\hat{\beta},\hat{\theta};y) - q\right\}, \tag{4.13}$$

where $\hat{\beta}$ is an estimate of β, $\hat{\theta}$ is an estimate of θ, q represents number of elements in $\hat{\theta}$. The smaller the AIC, the better the statistical model. And the Bayesian Information Criteria (BIC) is given by

$$BIC = -2\left\{l(\hat{\beta},\hat{\theta};y) - 0.5 \times q \times log(m)\right\}, \tag{4.14}$$

where m represents number of subjects. Here, again, the smaller the BIC, the better the statistical model.

We have fitted the linear mixed model considering different variance structures and adopted AIC and BIC criterion to select the best possible model. The smaller the AIC

and BIC, the better the fit. We have found that the model with unstructured G matrix with V matrix VC appears to be the best fit for the data under RMLE method of estimation to select the best model. Once the model that best fit the data has been identified, we estimated the parameters under MLE method. We perform likelihood ratio test of the proposed model which appears to be significant at the 5% level of significance and the test result is presented in Table 10, below.

Table 10: Test of the Model

Source	DF	Sum of square	Mean square	F-value	$Pr > F$
Model	214	1811887	8466.763	17.99	0.0001
Error	829	390173	470.655		
Total	1043	2202060			

We have observed that an interaction exists among the covariates under modeling, Figure 21, which strongly suggests that there is an interaction with time. The mean cholesterol levels for male and female appear to be different. We perform analysis of variance (ANOVA) test to identify if means of groups are equal of the population level. We are able to reject the null hypothesis at $\alpha = 7\%$ level of significance and conclude that there is a significance difference in population means in groups, Table 11.

Table 11: Test for Time and Gender Effect

Source	DF	Sum of square	Mean square	F-value	$Pr > F$
Time	5	106553.8	21310.76	45.28	0.0001
Gender	1	1399.88	1399.88	1.97	0.0694
Gender*time	5	10626.24	2125.247	4.52	0.0005
Age*time	6	157464.7	26244.12	55.76	0.0001
Id(gender)	197	1535843	7796.156	16.56	0.0001

We have considered time, age, gender, time*age, and time*gender as a fixed effects in the modeling. We test the hypothesis that fixed effects are zero and we were able to reject

the null hypothesis at $\alpha = 7\%$ level of significance and concluded that they are non zero. The test results are given by, Table 12.

Table 12: Test of Fixed Effect

Effect	Num DF	Den DF	F Value	$Pr > F$
Time	1	175	43.49	0.0001
Age	1	184	40.99	0.0001
Gender	1	192	3.34	0.0694
Time*age	1	176	18.87	0.0001
Time*gender	1	178	16.22	0.0001

The estimated parameters from the proposed model for fixed effects are displayed in Table 13.

Table 13: Estimated Parameters for the Model for Fixed Effect

Effect	Estimate	Standard Error	DF	t - Value	$Pr > \|t\|$
β_0	123.09	15.1493	181	8.13	0.0001
β_1	8.9267	1.2652	174	7.06	0.0001
β_2	9.8800	5.4101	192	1.83	0.0694
β_3	2.1761	0.3399	184	6.40	0.0001
β_4	-0.1221	0.02811	176	-4.34	0.0001
β_5	-1.8083	0.4491	178	-4.03	0.0001

Thus, the final form of the proposed statistical model is given by

$$\begin{aligned} cholestlevel_{ij} = 123.09 + 8.9267 * time_{i,j} + 9.88 * gender_i + 2.1761 * age_i \\ - 0.1221 * time_{i,j} * age_i - 1.8083 * time_{i,j} * gender_i \\ + b_{0,i} + b_{1,i} * time_{i,j} + \epsilon_{i,j}. \quad (4.15) \end{aligned}$$

The test of the model is significant at $\alpha = 5\%$ level of significance which indicates that we have a good fit model to the data. After identifying the best fitted model, we performed the residual analysis of the model. We perform Pearson residual check for residuals Figure

(24) and Figure (25) and studentized residual check for residuals of the model, Figure (22) and Figure (23). The residuals appear to be around zero without any pattern that will indicate there is no correlation. Also, most of the residuals are within two standard deviations away from mean zero suggesting that we have a good model for the subject data. Q-Q plot and summary of residuals with density curve further support this claim. Equation (4.16) provides the estimated unstructured variance covariance matrix in modeling where diagonal elements represent variance of cholesterol level in a particular time and off diagonal elements represent their corresponding covariances as given below in the equation (4.16).

$$V = \begin{bmatrix} Time1 & Time2 & Time3 & Time4 & Time5 & Time6 \\ 1435.92 & 1063.01 & 1125.05 & 1187.09 & 1249.14 & 1311.18 \\ 1063.01 & 1572.6 & 1212.28 & 1286.92 & 1361.56 & 1436.2 \\ 1125.05 & 1212.28 & 1734.47 & 1386.75 & 1473.98 & 1561.22 \\ 1187.09 & 1286.92 & 1386.75 & 1921.54 & 1586.41 & 1686.24 \\ 1249.14 & 1361.56 & 1473.98 & 1586.41 & 2133.79 & 1811.25 \\ 1311.18 & 1436.2 & 1561.22 & 1686.24 & 1811.25 & 2371.23 \end{bmatrix}. \tag{4.16}$$

The covariance matrices of the random effects is obtained for female (G_f) and male (G_m) respectively are given by:

$$Var(G_f) = \begin{bmatrix} 1380.39 & -2.2097 \\ -2.2097 & 1.4663 \end{bmatrix} \tag{4.17}$$

and

$$Var(G_m) = \begin{bmatrix} 1000.96 & 31.0216 \\ 31.0216 & 3.1487 \end{bmatrix}. \tag{4.18}$$

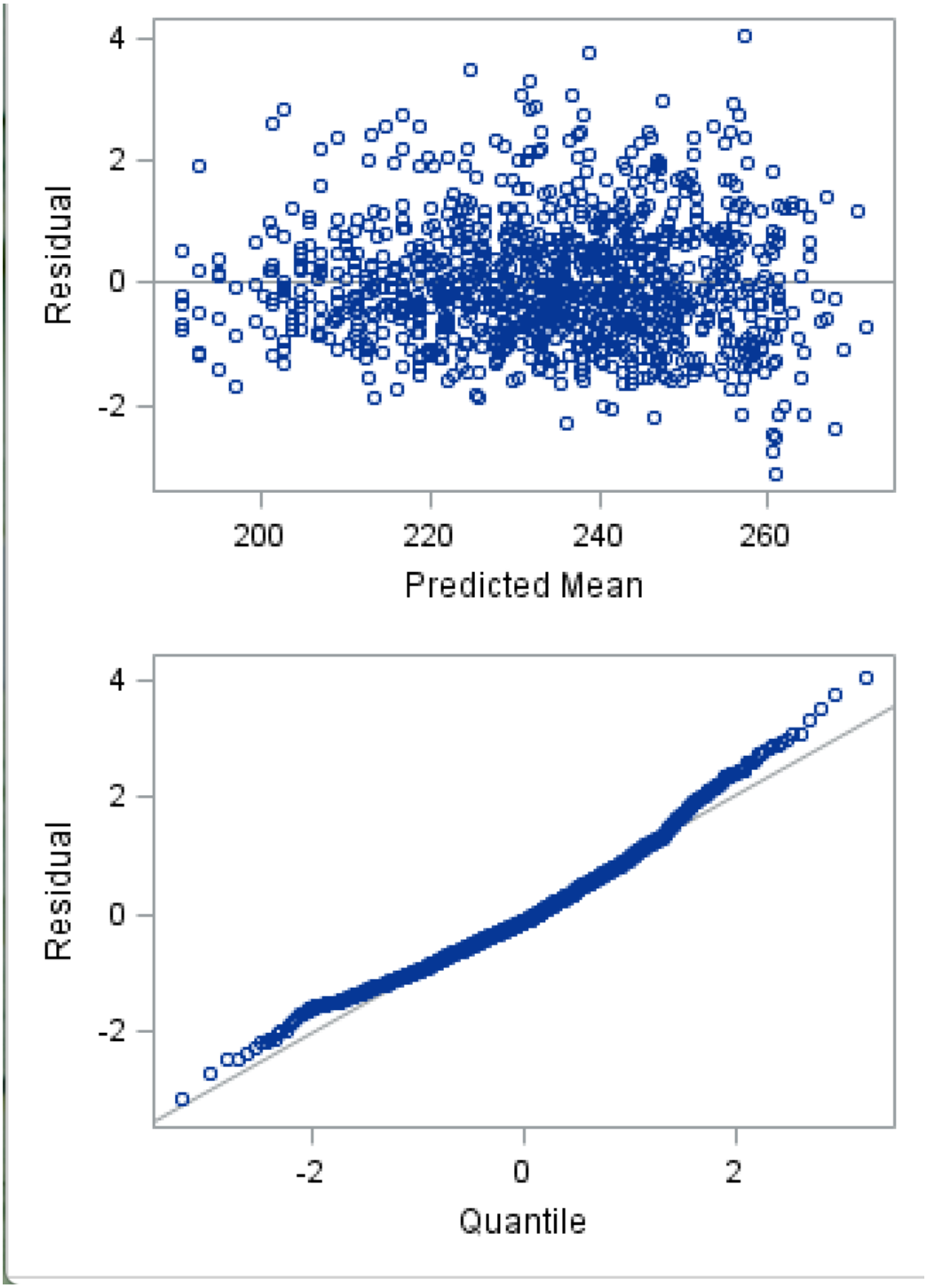
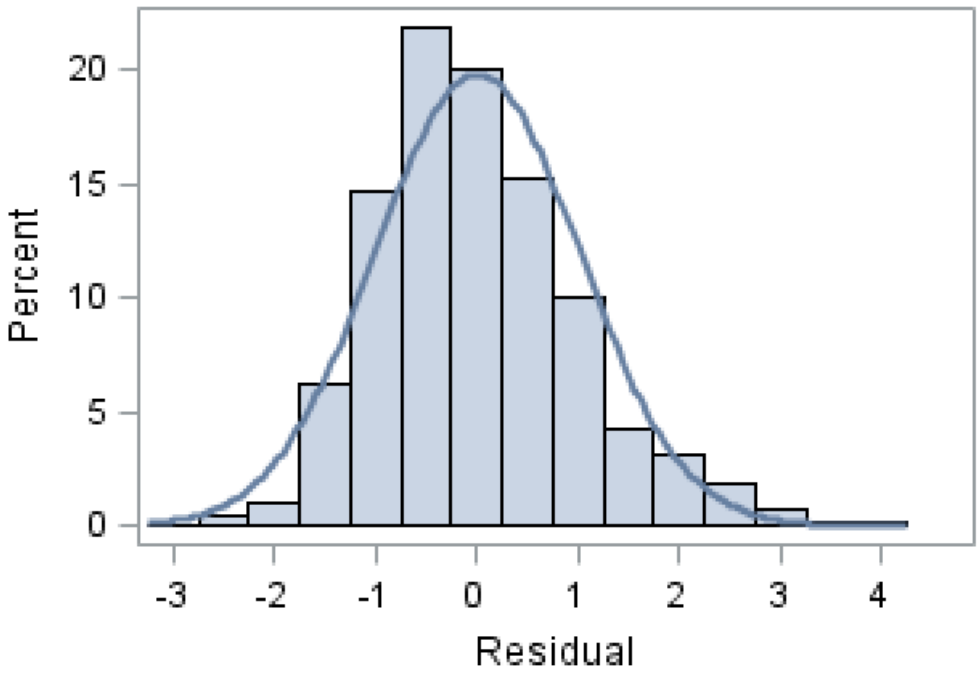


Figure 22.: Studentized Residuals of Cholesterol Level (I)

Then, we used this information to structure the final form of the proposed model, (4.15). The linear mixed models provide flexible tools for analysis of multiple correlated response



Residual Statistics	
Observations	1044
Minimum	-3.136
Mean	0.0031
Maximum	4.0433
Std Dev	1.0089

Fit Statistics	
Objective	9903.9
AIC	9929.9
AICC	9930.3
BIC	9972.8

Figure 23.: Studentized Residuals of Cholesterol Level (II)

variables. The model we adopted is based on fitting separate LME models for each response, while we combine them in a single model by imposing a joint multivariate distri-

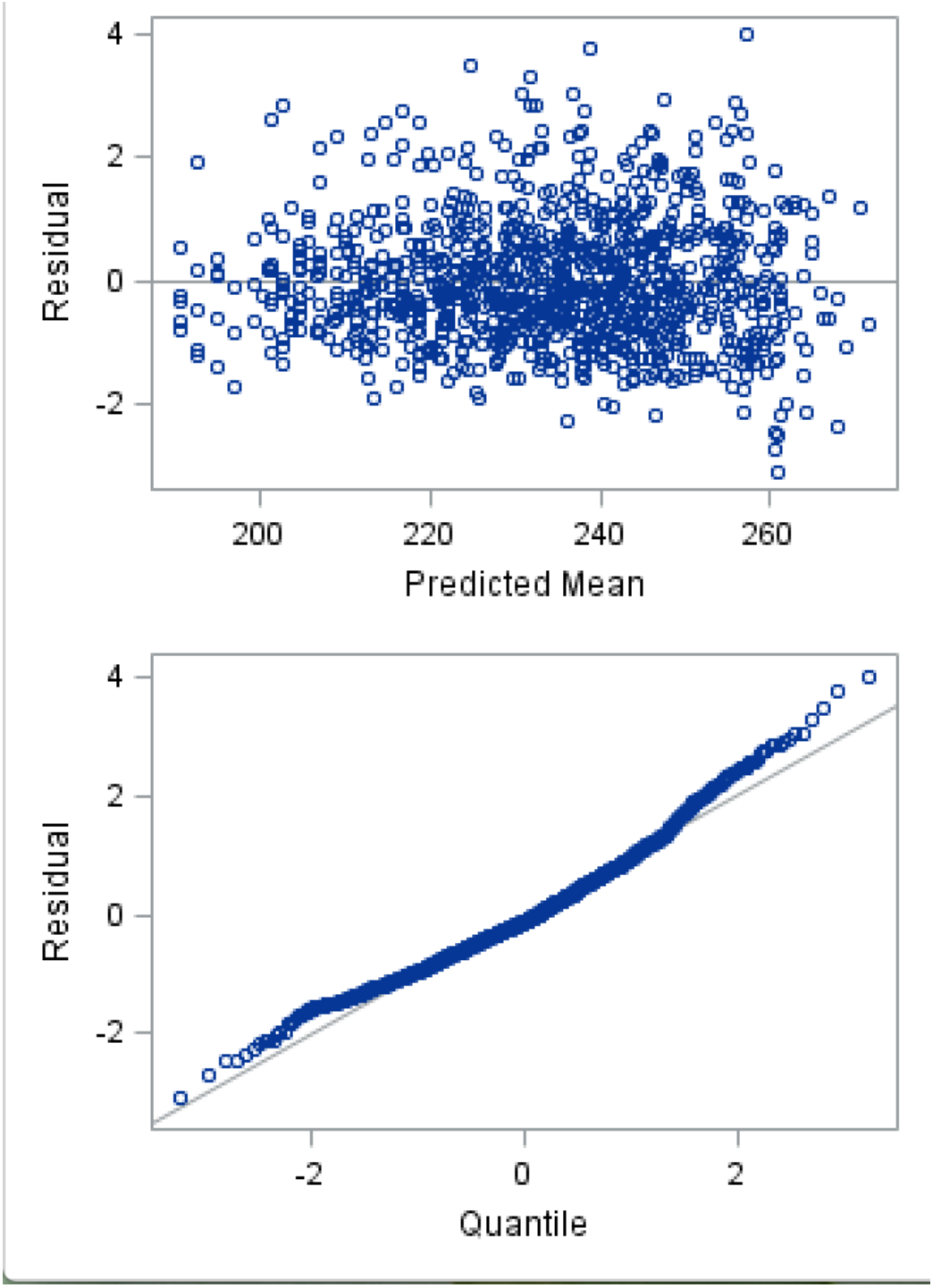
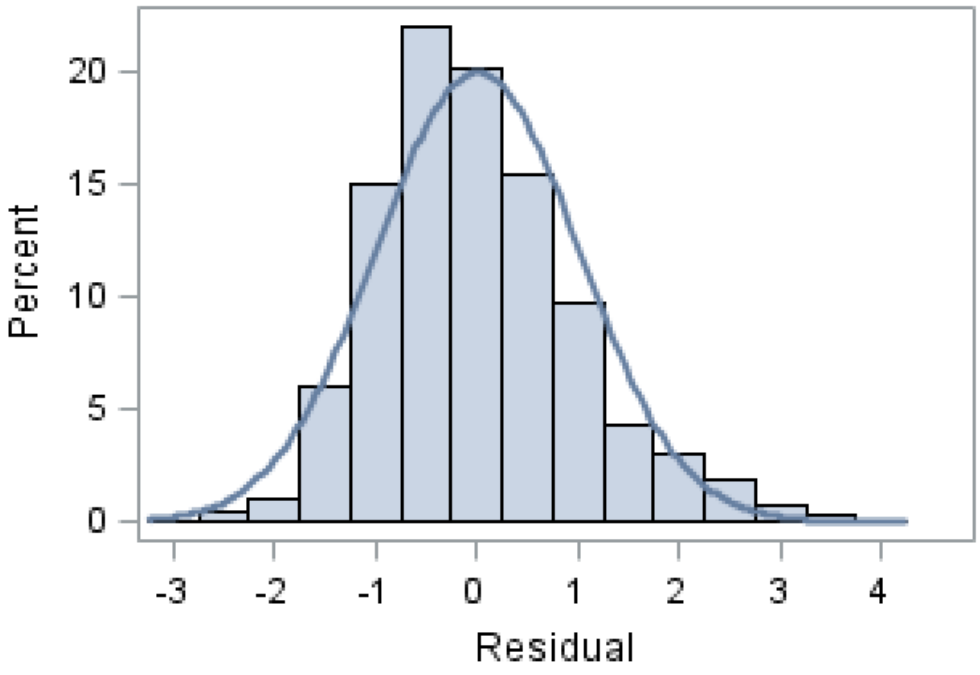


Figure 24.: Pearson Residuals of Cholesterol Level (I)

bution for error terms.



Residual Statistics	
Observations	1044
Minimum	-3.094
Mean	0.0033
Maximum	3.9945
Std Dev	1.0018

Fit Statistics	
Objective	9903.9
AIC	9929.9
AICC	9930.3
BIC	9972.8

Figure 25.: Pearson Residuals of Cholesterol Level (II)

Figure (26) shows that the fitted values represent the observed values quite well in general, which attempts to the quality of the model.

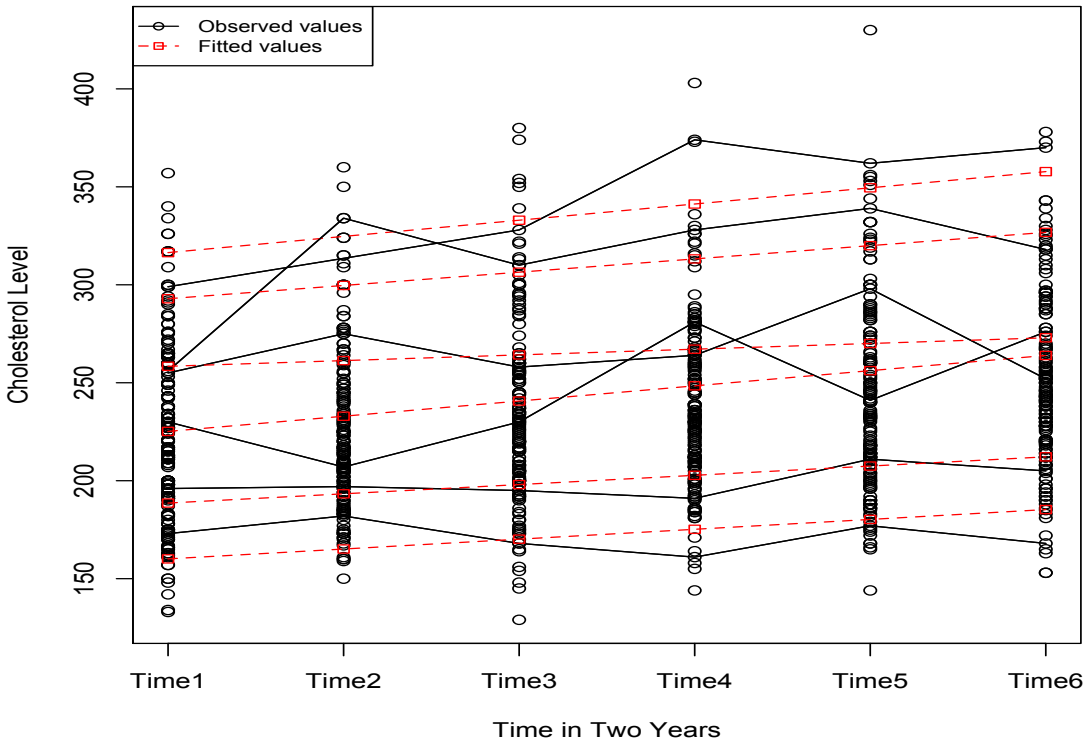


Figure 26.: Observed and Fitted Cholesterol Level for Six Random Subjects

Hence, the behavior of cholesterol level of patients can be explained through mixed statistical model approach is quite reasonable. In our work, we established that the proposed statistical model, equation (4.15), can be adapted once the baseline covariates are known. Note that the cholesterol level of a patient changes overtime and it changes within a patient as well as between patients. We also notice that a change in cholesterol level is different for male and female over time as patients get older [49]. The raise in total cholesterol level among healthy young men and women are significantly associated with age with obesity [92]. The National Institutes of Health Consensus Development Conference on Lipid Lowering has recommended that cholesterol levels to be reduced to 200 mg/dl in all persons [89].

We have found that interaction, between age and time are significant contributors. It suggests that age and time independently may not be contributing heavily. However, their joint effect could have a different impact in order to develop coronary heart disease. Our results are consistent with the result stated in similar study [90]. Similarly, age effect in conjunction with gender could lead to high risk of developing high cholesterol level since the mean trend cholesterol level for male and female are different, Figure 21. Since coronary heart disease developed at a much lower rate in young women than it did in men [89]. However, there could be a symptom complex in both genders with triglycerides which is assumed to be closely related to coronary heart disease.

The risk of coronary heart disease in persons younger than age 50 is highly related to total serum cholesterol levels. Increase in cholesterol level more than usual limit increases risk of coronary heart disease five times [93]. The high cholesterol level is also strongly related to an incidence of diabetes mellitus that is twice the rate in general population and are usually is due to obesity or overweight [89]. Cholesterol lowering interventions have helped to reduce coronary heart disease [94]. Since average cholesterol levels for male appears to be higher than female and so does an increased risk in proportional to antecedent serum cholesterol, male tend to be more at risk than female in order to suffer from cholesterol related issues [64]. The effect of interaction of cholesterol level with age and gender have strong association as we noticed in our study and similar to the findings in [95].

In particular, the physical and mental health components are highly affected by cholesterol level given to a person, for which we will be able to have prior information about the cholesterol level which can lead us to have some precautionary measure ahead of time to avoid any unforeseen situation, because of increased cholesterol level in the body. It is essential to study the effect of cholesterol level taking other factors like diabetes, triglyceride, and lipid into account.

4.5.6 Contribution

In this study, we have observed interesting behavior of average cholesterol level of an individual. We would like to list some of the important findings.

1. The average cholesterol level increases linearly as an individual get older posing the high risk of getting cholesterol related diseases.

2. Each individual has its own trajectory of increasing average cholesterol level as she/he gets older and it is true for both males and females.

3. There is a inter person variation of average cholesterol level which suggests that individuals' cholesterol level could be above or below from marginal cholesterol level.

4. Our study shows that there is an interaction between time with age and gender. That suggests that time with age and gender plays important role in identifying the effects of cholesterol in an individual.

5. Finally, we developed a statistical model that takes into consideration all significant attributable variables and interactions. Thus, one can use this model to estimate the cholesterol level of an individual, among other important information.

Chapter 5

Modeling Serum Cholesterol Levels by Functional Data Analysis Approach

5.1 Functional Data

Functional data is multivariate data with an ordering on the dimension (Muller, 2006). Functional data analysis is all about the analysis of information given by curves, [96, 97]. The observations are converted into a curve that is computable for any argument and the curve would be the input for functional data analysis. Functional data is represented by a set of basis functions that are the functional building block which are combined linearly to represent the statistical model. **Spline basis, Fourier basis, constant basis, and exponential basis**, etc, are some such examples. The basis functions are mathematical functions used to represent the observed data as a curve. If these observations are assumed to have no error, the process is interpolated otherwise the data needs to be separated by error and is called smoothing. The general purpose of the functional data analysis is to represent the data into different ways to produce relevant statistical analysis, that is, display the data so as to highlight its different characteristics and patterns it represents, and to explain the variation of an outcome with the help of the attributable variables, [98, 99].

5.2 Introduction

Coronary heart disease (CHD) is one of the leading cause of deaths in the USA, [49]. Smoking, high blood pressure, and high cholesterol level are the most clearly established risk factors that have been identified as being strongly associated with coronary heart disease, [48]. Total serum cholesterol level (SCL) is a major risk factor for CHD among them.

CHD is responsible for more deaths than all forms of cancer combined, [49]. A general criteria has been in practice that total SCL for adults should be below 200mg/dl and individuals with values between 200mg/dl to 239 mg/dl should be considered as borderline high risk; those with values more than 240 mg/dl should be regarded as high risk for CHD, [51, 52]. Hence, it is extremely important to understand the SCL behavior so as to reduce its associated risks.

It has been well established that diets and drugs play a crucial role in SCL, [53, 54]. The behavior of cholesterol under controlled diets and drugs are presently under intensive research, [61, 100], since diet and cholesterol level are highly correlated. Significant efforts have been also made to predict the cholesterol level based on the age and gender, [58, 59]. Males and females seem to have a different relationship in the prediction process. Very little work has been done about modeling the cholesterol level with respect to males and females, until the present study.

In this study, we have modeled the cholesterol data for males and females for USA, individually through functional data analysis approach to understand the behavior of cholesterol levels. Functional data analysis is an important analytical method that can be used for exploratory and hypothesis driven analyses. A primary advantage in using functional data analysis (FDA) is the ability to assess continuous data that change over time and it represents each curve that is given by a function. Our study provides us with important information about the relationship among gender, age, and average cholesterol level of individuals.

5.3 Data Source

The data utilized in the present study was made available by the inter-university Consortium for Political and Social Research and the data for the National Health and Nutrition Examination Survey (NHANS) II, 1976-1980: Serum Cholesterol was originally collected by United States Department of Health and Human Services.

NHANS II was conducted on a nationwide with approximately 28,000 persons. The target population for the survey was the civilian non-institutionalized population of the United States. The NHANES II serum cholesterol data files contain two parts of the extensive data available. One part consists of the demographic information obtained from household interviews and the other part is laboratory results. The survey started in February 1976 and was completed in February 1980. Samples were selected so that certain population groups thought to be at high risk of malnutrition (person with low incomes, preschool children, and the elderly) were oversampled. Adjusted sampling weights were then conducted for persons over the age of 76, sex, and race categories in order to inflate the sample in such a manner as to closely reflect the estimated civilian non-institutionalized U.S. population.

In addition to the general examination components, several more detailed examinations were performed on subsamples of the population. Our study includes 11,864 persons for SCL cases with 9,602 males and 2,262 females. The information relating to SCL in NHANS II survey considered codes 355-357. Primary site codes were 1 and 2 for males and females, respectively. The following diagram illustrates the data that we are using in the present study.

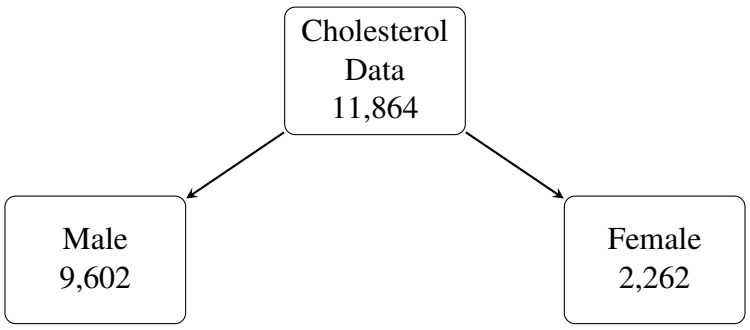


Figure 27.: Number of Individuals in the Study by Gender

5.4 Statistical Model

The functional data and parameters are usually assumed to be, at least implicitly, smooth or regular. This implies that there exist certain numbers of derivatives which are sufficiently smooth. The first step in functional data analysis (FDA) is to transform the available time series data into functions. The functions are represented as a linear combination of a set of basis functions $\phi_k, k = 1, 2, ..., K$ which are considered as a set of functional building blocks. The function $x(t)$ defined in this manner is expressed in mathematical notation as

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = c'\phi(t), \tag{5.1}$$

and is called a basis function expansion. The parameters $c_1, c_2, ..., c_k$ are the coefficients of expansion. The matrix expression in the last term of equation (5.1) uses c' to stand for the vector of K coefficients and ϕ denotes a vector of length K containing the basis functions. We have considered Fourier basis system during our modeling but we did not obtain satatisfactory results. As the number of basis functions increases the model can yield a better fit to the data.

The smoothing process is a statistical procedure that reduces the variance of the data without significantly affecting the structure of the data. Functions can be smoothed by minimizing the number of basis functions using the roughness penalty approach. This approach uses a large number of basis function but at the same time smoothness using penalizing process results in difficulty to measure the function complexity. More basis functions relative to data points can cause over fit whereas less basis functions may have a risk of losing functional features. In the present study we have smoothed the data using a harmonic acceleration roughness penalty that penalizes departures from a shift given by

$$x(t) = c_1 + c_2 Sin(2\pi t/365) + c_3 Cos(2\pi t/365), \tag{5.2}$$

where c_1, c_2, c_3 are Fourier coefficients.

A multiple of the smoothness penalty is applied to the error sum of squares (SSE) to produce the desired smooth curve. The generalized cross validation measure (GCV) developed by Craven and Wahba, [101], is designed to locate the best value for smoothing the parameter λ which is used to specify the degree of penalty. The fit of the function to the observed data improves as λ approaches zero. In our study, we have obtained a $\lambda = 0.1$. The selection of value of λ is based on achieving mean square error. Thus, the generalized cross validation measure is given by

$$GCV = (\frac{n}{n - df(\lambda)})(\frac{SSE}{n - df(\lambda)}). \tag{5.3}$$

Here λ is a smoothing parameter. This model has been selected based on root mean square values. The less the root mean square the better the fit. The selected model has root mean square of 1.7. We consider **penalized square error (PENSSE)** that will minimize the error so as to obtain a better fit function is given by

$$PENSSE = \Sigma_{i=1}^{n}(y_i - x(t_i)^2 + \lambda \int [D^2 x(t)]^2 dt. \tag{5.4}$$

Here $x(t)$ measures the roughness of x, λ is a continuous tuning (smoothing) parameter that trades off the fit to y_i and roughness and it should be chosen appropriately, and D is a differential operator. As λ increases, roughness is increasingly penalized and x(t) becomes smooth and vice versa.

5.5 Result

Cholesterol level by sex and race was investigated. Distinctively different cholesterol levels may have important implications in identifying the risk factors associated with it and its treatment. Pairwise comparisons was performed to determine if the mean response of each group is the same non-parametrically. The mean cholesterol levels of males and females

for overall behaviors were investigated. We used the Mann-Whitney-Wilcoxon, a non parametric test in order to identify whether or not the males and females cholesterol levels are the same. The hypothesis was rejected at $\alpha = 5\%$ level of significant that male and female cholesterol levels were coming from two identical populations (p-value=4.8e-09). Since non parametric statistical test suggests that male and female cholesterol data are from two different populations, it was therefore, crucial to investigate males and females separately.

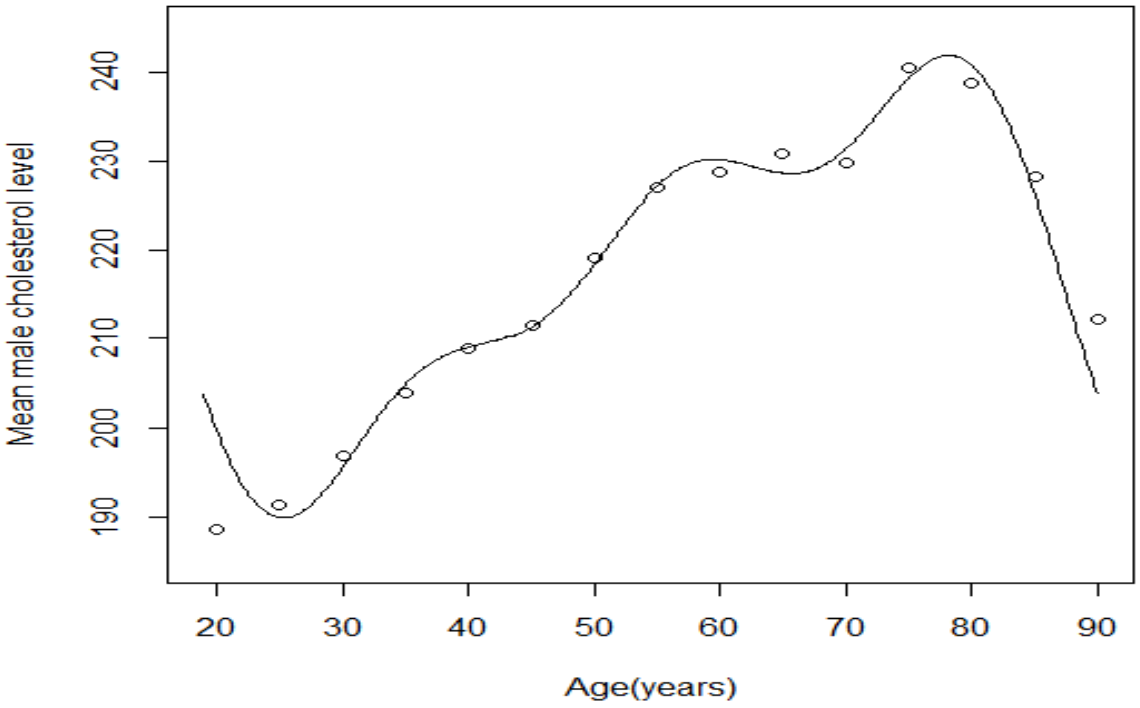


Figure 28.: Mean Cholesterol Level of Male with Respect to Ages (Years)

Figure (28), depicts the average cholesterol level of males that increases as age increases until the age of 60 years old and then decreases till the age of 70 years old. It again increases at the age of 80 years old. Notice that cholesterol levels seem to be increasing linearly from the late teen to 80 years of age. We also observed that after the age of 80, cholesterol levels are unstable and seems to decrease significantly to nearly 210 gm/dl. It appears that the cholesterol level seems to decrease rapidly for 80 years and above.

Figure (29) explains that the average cholesterol level behavior of females seems to fol-

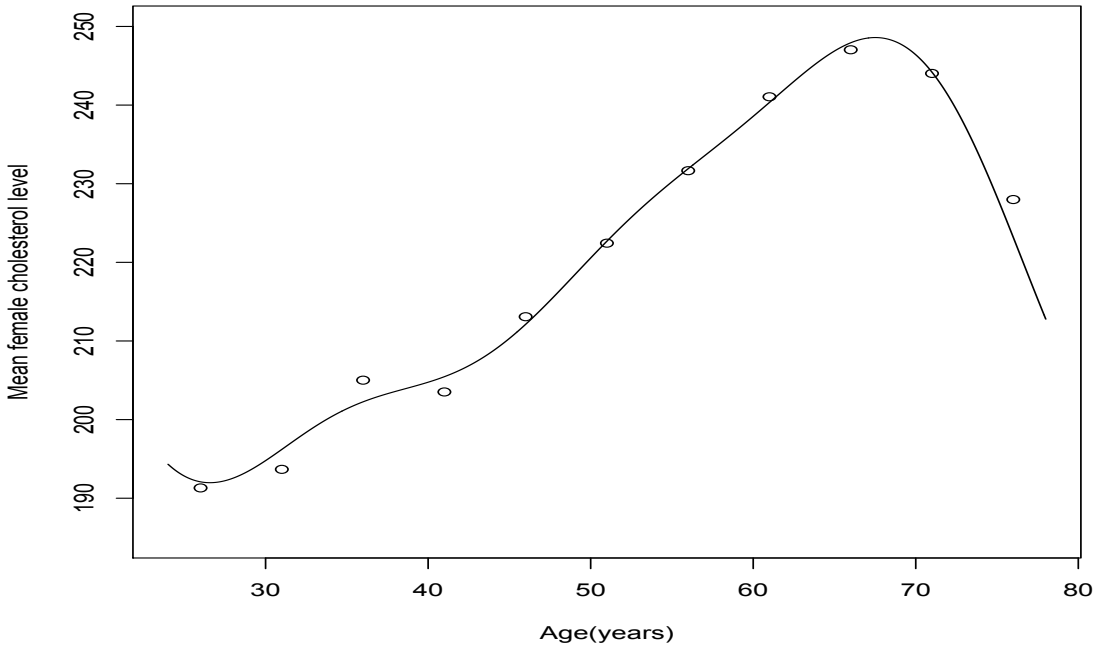


Figure 29.: Mean Cholesterol Level of Female with Respect to Ages (Years)

low the same pattern as males but at a higher rate. The average cholesterol level increases as age increases up to age of 60 years old and then decreases. It again increases slightly after the age of 80 years and then decreases continuously. The cholesterol level for females appears to be lower than that of males after the age of 80 years on the average. Average cholesterol levels of females appears to increase linearly from the late teen up to the age of 60. The average cholesterol levels for females appear to be volatile after the age of 70 years old. We also observed that the average cholesterol level changes slightly in males as well as females at the age of late 40's or early 50's years of age. However, it starts increasing afterwards indicating some kind of association with age in general.

5.6 Discussion

The cholesterol levels for both males and females seems to be well within the limits with respect to the health of an individual within the age of 30 years old. An individual is more

active physically and mentally might have played important role to maintain a balanced cholesterol despite the disparity of their dietary habits for males and females. We have observed that the cholesterol level becomes more volatile as people get older, most notably above the age of 60 years old. It has been found that cholesterol level develops from late teens to mid-40s [62, 63] in an individual. The risk of high cholesterol and heart disease increases with age, and men over 45 years old are most at risk. We have observed that the average cholesterol levels for males and females are significantly different throughout the ages, [63]. The higher the cholesterol level an individual has, the greater the risk of developing heart disease like coronary heart disease, heart attack, and stroke. The effect of cholesterol level is extensively discussed in a well-known study called Framingham Study, [64]. Since the average cholesterol level of females appears to be relatively more than males, it might be possible that females are more at risk in developing disease related to cholesterol levels. However, researchers have noticed that gender-related changes in insulin resistance is important in adolescents, and found that insulin resistance was more frequent in males than females. The insulin resistance was associated with a decrease in good cholesterol and an increase in triglycerides. This makes males more vulnerable than females.

Any recommendation based on males information can not be generalized to females and vice-versa. This study may assist in developing guidelines which will better provide physicians and public health practitioners, when determining how to treat their patients, dependent upon gender and age. The results of this study will inform individuals in preventing premature deaths, [49], possibly due to heart attacks and strokes. In our study we noticed the significant difference between the average cholesterol levels of males and females, that is also supported by [102]. More specific detailed study is required to identify the specific causes that lead to differences in cholesterol level in males and females. Since cholesterol level is highly dependent on eating habits, attitude, and life style, more careful study about these factors are essential to understand their actual associated risks, that were also stated

by [65, 66]. It is also recommended that we should pay attention to what we eat, especially our choice of oils and fats, watch our weight, and stay physically active. The inferences made without taking these factors into account may lead to misleading results and can have serious consequences, [67]. Also, data is needed to study the family history of heart diseases at younger age and cholesterol level as well. Our study may assist in designing of new experiments to collect a more appropriate data so as to define the attributable variables that drive the levels of cholesterol.

5.6.1 Contribution

In the present study we have presented a discussion on the role of cholesterol levels play in a person's health. More specifically, we can attest to the following.

1. Proved that male and female cholesterol levels are significantly different and must analyzed accordingly.

2. Have identified a method, harmonic acceleration procedure that uses a roughness penalty to smooth the function that characterizes the behavior of cholesterol levels as a function of age and gender.

3. Have developed a statistical model in term of basis functions that characterizes the behavior of the cholesterol level as a function of age and gender. A physician can use the basis function curve to identify his/her patients' behavior of cholesterol level as a function of his/her age and gender to take appropriate action.

Chapter 6

Future Work

In the first part of the dissertation, we have proposed the Bayesian Age - Period - Cohort Model for Lung Cancer Mortality that explains the mortality rate due to lung cancer based on age at death, period at death, and birth- cohort of an individual. It can explain the mortality for different age groups, different periods, and at the same time as birth-cohort. Histogram smoothing prior has been implemented to explain the mortality rate for different sub populations of our interest.

To understand lung cancer completely, all the possible contributing factors are essential to include in the analysis. We have intended to study lung cancer, including smoking and second hand smoking, as it is well established that more than 30% of lung cancer has been contributed by smoking. This study could provide more insight in understanding lung cancer behavior.

In conjunction to the histogram smoothing in Bayesian modeling, we intend to study different prior structure variance covariance matrix. The possible decomposition could provide more flexibility for assigning the prior and enhances the more precise estimates of parameters of interest. We have also planned to implement non parametric density estimates to assign a prior for covariance matrix. This approach could give a more accurate estimate for the parameters.

In Chapter Four, We analyzed the longitudinal data through linear mixed model approach, considering summary statistics of simple linear regression for each individual subject as a new response to understand the behavior of total cholesterol level. We intend to extend our study in non normal data as well.

It will provide more insight about the study if we can implement the analysis for non normal random and error assumption in modeling. Since the random effect follows stochastic properties, we could implement a stochastic approach to understand the longitudinal data. It would be interesting to consider a Bayesian approach of linear mixed model to study longitudinal data. Since we have observed that a handful of data missing, we intend to address missing values by implementing EM algorithm and see how the inference is affected by two different approaches. We are working on implementing the different approaches to include the missing values issue in modeling longitudinal data analysis.

Data analysis through functional data analysis is a relatively new approach. It has a lot of flexibility in modeling as it converts data into basis functions. There is an abundance of possibilities in exploring this area in data analysis. Functional data provides access to models of rates of change that can help to understand the rate of change of quantity in which we are interested. It has tremendous applications virtually in every area. All the statistical theories that have been developed in parametric analysis can possibly develop in functional data analysis. To identify the order of derivatives under a functional data approach modeling could be interesting to work on in the future.

References

[1] A. Jemal, M. J. Thun, L. A. Ries, H. L. Howe, H. K. Weir, M. M. Center, E. Ward, X.-C. Wu, C. Eheman, R. Anderson *et al.*, "Annual report to the nation on the status of cancer, 1975–2005, featuring trends in lung cancer, tobacco use, and tobacco control," *Journal of the National Cancer Institute*, vol. 100, no. 23, pp. 1672–1694, 2008.

[2] S. J. Henley, T. B. Richards, J. M. Underwood, C. R. Eheman, M. Plescia, and T. A. McAfee, "Lung cancer incidence trends among men and women–united states, 2005-2009." *MMWR. Morbidity and mortality weekly report*, vol. 63, no. 1, pp. 1–5, 2014.

[3] B. P. Tharu, R. C. Kafle, and C. P. Tsokos, "Bayesian age-period-cohort model of lung cancer mortality," *Epidemiology, Biostatistics and Public Health*, vol. 12, no. 3, 2015.

[4] B. P. Tharu and C. P. Tsokos, "Modeling serum cholesterol level by functional data analysis approach," *Dynamic Systems and Application*, no. 7, pp. 1–5, 2016.

[5] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 583–639, 2002.

[6] H. Akaike, "Maximum likelihood identification of gaussian autoregressive moving average models," *Biometrika*, vol. 60, no. 2, pp. 255–265, 1973.

[7] U. C. S. W. Group *et al.*, "United states cancer statistics: 1999–2011 incidence and mortality web-based report," *Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute*, 2014.

[8] D. Clayton and E. Schifflers, "Models for temporal variation in cancer rates. i: Age-period and age-cohort models," *Stat Med*, vol. 6, no. 4, pp. 449–467, 1987.

[9] S. H. Jee, S. Kim, D. Shin, L. J. Appe *et al.*, "Projected mortality from lung cancer in south korea, 1980–2004," *International journal of epidemiology*, vol. 27, no. 3, pp. 365–369, 1998.

[10] D. Clayton and E. Schifflers, "Models for temporal variation in cancer rates. ii: age–period–cohort models," *Statistics in medicine*, vol. 6, no. 4, pp. 469–481, 1987.

[11] T. R. Holford, "Understanding the effects of age, period, and cohort on incidence and mortality rates," *Annual review of public health*, vol. 12, no. 1, pp. 425–457, 1991.

[12] T. R. Holford, "Analysing the temporal effects of age, period and cohort," *Statistical methods in medical research*, vol. 1, no. 3, pp. 317–337, 1992.

[13] C. Osmond, "Using age, period and cohort models to estimate future mortality rates," vol. 14, no. 1, pp. 124–129, 1985.

[14] C. Robertson and P. Boyle, "Age, period and cohort models - the use of individual records," *Statistics in Medicine*, vol. 5, no. 5, pp. 527–538, 1986.

[15] T. Hakulinen and T. Dyba, "Precision of incidence predictions based on poisson distributed observations," *Statistics in medicine*, vol. 13, no. 15, pp. 1513–1523, 1994.

[16] T. Dyba, T. Hakulinen, and L. Päivärinta, "A simple non-linear model in incidence prediction," *Statistics in medicine*, vol. 16, no. 20, pp. 2297–2309, 1997.

[17] E. Negri, C. La Vecchia, F. Levi, A. Randriamiharisoa, A. Decarli, and P. Boyle, "The application of age, period and cohort models to predict swiss cancer mortality," *Journal of cancer research and clinical oncology*, vol. 116, no. 2, pp. 207–214, 1990.

[18] L. Knorr-Held and E. Rainer, "Projections of lung cancer mortality in west germany: a case study in bayesian prediction," *Biostatistics*, vol. 2, no. 1, pp. 109–129, 2001.

[19] B. J. S. C. M. Wong, Irene OL Cowling, "Vulnerability to diabetes in chinese: an age-period-cohort analysis," *Annals of Epidemiology*, vol. 25, no. 1, pp. 34–39, 2015.

[20] R. Clèries, J. M. Martínez, J. M. Escribà, L. Esteban, L. Pareja, J. M. Borràs, and J. Ribes, "Monitoring the decreasing trend of testicular cancer mortality in spain during 2005–2019 through a bayesian approach," *Cancer Epidemiology*, vol. 34, no. 3, pp. 244–256, 2010.

[21] T. R. Holford, "The estimation of age, period and cohort effects for vital rates," *Biometrics*, pp. 311–324, 1983.

[22] R. M. O?Brien, "The age–period–cohort conundrum as two fundamental problems," *Quality & Quantity*, vol. 45, no. 6, pp. 1429–1444, 2011.

[23] "Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) seer*stat database: Mortality - all cod, aggregated with state, total u.s. (1969-2009) (katrina/rita population adjustment), national cancer institute, dccps, surveillance research program, surveillance systems branch, released april 2012. underlying mortality data provided by nchs (www.cdc.gov/nchs)."

[24] "Cancer trends progress report-2011/2012 update, national cancer institute, nih, dhhs, bethesda, md, august 2012, http://progressreport.cancer.gov."

[25] "Centers for disease control and prevention/ national center for health statistics (www.cdc.gov/nchs)."

[26] B. Carstensen, "Age–period–cohort models for the lexis diagram," *Statistics in medicine*, vol. 26, no. 15, pp. 3018–3045, 2007.

[27] V. Jürgens, S. Ess, T. Cerny, and P. Vounatsou, "A bayesian generalized age–period–cohort power model for cancer projections," *Statistics in medicine*, vol. 33, no. 26, pp. 4627–4636, 2014.

[28] T. Leonard, "A bayesian method for histograms," *Biometrika*, vol. 60, no. 2, pp. 297–308, 1973.

[29] C. Berzuini and D. , "Bayesian analysis of survival on multiple time scales," *Statistics in medicine*, vol. 13, no. 8, pp. 823–838, 1994.

[30] U. H. W. K. H. B. J. S. G. F. F. S. H. B. G. Oberaigner, Willi Siebert, "Prostate-specific antigen testing in tyrol, austria: prostate cancer mortality reduction was supported by an update with mortality data up to 2008," *International Journal of Public Health*, vol. 57, no. 1, pp. 57–62, 2012.

[31] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 9–25, 1993.

[32] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 583–639, 2002.

[33] A. Jemal, K. C. Chu, and R. E. Tarone, "Recent trends in lung cancer mortality in the united states," *Journal of the National Cancer Institute*, vol. 93, no. 4, pp. 277–283, 2001.

[34] K. Shibuya, M. Inoue, and A. D. Lopez, "Statistical modeling and projections of lung cancer mortality in 4 industrialized countries," *International Journal of Cancer*, vol. 117, no. 3, pp. 476–485, 2005.

[35] S. S. Devesa, W. J. Blot, and J. F. Fraumeni, "Declining lung cancer rates among young men and women in the united states: a cohort analysis," *Journal of the National Cancer Institute*, vol. 81, no. 20, pp. 1568–1571, 1989.

[36] H. Witschi, "A short history of lung cancer," *Toxicological Sciences*, vol. 64, no. 1, pp. 4–6, 2001.

[37] M. J. Thun, B. D. Carter, D. Feskanich, N. D. Freedman, R. Prentice, A. D. Lopez, P. Hartge, and S. M. Gapstur, "50-year trends in smoking-related mortality in the united states," *New England Journal of Medicine*, vol. 368, no. 4, pp. 351–364, 2013

[38] W. H. Organization, *The world health report 2002: reducing risks, promoting healthy life*. World Health Organization, 2002.

[39] D. M. Burns, L. Lee, L. Z. Shen, E. Gilpin, H. D. Tolley, J. Vaughn, and T. G. Shanks, "Cigarette smoking behavior in the united states," *Changes in cigarette-related disease risks and their implication for prevention and control. Smoking and Tobacco Control Monograph*, vol. 8, pp. 13–42, 1997.

[40] J. E. Harris, "Cigarette smoking among successive birth cohorts of men and women in the united states during 1900–1980," *Journal of the National Cancer Institute*, vol. 71, no. 3, pp. 473–479, 1983.

[41] R. G. Stevens and S. H. Moolgavkar, "A cohort analysis of lung cancer and smoking in british males," *American Journal of Epidemiology*, vol. 119, no. 4, pp. 624–641, 1984.

[42] N. M. Van Der Hoff, "Cohort analysis of lung cancer in the netherlands," *International Journal of Epidemiology*, vol. 8, no. 1, pp. 41–48, 1979.

[43] C. C. Brown and L. G. Kessler, "Projections of lung cancer mortality in the united states: 1985–20251," *Journal of the National Cancer Institute*, vol. 80, no. 1, pp. 43–51, 1988.

[44] L. Reis, M. Eisner, C. Kosary, B. Hankey, B. Miller, L. Clegg, and B. Edwards, *SEER Cancer Statistics Review, 1973-1997*, 2000.

[45] L. S. S. J. O. R. B. M. C. E. Presson, Clark C Chassin, "Predictors of adolescents' intentions to smoke: age, sex, race, and regional differences," *Substance Use & Misuse*, vol. 19, no. 5, pp. 503–519, 1984.

[46] A. V. Peterson, K. A. Kealey, S. L. Mann, P. M. Marek, and I. G. Sarason, "Hutchinson smoking prevention project: Long-term randomized trial in school-based tobacco use prevention?results on smoking," *Journal of the National Cancer Institute*, vol. 92, no. 24, pp. 1979–1991, 2000.

[47] E. E. W. V. Z. S. M. S. R. M. D. T. D. P. Sherman, M. P. Aeberhard, "Effects of smoking marijuana, tobacco or cocaine alone or in combination on dna damage in human alveolar macrophages," *Life Sci*, vol. 56, no. 1, pp. 2201–7, 1995.

[48] N. I. of Health, "Lowering blood cholesterol to prevent heart disease." *NIH Consens Statement Online*, vol. 5, no. 7, pp. 1–11, 1984.

[49] "The lipid research clinics coronary primary prevention trial results. i. reduction in incidence of coronary heart disease," *JAMA*, vol. 251, no. 3, pp. 351–64, 1984.

[50] "The lipid research clinics coronary primary prevention trial results. ii. the relationship of reduction in incidence of coronary heart disease to cholesterol lowering," *JAMA*, vol. 251, no. 3, pp. 365–74, 1984.

[51] S. P. Caudill, S. J. Smith, and G. R. Cooper, "Cholesterol-based personal risk assessment in coronary heart disease," *Statistics in medicine*, vol. 8, no. 3, pp. 295–309, 1989.

[52] Centers for Disease Control (CDC) and others *et al.*, "Predicting future cholesterol levels for coronary heart disease risk assessment." *MMWR. Morbidity and mortality weekly report*, vol. 38, no. 20, p. 364, 1989.

[53] X. Mu, K. Wang, T. Chai, L. Zhu, Y. Yang, J. Zhang, S. Pang, C. Wang, and X. Li, "Sex specific response in cholesterol level in zebrafish (danio rerio) after long-term exposure of difenoconazole," *Environmental Pollution*, vol. 197, pp. 278–286, 2015.

[54] S. Y. H. S. M. N. S. Y. E. Ogawa, K. Hirose, "Synthesis of oolongtheanins and their inhibitory activity on micellar cholesterol solubility in vitro," *Bioorg Med Chem Lett*, vol. 25, pp. 749–52, 2015.

[55] C. J. Rodriguez, J. Cai, K. Swett, H. M. Gonzlez, G. A. Talavera, L. M. Wruck, S. Wassertheil?Smoller, D. Lloyd?Jones, R. Kaplan, and M. L. Daviglus, "High cholesterol awareness, treatment, and control among hispanic/latinos: Results from the hispanic community health study/study of latinos," *Journal of the American Heart Association*, vol. 4, no. 7, 2015.

[56] L. C. Tan, K. Methawasin, E.-K. Tan, J. H. Tan, W.-L. Au, J.-M. Yuan, and W.-P. Koh, "Dietary cholesterol, fats and risk of parkinson's disease in the singapore chinese health study," *Journal of Neurology, Neurosurgery & Psychiatry*, 2015.

[57] B. Schfer, E. Orbn, Z. Kele, and C. Tmbly, "Tritium labelling of a cholesterol amphiphile designed for cell membrane anchoring of proteins," *Journal of Labelled Compounds and Radiopharmaceuticals*, vol. 58, no. 1, pp. 7–13, 2015.

[58] S. J. Chung, "Relationship among age, serum cholesterol level and population percentile in adults," *International journal of bio-medical computing*, vol. 31, no. 2, pp. 99–116, 1992.

[59] S. Chung, "Formulas predicting the percentile of serum cholesterol levels by age in adults." *Archives of pathology & laboratory medicine*, vol. 114, no. 8, pp. 869–875, 1990.

[60] J. Lo and C. Fung, "Basic microcomputer program for generating percentile values, based on age and serum cholesterol levels." *Archives of pathology & laboratory medicine*, vol. 115, no. 2, pp. 106–107, 1991.

[61] B. K. Lauridsen, S. Stender, R. Frikke-Schmidt, B. G. Nordestgaard, and A. Tybjærg-Hansen, "Genetic variation in the cholesterol transporter npc1l1, ischaemic vascular disease, and gallstone disease," *European heart journal*, p. ehv108, 2015.

[62] D. S. Freedman, C. L. Shear, S. R. Srinivasan, L. S. Webber, and G. S. Berenson, "Tracking of serum lipids and lipoproteins in children over an 8-year period: the bogalusa heart study," *Preventive medicine*, vol. 14, no. 2, pp. 203–216, 1985.

[63] R. P. L. H. H. P. N. D. A. L. Orchard, T. J.Donahue, "Cholesterol screening in childhood: does it predict adult hypercholesterolemia? the beaver county experience," *J Pediatr*, vol. 103, pp. 687–91, 1983.

[64] W. B. Kannel, W. P. Castelli, T. Gordon, and P. M. McNamara, "Serum cholesterol, lipoproteins, and the risk of coronary heart disease: the framingham study," *Annals of Internal Medicine*, vol. 74, no. 1, pp. 1–12, 1971.

[65] J.-P. Liu, Y. Tang, S. Zhou, B. H. Toh, C. McLean, and H. Li, "Cholesterol involvement in the pathogenesis of neurodegenerative diseases," *Molecular and Cellular Neuroscience*, vol. 43, no. 1, pp. 33–42, 2010.

[66] L. De Lau, M. Bornebroek, J. Witteman, A. Hofman, P. Koudstaal, and M. Breteler, "Dietary fatty acids and the risk of parkinson disease the rotterdam study," *Neurology*, vol. 64, no. 12, pp. 2040–2045, 2005.

[67] R. Knopp, P. Paramsothy, B. Retzlaff, B. Fish, C. Walden, A. Dowdy, C. Tsunehara, K. Aikawa, and M. Cheung, "Sex differences in lipoprotein metabolism and dietary response: Basis in hormonal differences and implications for cardiovascular disease," *Current Cardiology Reports*, vol. 8, no. 6, pp. 452–459, 2006.

[68] N. Lange and L. Ryan, "Assessing normality in random effects models," *The Annals of Statistics*, pp. 624–642, 1989.

[69] B. T. West, "Analyzing longitudinal data with the linear mixed models procedure in spss," *Evaluation & the health professions*, vol. 32, no. 3, pp. 207–228, 2009.

[70] R. Wolfinger, "Covariance structure selection in general mixed models," *Communications in Statistics - Simulation and Computation*, vol. 22, no. 4, pp. 1079–1106, 1993.

[71] B. Engel and A. Keen, "A simple approach for the analysis of generalizea linear mixed models," *Statistica neerlandica*, vol. 48, no. 1, pp. 1–22, 1994.

[72] D. Zhang and M. Davidian, "Linear mixed models with flexible distributions of random effects for longitudinal data," *Biometrics*, vol. 57, no. 3, pp. 795–802, 2001.

[73] G. Verbeke and E. Lesaffre, "The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data," *Computational Statistics & Data Analysis*, vol. 23, no. 4, pp. 541–556, 1997.

[74] E. L. Geert Verbeke, "A linear mixed-effects model with heterogeneity in the random-effects population," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 217–221, 1996.

[75] R. Wolfinger and M. O'connell, "Generalized linear mixed models a pseudo-likelihood approach," *Journal of statistical Computation and Simulation*, vol. 48, no. 3-4, pp. 233–243, 1993.

[76] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White, "Generalized linear mixed models: a practical guide for ecology and evolution," *Trends in ecology & evolution*, vol. 24, no. 3, pp. 127–135, 2009.

[77] R. B. D. M. Lindstrom, Mary J an "Newton?raphson and em algorithms for linear mixed-effects models for repeated-measures data," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1014–1022, 1988.

[78] J. D. Hadfield *et al.*, "Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package," *Journal of Statistical Software*, vol. 33, no. 2, pp. 1–22, 2010.

[79] S. L. Zeger and K.-Y. Liang, "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, pp. 121–130, 1986.

[80] T. E. MaCurdy, "The use of time series processes to model the error structure of earnings in a longitudinal data analysis," *Journal of econometrics*, vol. 18, no. 1, pp. 83–114, 1982.

[81] J. Fan and R. Li, "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 710–723, 2004.

[82] M. Petri, C. Lakatta, L. Magder, and D. Goldman, "Effect of prednisone and hydroxychloroquine on coronary artery disease risk factors in systemic lupus erythematosus: a longitudinal data analysis," *The American journal of medicine*, vol. 96, no. 3, pp. 254–259, 1994.

[83] J. H. W. Nan M. Laird, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[84] Y. Fujikoshi and C. R. Rao, "Selection of covariables in the growth curve model," *Biometrika*, vol. 78, no. 4, pp. 779–785, 1991.

[85] A. Cnaan, N. Laird, and P. Slasor, "Tutorial in biostatistics: Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data," *Stat Med*, vol. 16, pp. 2349–2380, 1997.

[86] K. M. Anderson, W. P. Castelli, and D. Levy, "Cholesterol and mortality: 30 years of follow-up from the framingham study," *Jama*, vol. 257, no. 16, pp. 2176–2180, 1987.

[87] "Who cooperative trial on primary prevention of ischaemic heart disease with clofibrate to lower serum cholesterol: final mortality follow up: report of the committee of principle investigators," *The Lancet*, vol. 324, no. 8403, pp. 600 – 604, 1984, originally published as Volume 2, Issue 8403.

[88] B. E. Kreger, K. M. Anderson, A. Schatzkin, and G. L. Splansky, "Serum cholesterol level, body mass index, and the risk of colon cancer. the framingham study," *Cancer*, vol. 70, no. 5, pp. 1038–1043, 1992.

[89] W. P. Castelli and K. Anderson, "A population at risk: prevalence of high cholesterol levels in hypertensive patients in the framingham study," *The American journal of medicine*, vol. 80, no. 2, pp. 23–32, 1986.

[90] T. Gordon, W. P. Castelli, M. C. Hjortland, W. B. Kannel, and T. R. Dawber, "High density lipoprotein as a protective factor against coronary heart disease: the framingham study," *The American journal of medicine*, vol. 62, no. 5, pp. 707–714, 1977.

[91] S. M. Butler and T. A. Louis, "Random effects models with non-parametric priors," *Statistics in medicine*, vol. 11, no. 14-15, pp. 1981–2000, 1992.

[92] P. W. F. Wilson, K. M. Anderson, T. Harri, W. B. Kannel, and W. P. Castelli, "Determinants of change in total cholesterol and hdl-c with age: The framingham study," *Journal of Gerontology*, vol. 49, no. 6, pp. M252–M257, 1994.

[93] W. B. Kannel, W. P. Castelli, and T. Gordon, "Cholesterol in the prediction of atherosclerotic diseasenew perspectives based on the framingham study," *Annals of Internal Medicine*, vol. 90, no. 1, pp. 85–91, 1979.

[94] S. S. S. S. Group *et al.*, "Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival study (4s)," *The Lancet*, vol. 344, no. 8934, pp. 1383–1389, 1994.

[95] G. Jarvik, E. M. Wijsman, W. Kukull, G. Schellenberg, C. Yu, and E. Larson, "Interactions of apolipoprotein e genotype, total cholesterol level, age, and sex in prediction of alzheimer's disease a case-control study," *Neurology*, vol. 45, no. 6, pp. 1092–1096, 1995.

[96] J. Ramsay and B. Silverman, "Functional data analysis," 2005.

[97] J. O. Ramsay and C. Dalzell, "Some tools for functional data analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 539–572, 1991.

[98] J. Ramsay and B. Silverman, "Functional data analysis, second edition." *Springer*, 2005.

[99] J. Ramsay and B. Silverman, *Applied Functional Data Analysis.* Springer,New York, 2002.

[100] W. Kriengsinyos, A. Wangtong, and S. Komindr, "Serum cholesterol reduction efficacy of biscuits with added plant stanol ester," *Cholesterol*, vol. 2015, 2015.

[101] P. Craven and G. Wahba, "Smoothing noisy data with spline functions," *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1978.

[102] P. Bandosz, M. O'Flaherty, M. Rutkowski, C. Kypridemos, M. Guzman-Castillo, D. O. Gillespie, B. Solnica, M. J. Pencina, B. Wyrzykowski, S. Capewell *et al.*, "A victory for statins or a defeat for diet policies- cholesterol falls in poland in the past decade: A modeling study," *International journal of cardiology*, vol. 185, pp. 313–319, 2015.