ABSTRACT

Title of dissertation: ANALYSIS OF MODELS FOR EPIDEMIOLOGIC AND SURVIVAL DATA

Jiraphan Suntornchost,
Doctor of Philosophy, 2012

Dissertation directed by: Professor Eric V. Slud
Department of Mathematics

Mortality statistics are useful tools for public-health statisticians, actuaries and policy makers to study health status of populations in communities and to make plans in health care systems. Several statistical models and methods of parameter estimation have been proposed. In this thesis, we review some benchmark mortality models and propose three alternative statistical models for both epidemiologic data and survival data.

For epidemiologic data, we propose two statistical models, a Smoothed Segmented Lee-Carter model and a Smoothed Segmented Poisson Log-bilinear model. The models are modifications of the Lee-Carter (1992) model which combine an age segmented Lee-Carter parameterization with spline smoothed period effects within each age segment. With different period effects across age groups, the two models are fitted by maximizing respectively a penalized least squares criterion and a penalized Poisson likelihood. The new methods are applied to the 1971-2006 public-use mortality data sets released by the National Center for Health Statistics (NCHS).

Mortality rates for three leading causes of death, heart diseases, cancer and accidents, are studied.

For survival data, we propose a phase type model having features of mixtures, multiple stages or "hits", and a trapping state. Two parameter estimation techniques studied are a direct numerical method and an EM algorithm. Since phase type model parameters are known to be difficult to estimate, we study in detail the performance of our parameter estimation techniques by reference to the Fisher Information matrix. An alternative way to produce a Fisher Information matrix for an EM parameter estimation is also provided. The proposed model and the best available parameter estimation techniques are applied to a large SEER 1992-2002 breast cancer dataset.

ANALYSIS OF MODELS FOR EPIDEMIOLOGIC AND SURVIVAL DATA

by

Jiraphan Suntornchost

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Dr. Eric V. Slud, Chair/Advisor
Dr. Mei-Ling Ting Lee
Dr. Paul Smith
Dr. Rong Wei
Dr. Grace Yang

Dedication

This thesis is dedicated to my parents

Kumron Suntornchost

and

Kaneung Suntornchost

Acknowledgments

I owe my deepest gratitude to all the people who have made this thesis possible.

First and foremost, I would like to gratefully and sincerely thank Dr. Eric Slud for his excellent guidance, supervision, caring, and patience. He has always made himself available for help and advice through my whole doctoral study. His expertise and his suggestions have strengthened my knowledge and skills in statistics. He also carefully read and improved all of my writings not only the thesis, but also research manuscripts and my CV. I also would like to express my deep and sincere appreciation to his generosity during this year of his sabbatical year that he has sacrificed his personal time and made regular trips to the university for our meetings. Without his support, I would not able to make it this far. I hope that one day I would become as good an advisor to my students as Dr. Eric Slud has been to me.

I would also like to thank Dr. Rong Wei, my mentor at the National Center for Health Statistics, for her support, her guidance and thoughtful suggestions through my two years training at the National Center for Health Statistics. My thanks also go to Dr. Paul Smith and Meena Khare for introducing me to the National Center for Health Statistics which leads me to an exceptional experience. I also would like to thank the dissertation committee members, Drs. Paul Smith, Mei-Ling Ting Lee, Rong Wei and Grace Yang for their suggestions and comments on the thesis that improve the thesis significantly.

I would also like to thank Dr. Partha Lahiri for his support and his guidance during the final year of my Doctoral study. He introduced a new research area,

Small Area Estimation, to me which broadens my knowledge in statistics. I have learnt a lot from him during this final year of my doctoral study.

I owe my deepest thanks to my family in Thailand- my mother, father and my sister who have always stood by me and encourage me through my career. My family have been the greatest motivation for me to complete this Ph.D. Without their support, I would have given up many times. They have been cheering me up and encouraging me whenever I faced difficulties of life during many years being away from home.

I would like to acknowledge financial support from the Thailand's government under the project "The Higher Educational Strategic Scholarship for Frontier Research Network", and financial support during this final year of my doctoral study from the Department of Mathematics, University of Maryland .

Table of Contents

List of Tables

List of Figures

List of Notations and Abbreviations

Notations

$\rightsquigarrow$	converges in distribution
$\rightarrow_p$	converges in probability
i.i.d.	independent identically distributed
$X_n = o_P(1)$	a sequence of random vectors $\{X_n\}_{n\in\mathbb{N}}$ converges to zero in probability
$X_n = o_{a.s.}(1)$	a sequence of random vectors $\{X_n\}_{n\in\mathbb{N}}$ converges to zero almost surely
$X_n = O_P(1)$	a sequence of random vectors $\{X_n\}_{n\in\mathbb{N}}$ is bounded in probability, ($X_n = O_P(1)$ if $\forall\epsilon > 0 \ \exists C < \infty \ : \quad P(\lvert X_n\rvert > C) < \epsilon$ where C may depend on all other quantities and parameters in the expression X_n)

Abbreviations

APH	Acyclic Phase-Type
EM	Expectation-Maximization
HP model	Heiligman-Pollard eight parameter model
LC model	Lee-Carter model
MLE	Maximize Likelihood Estimate
OU	Ornstein-Uhlenbeck
PB model	Poisson Log-bilinear model
PH	Phase-Type
SLC model	Segmented Lee-Carter model
SPB model	Segmented Poisson Log-bilinear model
SSLC model	Smoothed Segmented Lee-Carter model
SSPB model	Smoothed Segmented Poisson Log-bilinear model
SVD	Singular Value Decomposition

Chapter 1

Introduction

Mortality statistics provide a useful basis for public health statisticians, actuaries and policy makers to study health status of populations in communities and to make plans in health care systems. Several statistical models and parameter estimations have been proposed, for instance, the Gompertz model, Heligman-Polland model, Lee-Carter model, Multihit models and First-Hitting time models. In Chapter 2, we review some of their important features, variants and applications.

One of the best known models is the mortality model proposed by Lee and Carter in 1992. The model was originally proposed for modeling and forecasting U.S. mortality. Since the Lee-Carter model is relatively simple and performs well in many applications, it has drawn interest from demographers and epidemiologists and has been a benchmark in modeling national mortality data in many countries worldwide. For instance, Wilmoth (1998) applied the model to Japanese mortality data for the period 1951-1995; Brouhns, Denuit and Vermunt (2002) applied the model to Belgian mortality data for the period 1960-1998; Lundstrom and Qvist (2004) fitted the model to Swedish mortality data for the period 1901-2001; and Booth and Tickle (2003) fitted the model to Australian mortality data for the period 1968-2000. In Chapter 2, we give extensive background references on modifications of the Lee-Carter model and its variants in techniques for parameter estimation.

Even though the Lee-Carter model fits well in many applications, we found that the Lee-Carter model's property of having a common time trend among different age groups is not suitable to some applications, for instance, with U.S. cause specific mortality data, we found that time trend varies by age groups. Therefore, in this thesis, we propose a modification of the Lee-Carter model for cause specific mortality data which combines an age segmented Lee-Carter model with spline smoothed period effects within each age segment. With different period effects across age groups, two parameter estimation methods are studied: respectively based on a penalized least squares criterion and a penalized Poisson likelihood.

In Chapter 3, we explore the feasibility of our age segmentation idea for cause-specific mortality by using the 1971-2006 public use mortality data sets released by the National Center for Health Statistics (NCHS). Mortality rates for three leading causes of death, heart diseases, cancer and accidents, are studied. The singular value decomposition technique in the original Lee-Carter paper is replaced by penalized least squares parameter estimation in this chapter. Our study suggests advantages of the age segmented model over the original Lee-Carter model. To increase the efficiency of our segmented model, we also propose two methods of age-group segmentations in Chapter 3 by applying techniques from clustering analysis. In this chapter, we further study properties of parameter estimates by a bootstrap method. The bootstrap is a simulation technique proposed by Efron in 1979 and has been used for many purposes, such as bias reduction and variance estimation and point-wise confidence interval construction. Although the bootstrap is known to be a computer intensive technique, it is very useful in the situation when theoretical cal-

culation of parameter estimates is too complex, as in the situation of the Lee-Carter model and its variants (Brouhns et al., 2005). In this chapter, we apply a Poisson bootstrap in comparing the original model and our proposed model, with detailed graphical results shown in Chapter 10.

While our study in Chapter 3 shows that the age-segmented model improves the original Lee-Carter model in capturing time trends for cause-specific mortality, we further explore the segmentation concept for age by sex mortality in Chapter 4. According to Alho (2000), the least squares Lee-Carter model is not quite suited to mortality data because the errors are assumed to be homoskedastic. Moreover, since the number of deaths follows a count random variable, the Poisson distribution is shown to be suited well to mortality analyses (Brillinger 1986 , Brouhns et al., 2005). Therefore, in this chapter, we apply a penalized Poisson likelihood method of parameter estimation instead of the penalized least squares used in Chapter 3. In Chapter 5, we discuss alternative methods of variance estimation and confidence interval construction and future research directions on the SSLC and SSPB models. We further study asymptotic properties of parameter estimates obtained from a penalized likelihood parameter estimation and a bootstrap of the penalized likelihood method by specializing theorems of Pakes and Pollard (1989) and Chen et al. (2003) in Chapter 9.

In Chapter 6, we study properties and applications of a phase type family to survival models. The phase type distribution is defined as the first hitting time distribution of a Markov process, which was introduced by Neuts in 1975 as a generalization of the Erlang distribution. The family of phase-type distributions

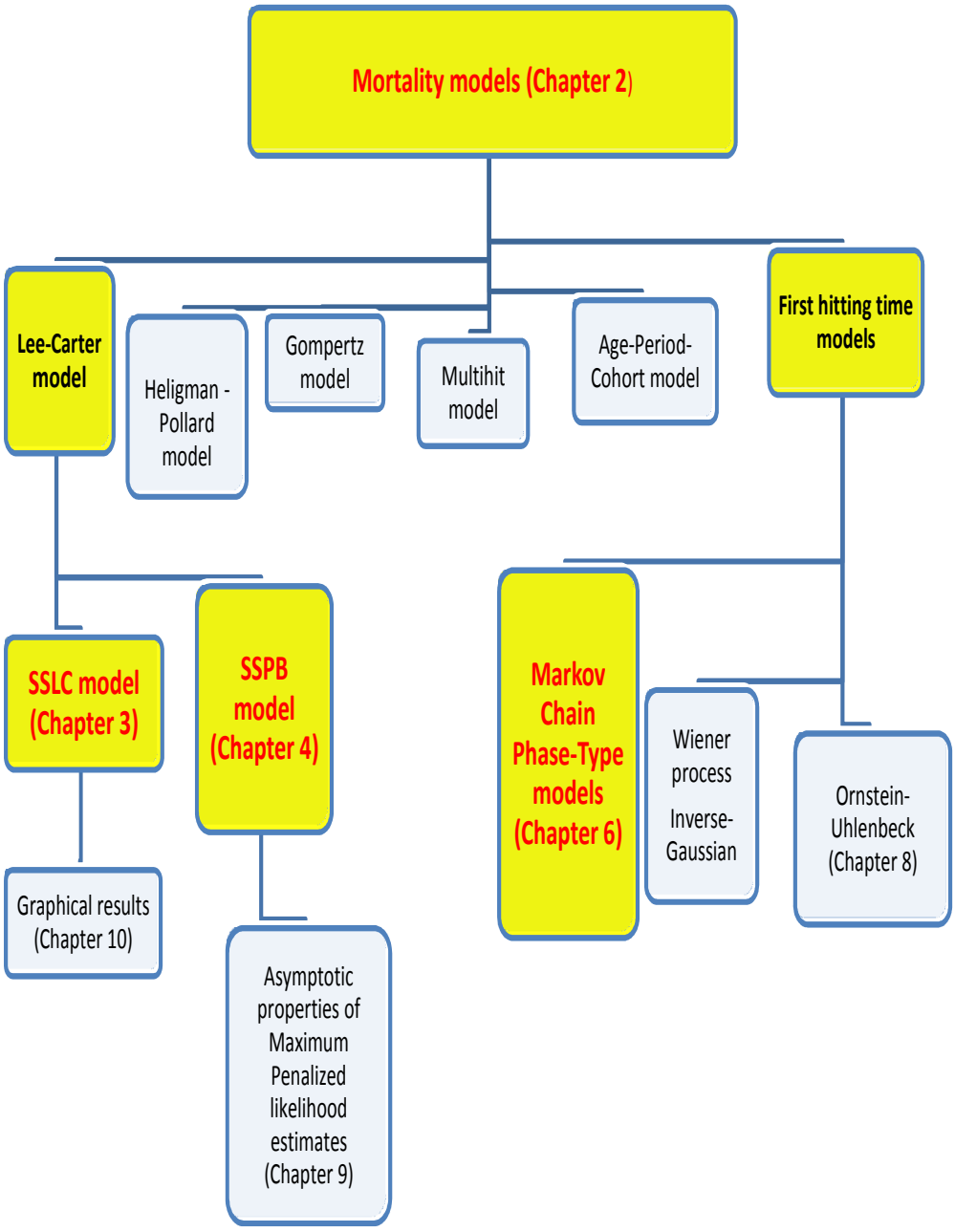
is known to be dense among all continuous distributions on the positive real line. Because the phase type distributions are mathematically tractable and have several useful features, they are widely applied in many fields of study, such as, in health care (Faddy and McClean 1999, Fackrell 2009, and Garg et al. 2011) and survival analysis (Aalen 1995, Olsson 1996).

Although the family of phase-type distributions is mathematically tractable, it is known that phase-type distributions do not have a unique representation (O'Cin neide, C.A. 1989) and the phase-type distributions are often over-parameterized. Parameter estimation of phase type distributions is difficult in practice. Therefore, many authors propose to restrict the phase type family to some subclasses of phase type distributions in order to avoid the problem. For example, Bobbio et al.(2003) restricted the phase type distributions to the subclass of Acyclic Phase Type (APH) distributions. Many parameter estimation methods have been proposed. One of the attractive parameter estimation methods is an EM algorithm proposed by Asmussen et al. (1996). In Chapter 6, we propose a phase type model for survival data having features of mixtures, multiple stages or "hits", and a trapping-state. Efficiencies of the Asmussen EM algorithm and a direct Newton-Raphson optimization method in phase type parameter estimation are studied by examining to the Fisher Information matrix. In this chapter, we also provide an alternative way to produce a Fisher Information matrix for an EM parameter estimation. The proposed model and the best parameter estimation are then applied to a large SEER 1992-2002 breast-cancer dataset.

Our new contributions to mortality statistics are distributed throughout this

thesis. In Chapter 3 we propose a new modification of the well known Lee-Carter model which is shown to have advantage over the original model in capturing time trends for U.S. cause specific mortality data. In this Chapter, we also provide a parameter estimation method and propose systematic methods for age group clustering. In Chapter 4, we extend our model to a penalized Poisson likelihood method and compare the two models. In Chapter 5, we suggest alternative methods of variance estimation and confidence interval construction. In Chapter 6, we propose a subclass of phase type models and propose a new method to estimate the Fisher Information matrix for EM parameter estimates in our proposed phase type model. This method is an application of the Oakes's formula (Oakes, 1999) for estimating Fisher Information matrix from general EM parameter estimates and a Runge Kutta numerical method. In Chapter 9, we specialize Consistency and Asymptotically normality conditions of Pakes and Pollard's conditions (1989) and a normality conditions for a nonparametric bootstrap estimates of Chen et al. (2003) to penalized likelihood parameter estimates.

Outline of Thesis



Chapter 2

Introduction to Mortality Models

Mortality data collected over time allow public health statisticians to provide current age specific mortality snapshots and also to model trends in age specific mortality. Forecasts of mortality trends, including trends in mortality from specific diseases, can play an important role in anticipating future costs and demands in the health care system. Several statistical modeling and estimation techniques have been developed to meet this challenge, for example, the Gompertz and Makeham models, the Heligman-Polland model, the Lee-Carter model, multihit models and first hitting time models. Recent discussions and reviews of mortality models are available in Alho et al. (2005), Schoen (2006), Girosi and King (2008) and Booth and Tickle (2008). In this section, we discuss some benchmark models that play important roles in mortality statistics. Among of the models discussed in this chapter, we will study modifications of Lee-Carter models in Chapters 3 and 4 and study phase type models in Chapter 6. To begin, we provide some common notations and concepts of mortality statistics.

2.1 Basic Notations and Concepts

Let T be a non-negative random variable, the waiting time until occurrence of an event of interest, such as death. The *survival function*, the probability that an

individual survives after a certain time t , is defined as

$$S(t) = P(T \geq t).$$

If T is a continuous random variable with a density f, then the survival function is the complement of the *cumulative distribution function* F which is also defined as

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(s)ds, \tag{2.1.1}$$

where f is the density function of T. Therefore $f(t) = -\frac{d}{dt}S(t)$.

Another quantity of interest in mortality models is the *hazard intensity*, which is also known as *the mortality rate* or *force of mortality* in demography and *age-specific failure rate* in epidemiology. The hazard function is defined as

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T \leq t + \delta | T \geq t)}{\delta}.$$

For a continuous random variable T with a density, the hazard function is

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln(S(t)). \tag{2.1.2}$$

The corresponding *cumulative hazard function* is defined by

$$H(t) = \int_0^t h(t)dt = -\ln(S(t)). \tag{2.1.3}$$

2.2 Life Table

A life table, or mortality table, is a table describing age-specific mortality rates and surviving rates of a population. There are two primary types of life table: (1) cohort life table representing mortality statistics of a particular birth cohort; (2)

period (or static) life table representing mortality statistics at a specific year or a short period of time. To produce a complete cohort life table, all individuals in the group of study are followed until death, which requires many years of study. Therefore the cohort life table is not feasible in general practice due to incompleteness of the dataset. In contrast, a period (static) life table provides mortality statistics based on what happened at a specific time assuming that the individuals who died within the year of study were followed from birth. It is more feasible to complete a period life table than a cohort table because it does not require long term study. Therefore, life tables in the statistics and demography literature generally refer to period life tables unless specifically cited as cohort life tables.

To construct a life table, we begin with an initial population of size l_0, "the life table radix", which is a large number and usually be set to 100,000.
The number of survivors at age x ($x = 1, 2, \ldots$), l_x, is defined as

$$l_x = l_{x-1}(1 - q_{x-1}),$$

where q_x is the probability of dying at age x, defined as the ratio of the number of deaths at age x to the number of survivors at age x,

$$q_x = \frac{d_x}{l_x}.$$

The number of person years lived between age x and $x + 1$, L_x is defined as

$$L_x = \int_0^1 l_{x+u} du.$$

In general practice, L_x is estimated by (Arias, 2006)

$$L_x = \frac{1}{2}(l_x + l_{x+1}).$$

The total person-years lived above age x, T_x, is defined as the sum of L_y for all $y \geq x$,

$$T_x = \sum_{y=x}^{\infty} L_y.$$

The expectation of life at age x is then defined as

$$e_x = \frac{T_x}{l_x}.$$

Table 2.1 shows a sample of a life table published by the National Center of Health Statistics [Arias, 2006] presenting the quantities described above.

Table 2.1: Life Table for the total population: United States, 2006

Age	Probability of dying in age $[x, x+1)$	Number surviving to age x	Number dying in age $[x, x+1)$	Person years lived in age $[x, x+1)$	Total number of person-years lived above age x	Expectation of life at age x
0-1	0.006713	100,000	671	99,409	7,770,850	77.7
1-2	0.000444	99,329	44	99,307	7,671,441	77.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
99-100	0.303810	2,494	758	2,115	6,024	2.4
100up	1.00000	1,737	1,737	3,909	3,909	2.3

Explanations of the columns in Table 2.1 are given below [Arias, 2006].

Column 1, Age [x to $x+1$), shows the age interval between the two ages (as

integers).

Column 2, Probability of dying (q_x), shows the probability of dying between ages x and $x+1$. This column forms the basis of life table calculations in other consequent columns.

Column 3, Number surviving (l_x), shows the number of persons from the origin of l_0 (100000) who survives to the beginning of each age interval.

Column 4, Number of dying (d_x), shows the number of dying in each age interval out of the origin l_0 of lives.

Column 5, Person years lived (L_x), shows the total time lived between the indicated birthdays by all those who reach the earlier birthday.

Column 6, Total number of person years lived (T_x), shows the total number of person-years that would lived after the beginning of the age interval x to $x+1$.

Column 7, Expectation of life (e_x), shows the average number of years remaining to be lived of those surviving to age x.

2.3 Gompertz model

Human mortality description in actuarial science began with life tables, the earliest of which is credited to John Graunt in 1662. Models describing the pattern of mortality by age would nowadays be described in terms of hazard rate, termed *force of mortality* by actuaries: the famous and influential model of Gompertz (1825) is an exponential

$$h(x) = bA^x$$

for hazard in the age variable x, where $b > 0$ and A is slightly greater than 1, and Makeham later (1864) added a constant term to h to allow location-scale shifts of possible limits,

$$h(x) = c + bA^x.$$

These models are presented in actuarial discussions either as formulas for hazard or for the *age-specific death-rates* $q_x = P(x < T < x + 1 \,|T \geq x)$ for one-year time intervals associated with continuous lifetime random variables T, but for most purposes these functions q_x and $h(x)$ are interchangeable except at advanced ages. For these and other historical references, see Bowers et al. (1997) and Lin and Liu (2007). The early actuarial models were intended to model qualitative features – convexly increasing shape at older ages, decreasing hazard in early childhood, the combination of which is classically called the "bathtub shape"– so as to facilitate the numerical calculation of expected present values under constant rates of compound interest.

The Gompertz model, which is closely related to the so-called *extreme value* distribution, is naturally combined with the Weibull model as being among the small class of distributions characterized by the fundamental Fisher-Tippett-Gnedenko (1927-1948) Theorem (Feller 1972, vol.2) as possible distributional limits of maxima of independent identically distributed sequences $\{X_i\}$. Brillinger (1961) highlighted the relevance of this theorem for actuarial science, including generalizations of this extreme value theory to limits of dependent and nonidentical sequences of random variables, and proposed general hazard expressions arising in this way as parametric

models for use in actuarial work.

2.4 Heligman-Pollard Eight Parameter Model (HP)

The Heligman-Pollard (1980) eight-parameter model is a generalization of the Gompertz model that allows variation of pattern of mortality curves among different age ranges: childhood, middle ages and old ages. The model is defined as

$$\frac{q_x}{1-q_x} = A^{(x+B)^C} + D\exp(-E\log^2\frac{x}{F}) + GH^x, \tag{2.4.4}$$

where x is the age variable.

The model consists of three terms representing different components of mortality. The first term, which is in exponential form represents the fall of mortality in childhood years. This term contains three parameters: A measuring the level of mortality; B a parameter accounting for infant mortality; and C measuring the rate of mortality decline in childhood. The second term represents accident mortality during middle age, which reflects "accident hump" in mortality curves. The three parameters in this term are F, E and D indicating respective location, spread and severity of the accident hump. The last term, which is the Gompertz exponential term, represents geometric increase of mortality rate in old age due to biological aging, called senescent mortality. The two parameters in this term are G, representing base level of senescent mortality, and H, representing the rate of increase of mortality rates. There are several further extensions of the HP model. For example, Kostaki (1992) extends the HP model by proposing a nine-parameter version of the HP model and Sherris and Njenga (2011) apply a Bayesian Vector Autoregressive

(BVAR) model for the parameters of the HP model to allow for dependence of parameters in the HP model. The HP model and its variants have been used in several contexts. For instance, Rogers and Gard (1991) applied the HP model to life table data for Australia, England and United States; Voulgaraki et al. (2008) applied the HP model to model mortality curve for small subpopulations; Sharrow et al. (2010) applied the HP model to study HIV mortality in South America, and Wei et al. (2011) applied the HP model to study US mortality in 1999-2001.

2.5 Age-Period-Cohort Model

The *age-period-cohort* (APC) model is a demographic mortality model that expresses the mortality rate $q_{a,p,c}$ as a superposition

$$\log(q_{a,p,c}) \;=\; \alpha_a + \beta_p + \gamma_c$$

of age effect (α_a, a function of age alone), period effect (β_p) and cohort effect (γ_c), where period refers to time at diagnosis and cohort refers to date of birth. By their definitions, the three factor-indices have linear dependence as described by $c = A - a + p$, where A is the number of age groups and a $(a = 1, ..., A)$, p $(p = 1, .., P)$, and c $(c = 1, .., C)$ denote indices of age-intervals, period intervals and cohort intervals, respectively. This relation shows that the APC model is nonidentifiable as we can

write the relation (Robertson et al. 1999)

$$\begin{aligned}\log(q_{a,p,c}) &= \alpha_a + \beta_p + \gamma_c \\ &= \alpha_a + \beta_p + \gamma_c + \lambda(A - a + p - c) \\ &= (\alpha_a + \lambda(A - a)) + (\beta_p + \lambda p) + (\gamma_c - \lambda c),\end{aligned}$$

where λ is an unidentifiable parameter. This non identifiability has led to many proposed constraints or side conditions on the three effects which can restore identifiability. For example, Clayton and Schifflers (1987) restrict the age-period-cohort models to age-period and age-cohort models. Fienberg and Mason (1979) and Holford (1983) suggest constraints $\sum_{a=1}^{A} \alpha_a = 0$, $\sum_{p=1}^{P} \beta_p = 0$ and $\sum_{c=1}^{C} \gamma_c = 0$. Rosenberg and Anderson (2010, 2012) suggest making the partition incorporate the constraint $c = A - a + p$ as the age-period form

$$\log(q_{a,p,c}) = \mu + (\alpha_L - \gamma_L)(a - \bar{a}) + \tilde{\alpha}_a + (\beta_L + \gamma_L)(p - \bar{p}) + \tilde{\beta}_p + \tilde{\gamma}_{p-a+A}$$

and the age-cohort form

$$\log(q_{a,p,c}) = \mu + (\alpha_L + \beta_L)(a - \bar{a}) + \tilde{\alpha}_a + (\beta_L + \gamma_L)(c - \bar{c}) + \tilde{\beta}_{c+a-A} + \tilde{\gamma}_c,$$

where $\tilde{\alpha}_a, \tilde{\beta}_p$, and $\tilde{\gamma}_c$ are age, period and cohort deviations, $\alpha_L + \beta_L$, $\alpha_L - \gamma_L$, and $\beta_L + \gamma_L$ are respective longitudinal age trend, cross-sectional age trend and net drift. Many other proposed solutions to achieve identifiability are studied in detail in Robertson et al. (1999) . A generalization of the APC model allowing for continuous age, period and cohort indices, where mortality rates are modeled by any function of the three effects, was proposed in Carstensen (2007), via cubic smoothing spline functions of the three indices.

2.6 The Lee-Carter model

Lee and Carter (1992) developed a method for modeling and forecasting mortality, including an age effect, a period effect, and another age by period component that explains the pattern of deviations from the main age effects through a secondary age effect multiplied by a period specific effect. The model is presented as follows.

For $a = 1, 2, 3, \ldots, A$ and $p = p_0 + 1,\ p_0 + 2,\ p_0 + 3, \ldots,\ p_0 + P$, let $\lambda_{a,p}$ denote the mortality rate from the disease of interest at age a in year p[1], where "age" indexes a single year of age, and the calender year of the observation is referred to as 'period'. The mortality rate is estimated by the proportion $\tilde{\lambda}_{a,p}$ of the observed number of deaths $D_{a,p}$ to the corresponding population size $N_{a,p}$. The LC model is defined as

$$\log(\tilde{\lambda}_{a,p}) = \alpha_a + \beta_a \gamma_p + \epsilon_{a,p}, \tag{2.6.5}$$

where $\sum_a \beta_a = 1$ and $\sum_p \gamma_p = 0$. Here α_a represents the fixed effect associated with age, β_a describes the pattern of slopes from the age profile as period p varies, γ_p is the time varying parameter and $\epsilon_{a,p}$ is the independent error term with mean 0 and variance σ_ϵ^2 independent of a and p. The Lee-Carter (LC) mortality model has been widely used worldwide since 1992. For instance, Wilmoth (1998) applied the model to Japanese mortality data for the period 1951-1995; Brouhns, Denuit and Vermunt (2002) applied the model to Belgian mortality data for the period 1960-

[1]Since a is an integer single age, this $\lambda_{a,p}$ has the same meaning as the death rate parameter q_a used in earlier sections.

1998; Lundstrom and Qvist (2004) fitted the model to Swedish mortality data for the period 1901-2001; and Booth and Tickle (2003) fitted the model to Australian mortality data for the period 1968-2000. Several variations of the LC models both in terms of parameter estimations and the features of the model itself have been proposed in the last two decades. In this section, we discuss in detail some well-known parameter estimations and modifications of the LC model. More discussion of applications of the LC model and its variants and extensions can be found in Bongaarts (2004), Booth et al. (2006), Koissi et al. (2006), Girosi and King (2007) and Booth and Tickle (2008).

2.6.1 The First Singular Value Decomposition

The first parameter estimation method to be discussed here is the first Singular Value Decomposition (SVD) which is the original estimation method used in Lee and Carter (1992). It is based on the Singular Value Decomposition Theorem presented below.

Theorem 2.1 (Singular Value Decomposition)**.** *Let* $\mathbf{A}$ *be an* $m \times n$ *matrix of rank* k*. Then there is an* $m \times m$ *orthogonal matrix* $\mathbf{U}$*,* $n \times n$ *orthogonal matrix* $\mathbf{V}$*, and an* $m \times n$ *diagonal matrix* $\mathbf{D}$ *such that*

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where the diagonal entries of $\mathbf{D}$*, called singular values of* $\mathbf{A}$*, can be arranged to be nonincreasing. The singular values are nonnegative and exactly* k *of them are strictly positive.*

Proof. See Lawson and Hanson (1974). □

The parameter estimates of α_a for $a = 1, \ldots, A$, are simply the least squares estimates $\hat{\alpha}_a = \dfrac{1}{P} \displaystyle\sum_{p=p_0+1}^{p_0+P} \big(\log(\tilde{\lambda}_{a,p})\big)$, $a = 1, \ldots, A$. The parameter estimates of $\beta_a, a = 1, \ldots, A$ and γ_p , $p = p_0 + 1, \ldots, p_0 + P$, can be obtained by applying the Singular Value Decomposition (SVD) to the matrix $\mathbf{T} = [T_{a,p}]_{a,p}$, where $T_{a,p} = \log(\tilde{\lambda}_{a,p}) - \hat{\alpha}_a$, $a = 1, \ldots, A$, $p = p_0 + 1, \ldots, p_0 + P$. According to Theorem 2.1, there is an $A \times A$ orthogonal matrix $\mathbf{U}$, an $P \times P$ orthogonal matrix $\mathbf{V}$ and an $A \times P$ diagonal matrix $\mathbf{D}$ containing singular values $\lambda_1, \ldots, \lambda_r$ of $\mathbf{T}$, where r is the rank of $\mathbf{T}$ such that

$$
\begin{aligned}
\mathbf{T} &= \mathbf{U}\mathbf{D}\mathbf{V}^T \\
&= \begin{pmatrix} u_{1,1} & \cdots & \cdots & u_{1,A} \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ u_{A,1} & \cdots & \cdots & u_{A,A} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ \cdots & \ddots & 0 & \ddots & 0 \\ \vdots & 0 & \lambda_r & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \begin{pmatrix} v_{1,1} & \cdots & \cdots & v_{P,1} \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ v_{1,P} & \cdots & \cdots & v_{P,P} \end{pmatrix} \\
&= \lambda_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots \lambda_r \mathbf{u}_r \mathbf{v}_r^T,
\end{aligned}
$$

where $\mathbf{u}_j$ is the j^{th} column of $\mathbf{U}$ and $\mathbf{v}_j$ is the j^{th} column of $\mathbf{V}$.

Therefore the first eigenvalue decomposition of $\mathbf{T}$ is $\lambda_1 \mathbf{u}_1 \mathbf{v}_1^T = \beta \gamma^T$. Applying the constraint $\displaystyle\sum_{a=1}^{A} \beta_a = 1$, the parameter estimates of β_a , $a = 1, \ldots, A$ and γ_p , $p = p_0 + 1, \ldots, p_0 + P$ are defined by

$$\hat{\beta}_a = \frac{u_{a,1}}{\displaystyle\sum_a u_{a,1}} \ : \ a = 1, \ldots, A$$

and

$$\hat{\gamma}_p = \lambda_1 v_{p,1} \big(\sum_a u_{a,1} \big) \ : \ p = p_0 + 1, \ldots, p_0 + P.$$

2.6.2 Expanded Singular Value Decomposition

While the LC model uses only the first singular value components, Renshaw and Haberman (2003) extended the Lee-Carter model by adding additional singular value components;

$$\log(\tilde{\lambda}_{a,p}) = \alpha_a + \sum_{i=1}^{r} \beta_a^{(i)} \cdot \gamma_p^{(i)} + \epsilon_{a,p},$$

where $\sum_a \beta_a^{(i)} = 1$ and $\sum_p \gamma_p^{(i)} = 0$ for $i = 1, \ldots, r \leq \min(A, P)$. Theorem 2.1 also implies that $\beta^{(i)}$ and $\beta^{(j)}$ are orthogonal for all $i \neq j$ and the same property applies to $\gamma^{(k)}$ and $\gamma^{(l)}$ for all $k \neq l$. The authors studied when $r = 1, \ldots, 5$ and suggested that there are significant improvements after adding the second components but not after adding the 3^{rd}, 4^{th} and 5^{th} components. Therefore the authors recommended the expanded Lee-Carter model with the first two singular value components. Detailed study of numerical comparisons between different numbers of singular value decomposition components can be found in Ranshaw and Haberman (2003).

2.6.3 Weighted Least Square approach

Wilmoth (1993) extended the Lee-Carter model to take care of the "zero cell" problem by applying weighted least squares (WLS) with weights equal to the observed number of deaths in each cell of data matrix. The author also found that the weighted least squares approach fits data better than the original singular value decomposition approach for Japanese women in 1951-1990 (Wilmoth, 1993). To apply the weighted least squares approach , we minimize the objective function

$$\sum_{a,p} D_{a,p} \big(\log(\tilde{\lambda}_{a,p}) - (\alpha_a + \beta_a \gamma_p) \big)^2, \tag{2.6.6}$$

where $D_{a,p}$ is the observed number of deaths at age a in year p. The corresponding weighted least squares estimates are given as

$$\hat{\alpha}_a = \frac{\sum_p D_{a,p}\big(\log(\tilde{\lambda}_{a,p}) - \hat{\beta}_a\hat{\gamma}_p\big)}{\sum_p D_{a,p}}, \tag{2.6.7}$$

$$\hat{\beta}_a = \frac{\sum_p D_{a,p}\hat{\gamma}_p\big(\log(\tilde{\lambda}_{a,p}) - \hat{\alpha}_a\big)}{\sum_p D_{a,p}\gamma_p^2}, \tag{2.6.8}$$

$$\hat{\gamma}_p = \frac{\sum_a D_{a,p}\hat{\beta}_a\big(\log(\tilde{\lambda}_{a,p}) - \hat{\alpha}_a\big)}{\sum_a D_{a,p}\beta_a^2}, \tag{2.6.9}$$

which can be solved by an iterative procedure after choosing a set of initial values.

2.6.4 Poisson Log-bilinear model

The original least squares approach used by Lee and Carter in fitting their model has a drawback of having to assume homoscedastic errors (Alho, 2000). Therefore the Poisson likelihood version of the Lee-Carter model proposed by Wilmoth (1993) seems to be better suited to the problem than the original least squares approach. To apply a Poisson likelihood, we assume that the number $D_{a,p}$ of death at age a in year p, follows a Poisson distribution with mean $\lambda_{a,p}N_{a,p}$, where $\lambda_{a,p}$ and $N_{a,p}$ are the corresponding mortality rate and population size . The mortality rate is assumed to satisfy the following equation

$$\lambda_{a,p} = \exp(\alpha_a + \beta_a\gamma_p).$$

Therefore the log-likelihood is given by

$$\mathbf{L}(\alpha,\beta,\gamma)=\sum_{a,p}\Big(D_{a,p}\big(\alpha_a+\beta_a\gamma_p\big)-N_{a,p}\exp\big(\alpha_a+\beta_a\gamma_p\big)\Big)+\text{constant}. \quad (2.6.10)$$

To solve for parameter estimates, ones need to solve the following equations:

$$\sum_p D_{a,p}-\sum_p N_{a,p}\exp(\alpha_a+\beta_a\gamma_p)=0 \quad (2.6.11)$$

$$\sum_p D_{a,p}\gamma_p-\sum_p N_{a,p}\gamma_p\exp(\alpha_a+\beta_a\gamma_p)=0 \quad (2.6.12)$$

$$\sum_a D_{a,p}\beta_a-\sum_a N_{a,p}\beta_a\exp(\alpha_a+\beta_a\gamma_p)=0. \quad (2.6.13)$$

Solving equation 2.6.11, we have $\hat{\alpha}_a=\log\left(\dfrac{\sum_p D_{a,p}}{\sum_p N_{a,p}\exp(\beta_a\gamma_p)}\right)$. Parameter estimates of β_a and γ_p can be obtained by applying an iterative method after choosing a set of initial values. The convergence rate depends on the iteration technique used in solving the system of equations and the starting values.

An alternative method to solve the system of equations by using an iterative Newton-Raphson method is also provided in Brouhns et al. (2002) as follows. Given the initial values $\hat{\alpha}_a^{(0)}=0$, $\hat{\beta}_a^{(0)}=1$, and $\hat{\gamma}_p^{(0)}=0$,

$$\hat{\alpha}_a^{(k+1)}=\hat{\alpha}_a^{(k)}+\frac{\sum_p(D_{a,p}-\hat{D}_{a,p}^{(k)})}{\sum_p \hat{D}_{a,p}^{(k)}},\quad \hat{\beta}_a^{(k+1)}=\hat{\beta}_a^{(k)},\quad \hat{\gamma}_p^{(k+1)}=\hat{\gamma}_p^{(k)}$$

$$\hat{\gamma}_p^{(k+2)}=\hat{\gamma}_p^{(k+1)}+\frac{\sum_p(D_{a,p}-\hat{D}_{a,p}^{(k+1)})\hat{\beta}_a^{(k+1)}}{\sum_p \hat{D}_{a,p}^{(k)}(\hat{\beta}_a^{(k+1)})^2},\quad \hat{\alpha}_a^{(k+2)}=\hat{\alpha}_a^{(k+1)},\quad \hat{\beta}_a^{(k+2)}=\hat{\beta}_a^{(k+1)}$$

$$\hat{\beta}_a^{(k+3)}=\hat{\beta}_a^{(k+2)}+\frac{\sum_p(D_{a,p}-\hat{D}_{a,p}^{(k+2)})\hat{\gamma}_p^{(k+2)}}{\sum_p \hat{D}_{a,p}^{(k+2)}(\hat{\gamma}_p^{(k+2)})^2},\quad \hat{\alpha}_a^{(k+3)}=\hat{\alpha}_a^{(k+2)},\quad \hat{\gamma}_p^{(k+3)}=\hat{\gamma}_p^{(k+2)},$$

where $\hat{D}_{a,p}^{(k)} = N_{a,p} \exp(\hat{\alpha}_a^{(k)} + \hat{\beta}_a^{(k)} \hat{\gamma}_p^{(k)})$. The criterion used to stop the iteration procedure is the small increase of the log-likelihood function; 10^{-6} was used in Brouhns et al. (2002). By using this method, parameter estimation could be performed by using the LEM program [Vermunt, 1979a,b]. This method is based on a Newton-Raphson-type equation applied successively one coordinate at a time and a convergence is guaranteed if the starting points are close enough to the true values. A reasonable set of starting points can be obtained by using least square estimates.

Later in 2007, Delwarde et al. found, from their empirical studies, that the estimated β_a's from the LC and PB models exhibit an irregular pattern of mortality curves that yields an irregular life table pattern. This could be undesirable from an actuarial point of view. Therefore, they recommended smoothing the age-specific component $\beta_a : \ a = 1, ..., A$ by applying a penalty term $\sum_a (\beta_a - 2\beta_{(a-1)} + \beta_{(a-2)})^2$ to the log-likelihood function. Their results suggest that this additional step yields smoother mortality curves.

2.7 Multihit Model

The very fruitful *multihit model* of cancer incidence was formulated by Armitage and Doll (1954) based on their observation that cancer incidence for many different sites and populations approximately follows a power law as a function of age. The multihit model essentially says that before a malignant tumor becomes clinically observable, its precursor cell must have passed successively through a series of independent stages, conceptualized as mutations or newly initiated developmental

events. The model is defined (Armitage and Doll 1954) as

$$q(t) = N\frac{p_1 p_2 \cdots p_{(r-1)} t^{r-1}}{(r-1)!},$$

where $q(t)$ is the incidence rate per person at time t, N is the mean number of cells at risk per person, p_i is the probability of occurrences of the $i-th$ mutation per unit time, and r is the number of mutations. The key contribution of this model was a mechanism "explaining"the observed power law: when the r successive transition rates λ are identical, the power dependence on age is the term t^{r-1} in the Gamma(r, λ) density for the sum of r Expon(λ) waiting times, and $r = 7$ was proposed in Armitage and Doll (1954).

This model already displays the key features that later characterize phase type models for mortality: independent, latent stages with exponentially distributed durations. For example, Knudson (1971) advanced a Markovian model (with 6 states and 7 transition rate parameters) for retinoblastoma development which was later substantially validated. See Moolgavkar (2004) for references and background on the 50 years of subsequent development of the multihit idea, which showed a rare concordance between conceptualized latent stages and mutations described in terms of molecular genetics. Moolgavkar (2004) explains that multistage cancer causation models are now explanatory, supported by genetic and other biological evidence, but that more accurate descriptive transition models must still be developed. Other Markov chain models of cancer incidence times or death times following diagnosis and initial treatment have been introduced by many different authors for several different cancers, such as Manton and Stallard's (1980) model of breast cancer mor-

tality. (See additional references to Moolgavkar and to Manton and Stallard for other examples.)

2.8 First hitting time models

A first hitting time model is a stochastic model having two main components (Lee and Whitmore, 2006): (1) a parent stochastic process $\{X(t)\}$ and (2) a threshold or a boundary set. The parent stochastic process is a stochastic process $\{X(t) : t \in \mathcal{T}, X \in \mathcal{X}\}$, where $X(t)$ is right continuous on the time space $\mathcal{T}$ and $\mathcal{X}$ is the state space. The boundary set is any closed set $\mathcal{B}$ such that $\mathcal{B} \subset \mathcal{X}$. The first hitting time is defined as the first time that the process reaches the threshold or the boundary set,

$$T = \inf\{t : X(t) \in \mathcal{B}\}.$$

Specific classes of stochastic processes X crossing a constant threshold determine well known failure time distributions. The best known example, when X is the Wiener process with drift, is the 2-parameter *Inverse Gaussian* distribution. Lee and Whitmore (2006) discuss several other such models, for example, the Bernoulli process, Poisson process, Gamma process, Ornstein-Uhlenbeck process and more general Markov processes. Their approach to survival data analysis is to choose a tractable process X and model survival times through regression models for the threshold a or initial point $x_0 = X(0)$ in terms of observable covariates. More ambitiously, Aalen and Gjessing (2001) study threshold-crossing times for a much wider class of continuous-time continuous-state Markov processes with the objective

of deriving qualitative properties of the hazard functions for crossing times from the underlying process properties. Among these stochastic models, there are three classes of processes that draw most attention among researchers: Wiener processes, Ornstein-Uhlenbeck processes and Markov processes. We will discuss the first two in this section and the third in Chapter 6.

2.8.1 Wiener Process

A standard Wiener process $W(t)$ is a continuous stochastic process such that $W(t)$ has independent increments and for $s > 0$,

$$E(W(s+t) - W(t)) = 0, \quad \text{and} \quad \text{Var}(W(s+t) - W(t)) = s.$$

To increase flexibility of the Wiener process for applications, we study the process

$$X(t) = x_0 + \mu t + \sigma W(t),$$

which is called Wiener process with initial value x_0, drift coefficient μ and diffusion coefficient σ. The density of the time T to absorption at zero is found to be (Aalen et al., 2008)

$$f(t) = \frac{x_0}{\sigma\sqrt{2\pi}} t^{-3/2} \exp\big(-\frac{(x_0 - \mu t)^2}{2\sigma^2 t}\big),$$

which is called an "Inverse Gaussian distribution". The corresponding survival function is defined as

$$S(t) = \Phi\left(\frac{x_0 - \mu t}{\sigma\sqrt{t}}\right) - \exp\left(\frac{2x_0\mu}{\sigma^2}\right)\Phi\left(\frac{-x_0 - \mu t}{\sigma\sqrt{t}}\right).$$

The Wiener process and inverse Gaussian distribution have wide applications in mortality statistics and survival analysis, such as Lancaster (1972), Whitmore(1998),

Weitz and Fraser (2001), Lee and Whitmore (2006), Aalen el at. (2008), and Balka et al. (2009).

2.8.2 Ornstein-Uhlenbeck Process

A natural extension of the Wiener process is to allow a random drift coefficient. The resulting stochastic process is the so-called Ornstein-Uhlenbeck process, which satisfies the differential equation

$$dX(t) = (a - bX(t))dt + \sigma dW(t). \tag{2.8.14}$$

Solving the equation (2.8.14), the Ornstein-Uhlenbeck process is also expressed as :

$$X(t) = \frac{a}{b} + \big(X(0) - \frac{a}{b}\big)e^{-bt} + \sigma \int_0^t e^{-b(t-s)} dW(s), \tag{2.8.15}$$

where $W(s)$ is the standard Wiener process.

The mean and variance of the Ornstein-Uhlenbeck process are given as

$$E(X(t)) = \frac{a}{b} + \big(E(X(0)) - \frac{a}{b}\big)e^{-bt} \tag{2.8.16}$$

$$Var(X(t)) = e^{-2bt} Var(X(0)) + \frac{\sigma^2}{2b}(1 - e^{-2bt}) \tag{2.8.17}$$

The term $(a - bX(t))$ represents a force pulling the process $X(t)$ back toward $\frac{a}{b}$ which is the asymptotic mean of the process, called "mean reversion" property. This property is associated with the strict stationarity of the Ornstein-Uhlenbeck process, which motivated many researchers to apply this process to phenomena that remain stable over time, such as interest rates and currency exchange rates in financial contexts, where most applications of the Ornstein-Uhlenbeck appear. However, the

phenomenon also appears in the context of mortality statistics and survival analysis. For example, AIDS and other chronic diseases can not be cured but can stay stable with some treatments. Therefore, the Ornstein-Uhlenbeck process has also attracted interest in biostatistics. For example, Taylor et al. (1994) applied an Ornstein-Uhlenbeck process to model longitudinal AIDS data, and Aalen and Gjessing (2004) introduced the Ornstein-Uhlenbeck process to the context of survival analysis and biology. Trost et al. (2010) discussed application of the Ornstein-Uhlenbeck process in a study of liver homeostasis, which is suggested to have stationary behavior around an equilibrium.

Even though the Ornstein-Uhlenbeck has wide application in many fields of interest, the first hitting time of the Ornstein-Uhlenbeck process is known to be available only in the special case where $a = 0$ and the threshold state is 0 (independently derived by Pitman and Yor (1981) and Ricciardi and Sato (1988)). The special case when $a = 0, b = 1, \sigma^2 = 2$ [Aalen et al. (2008)] is presented below:

$$f(t) = \sqrt{\frac{2}{\pi}} x_0 \frac{e^{2t}}{(e^{2t}-1)^{3/2}} \exp\left(-\frac{x_0^2}{2(e^{2t}-1)}\right). \tag{2.8.18}$$

This has led to an open problem of searching for the first hitting of the Ornstein-Uhlenbeck in the general case, and alternative indirect approaches have been proposed. For example, Ricciardi and Sato (1988) derived the first hitting time in the form of parabolic cylinder function and moment functions. Buonocore et al. (1987) presented the first hitting time in the form of integral equations, while Nobile et al. (1985) provided an asymptotic exponential approximation to the hitting time density. However these expressions are not tractable in application. We

also attempted to study a tractable form of the first hitting time of the Ornstein-Uhlenbeck mentioned in (2.8.14) for a general threshold stage c which is given in (2.8.19). The formula is an approximation of the first hitting time of the Ornstein-Uhlenbeck process starting at an initial state x_0 and the threshold state is at point $c \leq x_0$. The derivation of the density function is a direct application of density approximation of general Gaussian processes discussed in Durbin (1985). Lachaud (2004) applied Durbin's density approximation to a special case of the Ornstein-Uhlenbeck process when $a = c = 0$, which is the case where the equilibrium point and the threshold state are the same. We extend the formula of Lachaud (2004) to a general case where the equilibrium point and the threshold state are different. Our derivation of the formula is a direct application of Durbin (1985) and Lachaud (2004) and we refer to Chapter 8 for more details in deriving the formula.

$$f(t) = \frac{e^{bt}\big(a/b(e^{bt}-1)^2 - c(e^{2bt}+1) + 2x_0e^{bt}\big)}{\sigma\sqrt{\pi}} \left(\frac{b}{e^{2bt}-1}\right)^{3/2} \times \exp\left(\frac{-b[(a/b-c)e^{bt} + (x_0 - a/b)]^2}{\sigma^2(e^{2bt}-1)}\right). \tag{2.8.19}$$

The special case when $a = 0, b = 1, c = 0$, this formula coincides with the exact density given in (2.8.18).

Chapter 3

Smoothed Segmented Lee-Carter Model (SSLC)

3.1 Introduction

In many demographic and public-health applications, it is important to summarize mortality curves and time trends from population-based age-specific mortality data collected over successive years, and this is often done through the well-known model of Lee and Carter (1992). Because of its simplicity and fairly good accuracy, the LC model has been a benchmark mortality model since 1992. There are several applications of the LC model in modeling and forecasting mortality of many countries around the world as mentioned in Section 2.6. However, not all mortality dataset are perfectly described by the LC model. Therefore, several improvements and modifications have been proposed within the last two decades. For example, Wilmoth (1993) suggested fitting the LC model by using weighted least squares and Poisson maximum likelihood methods. A further study of the Poisson maximum likelihood method is found in Brouhns et al. (2002). Booth et al. (2002) suggested adjusting the period effect terms by applying a Poisson regression model to the annual number of deaths at each age while leaving the age effect terms unchanged. Ranshaw and Haberman (2003a) proposed a generalized linear modeling approach to the LC model. Delwarde et al. (2007) fitted the LC model by using penalized least squares and a penalized log-likelihood.

All of these methods were developed for the LC model with only one set of period effects. We found in our data sets that the appropriate period effects across ages for cause-specific mortality data seemed to differ. This observation was also observed by other authors, for example, Renshaw and Haberman (2003b), Hyndman and Ullah (2007), and Girosi and King (2008). They suggested, in their studies, to keep more than one set of singular value decomposition vectors in the LC model. In this chapter, we propose another method to accommodate the variation of period effect among different age groups. Our new modified LC model combines an age-segmented LC model with a spline smoothed period effect within each age segment. The segmented Lee-Carter model is fitted by using an iterative penalized least squares method. The new method is applied to the 1971-2006 public-use mortality data sets released by the National Center for Health Statistics (NCHS). Mortality rates for three leading causes of death, heart diseases, cancer and accidents, are studied in this research. The results from data analysis suggest that the age-segmented method improves the performance of the Lee-Carter method in capturing period effects across ages.

This chapter is organized as follows. Section 3.2 describes background on U.S. mortality from three selected causes of death: heart diseases, cancer, and accidents. Section 3.3 explains the details and fitting procedure for the new age-segmented model. In Section 3.4, the age-segmented method is compared to the original LC model using the 1971-2006 NCHS public use U.S. mortality data for each of the three leading causes of death. Bootstrap studies comparing the two models are

implemented in Section 3.5. Section 3.6 discusses age-segmentation.

3.2 Background on U.S. data for the three leading causes of mortality

The U.S. mortality data sets used in this research are public-use mortality data files from 1971-2006 released by the National Center for Health Statistics[1]. The population data files are drawn from the U.S. Census Bureau. Underlying causes of death were classified according to the International Classification of Diseases (ICD). The causes of death in this data period are coded according to three ICD revisions which are ICD8, ICD9 and ICD10, and the cause-specific mortality curves show discontinuities between two consecutive ICD revisions. These discontinuities are caused by coding differences among ICD revisions. To smooth the mortality curves, we apply comparability ratios[2], the ratio of the number of deaths classified by the new revision to the number of deaths classified by the previous revision, published by the National Center for Health Statistics for each data set. Table 3.1 shows ICD codes and comparability ratios for the three leading causes of death used in this study[3]. A comparability ratio for each cause of death is used to multiply the number of deaths from the previous revision to produce an updated numbers of deaths to become comparable to the new revision. The process is then repeated until the numbers of deaths comparable to the most current revision is obtained. For example,

[1] http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm

[2] http://www.cdc.gov/nchs/data/nvsr/nvsr49/nvsr49_02.pdf

[3] http://www.cdc.gov/nchs/data/dvs/comp2.pdf

the comparability ratio between ICD8 and ICD9 is used to multiply the numbers of deaths in years 1971-1978 to obtain updated numbers of deaths comparable to the numbers of deaths in ICD 9. The numbers are then multiplied by the comparability ratios between ICD9 and ICD10 to get numbers of deaths comparable to ICD 10.

Table 3.1: ICD codes and Comparability ratios for the three selected causes of death: heart diseases, cancer and accidents.

Causes of death	Heart diseases	Cancer	Accidents
ICD 8 (1971-1978)	390-398, 402-404, 410-429	140-209	E800-E949
Comparability ratios between ICD 8 and ICD 9	1.0126	1.0026	0.9970
ICD 9 (1979-1998)	390-398, 402-404, 410-429	140-280	E800-E949
Comparability ratios between ICD 9 and ICD 10	0.9852	1.0093	1.0251
ICD 10 (1999-2006)	I00-I09, I11-I13, I20-I51	C00-C97	V01-X59, Y85-Y86

3.3 Age-Segmented Modification of the Lee-Carter Model

3.3.1 Motivation

Within the LC model (2.6.5), we can see that γ_p is a common parameter of variation for all ages. When we apply the sequence γ_p to all ages, we force predicted values of time trends for all ages to be proportional. Time trend, $T_{a,p}, a = 1, 2, 3, \ldots, 84$ and $p = 1971, \ldots, 2006$, is the log mortality rate at age a in period p after subtracting the period average, $T_{a,p} = \log(\tilde{\lambda}_{a,p}) - \frac{1}{36}\sum_{p=1971}^{2006} \log(\tilde{\lambda}_{a,p})$. Figures 3.1-3.3 show that the proportionality assumption might be true for some age groups but not across all ages. Therefore, instead of having only one sequence of γ_p applied to all ages, it might be a good idea to have a few such sequences. Each sequence is applied to a specific age-group within which time trends have similar patterns. These age groups are intervals of consecutive ages which can be categorized by finding the time trends of log mortality rates in the data sets. Age-segmentation allows flexibility of period effect patterns by varying γ_p across age groups. However, having many sequences of time varying parameters for different age ranges would enormously increase the number of parameters. To avoid this problem, we use only a few age groups and then apply a smoothing spline method to smooth out the differences between the sequences of time varying parameters.

To specify age groups, we consider the smoothed curves of time trends, $\{T_{a,p} : p = 1971, .., 2006\}$, at ages $a = 1, 2, 3, ..., 84$. The curves that have similar trends are grouped into the same age group. By considering the patterns of time trends in

Figures 3.1-3.3, the appropriate age groups are 1-12, 13-36, 37-52 and 53-84 for heart diseases; 1-36, 37-60 and 61-84 for cancer; and 1-17, 18-34, 35-55 and 56-84 for accidents.

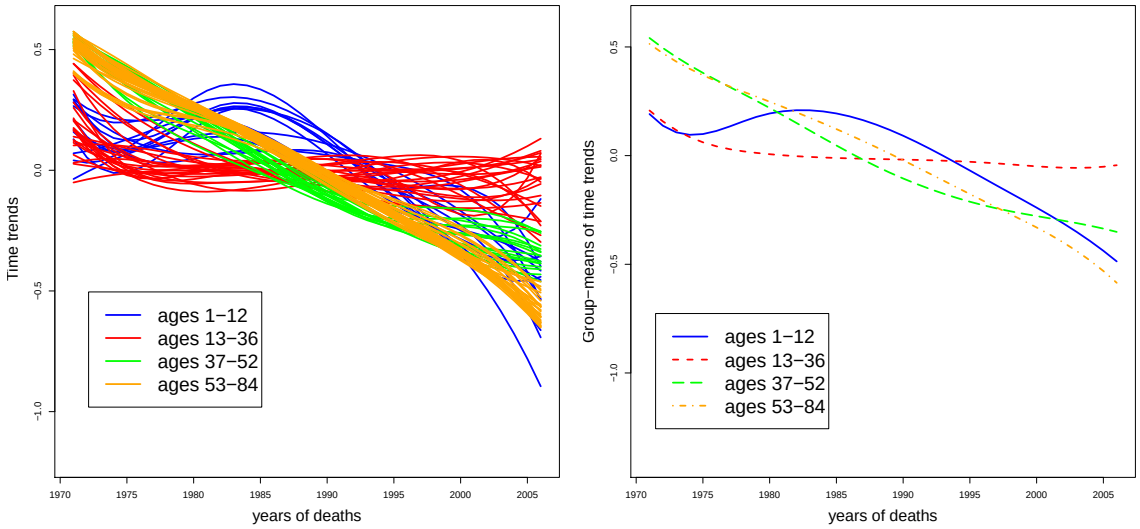


Figure 3.1: (Left) Smoothed time trends of log mortality rates from heart diseases at ages 1-84 years; (right) smoothed trends by period, averaged within age groups.

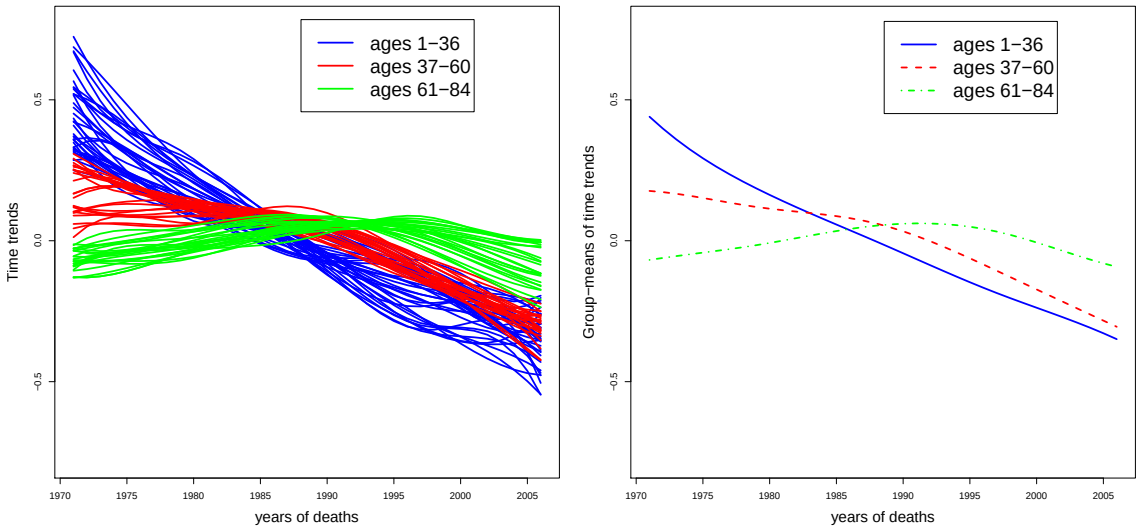


Figure 3.2: (Left) Smoothed time trends of log mortality rates from cancer at ages 1-84 years; (right) smoothed trends by period, averaged within age groups.

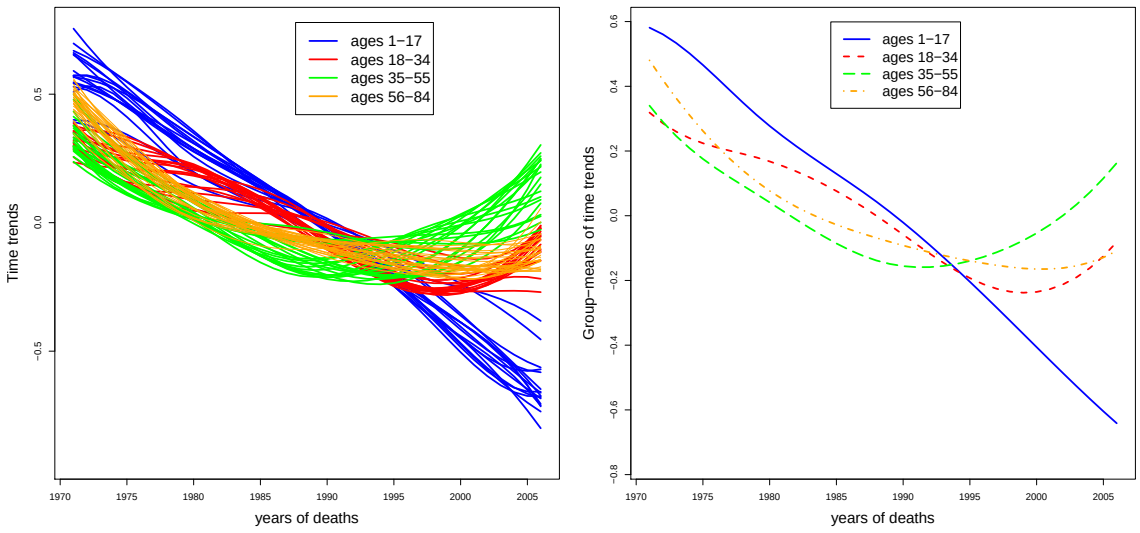


Figure 3.3: (Left) Smoothed time trends of log mortality rates from accidents at ages 1-84 years; (right) smoothed trends by period, averaged within age groups.

For each age-group $A_i = (a_{i-1}, a_i]$, where $i = 1, 2, 3..., I$, we define the mean of time trends within age-group by $k_p^{(i)} = \dfrac{1}{n(A_i)} \sum_{a \in A_i} T_{a,p}$, when $n(A_i)$ is the number of ages in age-group A_i. The smoothed curves of means of time trends within age groups are presented in the right panel of Figures 3.1-3.3. To see if a formal choice of age groups would improve the fit of the $T_{a,p}$ by age-group means $k_p^{(i)}$'s , we define the sum of squared differences (SSD) between time-trends and their group means by $\sum_{i=1}^{I} \sum_{a \in A_i} \sum_{p} \left(T_{a,p} - k_p^{(i)}\right)^2$. The SSD with a few age groups are 29.56, 15.42, and 11.41, definitely smaller than the respective SSD from a single age-group, 70.71, 44.40, and 50.50, for heart diseases, cancer and accidents. These comparisons are essentially the same as comparisons of sums of squared errors (SSE) in the context of Cluster Analysis (Tan et al. 2005, p. 499) . This suggests that a few segmented age groups can capture time trends better than a single age-group.

3.3.2 The Age-Segmented Lee-Carter model (SLC)

For $a = 1, 2, ..., A$ and $p = p_0+1,\ p_0+2, ...,\ p_0+P$, assume that the proportion $\tilde{\lambda}_{a,p}$ of the observed number of deaths $D_{a,p}$ to the corresponding population size $N_{a,p}$ satisfies the model:

$$\log(\tilde{\lambda}_{a,p}) = \alpha_a + \beta_a \gamma_{p,G(a)} + \epsilon_{a,p}, \tag{3.3.1}$$

where $\sum_a \beta_a = 1$, $\sum_p \gamma_{p,i} = 0$ and $G(a) = i$ if $a \in A_i = (a_{i-1}, a_i]$ for $i = 1, 2, ..., I$ and $\epsilon_{a,p}$ is the independent error term with mean 0 and variance σ_ϵ^2 independent of a and p.

3.3.3 Fitting the model

As mentioned in Delwarde et al. (2007), the estimated $\hat{\alpha}_a$'s are usually smooth since they represent an average of mortality at age a over the data periods. No further smoothing of the $\hat{\alpha}_a$'s is needed. Therefore, we need to smooth only the $\hat{\beta}_a$'s and $\hat{\gamma}_{p,G(a)}$'s. We use penalized least squares to smooth $\hat{\beta}_a$'s and obtain preliminary (unsmoothed) estimates for $\hat{\gamma}_{p,G(a)}$. The sequences of $\hat{\gamma}_{p,G(a)}$ for fixed a are smoothed by applying a cubic smoothing spline method with the number of knots varying by age group.

3.3.3.1 Fitting and smoothing $\hat{\beta}_a$'s via iterative penalized least squares

To smooth the $\hat{\beta}_a$'s, we apply penalized least squares as suggested by Delwarde et al. (2007) by minimizing

$$\sum_{a=1}^{A}\sum_{p=p_0+1}^{p_0+P}(\log(\tilde{\lambda}_{a,p})-(\alpha_a+\beta_a\cdot\gamma_{p,G(a)}))^2+\sigma\sum_{a=3}^{A}(\beta_a-2\beta_{a-1}+\beta_{a-2})^2. \qquad (3.3.2)$$

The penalized least squares estimate of α_a is simply the ordinary least squares estimate $\hat{\alpha}_a = \frac{1}{P}\sum_p \log(\tilde{\lambda}_{a,p})$, and the penalized least squares estimates $\hat{\beta}_a$ and $\hat{\gamma}_{p,G(a)}$ can be obtained by using an iterative algorithm as follows, with starting value $\hat{\gamma}^{(0)}_{p,G(a)} = p - \frac{1}{P}\sum_{p=p_0+1}^{p_0+P} p = (p-p_0) - \frac{(P+1)}{2}$.

For each $k = 0, 1, 2, \ldots$, we solve successively the following equations setting the gradients $\nabla_\beta{}^{(k)}$ and $\nabla_\gamma{}^{(k)}$ of (3.3.2) to $\mathbf{0}$, to obtain $\hat{\beta}^{(k)}_a$ and $\hat{\gamma}^{(k)}_{p,G(a)}$,

$$\mathbf{A}^{(k)} = [\mathbf{X}^{(k)} + \sigma\Delta'\Delta]\cdot\mathbf{B}^{(k)},$$

$$\Big(\sum_{a\in A_i}(\hat{\beta}^{(k)}_a)\Big)(\mathbf{\Gamma}^{(k)}_i) = (\mathbf{B}^{(k)}_i)\mathbf{T}_i,$$

where $\mathbf{A}^{(k)}$ and $\mathbf{B}^{(k)}$ are column vectors with dimension A, $\mathbf{\Gamma}^{(k)}_i$ and $\mathbf{B}^{(k)}_i$ are column vectors with dimensions P and $n(A_i)$, respectively; $\mathbf{T}_i$ is a matrix with dimension $n(A_i)\times P$, $\mathbf{X}^{(k)}$ is a diagonal matrix with dimension $A\times A$ and Δ is a matrix of

dimension $(A-2)\times A$. These matrices contain the following elements:

$$\mathbf{A}^{(k)} = \left[\sum_{p=p_0+1}^{p_0+P}\big(\log(\tilde{\lambda}_{a,p})\hat{\gamma}_{p,G(a)}^{(k-1)}\big)\right]_{a=1,2,3,\ldots,A},$$

$$\mathbf{B}^{(k)} = \big[\hat{\beta}_a^{(k)}\big]_{a=1,2,3,\ldots,A},$$

$$\mathbf{B}_i^{(k)} = \big[(\hat{\beta}_a^{(k)})^2\big]_{a\in A_i},$$

$$\mathbf{T}_i = \big[\log(\tilde{\lambda}_{a,p}) - \hat{\alpha}_a\big]_{a\in A_i},$$

$$\mathbf{\Gamma}_i^{(k)} = \big[\hat{\gamma}_{p,i}^{(k)}\big]_{p=p_0+1,\ p_0+2,\ p_0+3,\ldots,\ p_0+P},$$

$$\mathbf{X}^{(k)} = \mathrm{diag}\big((\hat{\gamma}_{p,G(a)}^{(k-1)})^2\big),$$

and

$$\Delta = \begin{pmatrix} 1 & -2 & 1 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & \ldots & 0 & \vdots & 0 \\ 0 & \vdots & & \ddots & -2 & 1 & 0 \\ 0 & 0 & 0 & \ldots & 1 & -2 & 1 \end{pmatrix},$$

where $\sum_a \hat{\beta}_a^{(k)} = 1$, $\sum_p \hat{\gamma}_{p,i}^{(k)} = 0$ for all $i = 1, 2, 3, \ldots, I$ in the k^{th} iteration step, and σ is the smoothness parameter selected by cross-validation, which will be discussed in the next section.

3.3.3.2 Selecting the smoothness parameter by cross-validation

To select the smoothness parameter σ, we follow the cross-validation method for the LC model suggested by Delwarde et al. (2007). For each $a = 1, 2, \ldots, A$ and $p = p_0 + 1,\ p_0 + 2, ...,\ p_0 + P$, let

$$e_{a,p}(\sigma) = \log(\hat{\lambda}_{a,p}) - (\hat{\alpha}_{a,\sigma}^{-(a,p)} + \hat{\beta}_{a,\sigma}^{-(a,p)} \cdot \hat{\gamma}_{p,G(a),\sigma}^{-(a,p)}),$$

where $\hat{\alpha}_{a,\sigma}^{-(a,p)}$, $\hat{\beta}_{a,\sigma}^{-(a,p)}$, and $\hat{\gamma}_{p,G(a),\sigma}^{-(a,p)}$ are the penalized least squares estimates obtained by excluding the observation at age a in year p and $\log(\hat{\lambda}_{a,p}) = \hat{\alpha}_a + \hat{\beta}_a \cdot \hat{\gamma}_{p,G(a)}$. The selected smoothing parameter is the one that minimizes $\sum_{a=1}^{A} \sum_{p=p_0+1}^{p_0+P} e_{a,p}^2(\sigma)$.

3.3.3.3 Parameter reduction: Smoothing $\hat{\gamma}_{p,G(a)}$'s using a penalized spline method

The SLC model having more than one sequence of period effect terms increases the number of parameters of the LC model. Therefore, for fixed SLC parameters $\hat{\alpha}_a$, $\hat{\beta}_a$ and smoothness parameter σ, we fit a cubic penalized spline to each sequence of period effects $\hat{\gamma}_{p,i}$'s, $i = 1, \ldots, I$, to reduce the number of parameters. This spline fitting can be done by applying the function "smooth.spline" in the R-package "stats" for each sequence of period effect terms. For each sequence of period effect terms, we compute a generalized cross-validation criterion (GCV) for the number of knots $K = 1, 2, \ldots, 10$, minimized over the penalty parameter. The number K that minimizes GCV is selected. Having completed the parameter reduction, the number of parameters for each sequence of $\gamma_{p,i}$'s, $i = 1, \ldots, I$, is therefore reduced from the number of period effect terms minus one, $P-1$, to the number of knots plus two, $K+2$. The SLC model with smoothed period effects is called the *"Smoothed Segmented Lee-Carter"* model, or SSLC model.

3.4 Data Analysis

In this section, we apply the age-segmented LC models to the data for three specific causes of mortality, namely heart diseases, cancer, and accidents. Statistical comparisons between the smoothed age-segmented (SSLC) and LC models are shown based on MSEs, SSEs, and R-Squared. Since corresponding results from the nonsmoothed age-segmented model (SLC) almost coincide with those of the SSLC model, while the SLC model requires much larger numbers of parameters, the SLC versus LC comparisons are not shown.

3.4.1 Heart diseases

Figure 3.4 shows that the curve of the $\hat{\alpha}_a$, $a = 1, 2, 3, \ldots, 84$, is reasonably smooth. Figure 3.5 shows the plot of $\hat{\beta}_a$, $a = 1, 2, 3, \ldots, 84$ with various values of the smoothing parameter: 0, 6000, 8000, 10000, and 30000. The figure suggests that results for all positive σ are similar. The optimal $\hat{\sigma}$ selected by the cross-validation method is 8000. Figures 3.1 and 3.7 suggest that estimated time trends obtained from the SSLC model are similar to the raw ones. Table 3.2 shows comparisons of the number of parameters, SSEs, and MSEs between the LC and the SSLC models. The number of parameters used in the LC model, 202, is the sum of the number of α's, the number of β's minus one and the number of γ's minus one, which are 84, 83 and 35, respectively. The number of parameters used in the SSLC model is the sum of the number of α's, the number of β's minus one and the number of cubic smoothing parameters of $\hat{\gamma}_i$'s corresponding to the number of knots 7, 10,

10 and 10 for age group $i = 1, 2, 3, 4$, respectively. The number of parameters for the SSLC model is then the sum of 84, 83, 9, 12, 12 and 12, which is 212. Tables 3.2 and 3.3 suggest that the Mean Square Error (MSE) for the whole age range and for age groups are smaller for the SSLC model, in particular for the 37-52 age group, within which MSE is reduced by 70 % over the LC model. In the bar-plots (Figures 3.8-3.9) of R-Squared for ages 1-84 years, where $R_a^2 = 1 - \dfrac{\mathrm{Var}(\hat{\epsilon}_{a,p})}{\mathrm{Var}(\log(\tilde{\lambda}_{a,p}))}$, and $\hat{\epsilon}_{a,p} = \log(\tilde{\lambda}_{a,p}/\hat{\lambda}_{a,p})$, the SSLC model yields higher R_a^2 for lower ages ($a \leq 50$) but that the values R_a^2 are similar in SSLC and LC for older ages ($a \geq 60$). The data analysis suggests that the SSLC model slightly improves the performance of the LC model in the young age group in the case of heart diseases. However, neither model fits very well for age group 13-36 years.

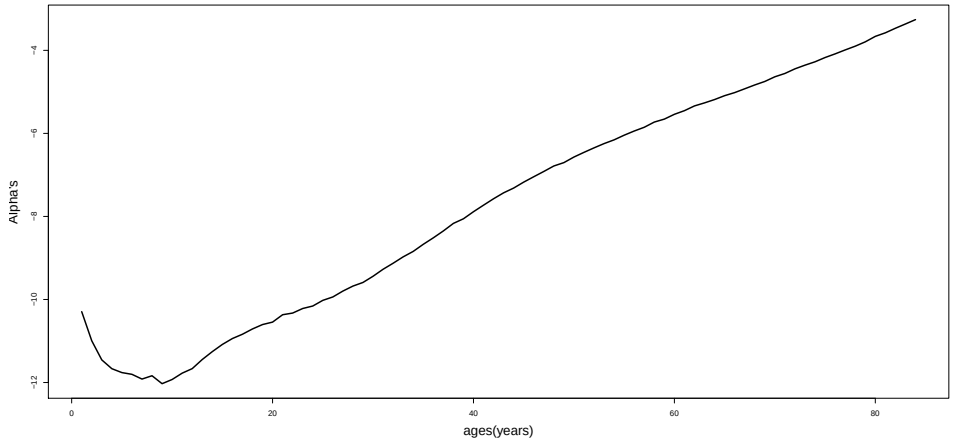


Figure 3.4: Plots of estimated $\hat{\alpha}$ for heart disease.

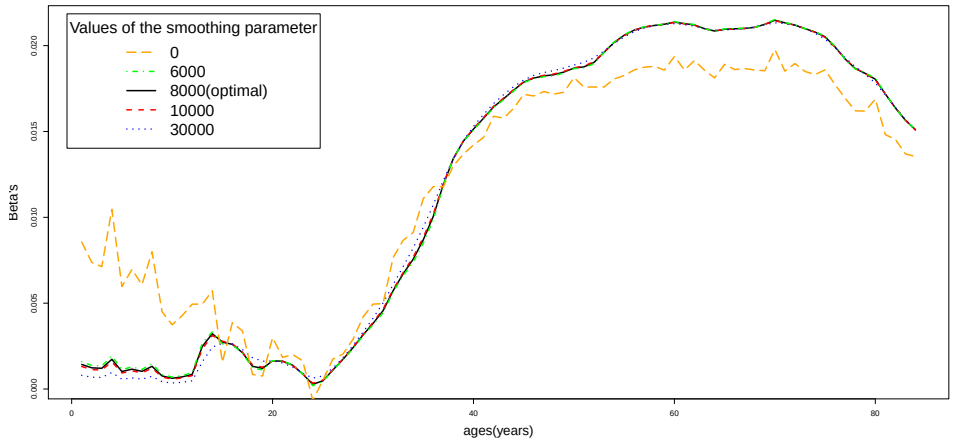


Figure 3.5: Plots of estimated $\hat{\beta}_a$ for heart disease for various values of the smoothing parameters σ. The optimal $\hat{\sigma}$, selected by cross-validation, is 8000.

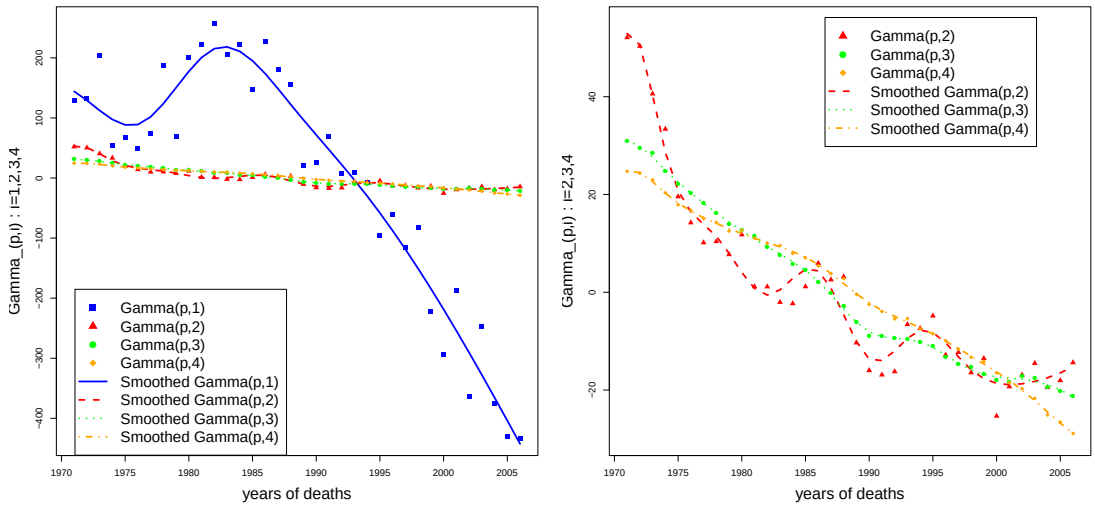


Figure 3.6: Period effect terms ($\hat{\gamma}_{p,i}$'s , $p = 1971, 1972, \ldots, 2006;\ i = 1, 2, 3, 4$) and their smoothed values for heart disease obtained from the SSLC model.

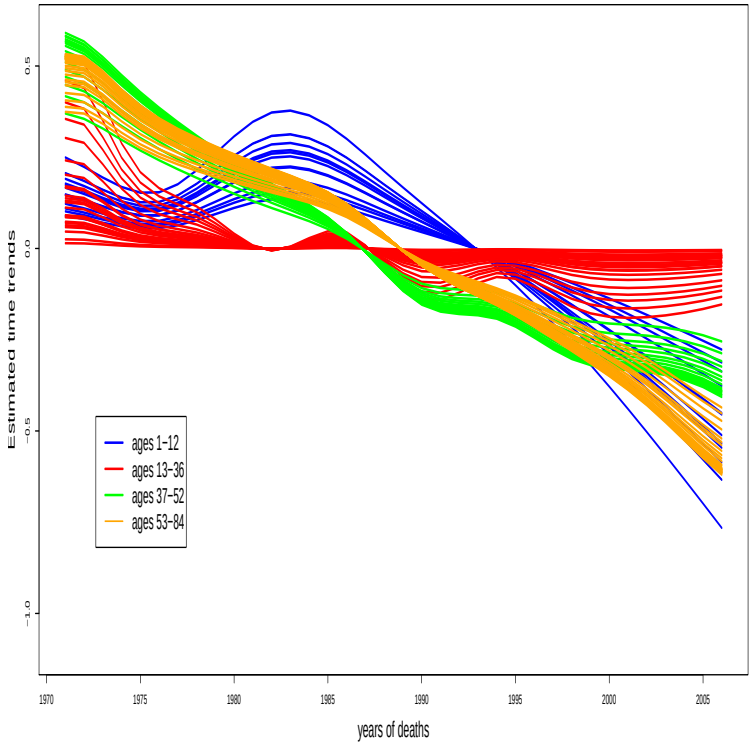


Figure 3.7: Groupwise estimated time trends of log mortality rates.

Table 3.2: Comparisons of Mean Square Errors and Sum of Square Errors of the LC model and the SSLC model for heart diseases.

Model	Number of parameters	SSE of log rates	MSE of log rates
The LC model	202	31.6438	0.0105
The SSLC model	212	23.7874	0.0079

Table 3.3: Comparisons of Mean Square Errors within age groups of the LC model and the SSLC model for heart diseases.

Model	Ages 1-12	Ages 13-36	Ages 37-52	Ages 53-84
The LC model	0.0419	0.0107	0.0052	0.0008
The SSLC model	0.0317	0.0100	0.0016	0.0007

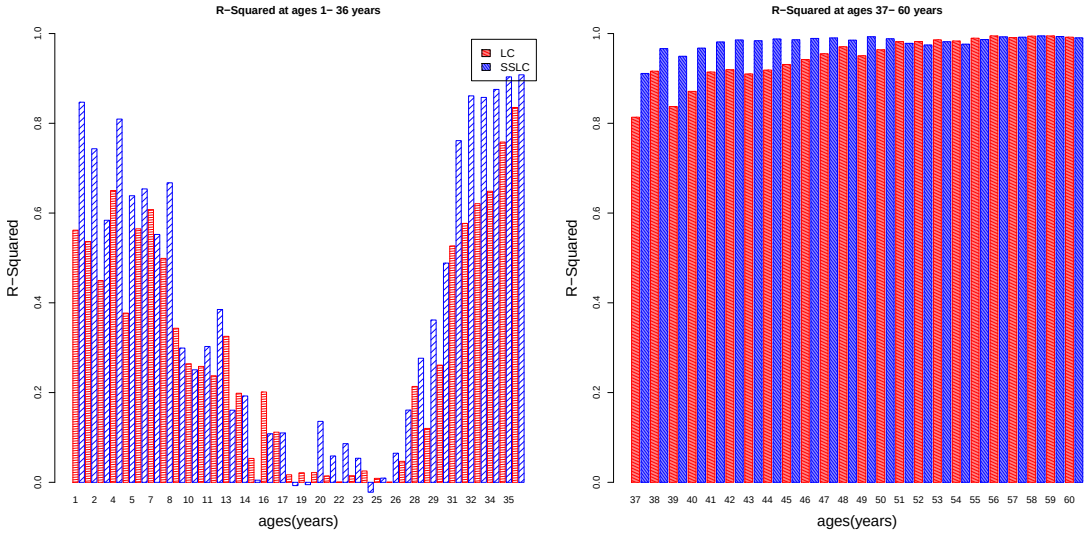


Figure 3.8: Bar plots of R-Squared, $R_a^2 = 1 - \text{Var}(\hat{\epsilon}_{a,p})/\text{Var}(\log(\tilde{\lambda}_{a,p}))$: $a = 1, .., 60$, of the LC model (red) and the SSLC model (blue) for heart diseases.

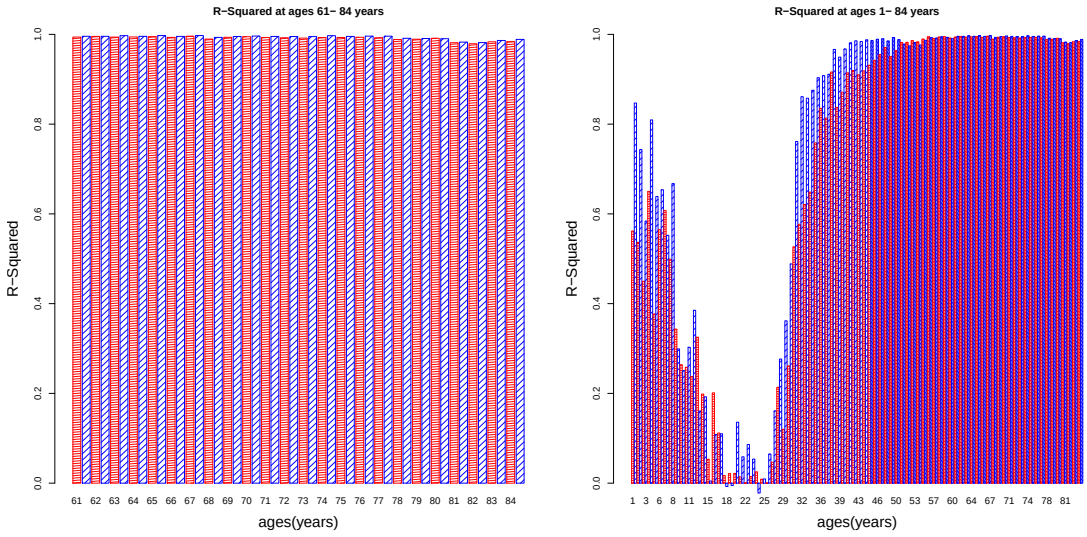


Figure 3.9: Bar plots of R-Squared, $R_a^2 = 1 - \text{Var}(\hat{\epsilon}_{a,p})/\text{Var}(\log(\tilde{\lambda}_{a,p}))$, of the LC model (red) and the SSLC model (blue) for heart diseases.

Assume that the number of deaths follows the Poisson distribution, $D_{a,p} \sim \text{Poi}(N_{a,p}\lambda_{a,p})$, which for large $N_{a,p}$ agrees closely with the approximate distribution from the Central Limit Theorem, $\mathcal{N}(N_{a,p}\lambda_{a,p}, N_{a,p}\lambda_{a,p})$. Then an estimate of the variance of the crude estimate of log mortality rate is $\widehat{\text{Var}}(\log(\tilde{\lambda}_{a,p})) = \dfrac{1}{D_{a,p}}$ and an approximate 95% pointwise confidence interval of the crude estimate of log mortality rate is given by

$$\log(\tilde{\lambda}_{a,p}) \pm \frac{1.96}{\sqrt{D_{a,p}}}.$$

Figures 3.10-3.11 show crude estimates of log mortality rates with estimated 95 % pointwise confidence intervals and the fitted curves from the LC and SSLC models at selected ages.

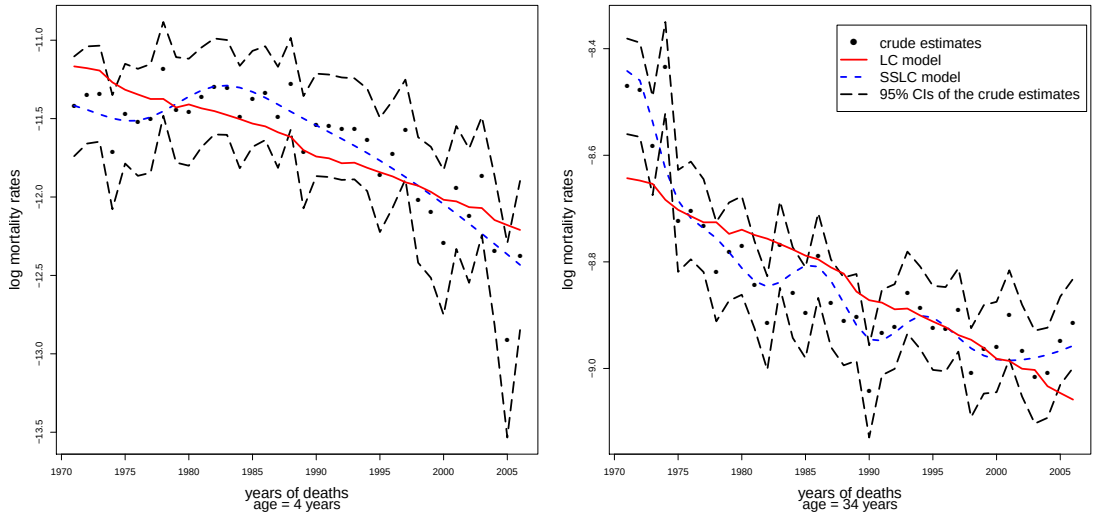


Figure 3.10: Crude estimates of log mortality rates from heart diseases with 95% pointwise confidence intervals and the fitted curves from the LC model and the SSLC model at ages 4 and 34 years.

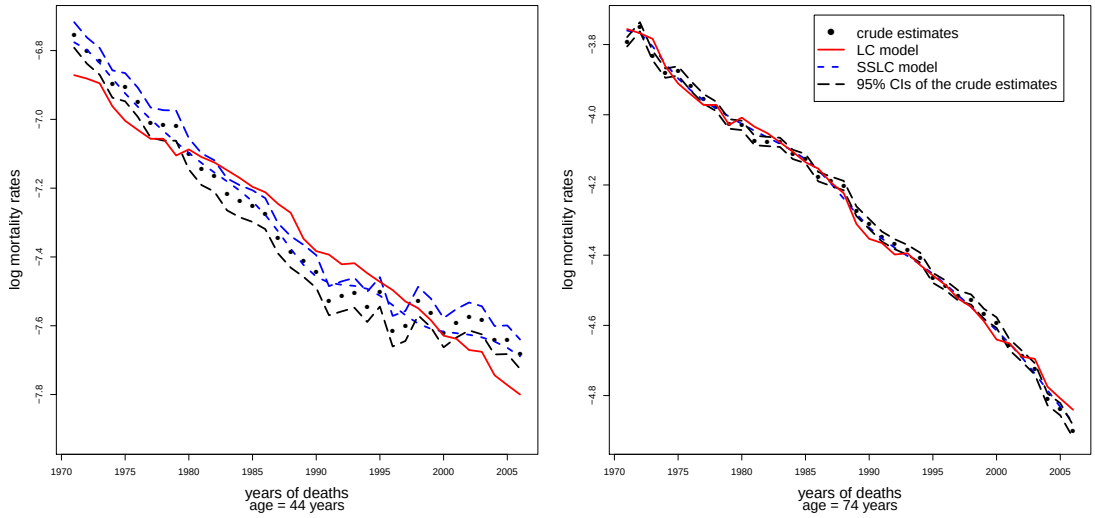


Figure 3.11: Crude estimates of log mortality rates from heart diseases with 95% pointwise confidence intervals and the fitted curves from the LC model and the SSLC model at ages 44 and 74 years.

We can see from Figures 3.10-3.11 that both models produce similar curves,

but the fitted curves from the SSLC model fit into the estimated 95% pointwise confidence intervals of crude estimates better than the fitted curves from the LC model.

3.4.2 Cancer

The plots of the $\hat{\alpha}_a$'s and $\hat{\beta}_a$'s $, a = 1, 2, 3, \ldots, 84$, with various smoothing coefficients are shown in Figures 3.12-3.13. The optimal value of the smoothing coefficient is 1000. Tables 3.4 and 3.5 show that the MSEs from the SSLC model are much smaller than from the LC model, in particular for older age groups 37-60 and 61-84 years, in both of which SSLC reduces MSE by approximately one-half. The number of parameters presented in Table 3.4 are calculated in the same ways as in Section 3.4.1, where the numbers of knots for the cubic smoothing spline of the $\hat{\gamma}_{p,i}$'s $i = 1, 2, 3$ are 10, 6 and 6, respectively. Figure 3.14 shows plots of estimated time trends which are similar to the smoothed curves of time trends shown in Figure 3.2. Figures 3.15-3.16 show that the SSLC model captures the patterns of age-specific log mortality rates very well, while the LC model fails to do so, for example at ages 64 and 74 years. As can be seen in Figures 3.17-3.18, correlations between fitted and raw log mortality rates remain high in the oldest age group for the SSLC model but low for the LC model, except at ages 77-83 years, where the correlations for the LC model are higher than for the SSLC model.

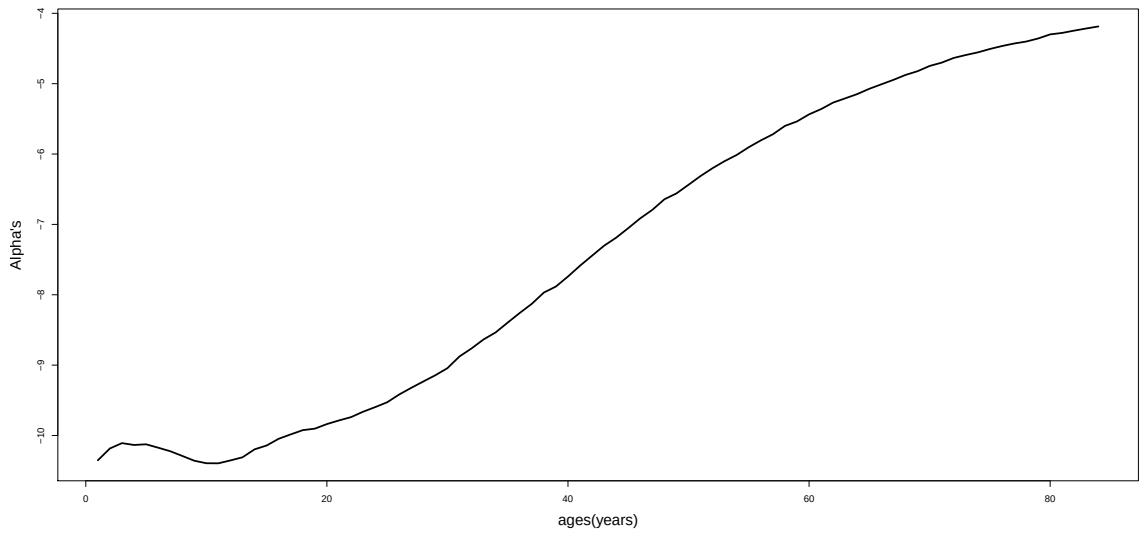


Figure 3.12: Plots of estimated $\hat{\alpha}$ for cancer.

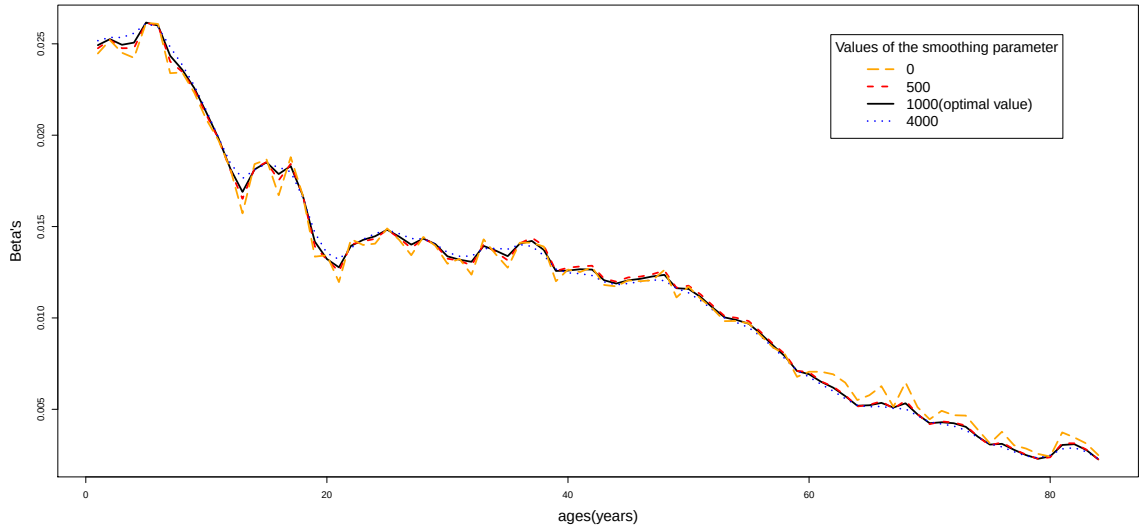


Figure 3.13: Plots for cancer of estimated of $\hat{\beta}_a$ for various values of the smoothing parameter σ. The optimal $\hat{\sigma}$, selected by cross-validation, is 1000.

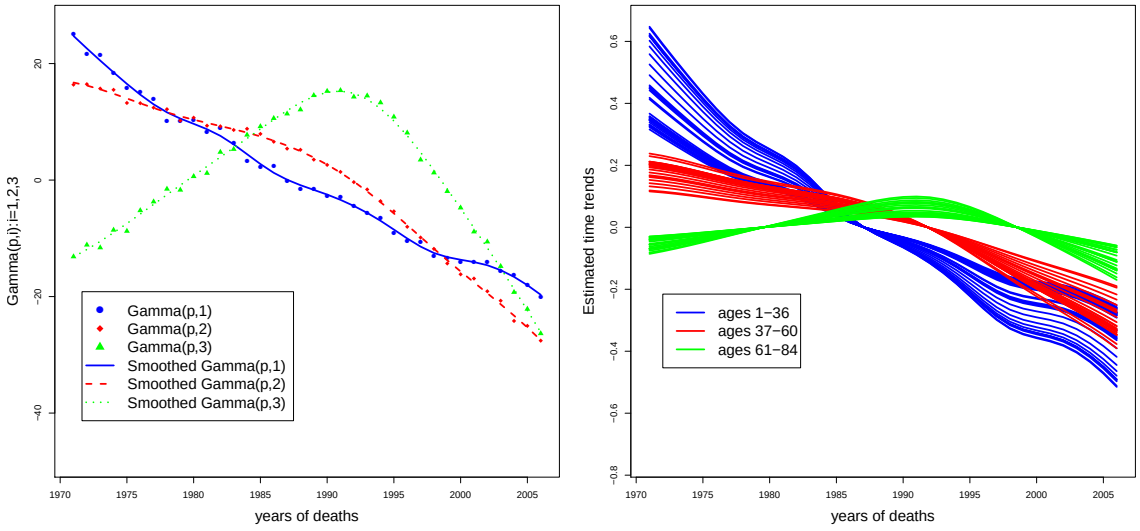


Figure 3.14: The left panel shows period effect terms ($\hat{\gamma}_{p,i}$'s , $p = 1971, 1972, \ldots, 2006$; $i = 1, 2, 3, 4$) and their smoothed values for cancer obtained from the SSLC model; the right panel shows the corresponding estimated time trends of log mortality rates.

Table 3.4: Comparisons of Mean Square Errors and Sum of Square Errors of the LC model and the SSLC model.

Model	Number of parameters	SSE of log rates	MSE of log rates
The LC model	202	12.8338	0.0042
The SSLC model	195	10.3700	0.0034

Table 3.5: Comparisons of Mean Square Errors within age groups of the LC model and the SSLC model.

Model	Ages 1-36	Ages 37-60	Ages 61-84
The LC model	0.0062	0.0028	0.0028
The SSLC model	0.0061	0.0013	0.0015

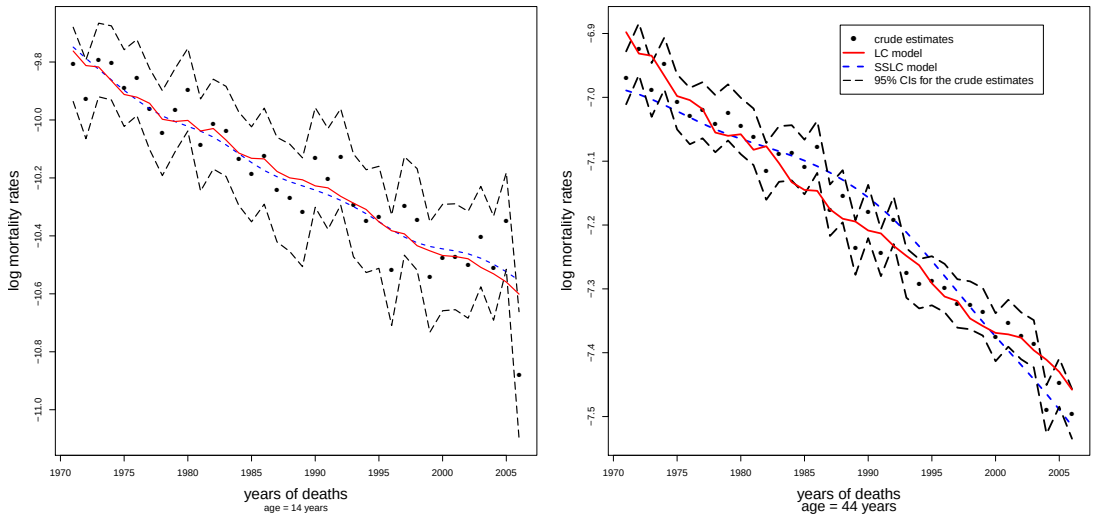


Figure 3.15: Crude estimates of log mortality rates from cancer with 95% pointwise confidence intervals and the fitted curves from the LC model and the SSLC model at ages 14 and 44 years.

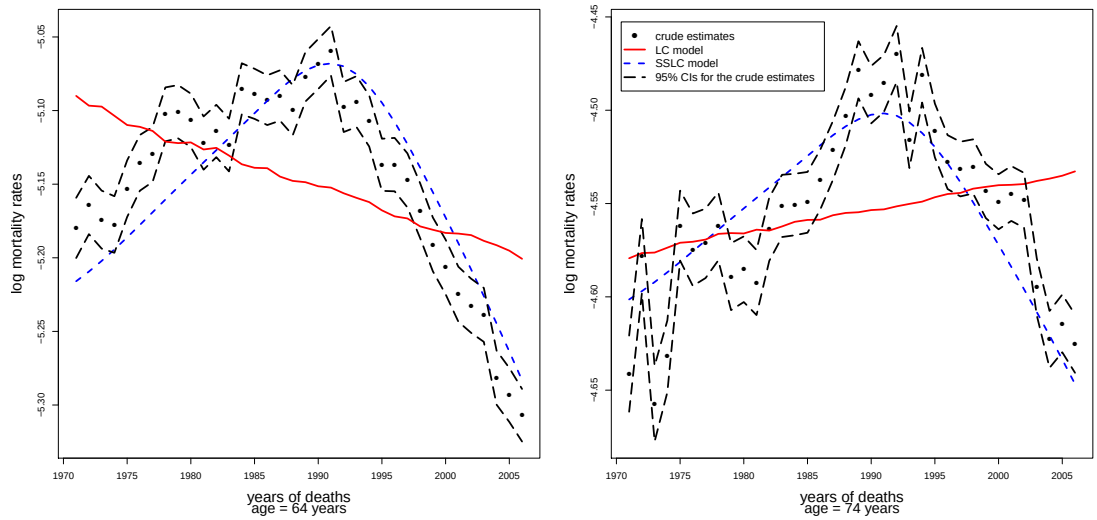


Figure 3.16: Crude estimates of log mortality rates from cancer with 95% pointwise confidence intervals and the fitted curves from the LC model and the SSLC model at ages 64 and 74 years.

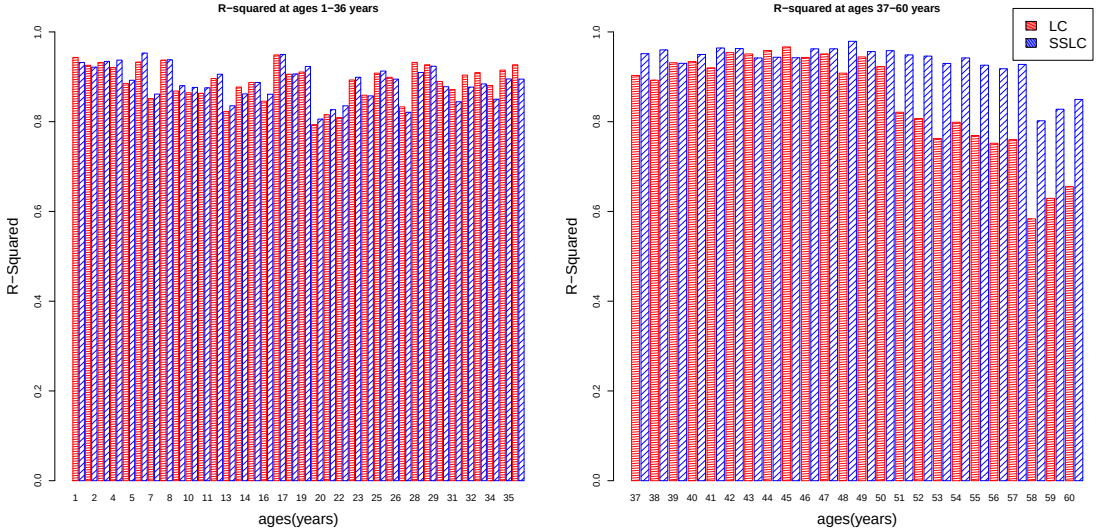


Figure 3.17: Bar plots of R-Squared, $R_a^2 = 1 - \text{Var}(\hat{\epsilon}_{a,p})/\text{Var}(\log(\tilde{\lambda}_{a,p}))$, $a = 1, \ldots, 60$, of the LC model (red) and the SSLC model (blue) for cancer.

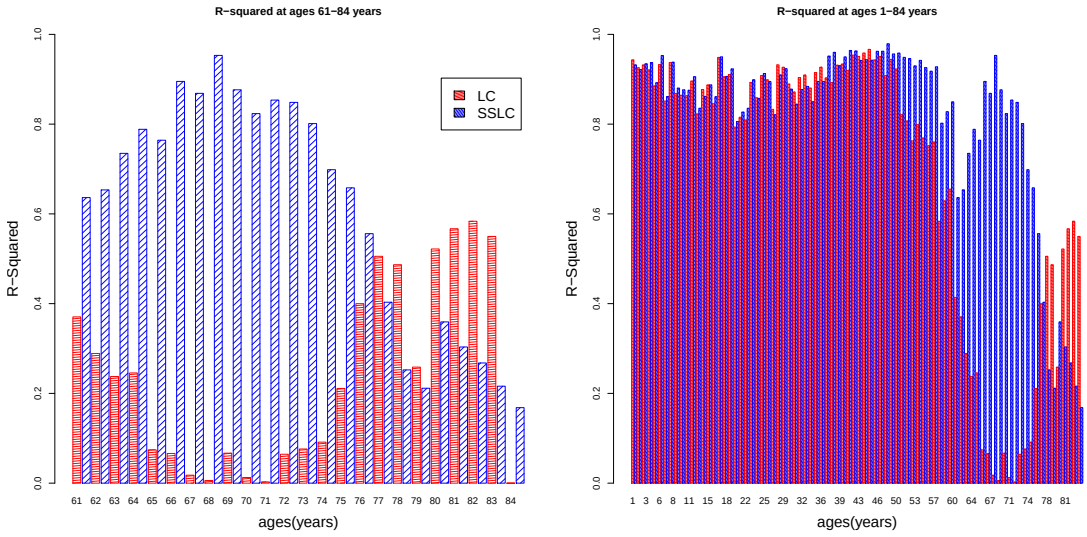


Figure 3.18: Bar plots of R-Squared, $R_a^2 = 1 - \text{Var}(\hat{\epsilon}_{a,p})/\text{Var}(\log(\tilde{\lambda}_{a,p}))$, of the LC model (red) and the SSLC model (blue) for cancer.

3.4.3 Accidents

The plots of parameter estimates are shown in Figures 3.19- 3.21. Tables 3.6 and 3.7 show that the SSLC model gives substantially smaller MSEs and SSEs for the whole age range and for each specific age group than the LC model. The numbers of parameters presented in Table 3.6 are calculated in the same ways as in Section 3.4.1 where the numbers of knots for the cubic smoothing spline of the $\hat{\gamma}_{p,i}$'s $i = 1, 2, 3, 4$, are all 9. The SSLC model reduces MSEs by approximately one-third for the age groups 1-17 and 35-55 years and approximately one-half for the age groups 18-34 and 56-84 years. Figures 3.22-3.23 show that the LC model gives approximately linear patterns of log mortality rates for all ages but the SSLC model gives different patterns for different age groups. The predicted pattern obtained from the SSLC

model is approximately linear for the age group 1-17 years, approximately quadratic for the age groups 18-34, 35-55 and 56-84 years with different curvatures which agree closely with the patterns of raw data. The SSLC model captures the patterns of age-specific log mortality rates better than the LC model, especially, in age group 35-55 years. Figures 3.24-3.25 confirm the result as we can see that the SSLC model substantially improves R-squared in age group 35-55 years.

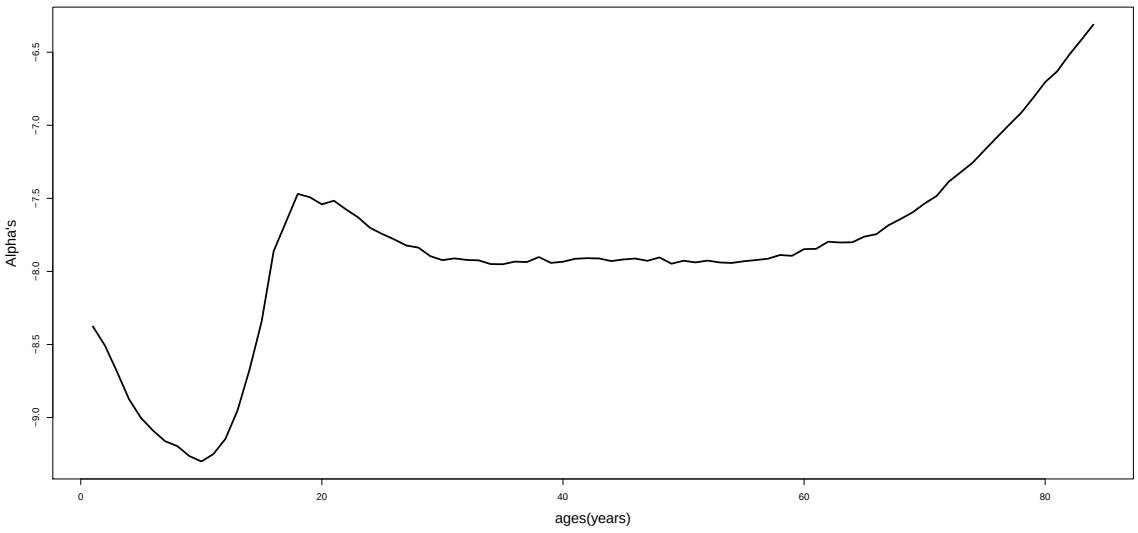


Figure 3.19: Plots of estimated $\hat{\alpha}$ for accidents.

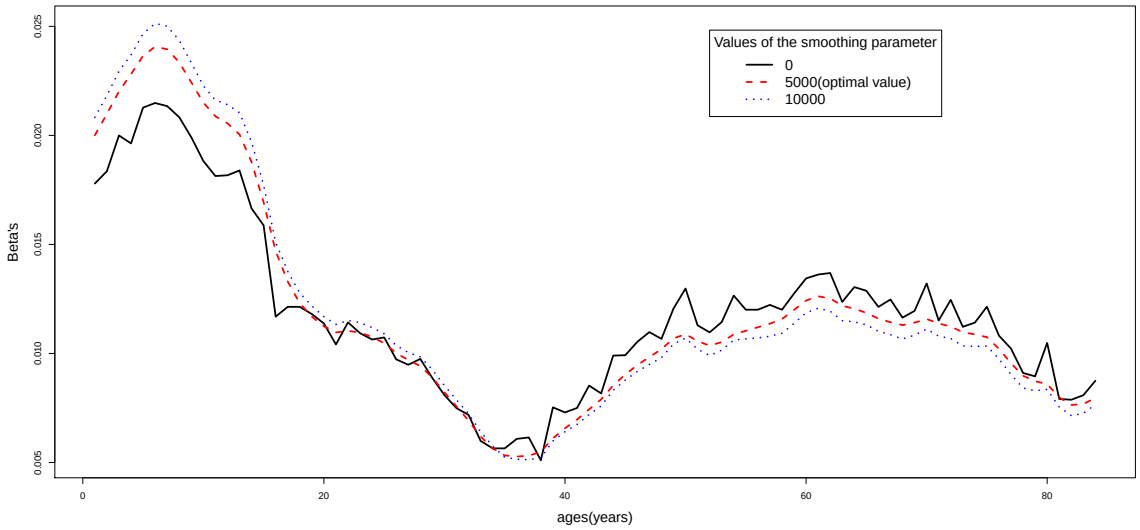


Figure 3.20: Plots for accidents of estimated of $\hat{\beta}_a$ for various values of the smoothing parameter σ. The optimal $\hat{\sigma}$, selected by cross-validation, is 1000.

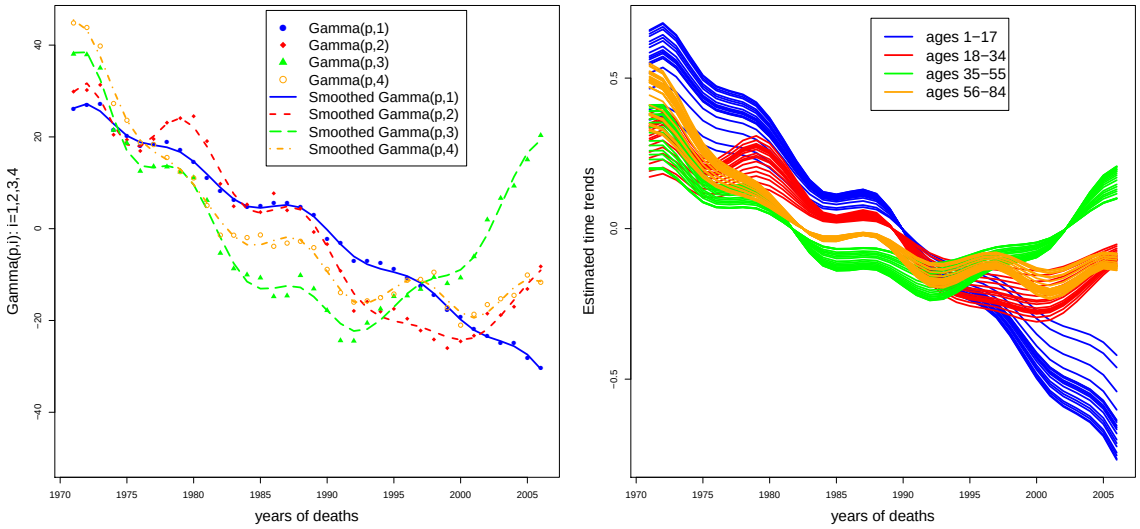


Figure 3.21: (Left) Period effect terms ($\hat{\gamma}_{p,i}$'s , $p = 1971, 1972, \ldots, 2006$; $i = 1, 2, 3, 4$) and their smoothed values for accidents obtained from the SSLC model: (right) Groupwise estimated time trends of log mortality rates.

Table 3.6: Comparisons of Mean Square Errors and Sum of Square Errors of the LC model and the SSLC model.

Model	Number of parameters	SSE of log rates	MSE of log rates
The LC model	202	24.7996	0.0082
The SSLC model	211	7.7678	0.0026

Table 3.7: Comparisons of Mean Square Errors within age groups of the LC model and the SSLC model.

Model	Ages 1-17	18-34	Ages 35-55	Ages 56-84
The LC model	0.0095	0.0041	0.0153	0.0047
The SSLC model	0.0032	0.0019	0.0034	0.0020

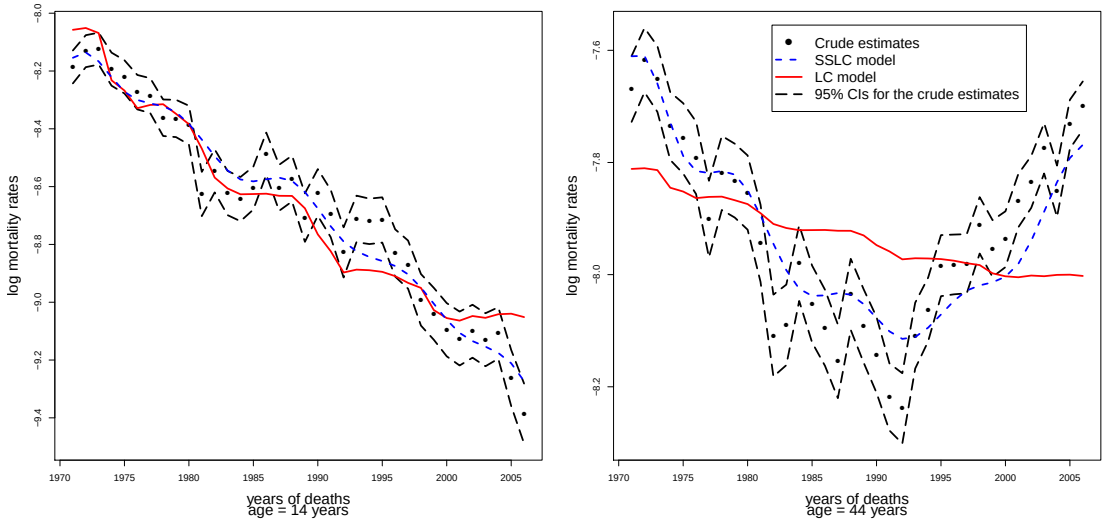


Figure 3.22: Crude estimates of log mortality rates from accidents with 95% pointwise confidence intervals and the fitted curves from the LC model and the SSLC model at ages 14 and 44 years.

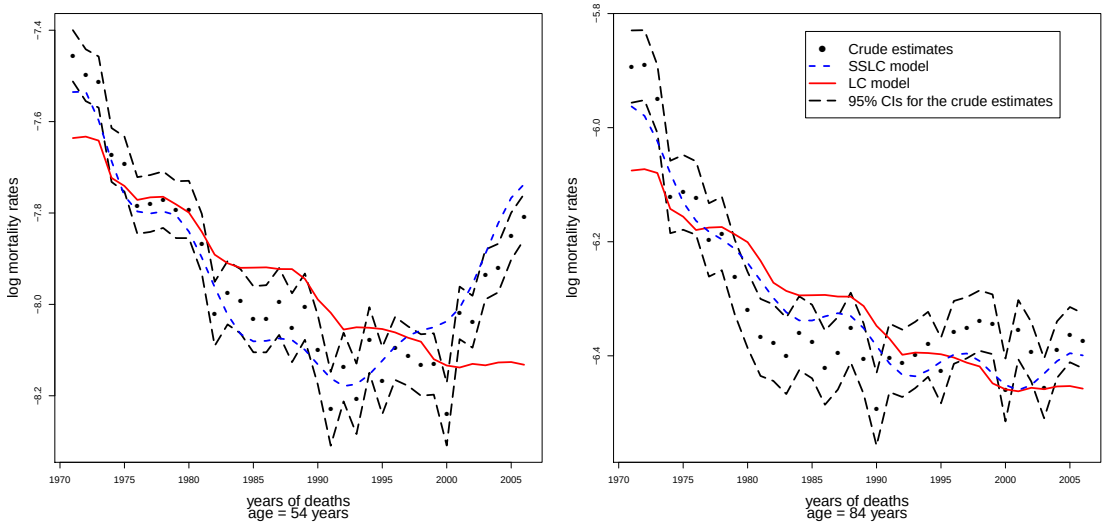


Figure 3.23: Crude estimates of log mortality rates from accidents with 95% pointwise confidence intervals and the fitted curves from the LC model and the SSLC model at ages 54 and 84 years.

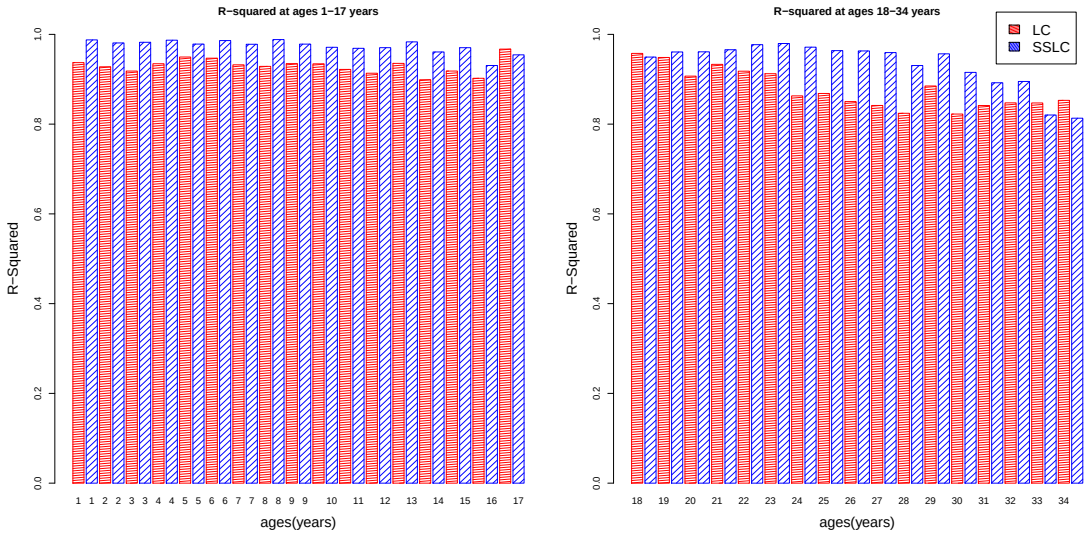


Figure 3.24: Bar plots of R-Squared, $R_a^2 = 1 - \text{Var}(\hat{\epsilon}_{a,p})/\text{Var}(\log(\tilde{\lambda}_{a,p}))$, $a = 1, \ldots, 34$, of the LC model (red) and the SSLC model (blue) for accidents.

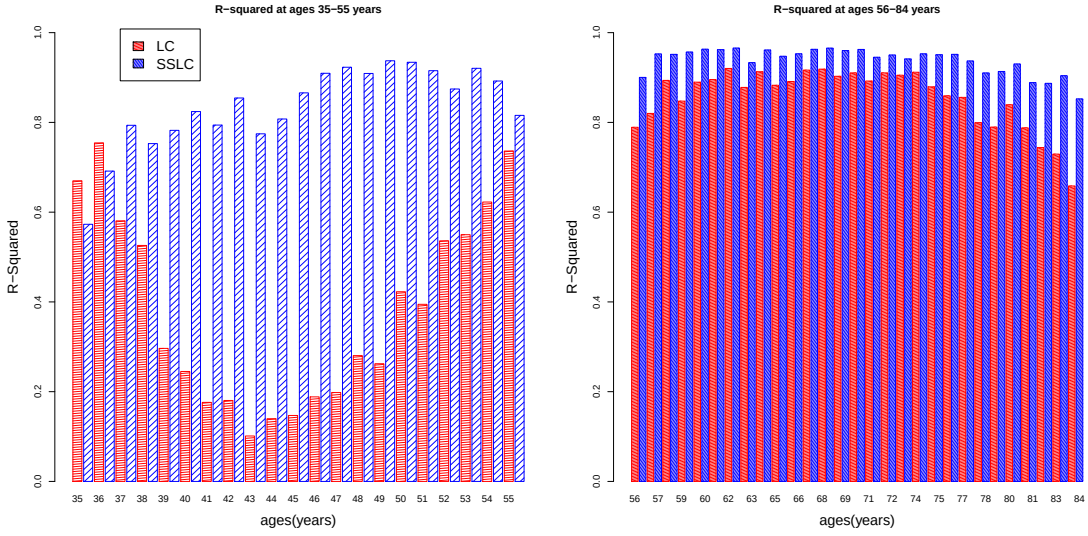


Figure 3.25: Bar plots of R-Squared, $R_a^2 = 1 - \text{Var}(\hat{\epsilon}_{a,p})/\text{Var}(\log(\tilde{\lambda}_{a,p}))$: $a = 35, \ldots, 84$, of the LC model (red) and the SSLC model (blue) for accidents.

3.5 A Bootstrap Study

Bootstrapping is a computer intensive method but is very useful when theoretical calculation is too complex, as in the situation of Lee-Carter parameter estimation (Brouhns et al., 2005). Therefore, the only available technique to study properties of estimates in the family of Lee-Carter models is the bootstrap. Two bootstrap techniques are used in this family of models: Residual bootstrap and Poisson bootstrap. However, the Poisson bootstrap seems to have received more attention and it provides reasonable results in the original Lee-Carter model. In this section, we apply a Poisson bootstrap to obtain estimated biases, estimated variances and pointwise confidence intervals for estimated log mortality rates. Comparisons of these estimates among the LC, the SLC and the SSLC models are studied. More detailed graphical results can be found in Chapter 10.

An Algorithm for Poisson Bootstrap

Given an $A \times P$ matrix of observed number of deaths $D_{a,p}$, the Poisson bootstrap algorithm proceeds as follows:

- Generate B (=1000) replications $\{D_{a,p}^{(b)}, b = 1, ..., B\}$, such that for each a, p and b, $D_{a,p}^{(b)}$ is sampled from the Poisson distribution with mean $N_{a,p}\widetilde{\lambda}_{a,p} = D_{a,p}$.

- Compute $\hat{\alpha}_a^{(b)}$, $\hat{\beta}_a^{(b)}$ and $\hat{\gamma}_{p,G(a)}^{(b)}$, and $\log(\hat{\lambda}_{a,p}^{(b)})$, for each bootstrap sample $\{D_{a,p}^{(b)}, b = 1, ..., B\}$. The smoothness parameter, $\hat{\sigma}$, for each bootstrap replication is fixed to be the smoothness parameter obtained by cross-validation of the original data and is not allowed to vary in the bootstrap replications.

Bootstrap Estimations of Bias and Variance

For each model, LC, SLC, SSLC, and all (a, p), the bootstrap estimates of $\log(\lambda_{a,p})$ are defined as

$$\log(\hat{\lambda}^{(*)}_{a,p}) \;=\; \frac{1}{B}\sum_{b=1}^{B} \log(\hat{\lambda}^{(b)}_{a,p}).$$

In terms of these, the estimates of bias and variance as in Efron and Tibshirani (1993) are calculated as

$$\widehat{\mathrm{Bias}}_{(B)}(a,p) = \log(\hat{\lambda}^{(*)}_{a,p}/\tilde{\lambda}_{a,p}) \;\;,\quad \widehat{\mathrm{Var}}_{(B)}(a,p) \;=\; \frac{1}{B-1}\sum_{b=1}^{B}\Big[\log(\hat{\lambda}^{(b)}_{a,p}/\hat{\lambda}^{(*)}_{a,p})\Big]^2.$$

The estimated MSE is then defined by

$$\widehat{\mathrm{MSE}}_{(B)}(a,p) = \widehat{\mathrm{Bias}}^{\,2}_{(B)}(a,p) + \widehat{\mathrm{Var}}_{(B)}(a,p).$$

Our summary figures display root-mean-square biases, averages of the variances and MSEs across period p for each $a = 1, \ldots, 84$. That is,

$$\widehat{\mathrm{Bias}}_{(B)}(a) \;=\; \sqrt{\frac{1}{36}\sum_{p=1971}^{2006} \widehat{\mathrm{Bias}}^{\,2}_{(B)}(a,p)} \;\;,\quad \widehat{\mathrm{Var}}_{(B)}(a) \;=\; \frac{1}{36}\sum_{p=1971}^{2006} \widehat{\mathrm{Var}}_{(B)}(a,p),$$

$$\widehat{\mathrm{MSE}}_{(B)}(a) \;=\; \frac{1}{36}\sum_{p=1971}^{2006} \widehat{\mathrm{MSE}}_{(B)}(a,p).$$

Figures 3.26- 3.28 display these root-mean-square biases (top left panel), period-averaged variances (top right panel), and period-averaged MSEs (bottom), respectively for cause specific crude mortality estimates due to heart diseases, cancer and accidents. Within each panel of each figure, different line types show the comparative results for the LC, SLC, and SSLC models. It appears in the MSE plots that the squared biases dominate the variances within the overall MSEs.

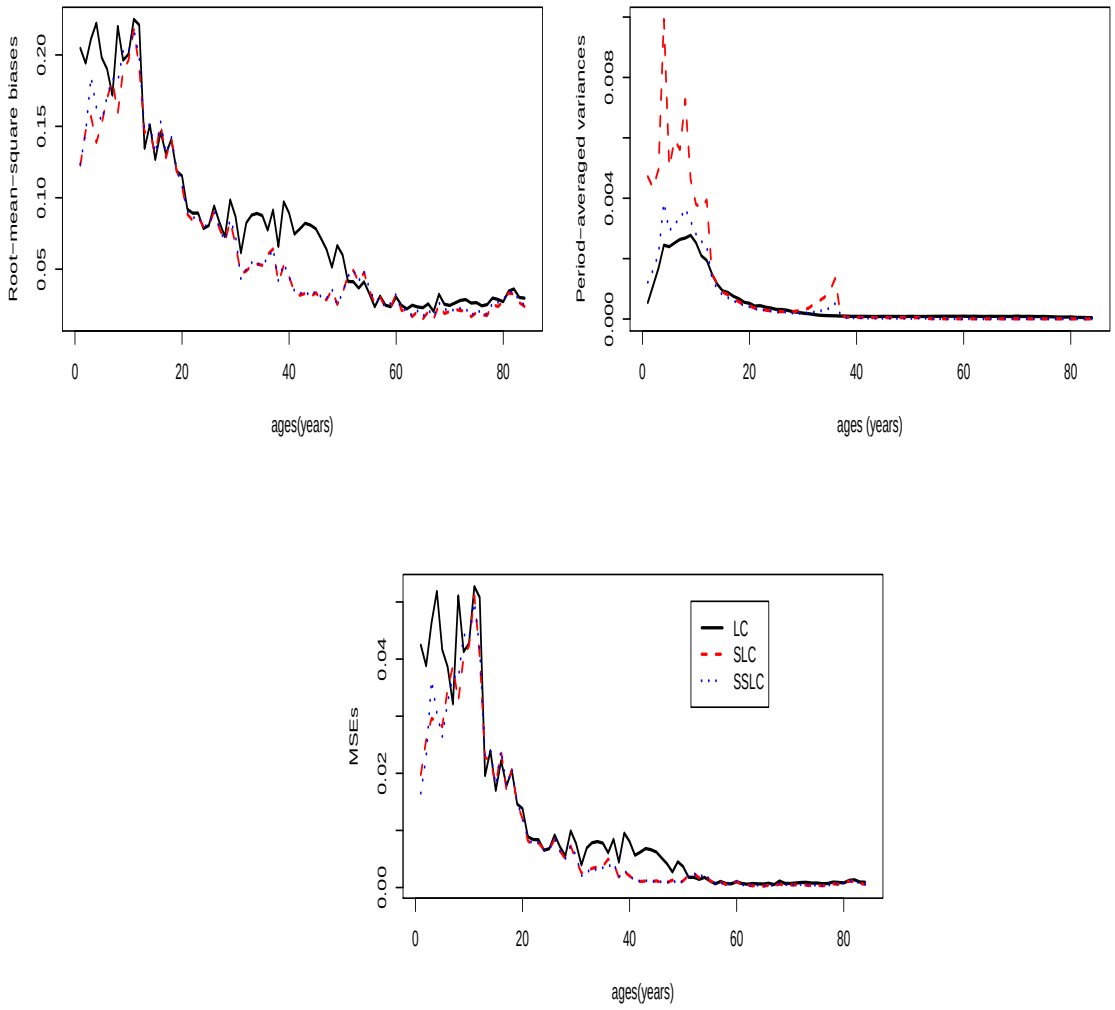


Figure 3.26: Heart diseases: The top left panel shows comparisons of root-mean-square biases among the LC, the SLC and the SSLC models of log mortality rate estimates at ages 1-84 years; the top right panel shows comparisons of the corresponding period-averaged variances; the bottom panel shows comparisons of corresponding period-averaged MSEs.

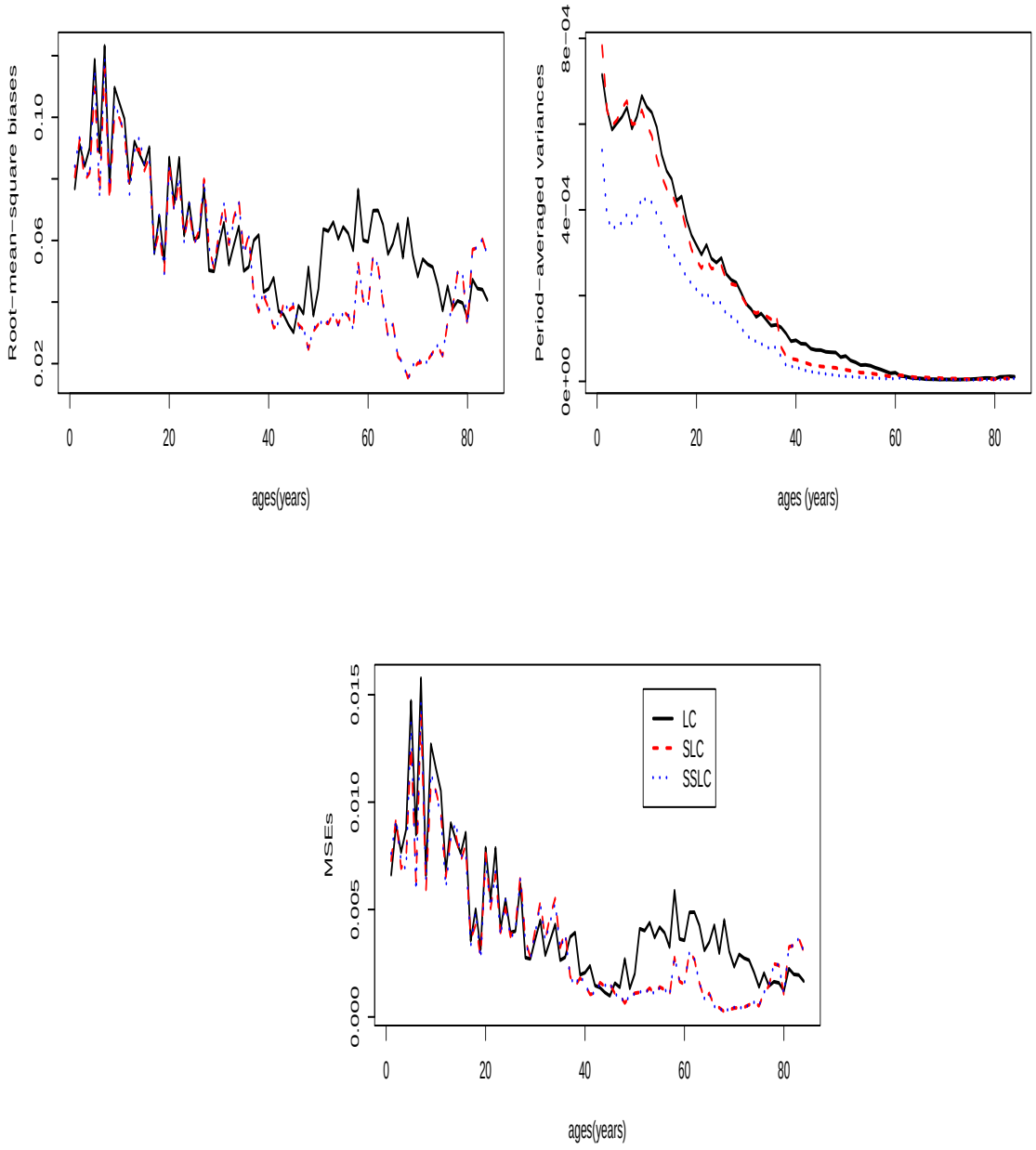


Figure 3.27: Cancer: The top left panel shows comparisons of root-mean-square biases among the LC, the SLC and the SSLC models of log mortality rate estimates at ages 1-84 years; the top right panel shows comparisons of the corresponding period-averaged variances; the bottom panel shows comparisons of corresponding period-averaged MSEs.

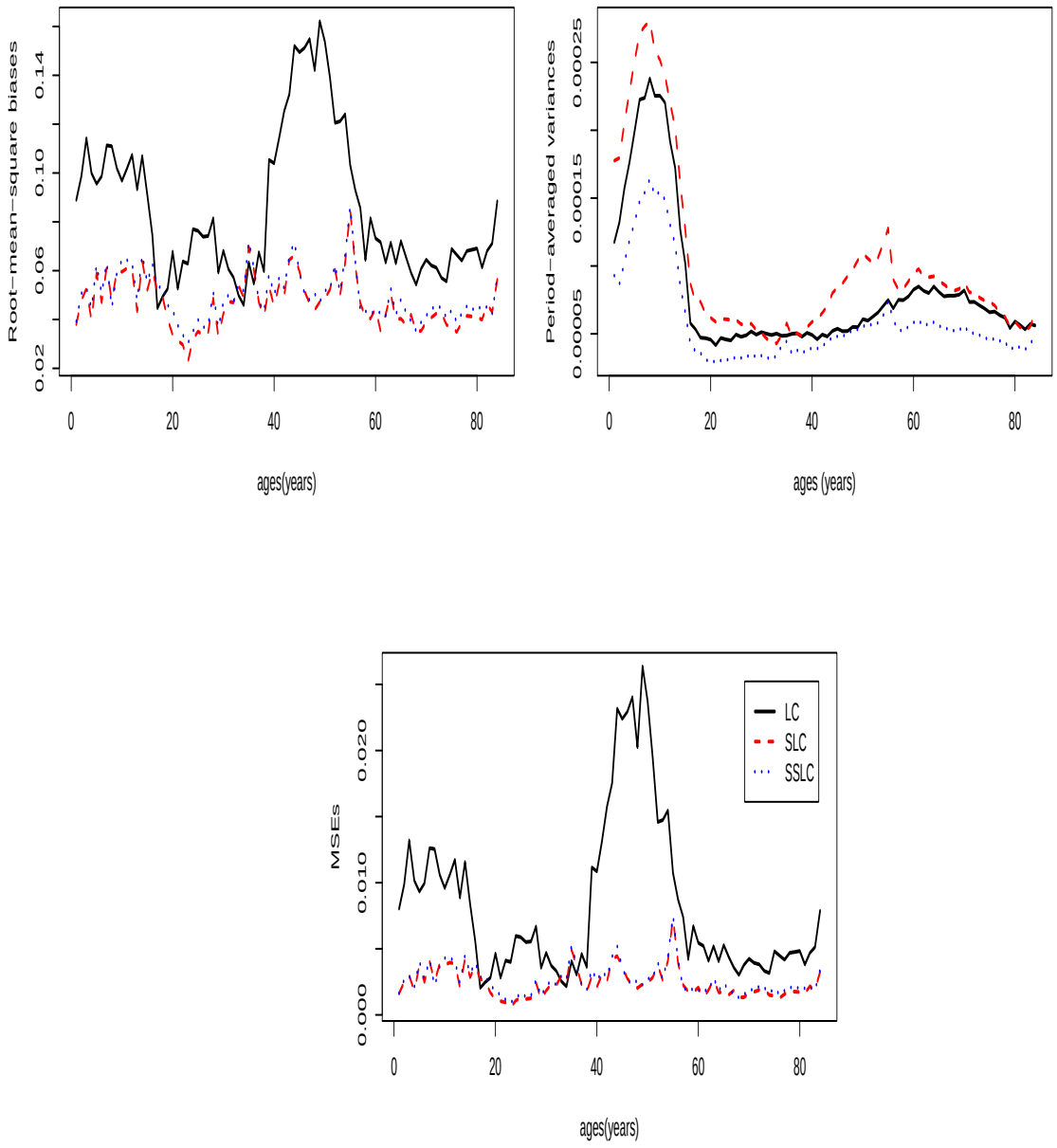


Figure 3.28: Accidents: The top left panel shows comparisons of root-mean-square biases among the LC, the SLC and the SSLC models of log mortality rate estimates at ages 1-84 years; the top right panel shows comparisons of the corresponding period-averaged variances; the bottom panel shows comparisons of corresponding period-averaged MSEs.

The left panels of Figure 3.26-3.28 suggest that the SLC and SSLC models reduce biases of estimated log mortality rates for the three causes of deaths and at most ages. They show moderate improvement in bias across the three causes of deaths, with the maximum reduction of 63%, 79% and 71% for heart diseases, cancer and accidents, respectively. The right panels showing period-averaged variances for the three models demonstrate different results for the three causes of deaths. Figure 3.26 shows that the SLC and SSLC models produce higher period-averaged variances than the LC model at earlier ages (1-37 years) with high peaks around age cut-point 36 (ages 29-38 years). The period-averaged variances under the LC model are relatively smaller at young ages and larger at old ages. Figure 3.27 suggests high period-averaged variance reductions in all ages for the SSLC model. The SLC model produces comparable period-averaged variances to the LC model at young and old ages but substantially smaller period-averaged variances at middle ages. Figure 3.28 shows that the SLC model produces substantially higher period-averaged variances than the LC model across ages while the variances are smaller for SSLC than for the other models. Common observations from Figure 3.26-3.28 are that the SLC and SSLC models produce high variances around age cut-points, and that the SSLC model produces smaller variance ratios than the SLC model by approximately $2-75\%$, $11-57\%$ and $15-53\%$, respectively, for heart diseases, cancer, and accidents.

Bootstrap Confidence Intervals

Two types of pointwise confidence intervals are studied: standard normal confidence interval and percentile confidence interval. Let $\hat{\theta}$ be an estimate of a parameter θ. The $(1-\alpha)100\%$ standard normal confidence interval is given by

$$[\hat{\theta} - z_{\frac{\alpha}{2}} \cdot se(\hat{\theta}), \hat{\theta} + z_{\frac{\alpha}{2}} \cdot se(\hat{\theta})],$$

where $se(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$. The $(1-\alpha)100\%$ percentile confidence interval is defined by

$$[\hat{\theta}_B^{\frac{\alpha}{2}}, \hat{\theta}_B^{(1-\frac{\alpha}{2})}],$$

where $\hat{\theta}_B^{\alpha}$ is the $(B\alpha)^{th}$ value in the ordered list of the B replications OF $\hat{\theta}$.

The two confidence intervals are compared by considering Percent Error which is defined by

$$\text{Percent Error} = \Big|\frac{\hat{\theta}_B - \hat{\theta}_N}{\hat{\theta}_N}\Big| \times 100\%$$

where $\hat{\theta}_B$ is the estimate computed from Percentile Interval, and $\hat{\theta}_N$ is the estimate computed from Standard Normal Interval.

Table 3.8: Maximum of Percent Error of Confidence Intervals

Causes of Deaths	Maximum Percent Error	
	Lower CIs	Upper CIs
heart diseases	0.560	0.641
cancer	1.482	1.661
accidents	0.079	0.071

Table 3.8 shows the Percent Errors of the lower and upper bounds of the Confidence intervals of log mortality rates from heart diseases, cancer and accidents. We can notice that the percent errors of the percentile confidence intervals from the standard normal confidence intervals are under 0.7% for heart diseases, under 2% for cancer, and 0.1% for accidents. These results suggest that the normality assumptions of log mortality rates are satisfied for the three causes of deaths, and the two different approaches of conducting confidence intervals give similar results.

The two pointwise confidence intervals coincide in most cases, in particular the pointwise confidence intervals for the α_a's and the log mortality rates, $\log(\lambda_{a,p})$'s. Our results indicate that the pointwise confidence interval widths for the parameter α_a's and the log mortality rates, $\log(\lambda_{a,p})$'s at old ages are much narrower than at young ages. These observations conform to a normal distributional behavior for $\log \tilde{\lambda}_{a,p}$ with variance $D_{a,p}^{(-1)}$, since the numbers of deaths are much higher at old ages than at young ages. Some differences between the two types of pointwise confidence intervals appear for the parameters β and $\gamma_{p,i}$ as the corresponding histograms of 1000 bootstrapped replications deviate from normal curves. See Chapter 10 for more details. Although the two pointwise confidence intervals are mostly comparable in our study, the percentile interval is preferred in general because it has a transformation-respecting property (Efron and Tibshirani, 1993) and avoids normality assumptions.

3.6 Discussion

This chapter proposes a new modification of the LC model in modeling historical data, by segmenting ages at death into a few age categories found by clustering age-specific mortality patterns over periods. The proposed model has advantages over the LC model in capturing variations of time trend between different age groups. The variation of time trend is not clearly seen for all-cause mortality since (Lee and Carter 1992) time trends for all-cause combined mortality are roughly linear in age. However, this approximate linearity does not seem to be valid for cause-specific mortality. Therefore our data analyses show our smoothed age-segmented model to be a superior alternative to the LC model. A further study evaluating the forecasting performance of the SSLC model could be made in the future.

The main idea of our proposed model is age-segmentation where the age groups specified in this chapter were chosen by using graphical judgement. More formal age clustering could be done. For example, we could find the number of age groups and the set of cut-points that (1) minimize the within-groups sum of squares (SSW) or (2) minimize the ratio of within-groups mean square (MSW) to between-groups mean square (MSB). The within-groups sum of squares (SSW) and the between-groups sum of squares (SSB) are defined by

$$SSW = \sum_{i=1}^{I} \sum_{a \in A_i} \sum_{p} (T_{a,p} - k_p^{(i)})^2,$$

and

$$SSB = \sum_{i=1}^{I} \sum_{p} n(A_i)(k_p^{(i)} - k_p)^2,$$

where $k_p^{(i)} = \dfrac{1}{n(A_i)} \sum_{a \in A_i} T_{a,p}$ and $k_p = \dfrac{1}{84} \sum_{a=1}^{84} T_{a,p}$, respectively. The within-groups mean square (MSW) and the between-groups mean square (MSB) are given by $MSW = \dfrac{SSW}{36(84 - I)}$, $MSB = \dfrac{SSB}{36(I - 1)}$, respectively.

The algorithm for method (2) is explained as follows:

(a) Select some candidates for the number I of age groups, and for their break-points, e.g., by considering the plots of time trends (Figures 3.1-3.3).

(b) Minimize MSW/MSB over all candidates for I and the age group cut-points.

For method (1), the algorithm is the same except that SSW is minimized over cut-points for a fixed number of age group intervals. Since SSW decreases by definition as the number of age groups increases, SSW cannot be used as a criterion select the number of age groups. Comparisons between the two methods and a method based only on graphical judgement are presented in Table 3.9.

Table 3.9: Age group specifications obtained by Minimizing the SSW, Minimizing the ratio MSW/MSB and a Graphical judgement

Diseases/Criteria	The number of age groups	Cut-points	SSW	MSW/MSB
1. Heart diseases				
- Minimizing the ratio MSW/MSB	3	12,35	11.571	0.0072
- Minimizing SSW	4	12,34,46	8.424	0.0074
- Graphical judgement	4	12,36,52	9.171	0.0082
2. Cancer				
- Minimizing the ratio MSW/MSB	3	18,60	5.829	0.0046
- Minimizing SSW	3	18,58	5.829	0.0046
- Graphical judgement	3	36,60	8.155	0.0069
3. Accidents				
- Minimizing the ratio MSW/MSB	4	15,34,53	6.168	0.0059
- Minimizing SSW	4	15,34,53	6.168	0.0059
- Graphical judgement	4	17,34,55	7.133	0.0070

Table 3.9 shows that the cut-points obtained from the three methods are quite similar, but there are some differences. For instance, the method of minimizing the ratio MSW/MSB suggests that we merge the last two age groups for heart diseases together so that there are only three age groups with cut-points 12 and 35. The number of age groups obtained from minimizing the ratio MSW/MSB is the same as that obtained by graphical judgement; the cut-points chosen by methods (1) and (2) are roughly the same as those chosen graphically, except for slight differences in the set of cut-points for cancer and accidents.

Chapter 4

The Smoothed Segmented Log-Bilinear model (SSPB)

4.1 Introduction

Sex differences play an important role in mortality trends due to differences in genes, biology and behavior between males and females [Lawlor, et al., 2001; Molarius and Johnson, 2002 ; Verbrugge, 1989, and Case and Paxson, 2005]. For instance, males have higher smoking rates than females, therefore, males have higher risk of smoking related mortality [Pampel, 2002], whereas females have higher risk of breast cancers than males. Studies on both sexes combined cannot provide sufficient information for future population planning. Therefore, research in sex specific mortality has stimulated interest from epidemiologists, actuaries, and policy makers in the last few decades. Studies on sex differences in mortality can be found in Pampel (2002) and Case and Paxson (2005). Many methodologies have been proposed for studying differences in mortality rates between males and females both for all causes of death combined and for cause specific mortality. Among these models, the Lee Carter model and its variants seem to get the most attention from demographers and policy makers. For instances, Carter and Lee (1992) applied the Lee-Carter (LC) mortality model to study sex differences in U.S. mortality using all causes of death combined data from 1933 to 1988 and drew forecasted differentials from 1990 to 2065. Booth and Tickle (2003) applied a modified LC model to study age-sex

Australian mortality using 1968-2000 data to forecast mortality to 2031. Wang and Preston (2009) applied the LC model to study sex-differences in U.S. mortality using cohort smoking histories from 1971 to 2004 to forecast mortality to 2034.

In Chapter 3, we claimed that the assumption of Lee and Carter in having only one pattern of time trends for all ages does not hold for cause-specific mortality and proposed a modification to the LC model, the Smoothed Segmented Lee-Carter model (SSLC), by segmenting ages at death into a few age categories found by clustering age-specific mortality patterns over period. That study showed improvement over the LC model in capturing time trends of mortality for cause-specific mortality data with both sexes combined. In this chapter, we study age-by-sex cause-specific mortality using U.S. cancer mortality data from 1971 to 2006 released by the National Center for Health Statistics. A Penalized Poisson Likelihood based estimation is applied to the age-segmented model in this study instead of using a penalized least squares method as used in the SSLC model to avoid the drawback of having to assume homoscedastic errors [Alho, 2000]. The Poisson likelihood version of the (Smoothed Segmented) Lee Carter model, (SS)LC, is then referred to as the (Smoothed Segmented) Poisson Log-Bilinear model, (SS)PB. Our study suggests that variations in time trends across age groups also occur in age-by-sex cause specific mortality data. Therefore statistical comparisons show improvements of the SSPB model over the PB model in capturing time trends for cancer age-sex specific mortality in both males and females. In this chapter, we also perform a study comparing the SSLC and SSPB models in Section 4.7 by using simulated datasets

where our results suggest that the two models are compatible.

Section 4.2 of this chapter describes background on cancer mortality data. Section 4.3 introduces the SSPB model and explains our fitting procedure for it. Section 4.4 gives details of a bootstrap study. Comparisons between the SSPB and PB models for age-sex mortality by using cancer mortality from 1971 to 2006 are performed in Section 4.5. Section 4.6 discusses sex-differences in U.S. cancer mortality. Section 4.7 compares SSLC and SSPB models. Section 4.8 summarizes the research.

4.2 Cancer Mortality Data

The cancer mortality data used in this study are public use mortality data files from 1971 to 2006 released by the National Center for Health Statistics[1]. The corresponding population data files are drawn from the U.S. Census Bureau to compute age specific mortality rates. The cancer mortality data in this data period are coded according to the International Classification of Diseases (ICD) revisions 8, 9, and 10, with codes 140-209, 140-180, and C00-C97, respectively. The cause-specific mortality curves show discontinuities between two consecutive ICD revisions caused by coding differences between the two ICD revisions. To smooth the mortality curves, we apply comparability ratios[2], the ratios of the numbers of deaths classified by the new revision over the numbers of deaths classified by the previous revision, published by the National Center for Health Statistics. Comparability ratios of ICD9/ICD8

[1]http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm

[2]http://www.cdc.gov/nchs/data/nvsr/nvsr49/nvsr49_02.pdf

and ICD10/ICD9 are 1.0026 and 1.0093, respectively.

4.3 An Age-Segmented Poisson Log-Bilinear Model

4.3.1 The Model

For $a = 1, 2, 3, \ldots, A$ and $p = p_0+1,\ p_0+2,\ p_0+3, \ldots,\ p_0+P$, let $D_{a,p}$ denote the number of deaths from the disease of interest at age a in year p. The year of the observation is referred as period throughout this chapter. The death counts $D_{a,p}$ are assumed to be Poisson distributed with mean $N_{a,p} \cdot \lambda_{a,p}$, where $N_{a,p}$ and $\lambda_{a,p}$ are corresponding population size and mortality rate, respectively. A direct estimator of the rate $\lambda_{a,p}$ is the ratio $\tilde{\lambda}_{a,p} = \dfrac{D_{a,p}}{N_{a,p}}$. The mortality rate for the Segmented Poisson Log-Bilinear model (SPB) is assumed to have the same form as the SLC model in Chapter 3, that is,

$$\tilde{\lambda}_{a,p} = \exp(\alpha_a + \beta_a \cdot \gamma_{p,G(a)}) \tag{4.3.1}$$

where $\sum_a \beta_a = 1$, $\sum_p \gamma_{p,i} = 0$ and $G(a) = i$ if $a \in A_i$ for $i = 1, 2, ..., I$.
The SSPB model is the SPB model with period effect terms $\gamma_{p,G(a)}$ smoothed over p. The PB model is the SPB model with only one age group, $I = 1$.

4.3.2 Age Group Segmentation

Age group segmentation is performed by minimizing the ratio of Within-Group Mean Square (MSW) to Between-Group Mean Square (MSB) of time trends in the same way as in Chapter 3, Section 3.6. The time trend $T_{a,p}$ for $a = 1, 2, 3, \ldots, 84$, and $p = 1971, \ldots, 2006$, is the log mortality rate at age a in period p after subtracting

the period averages, $T_{a,p} = \log(\tilde{\lambda}_{a,p}) - \frac{1}{36}\sum_{p=1971}^{2006} \log(\tilde{\lambda}_{a,p})$. The within-group sum of squares (SSW) and the between group sum of squares (SSB) are defined as

$$SSW = \sum_{i=1}^{I} \sum_{a \in A_i} \sum_{p} (T_{a,p} - k_p^{(i)})^2,$$

and

$$SSB = \sum_{i=1}^{I} \sum_{p} n(A_i)(k_p^{(i)} - k_p)^2,$$

where $k_p^{(i)} = \frac{1}{n(A_i)} \sum_{a \in A_i} T_{a,p}$ and $k_p = \frac{1}{84} \sum_{a=1}^{84} T_{a,p}$, respectively. The within-groups mean square (MSW) and the between-groups mean square (MSB) are given as $MSW = \frac{SSW}{36(84-I)}$ and $MSB = \frac{SSB}{36(I-1)}$, respectively. The optimal set of age cut points for sex specific mortality due to cancer are $\{18, 57\}$ and $\{18, 60\}$ with the ratios MSW/MSB of 0.00629 and 0.00458 for males and females, respectively. Detailed plots which are not shown here indicate that time trends between males and females are different for some age groups. Therefore the two sexes are studied separately and their cut points are also optimized separately in order to get better fit to the raw data and yield better results in future forecasting. If males and females are restricted to have the same cut points, then the minimized ratio $MSW/MSB = 0.00547$ is attained with the cut points $\{18, 58\}$. Using these cut points, Figures 4.1-4.2 show smooth curves of time trends of log mortality rates for males and females, respectively.

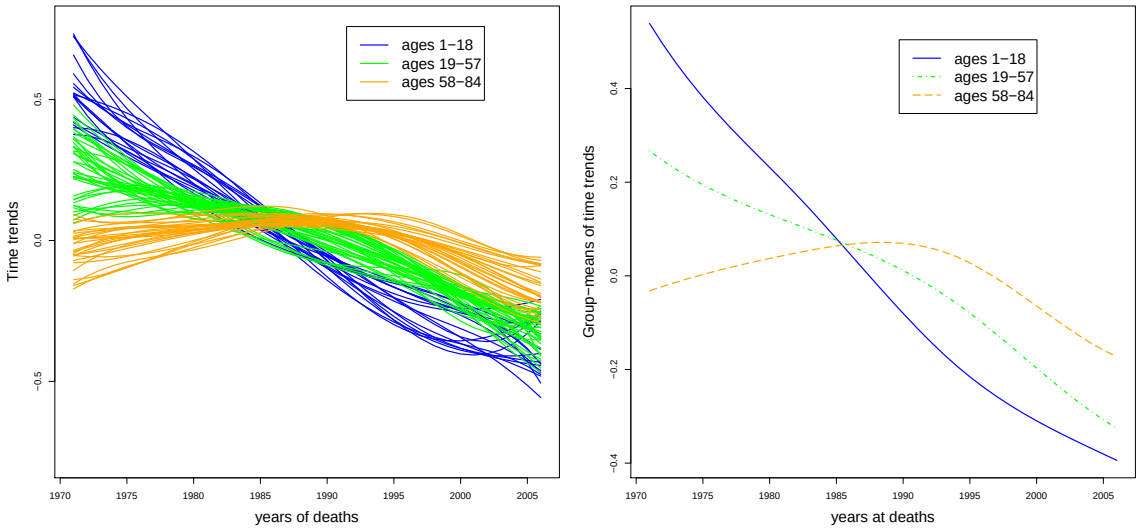


Figure 4.1: (Left) Smoothed time trends of log mortality rates from cancer for males at ages 1-84 years; (right) smoothed log mortality rates by period, averaged within age groups.

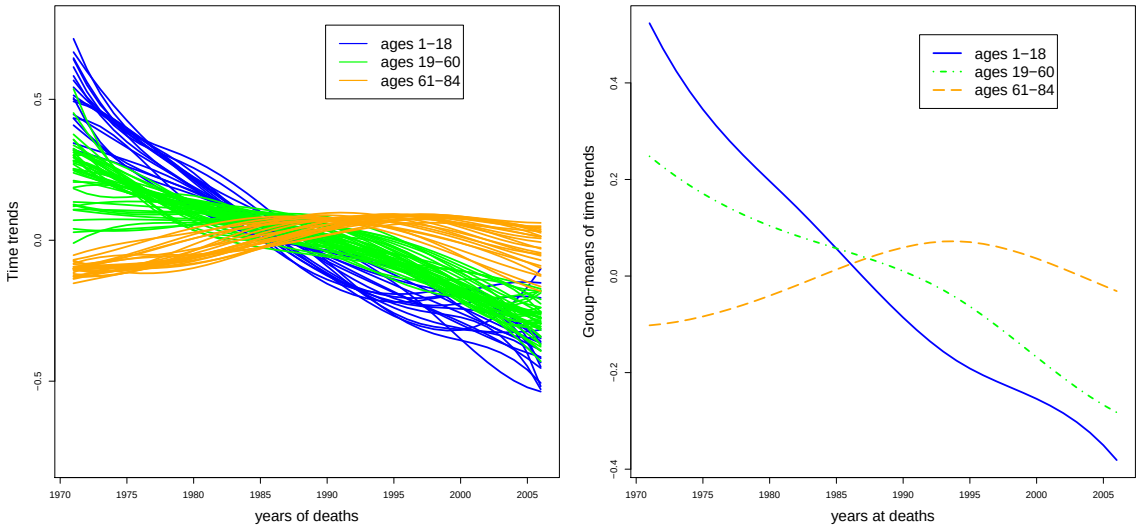


Figure 4.2: (Left) Smoothed time trends of log mortality rates from cancer for females at ages 1-84 years; (right) smoothed log mortality rates by period, averaged within age groups.

4.3.3 Fitting the model

As mentioned in Delwarde et al. (2007), the estimated $\hat{\alpha}_a$'s are usually smooth since they represent an average of mortality at age a over the data periods. No further smoothing of the $\hat{\alpha}_a$'s is needed. Therefore, we need to smooth only $\hat{\beta}_a$'s and $\hat{\gamma}_{p,G(a)}$'s. We use a penalized log-likelihood method to smooth the sequence of $\hat{\beta}$ to avoid sudden changes near the cut points of $\hat{\beta}$'s and to obtain preliminary (unsmoothed) estimates for $\hat{\gamma}_{p,G(a)}$. The sequences of $\hat{\gamma}_{p,G(a)}$ for fixed a are smoothed by using cubic smoothing splines with a restricted number of knots to reduce the number of parameters, as will be discussed in detail in Section 4.3.3.2.

4.3.3.1 Fitting and Smoothing $\hat{\beta}_a$'s using Poisson Log-Likelihood

To smooth the $\hat{\beta}_a$'s, we maximize a Penalized Poisson Log-likelihood function (PL), which is given as :

$$PL = \sum_{a,p} \left\{ D_{a,p}(\alpha_a + \beta_a \cdot \gamma_{p,G(a)}) - N_{a,p} \exp(\alpha_a + \beta_a \cdot \gamma_{p,G(a)}) \right\} - \sigma \sum_{a} (\beta_a - 2\beta_{(a-1)} + \beta_{(a-2)})^2, \tag{4.3.2}$$

where $\sum_a \beta_a = 1$, and $\sum_{p,G(a)=i} \gamma_{p,G(a)} = 0, \quad i = 1, 2, 3, ..., I.$

4.3.3.2 Selection of the Smoothing Parameters by Cross-Validation

To select the smoothness parameter σ, we follow the cross-validation method for the PB model suggested by Delwarde et al. (2007). For each $a = 1, 2, ..., A$, and

$p = p_0 + 1,\ p_0 + 2, ...,\ p_0 + P$, let

$$e_{a,p}(\sigma) = \sqrt{2} \cdot \text{sign}\big(D_{a,p} - \hat{\delta}_{a,p}^{-(a,p)}\big)\sqrt{D_{a,p} \ln\left(D_{a,p}/\hat{\delta}_{a,p}^{-(a,p)}\right) - \big(D_{a,p} - \hat{\delta}_{a,p}^{-(a,p)}\big)},$$

where

$$\hat{\delta}_{a,p}^{-(a,p)} = N_{a,p} \exp\left(\hat{\alpha}_{a,\sigma}^{-(a,p)} + \hat{\beta}_{a,\sigma}^{-(a,p)} \cdot \hat{\gamma}_{p,G(a),\sigma}^{-(a,p)}\right)$$

is the prediction of the number of deaths at age a in year p obtained by excluding the observation at age a in year p. The selected parameter is the minimizer of $\sum_{a=1}^{A} \sum_{p=p_0+1}^{p_0+P} e_{a,p}^2(\sigma)$ within a search domain.

4.3.3.3 Parameter Reduction: Smoothing $\hat{\gamma}_{p,G(a)}$ via Penalized Splines

The SPB model having more than one sequence of period effect terms increases the number of parameters of the PB model. Therefore, for fixed SPB parameters $\hat{\alpha}_a$, $\hat{\beta}_a$ and smoothing parameter σ obtained from the Penalized Poisson Log-likelihood estimation, we fit a cubic penalized spline to each sequence of period effects $\hat{\gamma}_{p,i}$'s, $i = 1, .., I$, to reduce the number of parameters. The fitting can be done with the function "smooth.spline " in the R-package "stats"[R, 2008] for each sequence of period effect terms. For each sequence of period effect terms, we compute a generalized cross-validation criterion (GCV) for the number of knots $K_i = 1, 2, ..., 10$, minimized over the penalty parameter. The number K_i that minimizes the GCV is selected. Having completed the parameter reduction, the number of parameters for each sequence of $\gamma_{p,i}$'s, $i = 1, ..., I$, is therefore reduced from the number of period effect terms minus one, $P - 1$, to the number of knots plus two, $K_i + 2$. The SPB model with smoothed period effects is then referred to as the

SSPB model.

4.4 A Bootstrap Study

In this section, we apply a Poisson bootstrap to obtain estimated MSEs and pointwise confidence intervals for estimated mortality rates.

4.4.1 An algorithm for a Poisson Bootstrap

Given an $A \times P$ matrix of observed numbers of deaths $D_{a,p}$, the Poisson bootstrap algorithm proceeds as follows:

- Generate B (=1000) replications $\{D_{a,p}^{(b)}, b = 1, \ldots, B\}$, such that for each a, p and b, $D_{a,p}^{(b)}$ is sampled from the Poisson distribution with mean $D_{a,p} = N_{a,p}\widetilde{\lambda}_{a,p}$.

- Compute $\hat{\alpha}_a^{(b)}$, $\hat{\beta}_a^{(b)}$ and $\hat{\gamma}_{p,G(a)}^{(b)}$, and $\hat{\lambda}_{a,p}^{(b)}$, for each bootstrap $\{D_{a,p}^{(b)}, b = 1, ..., B\}$. Because the cross-validation step mentioned in Section 4.3.3.1 is computationally burdensome, the smoothness parameter, $\hat{\sigma}$, for each bootstrap replication is fixed to be the smoothness parameter obtained by the cross-validation in the original data and is not allowed to vary in the bootstrap replications.

4.4.2 Bootstrap Estimation of MSEs and Confidence Intervals

Bootstrap Estimates of MSEs

For each $a = 1, 2, \ldots, A$ and $p = p_0 + 1, \ldots, p_0 + P$, the estimated MSE is defined as

$$\widehat{\mathrm{MSE}}_{(B)}(a,p) = \frac{1}{B}\sum_{b=1}^{B}\big(\hat{D}_{a,p}^{(b)} - \tilde{D}_{a,p}\big)^2.$$

Bootstrap Confidence Intervals

The $(1-\alpha)100\%$ percentile confidence interval is defined as

$$[\hat{\theta}_B^{(\alpha/2)}, \hat{\theta}_B^{(1-\alpha/2)}],$$

where $\hat{\theta}^{(\gamma)}$ is the $(B\gamma)$'th value of $\hat{\theta}^{(b)}$in the ordered list of the B iterations.

4.5 Data Analysis

In this section, we apply the SSPB model to U.S. cancer age-sex specific mortality for males and females separately. To obtain maximizers for the penalized log-likelihood function (4.3.2) subject to the corresponding constraints, the function is transformed to the following equivalent unconstrained optimization problem :

maximize :

$$\begin{aligned} NPL :=& \sum_{a,p}\big\{D_{a,p}(\alpha_a + \beta_a\cdot\gamma_{p,G(a)}) - N_{a,p}\exp(\alpha_a + \beta_a\cdot\gamma_{p,G(a)})\big\} \\ & - \sigma\sum_{a}(\beta_a - 2\beta_{(a-1)} + \beta_{(a-2)})^2 + r(\sum_{a}\beta_a - 1) + \sum_{i=1}^{I} r_i\big(\gamma_{p,i}\big), \end{aligned}$$

where r and r_i for $i = 1, \ldots, I$, are Lagrange multipliers. Numerical optimization of the Penalized Poisson Log-likelihood can be performed by using well-documented optimization functions in any well-tested software such as MATLAB or R (2008). Our analyses are performed by using a combination of the optimization functions "nlm" and "optim" in the R-package "stats"[R, 2008]. Some statistical summaries such as sum of squared deviance residuals, sum of squared Pearson residuals, sum of absolute errors and root mean squares comparing between the SSPB and PB models are shown. For each $a = 1, ..., A$ and $p = 1, ..., P$, the deviance residual is defined as

$$\text{sign}(D_{a,p} - \hat{D}_{a,p})\sqrt{D_{a,p} \cdot \ln(D_{a,p}/\hat{D}_{a,p}) - (D_{a,p} - \hat{D}_{a,p})},$$

and the Pearson residual is defined as

$$\frac{D_{a,p} - \hat{D}_{a,p}}{\sqrt{\hat{D}_{a,p}}},$$

where $\hat{D}_{a,p}$ is the estimated number of deaths under the model.

4.5.1 Male Mortality Data

Figure 4.3 shows plots of $\hat{\alpha}_a$'s and $\hat{\beta}_a$'s, $a = 1, \ldots, 84$ with varies smoothing coefficients. The optimal value of the smoothing coefficient selected from the cross-validation is 10^6. The left panel of Figure 4.4 shows a plot of period effect terms $(\hat{\gamma}_{p,i}\ ;\ p = 1971, ..., 2006\ ;\ i = 1, 2, 3)$ and their spline-smoothed curves with the numbers of knots, 8, 9, and 6, for groups 1, 2, and 3, respectively. The right panel of Figure 4.4 shows estimated time trends, $\hat{\beta}_a s(\hat{\gamma}_{p,G(a)})\ , a = 1, ..., 84$, where $s(\cdot)$ is referred to as a spline smoothed function. The figure suggests that the

estimated time trends obtained from the SSPB model are similar to the raw curves of time trends in Figure 4.1, but the estimated curves within the same group are more compressed than the raw curves. Table 4.1 shows that the SSPB model, by comparison to the PB model but with a smaller number of parameters, reduces the sum of squared deviance residuals, the sum of squared Pearson residuals, the sum of absolute errors $|D_{a,p} - \hat{D}_{a,p}|$, and root mean squares. The number of parameters used in the PB model, 202, is the sum of the number of α's (84), the number of β's minus one (83) and the number of γ's minus one (35). The number of parameters used in the SSPB model is the sum of the number of α's, the number of β's minus one and the number of cubic smoothing parameters of $\hat{\gamma}_i$'s corresponding to the number of knots, 8, 9 and 6 for age group $i = 1, 2, 3$, respectively. The number of parameters for the SSPB model is then the sum of 84, 83, 10, 11 and 8 which is 196. Table 4.2 shows comparisons of the mean absolute error within age groups. The table indicates that the SSPB model reduces the mean absolute errors by 47.75 %, 14.84 %, and 22.62 %, for age group 1,2, and 3, respectively. Figure 4.5 shows plots of raw estimates of mortality rates ages 14, 34, 44 and 74, their fitted values obtained from the PB, and SSPB models, and their corresponding 95 % bootstrap confidence intervals. The figures indicates that both models give similar results for middle age groups (for example, ages 34 and 44 years), but the SSPB model follows the patterns of raw data better for the young and old age groups (for example, ages 14, and 74 years). Figure 4.6 shows comparisons of the bootstrap MSEs, $\frac{1}{36}\sum_{p=1971}^{2006} \widehat{\text{MSE}}_{(B)}(a,p)$, at ages 1-84 years. The figure suggests that the SSPB reduces the MSEs for most ages except at ages 55-60 years, which are in the neighborhood of the cut point 57.

A detailed plot, which is not shown here shows that the increases in MSEs[3] at these ages occur because of the rapid changes in estimated time trend patterns between the two age groups while the raw patterns change gradually .

Table 4.1: Comparisons of sum of squared deviance residuals , sum of squared Pearson residuals, sum of absolute errors, and root mean squares of death counts between the PB and the SSPB models

Models	Number of parameters	Deviance residuals	Pearson residuals	Sum of absolute errors	Root mean squares
PB	202	133018.5	18472.45	278015.8	158.3407
SSPB	196	75250.9	11540.56	217733.3	127.0154

Table 4.2: Comparisons of mean absolute errors within age groups of the PB and the SSPB models.

Models	Group 1	Group 2	Group 3
PB	14.2603	45.1334	219.4528
SSPB	7.4508	38.4377	169.8063

[3]Since the smoothness parameter $\hat{\sigma}$ is assumed to be fixed for all bootstrap replications, the MSEs provided here could be underestimated.

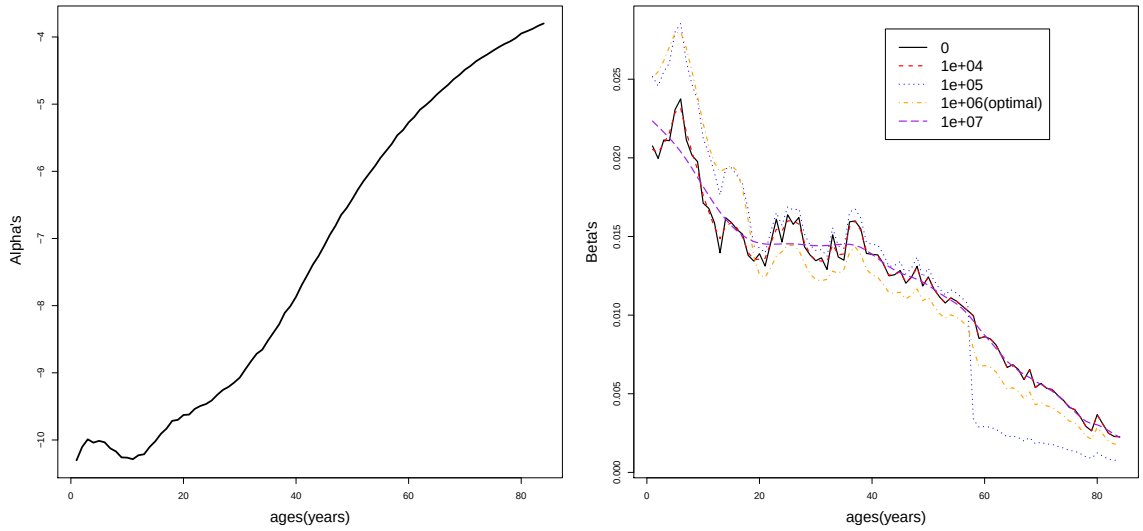


Figure 4.3: The left panel is the plot of estimates of the $\hat{\alpha}_a$'s for males ; the right panel shows curves of corresponding $\hat{\beta}_a$'s. The optimal $\hat{\sigma}$, selected by cross-validation, is 10^6.

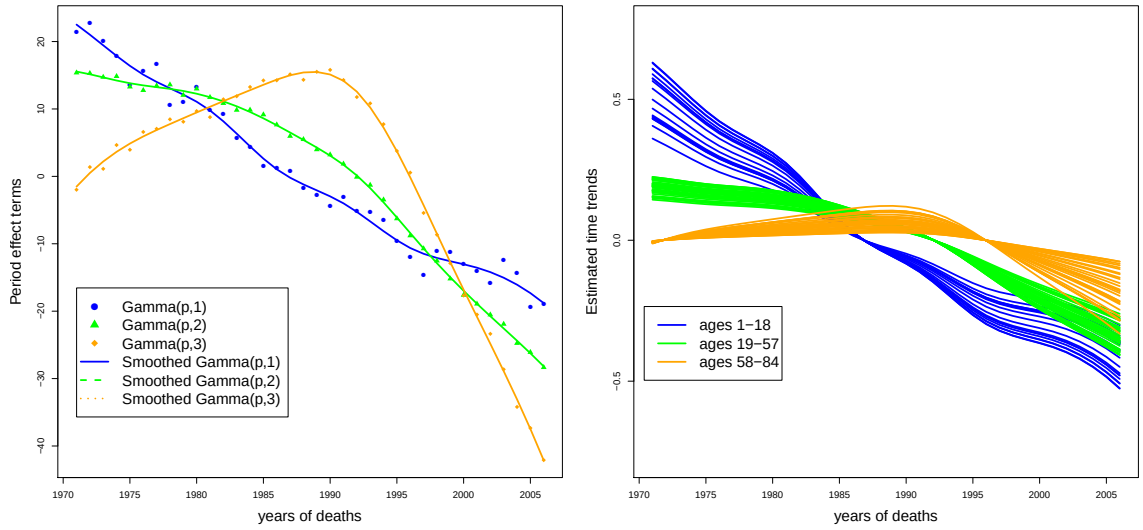


Figure 4.4: The left panel shows period effect terms($\hat{\gamma}_{p,i}$; $p = 1971, ..., 2006$; $i = 1, 2, 3$) for males and their smoothed values obtained from the SSPB model; the right panel shows the corresponding estimated time trends of log mortality rates

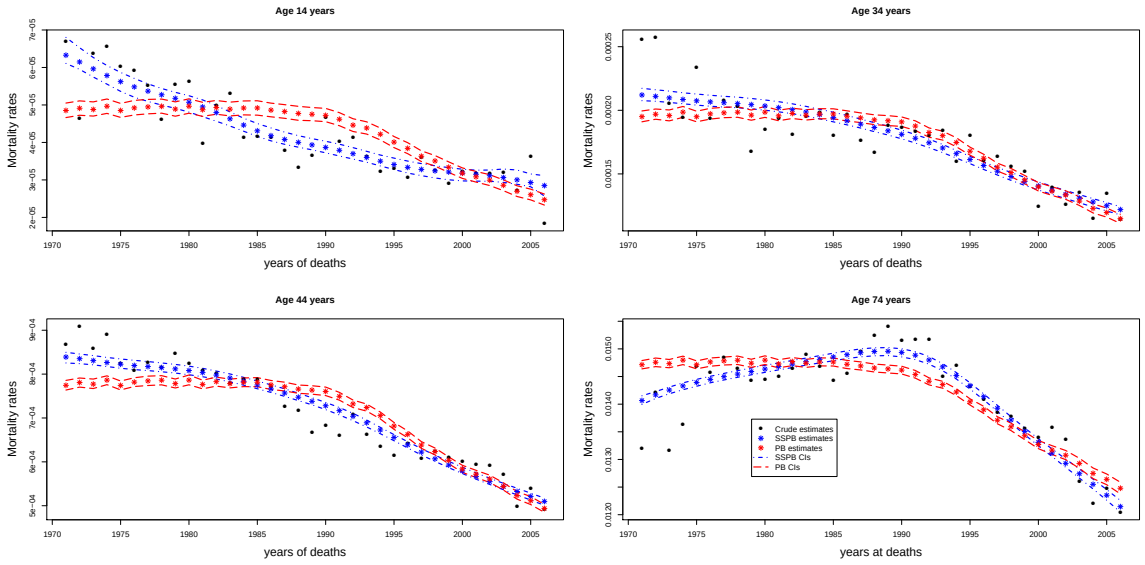


Figure 4.5: Cancer mortality rate estimates for males at selected ages obtained from PB (red) and SSPB (blue) models and corresponding 95% bootstrap percentile pointwise confidence intervals.

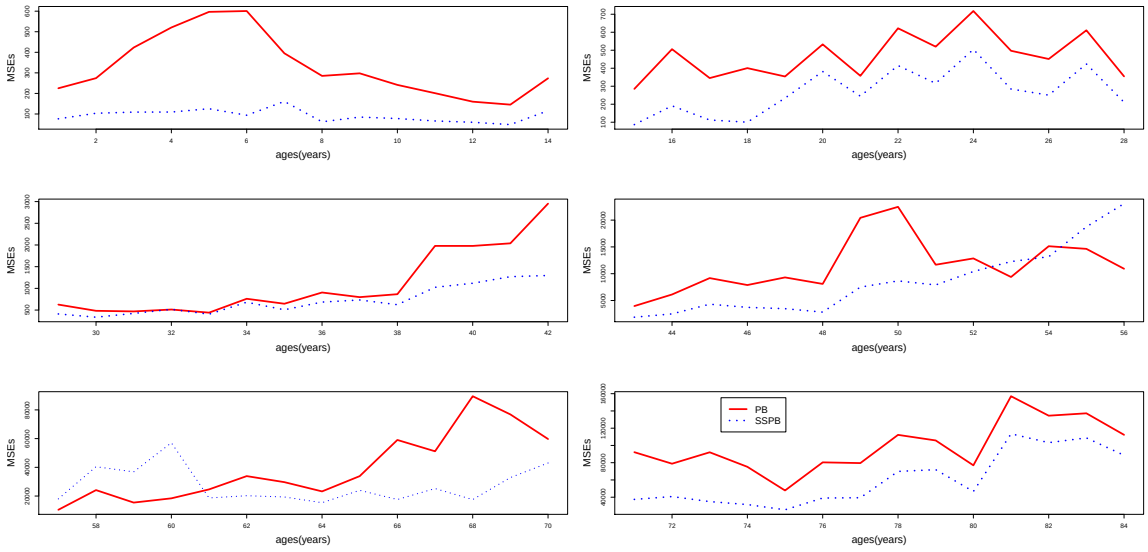


Figure 4.6: Comparisons of period-averaged MSEs of death counts for PB (red) and SSPB (blue) models.

4.5.2 Female Mortality Data

Figure 4.7 shows plots of $\hat{\alpha}_a$'s and $\hat{\beta}_a$'s $a = 1, \ldots, 84$ with different smoothing coefficients. The optimal value of the smoothing coefficient selected from the cross-validation is 1e+06. Figure 4.8 shows estimated period effect terms and their corresponding estimated time trends. Table 4.3 shows that the SSPB gives smaller values of the sum of deviance residual squared, the sum of Pearson residual squared, sum of absolute errors and root mean squares, than the PB model. The numbers of parameters presented in Table 4.3 are calculated in the same ways as in the previous section where the number of knots for the cubic smoothing spline of $\hat{\gamma}_{p,i}$'s for $i = 1, 2, 3$ are 10, 6, and 10, respectively. Table 4.4 shows the within-group mean absolute errors of the SSPB model are slightly smaller than of the PB model in groups 1 and 2, which are smaller by 13% and 16%, respectively. The SSPB model reduces the mean absolute error of the PB model in group 3 moderately, by about 32%. Figure 4.3 shows a plot of raw estimates of mortality rates at ages 14, 34, 64, and 74, their fitted values obtained from the PB and SSPB models, and their corresponding 95 % bootstrap confidence intervals. The figure indicates that fitted values from both models are similar in group 1 (for example, age 14 years) and 2 (for example, age 34 years). The SSPB model captures the patterns better than the PB model in group 3, where the raw curves are approximately quadratic but the PB model produces linear patterns with different slopes. Figure 4.10 shows comparisons of bootstrap MSEs [4]for all ages. The figure suggests that the SSPB model

[4]Since the smoothness parameter $\hat{\sigma}$ is assumed to be fixed for all bootstrap replications, the MSEs provided here may be underestimated.

gives lower MSEs at most ages in group 1 and 3, but slightly higher MSEs in group 2.

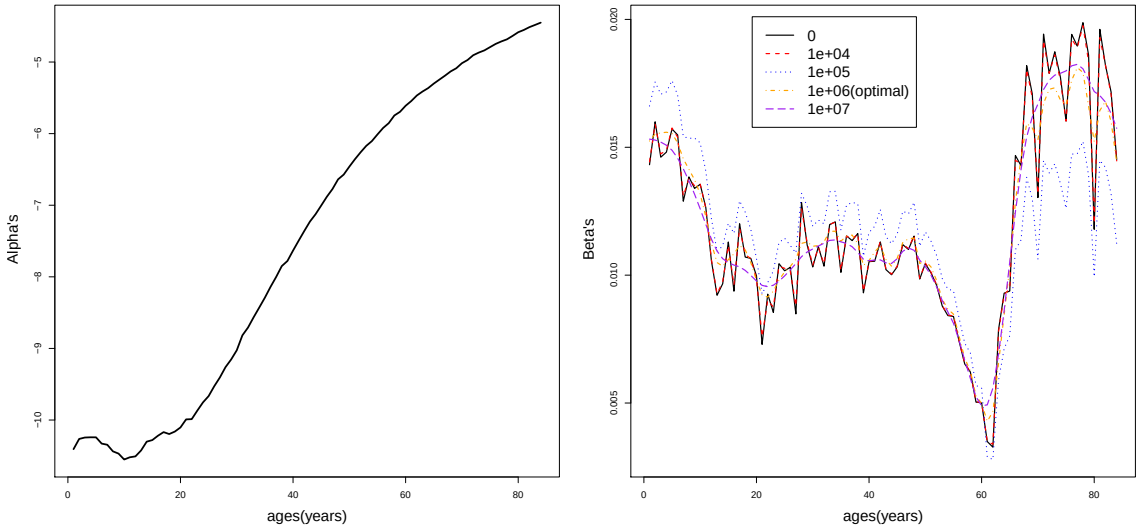


Figure 4.7: The left panel is the plot of estimates of the $\hat{\alpha}_a$'s for females ; the right panel shows curves of corresponding $\hat{\beta}_a$'s. The optimal $\hat{\sigma}$, selected by cross-validation, is 10^6.

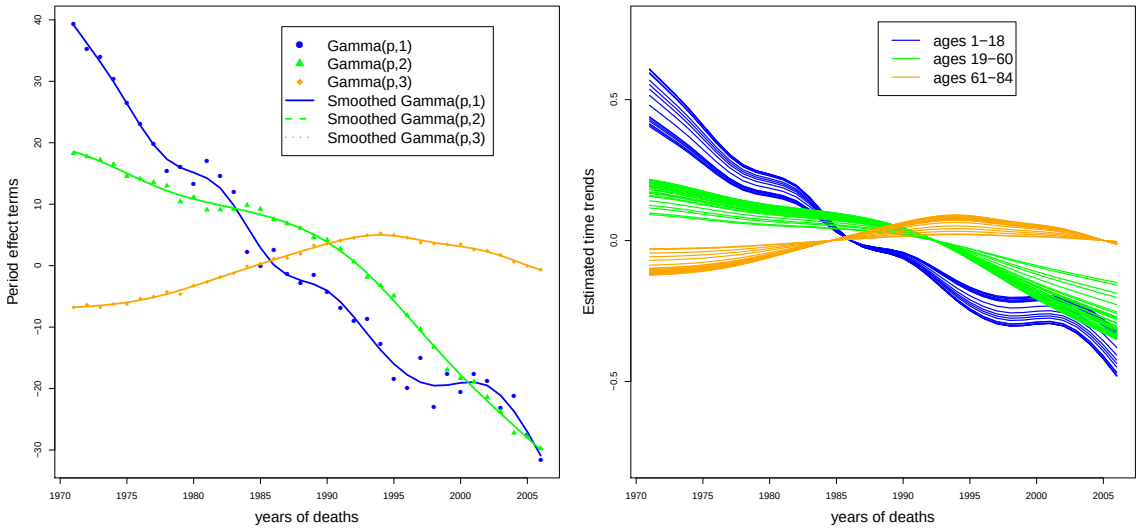


Figure 4.8: The left panel shows period effect terms($\hat{\gamma}_{p,i},\ p = 1971, \ldots, 2006;\ i = 1, 2, 3$) for females and their smoothed values obtained from the SSPB model; the right panel shows the corresponding estimated time trends of log mortality rates

Table 4.3: Comparisons of sum of squared deviance residuals , sum of squared Pearson residuals, sum of absolute errors, and root mean squares of death counts between the PB and the SSPB models

Models	Number of parameters	Deviance residuals	Pearson residuals	Sum of absolute errors	Root mean squares
PB	202	142223.6	19261.31	299118	174.4439
SSPB	199	72897.45	11611.98	217100	126.2908

Table 4.4: Comparisons of mean absolute errors within age groups of the PB and the SSPB models.

Models	Group 1	Group 2	Group 3
PB	6.9887	53.8389	246.7414
SSPB	6.0579	44.8214	168.2927

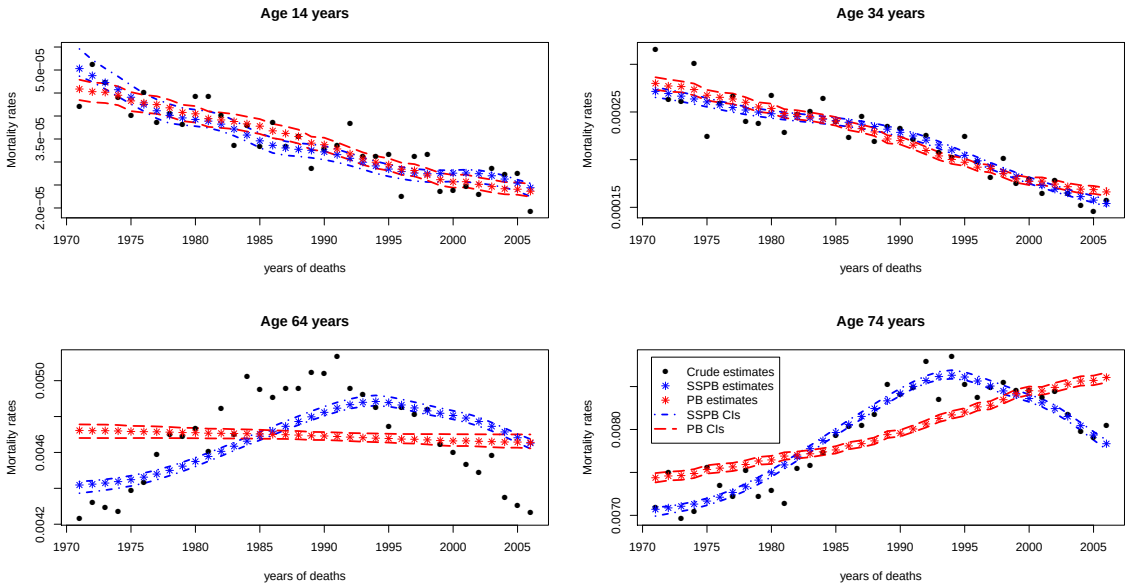


Figure 4.9: Cancer mortality rate estimates for females at selected ages obtained from PB (red) and SSPB (blue) models and corresponding 95% bootstrap percentile pointwise confidence intervals.

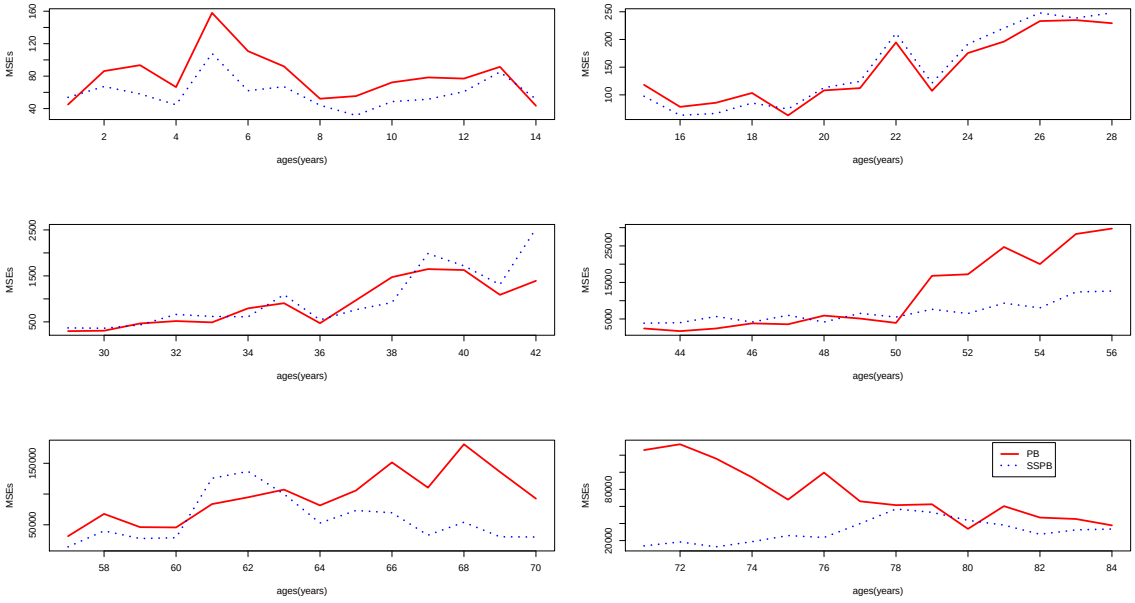


Figure 4.10: Comparisons of period-averaged MSEs of death counts between PB (red) and SSPB (blue) models.

4.6 A discussion on Sex differences in Cancer mortality

Figures 4.11 and 4.12 show that males have higher mortality rates than females at ages other than 30-47 years, an interval where the male mortality rates are lower. Our detailed data analysis which is not shown here suggests that the higher female mortality rates at ages 30-47 years are due to the much higher mortality of females than males from breast cancers. Figure 4.12 shows that differences in mortality rates between the sexes decrease as a function of time for most ages. Time trends of mortality for both sexes have a similar pattern, decreasing as a function of time in young and middle age groups. Time trends at old age groups are approximately unimodal concave for both sexes with different position of the highest peak. Males have peak mortality rates during 1985-1990, with a decreasing trend after the early 1990's, while female mortality peaks during 1990-1995 and decreases only after 1995. This lagged decrease in mortality for females could be caused by their later decreases in the percentage of smokers [Pampel, 2002]. More studies in sex-difference in smoking related-mortality can be found in Pampel (2002) and Preston and Wang (2006).

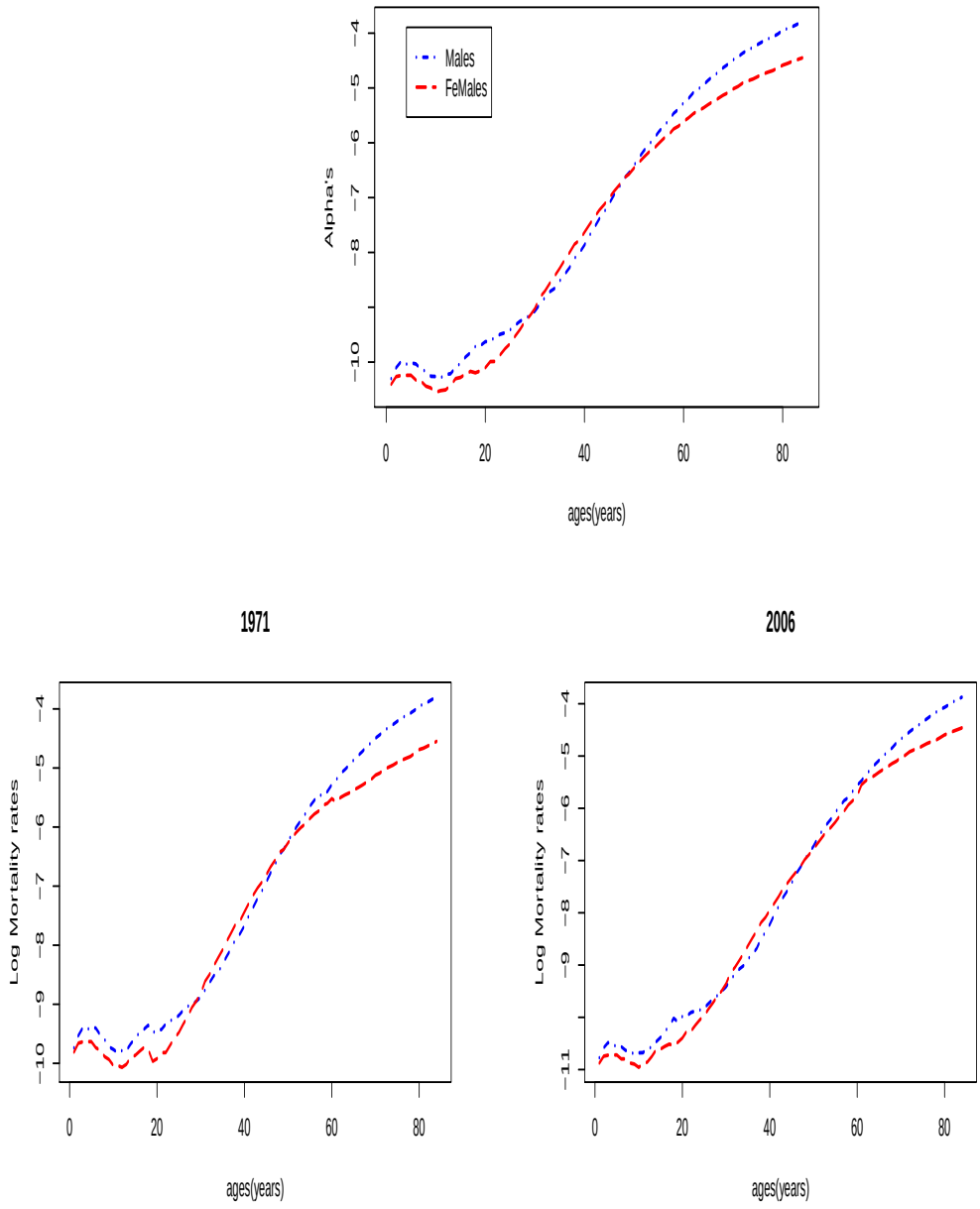


Figure 4.11: The top panel shows plots of estimates of the $\hat{\alpha}_a$'s for males and females; the bottom panels show plot of log mortality rates in 1971 (left) and 2006 (right), respectively.

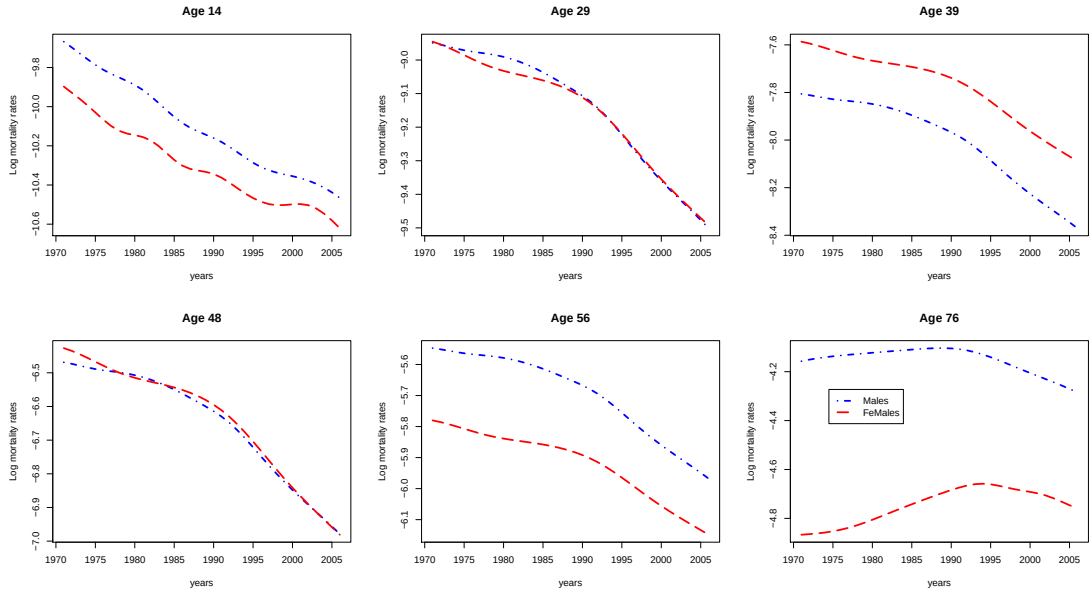


Figure 4.12: Comparisons of log mortality rates between males and females at selected ages.

4.7 Comparison of the SSLC and SSPB models

In this section, a simulation study comparing the SSLC and SSPB models and further comparisons using two mortality datasets used in Section 4.5 are discussed. Two simulated datasets used in this section are simulated from a given set of parameters, α_a^*, β_a^*, $\gamma_{p,G(a)}^*$, $a = 1, \ldots, 84$, and $p = 1, \ldots, 36$, and a given array of population sizes $N_{a,p}$, $a = 1, \ldots, 84$, and $p = 1, \ldots, 36$. To preserve the feature of cause-specific mortality data, these parameters are the parameter estimates obtained in the male cancer mortality studied in Section 4.5.1. The number of age groups and age cut points are chosen to be the same as in Section 4.5.1.

The simulation procedures are described below.

[A1] Dataset (1) is simulated from the SSLC formula: For $a = 1, \ldots, 84$ and $p = 1, \ldots, 36$,

1. generate i.i.d. $\epsilon^*_{a,p}$ from a normal distribution $\epsilon^*_{a,p} \sim \mathcal{N}(0, \sigma^2_\epsilon)$, where $\sigma_\epsilon = 0.06$, [5]

2. calculate $\log(\tilde{\lambda}_{a,p})$ from the formula:

$$\log(\tilde{\lambda}_{a,p}) = \alpha^*_a + \beta^*_a \gamma^*_{p,G(a)} + \epsilon^*_{a,p}, \tag{4.7.3}$$

3. generate the number of deaths at age a in year p, $D_{a,p} = N_{a,p}\tilde{\lambda}_{a,p}$.

[A2] Dataset (2) is simulated from the SSPB model: For $a = 1, .., 84$ and $p = 1, .., 36$,

1. calculate $\tilde{\lambda}_{a,p}$ from the formula:

$$\tilde{\lambda}_{a,p} = \exp(\alpha^*_a + \beta^*_a \gamma^*_{p,G(a)}), \tag{4.7.4}$$

2. generate the number of deaths at age a in year p, $D_{a,p}$, from the Poisson distribution with mean $N_{a,p}\tilde{\lambda}_{a,p}$, i.e., $D_{a,p} \sim \text{Poi}(N_{a,p}\tilde{\lambda}_{a,p})$.

Having simulated the datasets $\left[D_{a,p}, N_{a,p}\right]$, $a = 1, \ldots, 84$ and $p = 1, \ldots, 36$, as described above, each data set is then fitted by using SSLC and SSPB models described in Sections 3.3.3 and 4.3.3, respectively. The numbers of parameters, the

[5] This $\sigma_\epsilon = 0.06$ is the estimate $\hat{\sigma}_\epsilon = \dfrac{1}{3024} \displaystyle\sum_{a=1}^{84} \sum_{p=1971}^{2006} \left(\log(\hat{\lambda}_{a,p}) - \log(\tilde{\lambda}_{a,p})\right)^2$ obtained from the analysis in Section 4.5.1.

numbers of age groups and age cut points used in the two models are restricted to be the same as the numbers of parameters, the numbers of age groups and age cut points used to simulate the data.

In this section, we consider four statistics as estimated loss functions:

$$S_1 = \frac{1}{3024}\sum_{a,p}(\hat{D}_{a,p} - D_{a,p})^2 \tag{4.7.5}$$

$$S_2 = \sum_{a,p}(\log(\hat{D}_{a,p}) - \log(D_{a,p}))^2 \tag{4.7.6}$$

$$X_1^2 = \sum_{a,p}\frac{(\hat{D}_{a,p} - D_{a,p})^2}{\hat{D}_{a,p}}, \tag{4.7.7}$$

$$X_2^2 = \sum_{a,p}\hat{D}_{a,p}(\log(\hat{D}_{a,p}) - \log(D_{a,p}))^2. \tag{4.7.8}$$

The statistics S_1 and S_2 are special cases of the respectively statistics X_1^2 and X_2^2, respectively, where weights are equal among various a and p.

Under the assumption that $D_{a,p}$ $a = 1, \ldots, 84$ and $p = 1, \ldots, 36$ are independent Poisson residuals, $D_{a,p} \sim \text{Poi}(N_{a,p}\lambda_{a,p})$, the statistic X_1^2 is the sum of squares of $\frac{(\hat{D}_{a,p} - D_{a,p})}{\sqrt{\hat{D}_{a,p}}}$ which is approximately normal. Therefore, X_1^2 is approximately chi-squared distributed. If the estimated numbers of deaths, $\hat{D}_{a,p}$, were calculated from maximum log-likelihood estimates, we would expect X_1^2 to follow a chi-squared statistics with the degrees of freedom $n-k-1$ (Chernoff and Lehmann, 1954), where k is the number of parameter estimates and n is the number of cells in the (a, p) table.

By the delta method asymptotic approximation $\text{Var}(\log(D_{a,p})) = \frac{1}{N_{a,p}\lambda_{a,p}}$, which is estimated by $\frac{1}{\hat{D}_{a,p}}$. Therefore, X_2^2 is approximately chi-square distributed. Table 4.5 presents the four statistics calculated from the fitted values from SSLC

and SSPB models by using dataset (1) generated from the procedure given in (A1). Table 4.6 presents corresponding statistics by using dataset (2) generated from the procedure given in (A2). The tables show that the statistic S_1 slightly favors the SSPB model while the statistic S_2 slightly favors the SSLC model. The X_1^2 and X_2^2 favor the model that agrees to the assumption of dataset and produce smaller value for X_1^2 and X_2^2. For example, if the dataset was generated from the assumption of the SSLC model, the two statistics X_1^2 and X_2^2 calculated from the SSLC model will be smaller than from the SSPB model. However, the percentage of differences $\text{PD} = \dfrac{\mid S_T - S_W \mid}{S_T} \times 100$, where S_T and S_W are statistics from the true and wrong models are very small in most cases. For example, in dataset (1), PD for S_1, X_1^2 and X_2^2 are respectively 0.2%, 1.25%, and 1.7% which are very small, but PD for S_2 is quit significantly large in this dataset. In dataset (2), the PD for the four statistics are very negligible. This result and results from a few more iterations which are not shown here suggest that the two models could perform similarly and no statistically significant difference can be seen.

Table 4.5: Statistics derived from SSLC and SSPB models for a dataset (1) generated from (A1).

Statistic	SSLC	SSPB	PD (%)
S_1	58161.63	58028.66	0.2
S_2	10.57923	15.81677	49.5
X_1^2	27999.83	28321.96	1.2
X_2^2	27682.45	28156.65	1.7

Table 4.6: Statistics derived from SSLC and SSPB models for a dataset (2) generated from (A2).

Statistic	SSLC	SSPB	PD(%)
S_1	2441.144	2407.429	1.4
S_2	12.52677	12.69421	1.3
X_1^2	2859.789	2839.492	0.7
X_2^2	2858.763	2855.919	0.09

4.8 Conclusion

In this chapter, we applied an age-segmented Poisson Log-Bilinear model to U.S. cancer age-sex specific mortality data. Fitting and smoothing procedures have been described. Statistical comparisons based on deviance residuals, Pearson residuals, mean absolute errors, and root mean squared errors suggest advantages in cap-

turing time trends of age-segmentation for cancer age-sex specific mortality data. These advantages of the age-segmented model appear in the youngest age group for males and the oldest age group for females. Further studies comparing results between the two estimation methods: Penalized least squares and Penalized Log-likelihood methods are also discussed.

Chapter 5

Discussion and Future Research on SSLC and SSPB models

5.1 Discussion on Poisson Bootstrap

In Chapters 3 and 4, we applied a Poisson bootstrap to study biases and variances of parameter estimates. The Poisson bootstrap is a common method used in the Lee-Carter model and its variants in confidence interval construction because the Poisson distribution is believed to be an appropriate distribution of the numbers of deaths. However, other methods are also suggested in this family of models such as Monte Carlo variances suggested by Brouhns et al. (2005), residual bootstrap suggested by Koissi et al. (2006), or bootstrapping from other distributions of the numbers of deaths such as multinomial distribution (Brouhns et al., 2005). In this section, we review these alternative methods of variance estimation and confidence interval construction and suggest alternative methods such as bootstrapping from a binomial distribution.

5.1.1 Theoretical variance

In the context of the Poisson Log-bilinear model, we minimize the negative penalized likelihood:

$$NPL = -\frac{1}{N}\Big[\sum_{i=1}^{N} l(\theta|X_i) - \sigma p(\theta)\Big] \tag{5.1.1}$$

where $p(\theta)$ is a polynomial penalty function of degree two and the smoothness parameter is $\sigma = O(\sqrt{N})$. To obtain the asymptotic variances of parameters,

$$\nabla_\theta NPL = -\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta l(\theta|X_i) + \frac{\sigma}{N}\nabla_\theta p(\theta) \tag{5.1.2}$$

$$\nabla_\theta^{\otimes 2} NPL = -\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta^{\otimes 2} l(\theta|X_i) + \frac{\sigma}{N}\nabla_\theta^{\otimes 2} p(\theta) \tag{5.1.3}$$

Since the penalty function $p(\theta)$ is a second degree polynomial function and $\sigma = O(\sqrt{N})$, $\frac{\sigma}{N}\nabla_\theta^{\otimes 2} p(\theta)$ is negligible uniformly in θ as $N \to \infty$. Therefore $\nabla_\theta^{\otimes 2} NPL$ is asymptotically equally to $\nabla_\theta^{\otimes 2} NL$, where NL is denoted as the negative likelihood function:

$$NL = -\frac{1}{N}\sum_{i=1}^{N} l(\theta|X_i). \tag{5.1.4}$$

Therefore, under certain regularity conditions, the asymptotic variances of maximum penalized likelihood estimates are the same as of the maximum likelihood estimates, which is estimated by the observed Fisher Information matrix.

5.1.2 Monte Carlo

To obtain variances of parameter estimates via a Monte Carlo study, we follow the following algorithm given a set of parameters α_a, β_a and $\gamma_{p,G(a)}$ and an array $[N_{a,p}] : a = 1, .., A, p = p_0 + 1, ..., p_0 + P$ of the population size .

For $l = 1, .., L$,

- generate $(\alpha_a^{(l)}, \beta_a^{(l)}, \gamma_{p,G(a)}^{(l)})$ from a multivariate normal distribution, the covariance-covariance matrix can be obtained from an estimated Fisher Information matrix suggested in Brouhns et al (2002).

- for $a = 1, .., A$ and $p = p_0 + 1, ..., p_0 + P$, compute $\lambda_{a,p}^{(l)}$ from the formula:

$$\log(\lambda_{a,p}^{(l)}) = \alpha_a^{(l)} + \beta_a^{(l)} \gamma_{p,G(a)}^{(l)}.$$

- for $a = 1, .., A$ and $p = p_0 + 1, ..., p_0 + P$, generate the number of deaths $D_{a,p}^{(l)}$ by $D_{a,p}^{(l)} \sim \text{Poi}(\lambda_{a,p}^{(l)} N_{a,p}, \lambda_{a,p}^{(l)} N_{a,p})$.

- compute $(\hat{\alpha}_a^{(l)}, \hat{\beta}_a^{(l)}, \hat{\gamma}_{p,G(a)}^{(l)}$ for $a = 1, .., A$ and $p = p_0 + 1, ..., p_0 + P$) and estimated mortality rates $\left[\hat{\lambda}_{a,p}^{(l)}\right]$ from the model given in Section 4.3.3.

- the asymptotic variance of estimates $\theta \in \{\hat{\alpha}_a, \hat{\beta}_a, \hat{\gamma}_{p,G(a)}, \hat{\lambda}_{a,p}\}$ are calculated by

$$V(\theta) = \frac{1}{L} \sum_{l=1}^{L} \left(\theta^{(l)} - \bar{\theta}\right)^2. \tag{5.1.5}$$

5.1.3 Bootstrap

The bootstrapping we used in this study is a parametric bootstrap from a Poisson distribution, some alternative bootstrap techniques are

- Residual bootstrap: The residual bootstrap was applied to construct confidence intervals for parameters of the Lee-Carter and Poisson Log-bilinear models in Koissi et al. (2006). The algorithm is to generate B replications of residuals $\{r_{a,p}^{(b)}\} : b = 1, .., B$ by sampling with replacement and then compute the corresponding matrices of $\{D_{a,p}^{(b)}\}$. Then the parameter estimates are then estimated from the generated bootstrap samples.

- Multinomial bootstrap: a bootstrapping from a multinomial distribution is suggested in Brouhns et al (2005). To apply the method, the numbers of

deaths $D^{(b)}_{a,p}, D^{(b)}_{a+1,p+1}, ...$ can be generated from a multinomial with exponent $D_\bullet = \sum_{k\geq 0} D_{a+k,p+k}$ and parameters

$$\frac{D_{a,p}}{D_\bullet}, \frac{D_{a+1,p+1}}{D_\bullet}, ...$$

- Binomial bootstrap: an alternative bootstrapping is to consider the Poisson distribution as an approximation of the true Binomial distribution of death counts. The number of deaths $D_{a,p}$ can be seen as $\sum_{i=1}^{N_{a,p}} I_{i,a,p}$, where

$$I_{i,a,p} = \begin{cases} 1, & \text{if the } i^{th} \text{ individual in the risk group die at age } a \text{ in year } p; \\ 0, & \text{if the } i^{th} \text{ individual in the risk group survival year } p. \end{cases}$$

The random variable $I_{i,a,p}$, $i = 1, ..., N_{a,p}$ follows a Bernoulli distribution with $p = P(T_{i,a,p} = 1) = \lambda_{a,p}$ which is estimated by $\tilde{\lambda}_{a,p} = \frac{D_{a,p}}{N_{a,p}}$. Therefore the number of death $D_{a,p} \sim \text{Binomial}(N_{a,p}, \lambda_{a,p})$.

5.2 Future research

In chapters 3 and 4, we proposed segmented Lee-Carter models with two parameter estimation methods, penalized least squares and Poisson log-likelihood. The segmented models shown to improve to Lee-Carter in capturing time trends in mortality modeling. However, the Lee-Carter model was originally proposed for both modeling and forecasting. An important direction for future research in this area is to extend the age-segmented model in a random-effects framework to accommodate forecasting. A combination of our fitting procedure and an effective forecasting method that allows nonlinearity of time trends could dramatically improve the per-

formance of the original LC model in forecasting future age-specific mortality rates. Another important research direction is to study asymptotic theoretical properties of parameter estimates from SSLC and SSPB models. In Chapter 9 of this thesis, we study asymptotic theoretical properties of maximum penalized likelihood parameter estimates by specializing theorems of Pakes and Pollard (1989) and Chen et al. (2003). The results show consistency and asymptotic normality of parameter of SPB model under regularity conditions given in Chapter 9. However, time varying parameter estimates of SSPB ($\hat{\gamma}_{p,i}$ $i = 1, \ldots, I$) are calculated from two steps, maximum penalized likelihood and smoothing spline, which the results in Chapter 9 do not cover. Further theoretical studies on asymptotic properties of parameter estimates are needed.

Chapter 6

Phase Type Models

6.1 Introduction to Phase Type distributions

The phase type waiting time distributions introduced by Neuts in 1975 as a generalization of the Erlang distribution have been widely used in stochastic models in queueing and telecommunication (Sengupta 1989, Asmusssen 1992 , Ishay 2002, Ausin et al. 2004), traffic flow (Thümmler et al. 2006), actuarial science (Lin and Liu 2007 , Lin and Willmot 1999 2000 , Lee and Lin 2010), health care (Faddy and McClean 1999, Fackrell 2009, and Garg et al. 2011) and survival analysis (Aalen 1995, Olsson 1996).

The phase type distributions are known to be dense (in the sense of pointwise convergence of distribution functions) among all continuous distributions supported on the positive half line, and they have been fitted to many well known distributions. For example, Johnson (1993) fitted Mixtures of Erlang distributions to Lognormal, Weibull, and Uniform distributions. Thümmler et al. (2006)) fitted mixtures of Erlang distributions to Weibull, shifted exponential, Pareto II, and uniform distributions. Asmussen et al. (1996) fitted general phase type distributions to Weibull, Lognormal and uniform distribution. They are appealing because they include several of the most important constructions generally used by applied probabilists to describe realistically complex waiting time phenomena.

Even though the phase type (PH) distributions are flexible, it is known that the phase type distributions do not have unique respresentations as any given PH distribution can be represented by more than one Markov process (O'Cinneide, C.A. 1989) and the PH distributions are over-parameterized. Therefore fitting PH distributions and parameter estimation become challenging tasks. Most fitting methods avoid the problem by restricting to some specific subclasses of PH distributions. For example, Bobbio et al.(2003) restricted the PH distributions to the subclass of Acyclic Phase Type (APH) distributions, while Thümmler et al. (2006) proposed a subclass of mixtures of Erlang distributions. In this chapter, we propose a subclass of phase type models that has features of mixture and multiple states and we further study parameter estimations by two different parameter estimation techniques: direct quasi-Newton-Raphson optimization and an EM algorithm.

This chapter is organized as follows. Section 6.2 introduces the family of phase type distributions and discusses its properties. In Section 6.3, we discuss some examples of well-known phase type distributions. Section 6.4 explains our proposed class of phase type distributions. Parameter estimation of phase type distribution: direct quasi-Newton-Raphson optimization and an EM algorithm are studied in Section 6.5. In Section 6.6, we discuss our computational experience with the phase type parameter estimation. An application of the proposed class of phase type distributions to the SEER cancer dataset is presented in Section 6.7. Section 6.8 summarizes our study and discusses further research directions.

6.2 Definition and Properties of phase type Distributions

Consider a Markov process $\{Y_u\}$ with state space $\mathcal{E} = \{1, 2, 3, \ldots, m+1\}$, where states $1, 2, 3, \ldots, m$ are transient and state $m+1$ is an absorbing state. The corresponding infinitesimal generator matrix $\mathbf{Q}$ is given as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix},$$

where $\mathbf{T}$ is a $m \times m$ transition rate matrix for the transient states $1, 2, 3, \ldots, m$, and $\mathbf{t}$ is an $m \times 1$ exit rate vector to the absorbing state $m+1$. We assume further that the initial distribution is a probability vector π of length $m+1$. The random variable Y, defined as the time to absorption of the process Y_u is said to have a *phase type* distribution with a representation $(\mathbf{Q}, \pi)$. π is $(m+1)$ dimensional with $\pi_{m+1} = 0$ to avoid trivialities.

Theorem 6.1 (Neuts, 1981)**.** *Let Y have a phase type distribution with the representation $(\mathbf{Q}, \pi)$. Then*

(a) The probability distribution of Y is $F(y) = 1 - \pi \exp(\mathbf{T}y)\mathbf{e}$,

(b) the probability density function of Y is $f(y) = \pi \exp(\mathbf{T}y)\mathbf{t}$,

(c) the Laplace-Stieltjes transform of Y is $l(s) = \pi(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}$,

(d) the moment generating function of Y is $m(s) = \pi(-s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}$, and

(e) the n^{th} moment of $F(\cdot)$ is $\mu_n = (-1)^n n!(\pi \mathbf{T}^{-n}\mathbf{e})$,

where $\mathbf{e}$ is the vector of length m consisting of all 1's, $\mathbf{T}$ is nonsingular, and $\mathbf{t} = -\mathbf{T}\mathbf{e}$.

Proof. (a) Let $\mathbf{P}(t) = \{p_{ij}(t)\}_{i,j\in\mathcal{E}}$, where $p_{ij}(t) = P\big(X(t+s) = j | X(s) = i\big)$.

By Kolmogorov's differential equation,

$$\frac{d}{dt}P(t) = P(t)\mathbf{T} = \mathbf{T}P(t).$$

Define $\tau = \inf\{t \geq 0 : X_t = m+1\}$.

Then

$$\begin{aligned} P(\tau > y) &= P(X_y \leq m) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} P(X_0 = i)P(X_y = j | X_0 = i) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \pi_i p_{ij}(y) \\ &= \pi \exp(y\mathbf{T})\mathbf{e}. \end{aligned}$$

Hence $F(y) = P(\tau \leq y) = 1 - \pi \exp(y\mathbf{T})\mathbf{e}$.

(b) From (a), $f(y) = \dfrac{d}{dy}F(y) = -\pi \exp(y\mathbf{T})\mathbf{T}\mathbf{e} = \pi \exp(y\mathbf{T})\mathbf{t}$, since $\mathbf{t} = -\mathbf{T}\mathbf{e}$.

(c) The Laplace-Stieltjes transform of Y is

$$\begin{aligned} l(s) &= \int_0^\infty e^{-sy} f(y)dy \\ &= \int_0^\infty e^{-sy} \pi \exp(y\mathbf{T})\mathbf{t}dy \\ &= \pi \int_0^\infty \exp(-(s\mathbf{I} - \mathbf{T})y)\mathbf{t}dy \\ &= \pi\big(s\mathbf{I} - \mathbf{T}\big)^{-1}\mathbf{t}. \end{aligned}$$

(d) Similar to (c), $m(s) = \displaystyle\int_0^\infty e^{sy} f(y)dy = \pi\big(-s\mathbf{I} - \mathbf{T}\big)^{-1}\mathbf{t}$.

(e) For $n \in \mathbb{N}$, $\dfrac{d^n}{ds^n} m(s) = (-1)^{(n+1)}(n!)\pi\big(s\mathbf{I} + \mathbf{T}\big)^{-1-n}\mathbf{t}$.

Then $\mu_n = \dfrac{d^n}{ds^n} m(s)|_{s=0} = (-1)^{(n+1)}(n!)\pi\mathbf{T}^{-n-1}\mathbf{t} = (-1)^n(n!)\pi\mathbf{T}^{-n}\mathbf{e}$

□

Theorem 6.2 (Neuts, 1981)**.** *The states* $1, \ldots, m$ *are transient if and only if the matrix* $\mathbf{T}$ *is nonsingular.*

Proof. Let a_i for $1 \leq i \leq m$ denote the probability that the process is absorbed into the state $m+1$, starting at state i.

Therefore,

$$\begin{aligned} a_i &= p_{i,(m+1)} + \sum_{j \neq i} p_{ij} a_j, \\ &= \frac{t_i}{-T_{ii}} + \sum_{j \neq i} \frac{T_{ij}}{-T_{ii}} a_j, \end{aligned}$$

where $p_{ij} = P(X_{n+1} = j | X_n = i)$.

Therefore

$$0 = \frac{t_i}{-T_{ii}} + \sum_{j} \frac{T_{ij}}{-T_{ii}} a_j,$$

or, equivalently,

$$\mathbf{0} = \mathbf{t} + \mathbf{T}\mathbf{a}. \tag{6.2.1}$$

Since $\mathbf{t} = -\mathbf{T}\mathbf{e}$, we have

$$\mathbf{0} = \mathbf{T}\mathbf{x}, \tag{6.2.2}$$

where $\mathbf{x} = \mathbf{e} - \mathbf{a}$.

Therefore, if $\mathbf{T}$ is nonsingular, then $\mathbf{a} = \mathbf{e}$ and the probability of absorbtion at state $m+1$ is certain given that the process starts at state i for all $1 \leq i \leq m$. In

contrast, if $\mathbf{T}$ is singular, (6.2.2) has a non-zero and non-negative solution. That is there is at least one $1 \leq i \leq m$ such that $0 < a_i < 1$. Hence, by a contraposition, a probability of certain absorbtion implies that $\mathbf{T}$ is nonsingular. Therefore, the theorem is proved. □

Theorem 6.3 (Neuts, 1981)**.** *If $F(\cdot)$ and $G(\cdot)$ are both continuous PH-distributions with representations $(\mathbf{T}, \alpha)$ and $(\mathbf{S}, \beta)$ of orders m and n respectively, then their convolution $F * G(\cdot)$ is a PH-distribution with representation $(\mathbf{L}, \gamma)$, where γ is a row vector of length $m + n$:*

$$\gamma = (\alpha, \mathbf{0}_n),$$

and $\mathbf{L}$ is a $(m + n) \times (m + n)$ matrix:

$$\mathbf{L} = \begin{pmatrix} \mathbf{T} & \mathbf{tB}^0 \\ \mathbf{0} & \mathbf{S} \end{pmatrix},$$

where $\mathbf{tB}^0$ denotes the $m \times n$ matrix $\mathbf{t}\beta^T$ containing elements $t_i\beta_j$: $1 \leq i \leq m, 1 \leq j \leq n$.

Proof. From Theorem 6.1, the Laplace transform of $(\mathbf{L}, \alpha)$ is

$$i(y) = \gamma(y\mathbf{I}_{m+n} - \mathbf{L})^{-1}\mathbf{z},$$

where $\mathbf{z} = \begin{pmatrix} \mathbf{0}_m \\ \mathbf{s} \end{pmatrix}$ is a column vector of length $m + n$ and $\mathbf{s}$ is the exit rate vector of the phase type $(\mathbf{S}, \beta)$.

Therefore

$$
\begin{aligned}
i(y) &= \big(\alpha, \mathbf{0}_n\big)\big(y\mathbf{I}_{m+n} - \mathbf{L}\big)^{-1} \begin{pmatrix} \mathbf{0}_m \\ \\ \mathbf{s} \end{pmatrix} \\
&= \big(\alpha, \mathbf{0}_n\big) \begin{pmatrix} (y\mathbf{I}_m - \mathbf{T})^{-1} & (y\mathbf{I}_m - \mathbf{T})^{-1}\mathbf{t}\mathbf{B}^0(y\mathbf{I}_n - \mathbf{S})^{-1} \\ \\ \mathbf{0} & (y\mathbf{I}_n - \mathbf{S})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0}_m \\ \\ \mathbf{s} \end{pmatrix} \\
&= \begin{pmatrix} \alpha(y\mathbf{I}_m - \mathbf{T})^{-1} & \alpha(y\mathbf{I}_m - \mathbf{T})^{-1}\mathbf{t}\mathbf{B}^0(y\mathbf{I}_n - \mathbf{S})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0}_m \\ \\ \mathbf{s} \end{pmatrix} \\
&= \Big[\alpha\big(y\mathbf{I}_m - \mathbf{T}\big)^{-1}\mathbf{t}\Big]\Big[\beta\big(y\mathbf{I}_n - \mathbf{S}\big)^{-1}\mathbf{s}\Big] \\
&= l(y) \cdot k(y),
\end{aligned}
$$

where l, and k are Laplace transforms corresponding to F and G, respectively.

Hence $F * G(\cdot)$ is a PH-distribution with representation $(\mathbf{L}, \alpha)$. □

Theorem 6.4 (Neuts, 1981)**.** *A finite mixture of PH-distributions is a PH-distribution. If* $(p_1, p_2, ..., p_k)$ *is the mixing density and* $F_j(\cdot)$ *has the representation* $(\mathbf{T}(j), \alpha(j))$, $1 \leq j \leq k$, *then the mixure has the representation* $\alpha = (p_1\alpha(1), p_2\alpha_2, \ldots, p_k\alpha(k))$, *and*

$$
\mathbf{T} = \begin{pmatrix} T(1) & 0 & \cdots & 0 \\ \\ 0 & T(2) & \cdots & 0 \\ \\ \vdots & \vdots & \vdots & \vdots \\ \\ 0 & 0 & \cdots & T(k) \end{pmatrix}.
$$

Proof. Obvious by the statement. □

Theorem 6.5 (Neuts, 1981)**.** *If* $F(\cdot)$ *and* $G(\cdot)$ *are both continuous PH-distributions of random variables* X *and* Y *with representations* $(\mathbf{T}, \alpha)$ *and* $(\mathbf{S}, \beta)$ *of orders* m *and*

n respectively, then the distributions $F_1(\cdot) = F(\cdot)G(\cdot)$ *and* $F_2(\cdot) = 1 - [1 - F(\cdot)][1 - G(\cdot)]$*, corresponding to* $\min\big(X, Y\big)$ *and* $\max\big(X, Y\big)$ *are also PH-distributions, where* $F_1(\cdot)$ *has the representation* $(\mathbf{L}, \gamma)$ *of order* $mn + m + n$*, given by*

$$\gamma = \big[\alpha \otimes \beta, \mathbf{0}_m, \mathbf{0}_n\big],$$

$$\mathbf{L} = \begin{pmatrix} \mathbf{T} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{S} & \mathbf{I}_m \otimes \mathbf{s} & \mathbf{t} \otimes \mathbf{I}_n \\ \mathbf{0}_{m\times(mn)} & \mathbf{T} & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times(mn)} & \mathbf{0}_{n\times m} & \mathbf{S} \end{pmatrix}.$$

and $F_2(\cdot)$ *has the representation* $\big[\mathbf{T} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{S}, \alpha \otimes \beta\big]$*.*

We present alternative versions and proofs of Theorems 6.3-6.5 as follows.

Theorem 6.6. *Suppose that* $T_1, T_2, \ldots, T_m$ *are phase type waiting time random variables.*

(a) If $(p_1, \ldots, p_m)$ *is a probability vector, then the mixture random variable* T_* *with density* $\sum_{j=1}^m p_j\, f_{T_j}(x)$ *is also a phase type variable.*

(b) The sum $T_1 + \cdots + T_m$ *is a phase type random variable.*

(c) Both $\min\{T_j : \ j = 1, \ldots, m\}$ *and* $\max\{T_j : \ j = 1, \ldots, m\}$ *are phase type random variables.*

Proof. Let the state spaces, initial distributions, and transition intensities of the phase type Markov chains M_j whose absorption times are T_j be denoted respectively, for $1 \le j \le m$, by $s \in S_j$, by $\pi_j(s)$, and by $Q_j(s_1, s_2)$ for $s_1, s_2 \in S_j$. The S_j are disjoint. Denote the terminal (death) state in the j-th chain by D_j. In the first two parts of the proof, we define a Markov chain M with state spaces $\cup_{j=1}^m S_j$, after

identifying certain states and defining a suitable initial distribution, for which the absorption time into a designated death state D is the desired random variable.

(a) Now the initial distribution is defined for all $j = 1, \ldots, m$ and $s \in S$ by $\pi(s) = \sum_{j=1}^{m} p_j \, \pi_j(s) \, I_{[s \in S_j]}$. Define the state $D \equiv \cup_{j=1}^{m} \{D_j\}$ by lumping the death states of all the chains M_j into a single death state. The chain M (with intensity matrix Q) allows only the transitions $s \mapsto s'$ (for $s, s' \in S_j$ for some j) which can occurs in the component chains M_j. The intensity matrix of M has entries

$$Q(s, s') = \sum_{j=1}^{m} I_{[s, s' \in S_j]} \, Q_j(s, s') \qquad \text{for} \qquad s, \; s' \in S$$

All other transitions are impossible. That is, they have transition intensity 0. In this chain, the waiting time to absorption is exactly T_j if the initial state lies in S_j, which is an event of probability p_j. Therefore the unconditional absorption time is distributed according to the mixture with probabilities p_j of the distributions of the respective times T_j, as desired.

(b) Now the initial distribution is defined to be $\pi_1(\cdot)$ on S, and the overall death state for the new chain is defined as D_m. Moreover, in the newly defined chain, each transition $s \mapsto D_j$ for $j = 1, \ldots, m-1$ and $s \in S_j$ is disallowed (given intensity 0), and new transitions $(j, s) \mapsto (j+1, s')$ for all $s' \in S_{j+1}$ are included, with intensities

$$Q(s, s') = \sum_{j=1}^{m-1} I_{[s \in S_j, \, s' \in S_{j+1}]} \, Q_j(s, D_j) \cdot \pi_{j+1}(s')$$

That is, in this new chain the transitions to intermediate death-states D_j at the expiration of the successive waiting times T_j are replaced by transitions to

the starting states for the T_{j+1} chain, with probabilities according to the initial distribution for the $j+1$ chain.

(c) For each of the desired constructions in this part, the state space now consists of the cartesian product space $S' = S_1 \times S_2 \times \cdots \times S_m$; the initial distribution is defined by

$$\pi(s_1, s_2, \ldots, s_m) = \prod_{j=1}^{m} \pi_j(s_j)$$

and the allowed transitions are, for $s_k \in S_k,\ k = 1, \ldots, m,$ by

$$(s_1, s_2, \ldots, s_m) \mapsto (s_1, \ldots, s_{j-1}, s', s_{j+1}, \ldots, s_m) \qquad \text{for} \qquad s' \in S_j$$

with intensity equal to $Q_j(s_j, s')$. For this Markov chain definition, the absorbing terminal state-set is defined to be

$$D \equiv \cup_{j=1}^{m} \{(s_1, \ldots, s_m) : \ s_j = D_j \quad \text{for some} \quad j = 1, \ldots, m\}$$

in order to achieve $\min(T_1, \ldots, T_m)$ as overall absorption time; and the terminal state-set is defined as

$$D \equiv \cup_{j=1}^{m} \{(s_1, \ldots, s_m) : \ s_j = D_j \quad \text{for all} \quad j = 1, \ldots, m\}$$

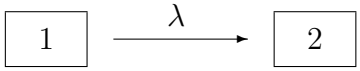
in order to achieve $\max(T_1, \ldots, T_m)$ as overall absorption time. □

Theorem 6.7 (Asmussen, 2000)**.** *The Class of phase type distributions is dense (in the sense of weak convergence) in the class of all distributions on* $(0, \infty)$.

6.3 Examples of common phase type distributions

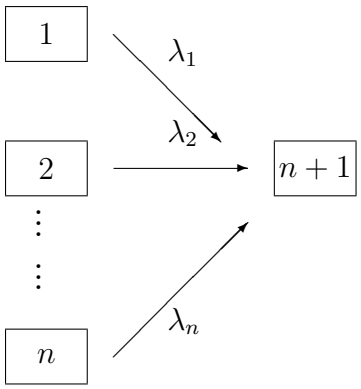
Exponential distribution

The exponential distribution is the simplest class of phase type distributions having two states: the initial state and the absorbing state:



Hyper Exponential distributions

A Hyper Exponential distribution or mixture of n Exponential distributions with intensity rates $\lambda_1, \lambda_2, \ldots, \lambda_n$ and initial distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$:

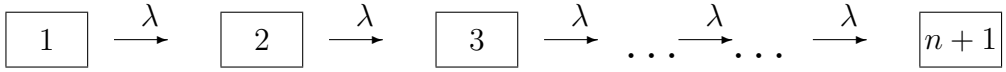


The corresponding density function is $f(s) = \sum_{i=1}^{n} \pi_i \lambda_i e^{-\lambda_i s}$. It has the representation $[\mathbf{T}, \pi]$ where

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & -\lambda_n \end{pmatrix}.$$

Erlang distributions

An Erlang distribution is a convolution of Exponential distributions with the same transition rate:

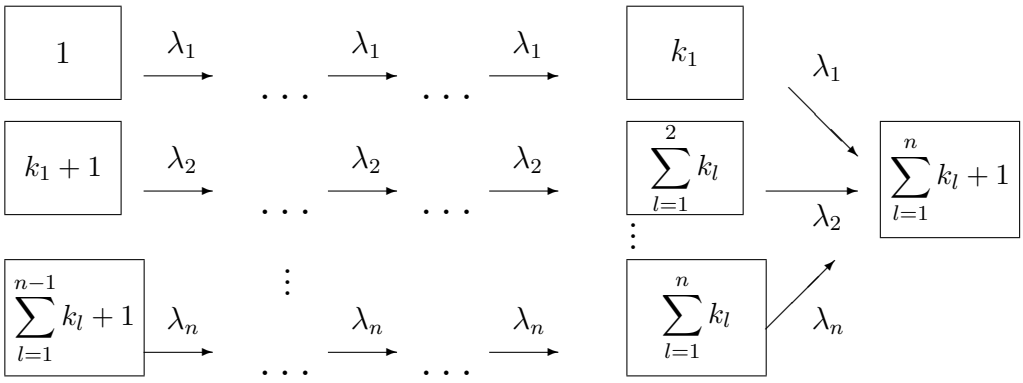


The corresponding density function is $f(s) = \dfrac{\lambda^n s^{(n-1)} e^{-\lambda s}}{(n-1)!}$ and the representation $[\mathbf{T}, -\mathbf{Te}]$ where

$$\mathbf{T} = \begin{pmatrix} -\lambda & \lambda & \cdots & 0 & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \lambda & 0 \\ 0 & 0 & \cdots & 0 & -\lambda \end{pmatrix}.$$

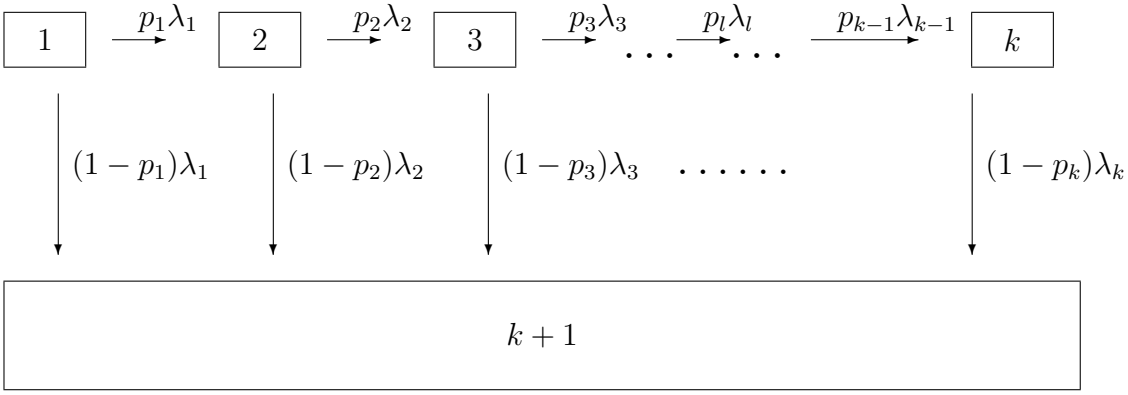
Mixture of Erlang distributions

A mixture of Erlang distributions with intensity rates $\lambda_1, \lambda_2, \ldots, \lambda_n$ and initial distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$:



Coxian distributions

A Coxian distribution is a generalization of an Erlang distribution which is allowed to reach the absorbing state $k+1$ from any transient state:



It can be represented as $[\mathbf{T}, \pi]$ where

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & p_1\lambda_1 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & p_2\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & -\lambda_k \end{pmatrix}.$$

6.4 A proposed class of phase type distributions

We propose to study statistical inference within a moderately parameterized phase type model family. The particular topology we consider, displayed in Figure 6.1 below and cited as `Model F`, seems to us particularly appropriate in a survival setting for which the time origin and initial state O correspond

to diagnosis and first treatment for a serious disease like a cancer. Immediately after treatment, direct transitions to death (state D) or a cure/quiescent state C are possible, but there may also begin a slower process of migration or mutation of existing diseased or precursor cells, along one or more pathways the selection of which might depend either on new internal biological events (e.g., mutations related to environmental or radiologic exposures) or genetics (alleles related to disease susceptibility). Because of our motivating data illustration involving breast cancer in the following Section, we also are interested in allowing the data to impose a model structure involving two separate disease paths, paths which are known (Anderson et al. 2006) to correspond to positive and negative Estrogen Receptor status in breast cancer. The Markov chain transition intensities are given in Figure 6.1, and can be understood more simply by saying that the chain begins by waiting in state O for a time $T_1 \sim \text{Expon}((1 + b_C + b_D)\,\mu)$, and then jumps to one of the states C, D, 1, or $k_1 + 1$, with respective probabilities

$$(p_C,\, p_D,\, p_1,\, p_2) \;=\; \frac{1}{1 + b_C + b_D}\,(b_C,\, b_D,\, p,\, 1 - p)$$

States C and D are absorbing. The chain may reach state D from state O in one step with probability $b_D/(1 + b_C + b_D)$. The chain reaches state 1 from state O with probability $p/(1 + b_C + b_D)$. It remains in state 1 for a random $\text{Expon}(\lambda_1 + \beta_1)$ waiting time T_1 and then either jumps to D with probability $\beta_1/(\beta_1 + \lambda_1)$ or to state 2 with probability $\lambda_1/(\beta_1 + \lambda_1)$. From state 2 it is

eventually absorbed in D after a random Gamma$(k_1 - 1, \lambda_1)$ waiting time G_1. The chain reaches state k_1+1 from state O with probability $(1-p)/(1+b_C+b_D)$. It remains in state $k_1 + 1$ for a random Expon$(\lambda_2 + \beta_2)$ waiting time T_{k_1+1} and either jumps to D with probability $\beta_2/(\beta_2 + \lambda_2)$ or to state $k_1 + 2$ with probability $\lambda_2/(\beta_2 + \lambda_2)$. From state $k_1 + 2$ it is eventually absorbed in D after a random Gamma$(k_2 - 1, \lambda_2)$ waiting time G_2. Note that if $\beta_1 = 0$, then the overall waiting time from state 1 to reach D is distributed as Gamma(k_1, λ_1). The decomposition into waiting times T_1 and G_1 accounts separately for the waiting time to leave state 1 and to progress from 2 to D on the event $1 \mapsto 2$.

In this description, the properties of Markov chains and exponential waiting times ensure that at all branches, the branching events are discrete trials independent of all waiting times. If either of the Gamma shape parameters k_j is equal to 1, then the corresponding intensity pair (β_j, λ_j) is unidentifiable and the two transition arcs with these intensities can be replaced by a single arc with transition intensity $\beta_j + \lambda_j$. Thus, if $k_j = 1$, without loss of generality $\beta_j = 0$.

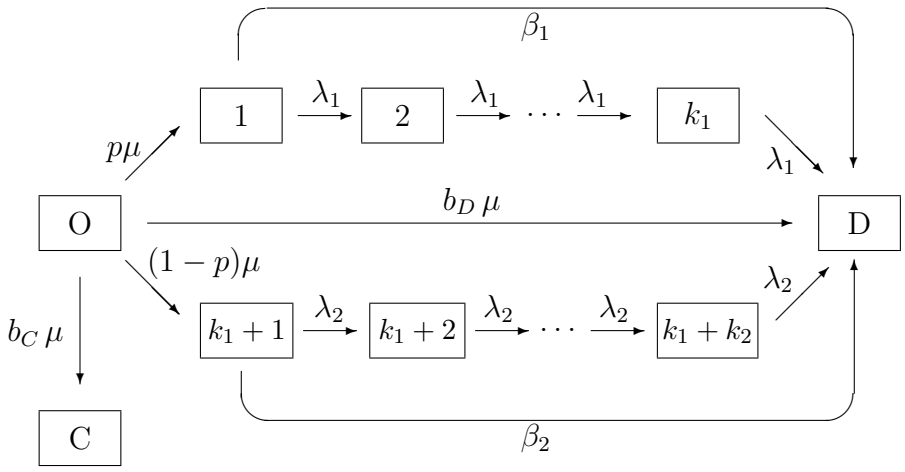


Figure 6.1: Markov transition diagram for `Model F` with immediate cures and failures, additional direct failures from states 1, $k_1 + 1$, and two failure pathways.

It is apparent from the foregoing paragraph that the absorption time density of the pictured `Model F` Markov chain is a mixture with weights p_D, p_1q_1, p_2q_2, $p_1(1-q_1)$, and $p_2(1-q_2)$ of the $\text{Expon}((1+b_C+b_D)\mu)$, $\text{Expon}((1+b_C+b_D)\mu) * \text{Expon}(\beta_1+\lambda_1)$, $\text{Expon}((1+b_C+b_D)\mu) * \text{Expon}(\beta_2+\lambda_2)$, $\text{Expon}((1+b_C+b_D)\mu) * \text{Expon}(\beta_1+\lambda_1) * \text{Gamma}(k_1-1,\lambda_1)$, and $\text{Expon}((1+b_C+b_D)\mu) * \text{Expon}(\beta_1+\lambda_1) * \text{Gamma}(k_1-1,\lambda_1)$ densities, where $*$ denotes convolution. The weights in this mixture add up to $1-p_C < 1$ because of the positive probability p_C with which the chain is absorbed at C and never hits D. The convolutions in these densities are in fact easy to write down in closed form, for positive integers k_1, k_2, which makes the densities and survival functions fully explicit and easy to compute in vectorial form in the likelihood for `Model F` based on right-censored survival data. A computing formula that allows

these calculations to be implemented simply is

$$
\begin{aligned}
P_{OD}(t) \;=\; & \frac{b_D}{1+b_C+b_D}\,(1-e^{-\mu(1+b_C+b_D)t}) \\
& + \sum_{j=1}^{2} \frac{p^{2-j}(1-p)^{j-1}}{1+b_C+b_D}\,\Big[q_j\,\mathrm{Exp}(\mu(1+b_C+b_D)) * \mathrm{Exp}(\beta_j+\lambda_j)(t) \\
& + \;(1-q_j)\,\mathrm{Exp}(\mu(1+b_C+b_D)) * \mathrm{Exp}(\beta_j+\lambda_j) * \mathrm{Gam}(k_j-1,\lambda_j)(t)\Big]
\end{aligned}
\tag{6.4.3}
$$

where for $S \sim \mathrm{Exp}(a)$, $T \sim \mathrm{Exp}(b)$, $U \sim \mathrm{Gam}(r,\lambda)$,

$$
f_{S+T}(t) \;=\; \frac{ab}{b-a}\,(e^{-at} \;-\; e^{-bt}) \;\;, \qquad f_{S+U}(t) \quad \text{also explicit.}
$$

The `Model F` Markov chains include a variety of cure models along with the Erlang-type multi-hit model considered by Armitage and Doll (1954), including special cases of that model with up to 3 distinct rates for successive mutation 'hits'. Models of these types can all be accommodated within cases of `Model F` for which $p = 0$ or $p = 1$, and we refer to the resulting phase type absorption times as 'single path Model F' densities. As a matter of notation, we refer to the single-path `model F` absorption density with $p = 1$ in Figure 6.1 as the $(b_C, b_D, \mu, \beta_1, \lambda_1)$ single path density, with shape parameter k_1 generally fixed. The `Model F` class was designed to include such single-path densities as well as a large class of two component mixtures of them, which we will find to be particularly useful in the data illustration of Section 6.7. The formal result justifying this idea is the following Lemma.

Lemma 6.1. *The mixture with weights p and $1-p$ of two single-path model F densities which have respective parameters $(b_C, b_D, \mu, \beta_1, \lambda_1)$ with shape k_1*

and $(\tilde{b}_C, \tilde{b}_D, \tilde{\mu}, \beta_2, \lambda_2)$ *with shape* k_2*, is again a* `Model F` *phase type density if and only if* $(1 + \tilde{b}_C + \tilde{b}_D)\,\tilde{\mu} = \bar{\mu} \equiv (1 + b_C + b_D)\,\mu$.

Proof. The stated condition is necessary because the two single-path models respectively have $\text{Expon}((1 + b_C + b_D)\,\mu)$ and $\text{Expon}((1 + \tilde{b}_C + \tilde{b}_D)\,\tilde{\mu})$ distributed waiting times until exit from the initial state. See the discussion immediately following Figure 6.1 to see that each of the phase type single-path densities is itself a mixture of an exponential density with other convolved density components; a mixture of two such mixtures cannot be of the same type unless the single exponential density term in both mixture components is the same.

Now suppose that the condition of the Lemma holds, and that $p \neq 1$. Then the expression of the `Model F` absorption time density with parameters

$$(p^*,\, b_C^*,\, b_D^*,\, \bar{\mu}/(1 + b_C^* + b_D^*),\, \beta_1,\, \beta_2,\, \lambda_1,\, \lambda_2)$$

as a mixture of an exponential density and convolutions is the same as the expression for the mixture with weights $p,\, 1-p$ of the two single-path model F densities as long as all three of the following equalities hold

$$\frac{p^*}{1 + b_C^* + b_D^*} = \frac{p}{1 + b_C + b_D} \quad , \qquad \frac{1 - p^*}{1 + b_C^* + b_D^*} = \frac{1 - p}{1 + \tilde{b}_C + \tilde{b}_D}$$

$$\frac{b_C^*}{1 + b_C^* + b_D^*} = \frac{p b_C}{1 + b_C + b_D} + \frac{(1-p)\tilde{b}_C}{1 + \tilde{b}_C + \tilde{b}_D}$$

.

We solve these equations explicitly for parameters $p^* \in [0,1],\ b_C^*,\ b_D^*$. First, taking ratios of the first two of these equations leads to the equality

$$\frac{p^*}{1-p^*} \;=\; \frac{1+\tilde{b}_C+\tilde{b}_D}{1+b_C+b_D} \;\cdot\; \frac{p}{1-p}$$

which uniquely determines $p^* \neq 1$. Next, substituting the first two equalities in the third shows that $b_C^* = p^* b_C + (1-p^*)\tilde{b}_C$. Also, subtracting the sums of the three equalities from 1 on each side shows that the third equality holds with C's and D's reversed, from which it follows that $b_D^* = p^* b_D + (1-p^*)\tilde{b}_D$. The proof of the Lemma is complete. □

6.5 Parameter Estimation of phase type distributions

Many fitting methods for general phase type distributions or subclasses have been proposed. Four main methods are moment matching (Bobbio et al. 2005), numerical nonlinear minimization (Johnson 1993), Expectation-Maximization (EM) algorithms (Asmussen et al. 1996, Olsson 1996), and Bayesian methods (Bladt et al. 2003, Ausin et al. 2004, and McGrory et al. 2009). In this section, we study two methods of parameter estimation which are a direct method by applying a numerical optimization and an EM algorithm proposed in Asmussen et al. (1996).

6.5.1 Direct numerical optimization

A direct numerical optimization study in this section is carried out by applying the R optimization functions "nlm" and "optim" to the density function mentioned in (6.4.3).

6.5.1.1 Simulation Results

In this section, we study a Monte Carlo simulation of the direct method by considering a specific set of parameters $(p, \mu, \beta_1, \beta_2, \lambda_1, \lambda_2) = (0.3, 2.0, 0.4, 0.6, 0.2, 0.3)$, with (b_C, b_D) fixed at $(0, 0)$, and $(k_1, k_2) = (4, 3)$, unless we specify otherwise.

Simulation method

- Generate B $(= 1000)$ replications of a sample of size 20,000 from the true parameters $(p, \mu, \beta_1, \beta_2, \lambda_1, \lambda_2) = (0.3, 2.0, 0.4, 0.6, 0.2, 0.3)$.
- Compute model-based estimators $\hat{p}^{(b)}$, $\hat{\mu}^{(b)}$, $\hat{\beta}_1^{(b)}$, $\hat{\beta}_2^{(b)}$, $\hat{\lambda}_1^{(b)}$ and $\hat{\lambda}_2^{(b)}$, for $b = 1, \ldots, 1000$.

The Monte Carlo empirical average of the (ML) parameter estimates $\hat{\theta}$ of θ is defined as

$$\hat{\theta}^{(*)} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)},$$

and the Monte Carlo estimated standard error of the ML estimator, $\widehat{SD}_M(\hat{\theta})$, is defined as

$$\widehat{\mathrm{SD}}_M(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}^{(*)})^2}.$$

Numerical Results

Parameter Estimates and Standard Errors

Table 6.1 shows the Monte Carlo average ML estimates and Monte Carlo standard errors of our case of study. Since our method finds MLE of parameters

in logit and log scales, Table 6.2 shows corresponding parameter estimates in logit and log scales comparing Monte Carlo estimates of standard errors to theoretical standard deviations, $\widehat{\mathrm{SD}_T}(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\big(-\widehat{H}(\hat{\theta})^{(b)}\big)^{-1}}$, where $\widehat{H}(\hat{\theta})$ denotes the Hessian matrix of the log-likelihood at the MLE. Its negative is called the observed information matrix.

Table 6.1: Monte Carlo Estimates and Standard Errors

sample size = 20,000, replicated B=1000 times

	True values	Estimates	Standard Errors
p	0.3	0.30258	0.09127
μ	2	1.95705	0.15710
β_1	0.4	0.39722	0.12143
β_2	0.6	0.59999	0.08375
λ_1	0.2	0.20050	0.00993
λ_2	0.3	0.30539	0.02455

Table 6.2: Monte Carlo Estimates and Standard Errors

sample size = 20,000, replicated B=1000 times

	True values	Estimates	$\widehat{\text{SD}_M}$	$\widehat{\text{SD}_T}$
logit(p)	-0.8473	-0.8740	0.4503	0.5077
$\log(\mu)$	0.6931	0.6681	0.0829	0.0858
$\log(\beta_1)$	-0.9163	-0.9777	0.3644	0.3620
$\log(\beta_2)$	-0.5108	-0.5206	0.1406	0.1640
$\log(\lambda_1)$	-1.6094	-1.6082	0.0500	0.0513
$\log(\lambda_2)$	-1.2040	-1.1894	0.0803	0.0867

Variation of Estimates with Sample Size and Model Complexity

In this section, we study performance of the parameter estimates and SE's as a function of sample size and of the number of unknown parameters in the phase type Model F specification.

(1) $(p, \mu, \beta_1, \beta_2, \lambda_1, \lambda_2) = (0.3, 2.0, 0.4, 0.6, 0.2, 0.3)$, and (b_C, b_D) were fixed at $(0, 0)$;

(2) $(p, \mu, \lambda_1, \lambda_2) = (0.3, 2.0, 0.2, 0.3)$, and $(b_C, b_D, \beta_1, \beta_2)$ were fixed at $(0, 0, 0, 0)$.

Tables 6.3 and 6.5 suggest that parameter estimates are more precise as the sample size increases, and that the sample sizes required for precise estimates depend strongly on the number of unknown parameters in the model. Tables 6.4 and 6.6 show that as expected, all eigenvalues of the observed information matrix increase as a function of sample size.

Table 6.3: Parameter estimates and Standard Errors as a function of sample size N in phase type (1) Model

	True values	N=100	N=1000	N=10000	N=20000	N=100000
logit(p)	-0.8473	2.6497	1.0691	-0.2915	-0.6784	-0.7075
(SD)		(12.7002)	(0.1324)	(0.5231)	(0.4443)	(0.3226)
log(μ)	0.6931	-0.2841	-0.2016	0.7879	0.7193	0.7257
(SD)		(0.3155)	(0.1972)	(0.0528)	(0.0362)	(0.0309)
log(β_1)	-0.9163	-0.0488	0.5981	-0.6760	-0.6298	-0.7500
(SD)		(0.7771)	(0.2739)	(0.2244)	(0.2115)	(0.1954)
log(β_2)	-0.5108	1.1044	-8.3523	-0.6643	-0.6472	-0.5876
(SD)		(5.7980)	(7.5667)	(0.1917)	(0.1203)	(0.0975)
log(λ_1)	-1.6094	-1.3040	-1.8545	-1.5847	-1.6571	-1.6281
(SD)		(0.1604)	(0.1584)	(0.0609)	(0.0537)	(0.0238)
log(λ_2)	-1.2040	-1.4369	-0.9871	-1.1050	-1.2057	-1.2180
(SD)		(3.7418)	(0.0897)	(0.1314)	(0.0845)	(0.0459)

Table 6.4: Eigenvalues of negative Hessian matrix of PH Model (1)

	N=100	N=1000	N=10000	N=20000	N=100000
1	63.6023	441.5916	2885.8493	6331.3067	32155.1794
2	39.9563	355.6560	2073.2537	2986.6660	16282.5302
3	5.6139	112.7793	1171.5814	2680.9130	12624.1922
4	0.0785	37.2936	383.8410	786.2287	4281.3955
5	0.0687	8.0810	26.9348	46.0570	139.1820
6	0.0055	0.0175	2.9148	4.0965	6.7649

Table 6.5: Parameter estimates and Standard Errors as a function of sample size N in phase type (2) Model

	Parameters	N=100	N=1000	N=10000	N=20000	N=100000
logit(p)	-0.8473	-0.4273	0.1693	-0.8167	-0.9354	-0.7541
(SD)		(0.5218)	(0.6391)	(0.1859)	(0.1469)	(0.0571)
log(μ)	0.6931	-1.1358	-0.9828	0.4574	0.7304	0.5782
(SD)		(0.9058)	(1.0260)	(0.2325)	(0.1878)	(0.0815)
log(λ_1)	-1.6094	-1.4574	-1.3981	-1.6000	-1.6227	-1.5952
(SD)		(0.0941)	(0.1538)	(0.0332)	(0.0256)	(0.0101)
log(λ_2)	-1.2040	-0.6957	-0.6187	-1.1835	-1.2213	-1.1769
(SD)		(0.2788)	(0.3469)	(0.0465)	(0.0313)	(0.0149)

Table 6.6: Eigenvalues of negative Hessian matrix of PH Model (2)

	N=100	N=1000	N=10000	N=20000	N=100000
1	154.1677	1074.7257	12862.7408	27056.9675	126073.3137
2	57.0719	775.6723	7411.2229	13426.5749	77029.7519
3	4.0446	27.0578	54.9117	88.4556	544.6693
4	0.7497	0.8771	13.6092	21.2520	119.6432

Fisher Information matrix

Let θ be the vector of k parameters of interest and let $\hat{\theta}$ be the vector of its maximum likelihood estimate. To study behavior of the observed Fisher information matrix $\hat{I}(\theta)$, we consider the precision of the linear combination $\hat{\mathbf{v}}_i^T\hat{\theta}$, where $\hat{\mathbf{v}}_i$ for $i = 1, \ldots, K$, are the unit eigenvectors of $\hat{I}(\theta)$, corresponding to the ordered eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$ of the observed Information matrix.

By the Central Limit Theorem, under the standard regularity conditions of $\sqrt{N}$-consistency and asymptotic normality of MLE's, for any vector $\mathbf{v}$ of length K,

$$\sqrt{N}\Big(\mathbf{v}^T(\hat{\theta}-\theta)\Big) \rightsquigarrow \mathcal{N}\Big(\mathbf{0}, \mathbf{v}^T[I(\theta)]^{-1}\mathbf{v}\Big), \tag{6.5.4}$$

where $I(\theta)$ is the per-observation Fisher information matrix.

By the Spectral Decomposition, the observed Fisher information matrix $\hat{I}(\hat{\theta})$

can be represented as

$$\hat{I}(\hat{\theta}) = \begin{pmatrix} \hat{\mathbf{v}}_1 & \hat{\mathbf{v}}_2 & \cdots & \hat{\mathbf{v}}_n \end{pmatrix} \begin{pmatrix} \hat{\lambda}_1 & 0 & 0 & 0 \\ 0 & \hat{\lambda}_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\lambda}_k \end{pmatrix} \begin{pmatrix} \hat{\mathbf{v}}_1^T \\ \hat{\mathbf{v}}_2^T \\ \vdots \\ \hat{\mathbf{v}}_k^T \end{pmatrix}.$$

Therefore,

$$[\hat{I}(\hat{\theta})]^{-1} = \begin{pmatrix} \hat{\mathbf{v}}_1 & \hat{\mathbf{v}}_2 & \cdots & \hat{\mathbf{v}}_k \end{pmatrix} \begin{pmatrix} \frac{1}{\hat{\lambda}_1} & 0 & 0 & 0 \\ 0 & \frac{1}{\hat{\lambda}_2} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{\hat{\lambda}_k} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{v}}_1^T \\ \hat{\mathbf{v}}_2^T \\ \vdots \\ \hat{\mathbf{v}}_k^T \end{pmatrix}.$$

Hence, for all $i = 1, \ldots, k$,

$$\begin{aligned} \widehat{\mathrm{Var}}\Big(\mathbf{v_i}^T(\hat{\theta} - \theta)\Big) &= \frac{1}{N}\hat{\mathbf{v}}_\mathbf{i}^T[\hat{I}(\hat{\theta})]^{-1}\hat{\mathbf{v}}_\mathbf{i} \\ &= \frac{1}{N}\hat{\mathbf{v}}_\mathbf{i}^T \begin{pmatrix} \hat{\mathbf{v}}_1 & \hat{\mathbf{v}}_2 & \cdots & \hat{\mathbf{v}}_k \end{pmatrix} \begin{pmatrix} \frac{1}{\hat{\lambda}_1} & 0 & 0 & 0 \\ 0 & \frac{1}{\hat{\lambda}_2} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{\hat{\lambda}_k} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{v}}_1^T \\ \hat{\mathbf{v}}_2^T \\ \vdots \\ \hat{\mathbf{v}}_k^T \end{pmatrix} \hat{\mathbf{v}}_\mathbf{i} \\ &= \frac{1}{N\hat{\lambda}_i}. \end{aligned} \tag{6.5.5}$$

By the Continuous mapping theorem, $\hat{\lambda}_i$ converges to λ_i, where $\lambda_i, i = 1, \ldots, k$ are the eigenvalues of the Fisher Information matrix $I(\theta)$.

In the phase type model specification (1), the per-observation observed Fisher Information $\hat{\mathbf{I}}(\hat{\theta})$, for sample size $N = 10^5$ has eigenvalues $3.2155\text{e}-01, 1.6283\text{e}-$

$01, 1.2624e-01, 4.2814e-02, 1.3918e-03, 6.7649e-05$. By (6.5.5), the standard errors obtaining by taking reciprocal square roots, $\sqrt{\frac{N}{\lambda}}$, are $\frac{1.7635}{\sqrt{N}}$, $\frac{2.4782}{\sqrt{N}}$, $\frac{2.8145}{\sqrt{N}}$, $\frac{4.8329}{\sqrt{N}}$, $\frac{26.8045}{\sqrt{N}}$, $\frac{121.5820}{\sqrt{N}}$, respectively for the linear combinations of the parameter estimates $v_1 \operatorname{logit}(\hat{p}) + v_2 \log(\hat{\mu}) + v_3 \log(\hat{\lambda}_1) + v_4 \log(\hat{\lambda}_2)) + v_5 \log(\hat{\beta}_1) + v_6 \log(\hat{\beta}_2)$ where $\mathbf{v} = (v_1, .., v_6)$ is successively replaced by each of the six unit eigenvectors of the Information matrix. For the moderate sample size $N = 1000$, the first eigenvector parameter combination $-0.0359 \operatorname{logit}(\hat{p}) - 0.3966 \log(\hat{\mu}) - 0.3224 \log(\hat{\lambda}_1) - 0.7823 \log(\hat{\lambda}_2)) + 0.2087 \log(\hat{\beta}_1) + 0.2863 \log(\hat{\beta}_2) = -0.2496$ with predicted standard error of 0.0558. The sixth eigenvector combination is $0.8331 \operatorname{logit}(\hat{p}) + 0.0565 \log(\hat{\mu}) + 0.4912 \log(\hat{\lambda}_1) - 0.2274 \log(\hat{\lambda}_2)) + 0.0265 \log(\hat{\beta}_1) + 0.0950 \log(\hat{\beta}_2) = -0.9420$ with predicted standard error of 3.8448, which shows ill-conditioning of the Information matrix. In phase type model specification (2), the observed Fisher Information $\hat{\mathbf{I}}(\hat{\theta})$, for large sample $N = 10^5$ has eigenvalues $1.2601, 0.770, 0.0054, 0.0012$. By (6.5.5), the standard errors obtaining by taking reciprocal square roots, $\sqrt{\frac{N}{\lambda}}$, are $\frac{0.891}{\sqrt{N}}$, $\frac{1.139}{\sqrt{N}}$, $\frac{13.550}{\sqrt{N}}$, $\frac{28.911}{\sqrt{N}}$, respectively for the linear combinations of the parameter estimates $v_1 \operatorname{logit}(\hat{p}) + v_2 \log(\hat{\mu}) + v_3 \log(\hat{\lambda}_1) + v_4 \log(\hat{\lambda}_2))$ for each of the four unit eigenvectors $\mathbf{v} = (v_1, .., v_4)$ of the Information matrix. For the moderate sample size $N = 1000$,the first eigenvector parameter combination $0.2155 \operatorname{logit}(\hat{p}) - 0.0749 \log(\hat{\mu}) - 0.4247 \log(\hat{\lambda}_1) - 0.8761 \log(\hat{\lambda}_2) = 1.5028$ with predicted standard error of 0.028. The fourth eigenvector combination is $0.482 \operatorname{logit}(\hat{p}) - 0.859 \log(\hat{\mu}) + 0.082 \log(\hat{\lambda}_1) + 0.152 \log(\hat{\lambda}_2)) = -1.170$ with predicted standard error of 0.814 which again indicates ill-conditioning of the Information matrix results.

Next , we study asymptotic properties of the Hessian matrix by comparing two estimates of the per-observation Fisher Information Matrix: (1) Fisher Information matrix based on one sample of size 200,000, $\hat{I}_1(\theta) = \frac{-\widehat{H(\theta)}}{200000}$; (2) Fisher Information matrix based on B (= 1000) replicated simulations of samples of size 20,000, $\hat{I}_2(\theta) = \frac{1}{B}\sum_{b=1}^{B}\frac{-\widehat{H(\theta)}^{(b)}}{20000}$. The specific phase type model studied is phase type (1). Table 6.7 and 6.8 show that $\hat{I}_1(\theta)$ and $\hat{I}_2(\theta)$ are very close which agrees well with the predicted large-sample convergence of parameter and Information estimates.

Table 6.7: Fisher Information matrix based on one iteration of 200000 simulated samples, $\hat{I}_1(\theta) = \frac{-\widehat{H(\theta)}}{200000}$.

	logit(p)	log(μ)	log(β_1)	log(β_2)	log(λ_1)	log(λ_2)
logit(p)	0.0148	-0.0048	-0.0179	-0.0046	-0.0229	-0.0315
log(μ)	-0.0048	0.0950	0.0337	0.1011	0.0041	-0.0191
log(β_1)	-0.0179	0.0337	0.0542	0.0859	0.0095	0.0065
log(β_2)	-0.0046	0.1011	0.0859	0.2160	-0.0430	-0.0621
log(λ_1)	-0.0229	0.0041	0.0095	-0.0430	0.1512	0.0099
log(λ_2)	-0.0315	-0.0191	0.0065	-0.0621	0.0099	0.1349

Table 6.8: Fisher Information matrix based on B (= 1000) iterations of 20000 simulated samples, $\hat{I}_2(\theta) = \dfrac{1}{B}\sum_{b=1}^{B} \dfrac{-\widehat{H(\theta)}^{(b)}}{20000}$.

	logit(p)	log(μ)	log(β_1)	log(β_2)	log(λ_1)	log(λ_2)
logit(p)	0.0193	-0.0061	-0.0179	-0.0106	-0.0240	-0.0296
log(μ)	-0.0061	0.1003	0.0316	0.1070	0.0038	-0.0213
log(β_1)	-0.0179	0.0316	0.0562	0.0733	0.0105	0.0101
log(β_2)	-0.0106	0.1070	0.0733	0.2203	-0.0411	-0.0631
log(λ_1)	-0.0240	0.0038	0.0105	-0.0411	0.1644	0.0071
log(λ_2)	-0.0296	-0.0213	0.0101	-0.0631	0.0071	0.1337

Histogram of parameter estimates (in log and logit scales)

In this section, we exhibit histograms of parameter estimates based on 1000 Monte Carlo iterations with sample size 20000. Parameter estimates are presented in transformed scales (logit for parameter p, and log for the other parameter arguments). A specific set of parameters $(p, \mu, \beta_1, \beta_2, \lambda_1, \lambda_2) = (0.3, 2.0, 0.4, 0.6, 0.2, 0.3)$, and (b_C, b_D) were fixed at $(0, 0)$, and k_1 and k_2 were respectively fixed at 3 and 2. The overlaid normal curves are centered at the true values and the standard errors are derived from $\widehat{\mathrm{SD}_T}(\theta) = \sqrt{\dfrac{1}{B}\sum_{b=1}^{B} \mathrm{diag}\big(-\widehat{H}(\theta)^{(b)}\big)^{-1}}$. Figures 6.2-6.7 show some skewness of Monte Carlo simulations of parameter estimates.

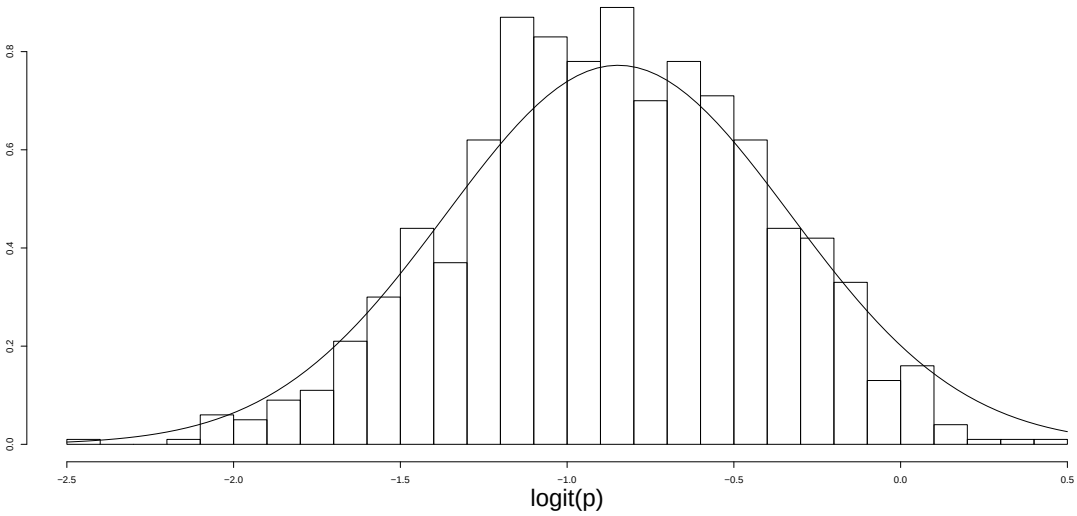


Figure 6.2: Monte Carlo histogram for logit(p).

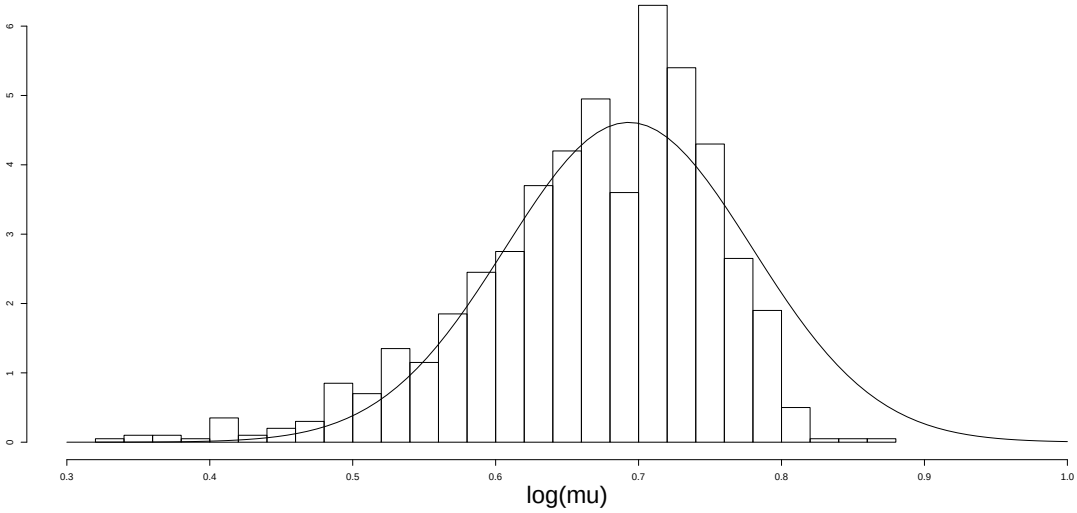


Figure 6.3: Monte Carlo histogram for $\log(\mu)$.

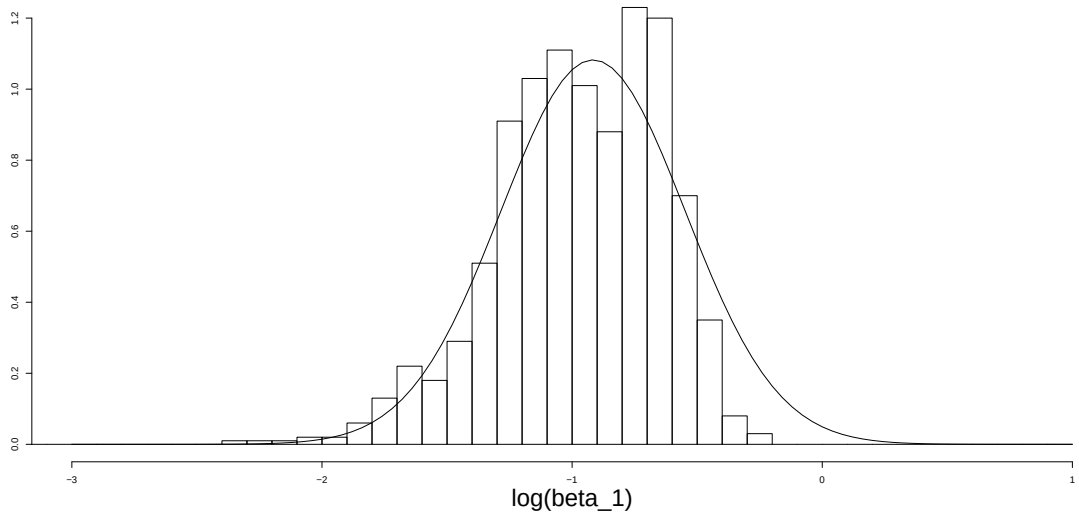


Figure 6.4: Monte Carlo histogram for $\log(\beta_1)$.

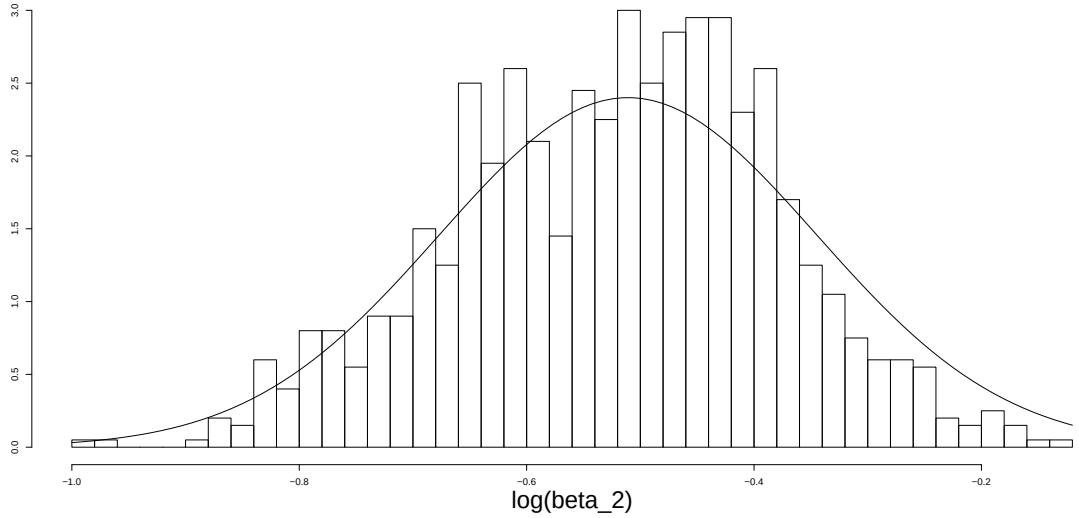


Figure 6.5: Monte Carlo histogram for $\log(\beta_2)$.

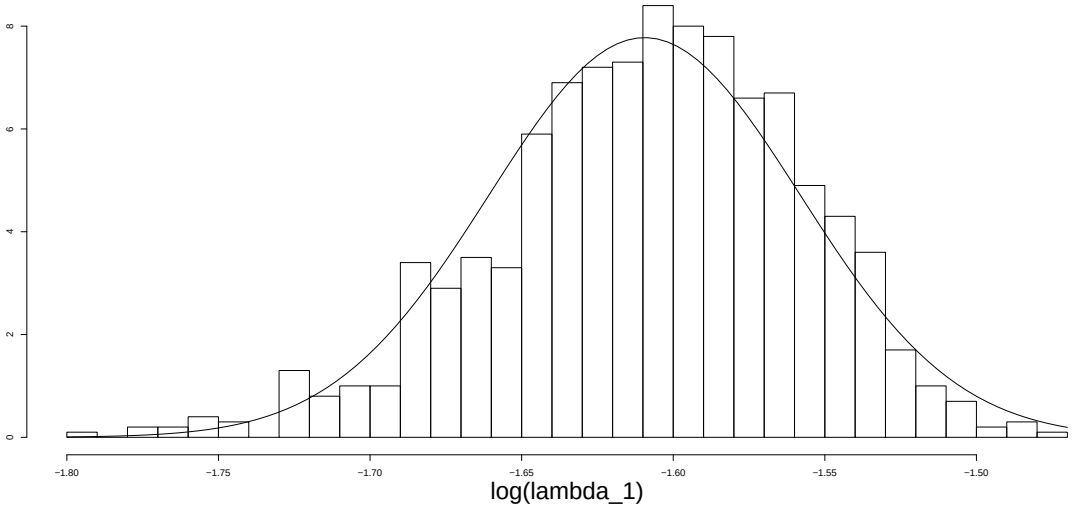


Figure 6.6: Monte Carlo histogram for $\log(\lambda_1)$.

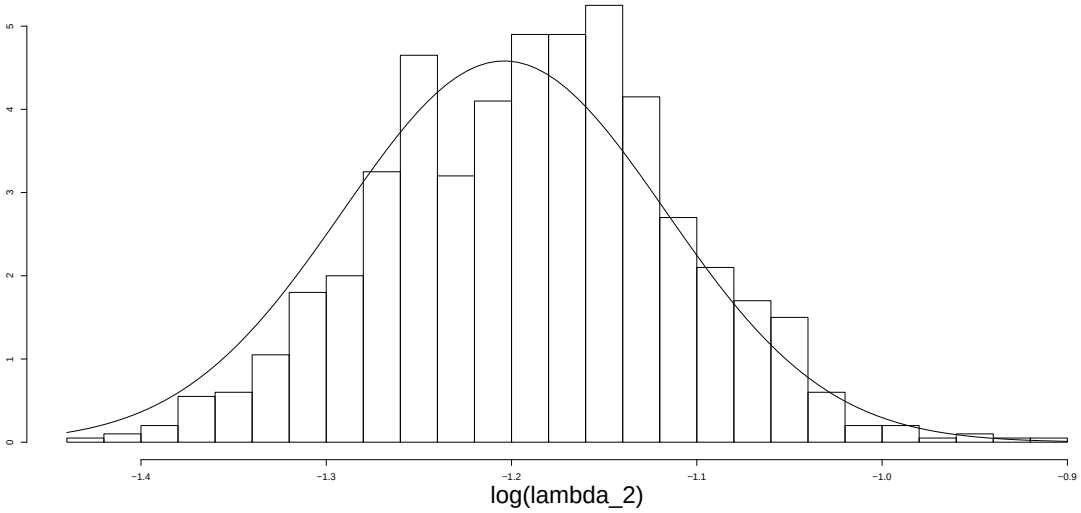


Figure 6.7: Monte Carlo histogram for $\log(\lambda_2)$.

6.5.2 EM algorithm

One of the most common methods of estimation of parameters in the literature, applicable in principle to general phase type models, is the EM approach which is

introduced to phase type parameter estimation by Asmussen et al. (1996) and Olsson (1996). The method has a companion well-documented and publicly available C program (Olsson, 1998) which is available for users.

Because of its generality and accessibility of its software, the method has been widely used in phase type parameter estimations. For instances, Ishay (2002) applied the method to analyze "Service Times and Customers' patience" at a call center of one of Israel's banks. Fackrell (2009) applied the method to healthcare datasets, Garg et al. (2011) applied the method to study phase type survival trees for clustering lengths of patients' hospital stays.

The general idea of the EM method is first to write down the log-likelihood function for the complete observations, i.e., the absorption-time dataset augmented as though all of the intermediate transition times had also been observed. This log-likelihood, as a function of the free parameter ϑ, is then replaced (the *E-step*) by its conditional expectation given the actually observed data, taken with respect to a hypothetical fixed parameter vector ϑ_k . Then the conditional expected log-likelihood given observed data is maximized over ϑ (the *M-step*), yielding the next iteration ϑ_{k+1} in the estimated-parameter sequence. The E and M steps are repeated until the sequence ϑ_k appears to have converged. The calculation of conditional expectations in the E-step is performed in the phase type model by setting up a system of differential equations related to the intensity matrix, for the unknown transition-intensity parameters, and these equations are solved numerically by the Runge-Kutta method.

In this section, we study the EM parameter estimation method and an application to a sub-class of our proposed model F. The diagram of interest is as follows.

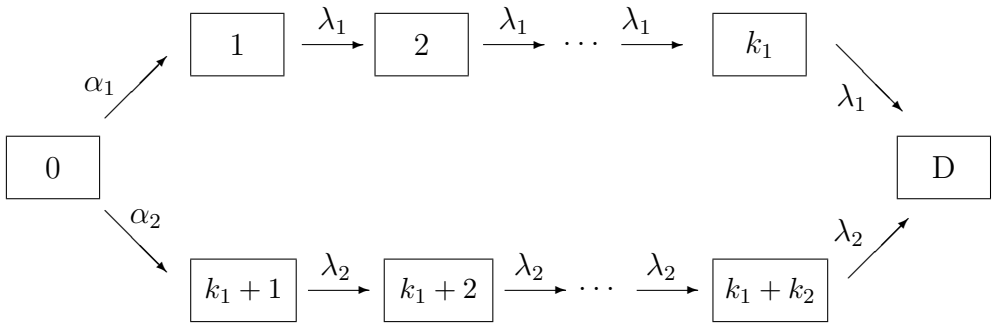


Figure 6.8: Markov transition diagram for `Model F` with two failure pathways.

To apply an EM algorithm to a general phase type distribution with state space $\{1, ..., k, D\}$, we consider the embedded Markov chain $I_0, ..., I_{M-1}, I_M = D$, where D is the absorbing state, and the sojourn times $S_0, ..., S_{M-1}$, where M is the number of jumps until the process reaches the absorbing state D. Define the transition probability p_{ij} as

$$p_{ij} = \mathbf{P}(I_{n+1} = j | I_n = i) = \begin{cases} \dfrac{t_{ij}}{\lambda_i}, & i, j = 1, .., k; \\ \dfrac{t_i}{\lambda_j}, & i = 1, .., k, \text{ and } j = D, \end{cases} \tag{6.5.6}$$

where λ_i is the intensity rate of leaving state i, which is $\lambda_i = -t_{ii} = -\big(t_i + \sum_{j \neq i} t_{ij}\big)$.

The density of a complete non-censored observation y is given (Asmussen et

al. (1996)) as

$$f(y;\pi,\mathbf{T}) = \pi_{i_0}\lambda_{i_0}\exp(-\lambda_{i_0}s_0)p_{i_0i_1}\ldots\lambda_{i_{(m-1)}}\exp(-\lambda_{i_{(m-1)}}s_{(m-1)})p_{i_{(m-1)}D}$$

$$= \pi_{i_0}\exp(-\lambda_{i_0}s_0)t_{i_0i_1}\ldots\exp(-\lambda_{i_{(m-1)}}s_{(m-1)})t_{i_{(m-1)}}.$$

The density of a complete right-censored observation y_c is given (Olsson 1996) as

$$f(y_c;\pi,\mathbf{T}) = \pi_{i_0}\exp(-\lambda_{i_0}s_0)t_{i_0i_1}\ldots\exp(-\lambda_{i_{(m_c-1)}}s_{(m_c-1)})t_{i_{(m_c-1)}i_{(m_c)}}\exp(-\lambda_{m_c}s_{m_c}), \tag{6.5.7}$$

where m_c is the observed value of the number of jumps on $(0,c]$ and c is the censoring time.

The joint density function of the complete observations is given as

$$f(\mathbf{x};\pi,\mathbf{T}_{n;n=1,..,N}) = \prod_{n=1}^{N}\Big(\prod_{i=1}^{k}\pi_i^{B_i^{(n)}}\prod_{i=1}^{k}\exp(t_{ii}^{(n)}Z_i^{(n)})\prod_{i=1}^{k}\prod_{j=1,j\neq i}^{k}(t_{ij}^{(n)})^{N_{ij}^{(n)}}\Big), \tag{6.5.8}$$

where

$B_i^{(n)} = I_{\{I_0^{(n)}=i\}}$,

$Z_i^{(n)} = \sum_{l=0}^{m(n)-1} I_{\{I_l^{(n)}=i\}}S_l^{(n)}$ is the total time the process n^{th} spent in state i, $i = 1,\ldots,k$,

$N_{ij}^{(n)} = \sum_{l=0}^{m(n)-1} I_{\{I_l^{(n)}=i,I_{l+1}^{(n)}=j\}}$ is the number of jumps from state i to j, for $i\neq j$, $i = 1,\ldots,k$, and $j = 1,\ldots,k$.

Therefore, the corresponding log-likelihood function for the diagram in Figure 6.8, assuming that the Markov chain always starts at state 0, is given as

$$\begin{aligned}
\text{Loglik} = &-(\alpha_1+\alpha_2)\sum_{n=1}^{N_1} Z_1^{(n)} - \lambda_1 \sum_{n=1}^{N_1}\sum_{i=2}^{k_1} Z_i^{(n)} - \lambda_2 \sum_{n=1}^{N_1}\sum_{i=k_1+1}^{k_1+k_2} Z_i^{(n)} \\
&+ \log(\alpha_1)\sum_{n=1}^{N_1} N_{12}^{(n)} + \log(\alpha_2)\sum_{n=1}^{N_1} N_{1(k_1+1)}^{(n)} \\
&+ \log(\lambda_1)\sum_{n=1}^{N_1}\Big[\sum_{i=2}^{k_1-1} N_{i,(i+1)}^{(n)} + N_{k_1 D}^{(n)}\Big] + \log(\lambda_2)\sum_{n=1}^{N_1}\Big[\sum_{i=k_1+1}^{k_1+k_2-1} N_{i,(i+1)}^{(n)} + N_{(k_1+k_2)D}^{(n)}\Big] \\
&- (\alpha_1+\alpha_2)\sum_{n=N_1+1}^{N_1+N_2} Z_1^{(n)} - \lambda_1 \sum_{n=N_1+1}^{N_1+N_2}\sum_{i=2}^{k_1-1} Z_i^{(n)} - \lambda_2 \sum_{n=N_1+1}^{N_1+N_2}\sum_{i=k_1+1}^{k_1+k_2-1} Z_i^{(n)} \\
&+ \log(\alpha_1)\sum_{n=N_1+1}^{N_1+N_2} N_{12}^{(n)} + \log(\alpha_2)\sum_{n=N_1+1}^{N_1+N_2} N_{1(k_1+1)}^{(n)} \\
&+ \log(\lambda_1)\sum_{n=N_1+1}^{N_1+N_2}\sum_{i=2}^{k_1-1} N_{i,(i+1)}^{(n)} + \log(\lambda_2)\sum_{n=N_1+1}^{N_1+N_2}\sum_{i=k_1+1}^{k_1+k_2-1} N_{i,(i+1)}^{(n)}, \qquad (6.5.9)
\end{aligned}$$

where N_1 is the number of uncensored observations and N_2 is the number of right censored observations.

The maximum likelihood estimates of the parameters $\{\alpha_1, \alpha_2, \lambda_1, \lambda_2\}$ are given as

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{\sum_{n=1}^{N_1+N_2} N_{12}^{(n)}}{\sum_{n=1}^{N_1+N_2} Z_1^{(n)}}, \qquad
\hat{\lambda}_1 = \frac{\sum_{n=1}^{N_1+N_2}\sum_{i=2}^{k_1-1} N_{i,(i+1)}^{(n)} + \sum_{n=1}^{N_1} N_{k_1,D}^{(n)}}{\sum_{n=1}^{N_1}\sum_{i=2}^{k_1} Z_i^{(n)} + \sum_{n=N_1+1}^{N_1+N_2}\sum_{i=2}^{k_1-1} Z_i^{(n)}} \\
\hat{\alpha}2 &= \frac{\sum_{n=1}^{N_1+N_2} N_{1(k_1+1)}^{(n)}}{\sum_{n=1}^{N_1+N_2} Z_1^{(n)}}, \qquad
\hat{\lambda}_2 = \frac{\sum_{n=1}^{N_1+N_2}\sum_{i=k_1+1}^{k_1+k_2-1} N_{i,(i+1)}^{(n)} + \sum_{n=1}^{N_1} N_{k_1+k_2,D}^{(n)}}{\sum_{n=1}^{N_1}\sum_{i=k_1+1}^{k_1+k_2} Z_i^{(n)} + \sum_{n=N_1+1}^{N_1+N_2}\sum_{i=k_1+1}^{k_1+k_2-1} Z_i^{(n)}}. \qquad (6.5.10)
\end{aligned}$$

In the E-step, the unknown values $N_{ij}^{(n)}, Z_i^{(n)}$ for uncensored observations ($n = 1, \ldots, N_1$) are replaced by their conditional expectations given observed data as the

following (Asmussen et al., 1996):

$$E_{(\pi,\mathbf{T})}\big(Z_i^{(n)}|Y=y_n\big)=\frac{c_i(y_n:i|\pi,\mathbf{T})}{\pi\mathbf{b}(y_n|\mathbf{T})} \tag{6.5.11}$$

$$E_{(\pi,\mathbf{T})}\big(N_{ij}^{(n)}|Y=y_n\big)=\frac{t_{ij}c_j(y_n:i|\pi,\mathbf{T})}{\pi\mathbf{b}(y_n|\mathbf{T})} \tag{6.5.12}$$

$$E_{(\pi,\mathbf{T})}\big(N_{i0}^{(n)}|Y=y_n\big)=\frac{t_i a_i(y_n|\pi,\mathbf{T})}{\pi\mathbf{b}(y_n|\mathbf{T})}, \tag{6.5.13}$$

where $\mathbf{a}(y_n|\pi,\mathbf{T})$, $\mathbf{b}(y_n|\pi,\mathbf{T})$, $\mathbf{c}(y_n;i|\pi,\mathbf{T})$, $i=1,\ldots,k$ are k-dimensional vector functions defined by

$$\mathbf{a}(y|\pi,\mathbf{T})=\pi\exp(\mathbf{T}y) \tag{6.5.14}$$

$$\mathbf{b}(y|\pi,\mathbf{T})=\exp(\mathbf{T}y)\mathbf{t} \tag{6.5.15}$$

$$\mathbf{c}(y;i|\pi,\mathbf{T})=\int_0^y \pi\exp(\mathbf{T}u)\mathbf{e}_i\exp(\mathbf{T}(y-u))\mathbf{t}du \quad i=1,\ldots,k. \tag{6.5.16}$$

The unknown values $N_{ij}^{(n)}, Z_i^{(n)}$ for right-censored observations ($n=N_1+1,\ldots,N_1+N_2$) are replaced by their conditional expectations given observed data as follows (Olsson, 1996):

$$E_{(\pi,\mathbf{T})}\big(Z_i^{(n)}|Y>c\big)=\frac{d_i(c:i|\pi,\mathbf{T})}{\pi\mathbf{h}(c|\mathbf{T})} \tag{6.5.17}$$

$$E_{(\pi,\mathbf{T})}\big(N_{ij}^{(n)}|Y>c\big)=\frac{t_{ij}d_j(c:i|\pi,\mathbf{T})}{\pi\mathbf{h}(c|\mathbf{T})} \tag{6.5.18}$$

where $\mathbf{h}(y_n|\pi,\mathbf{T}), \mathbf{d}(y_n;i|\pi,\mathbf{T})$ $, i=1,...,k$ are k- dimensional vector functions defined by

(6.5.19)

$$\mathbf{h}(c|\pi,\mathbf{T})=\exp(\mathbf{T}c)\mathbf{e} \tag{6.5.20}$$

$$\mathbf{d}(c;i|\pi,\mathbf{T})=\int_0^c \pi\exp(\mathbf{T}u)\mathbf{e}_i\exp(\mathbf{T}(c-u))\mathbf{e}du, \quad i=1,\ldots,k. \tag{6.5.21}$$

Having completed the E-step, the M-step is the performed by replacing the unknown term in (6.5.10) with conditional expectation derived from the E-step and maximizing with respect to the parameters. The E- and M- steps are then repeated until convergence is achieved. The convergence criterion generally used is a small increase of the log-likelihood for the successive iteration steps. Dempster et al. (1977) showed that the the log-likelihood function is always at least as large after as before each EM iteration. That is,

$$\text{Loglik}(\theta^{(k)}) \geq \text{Loglik}(\theta^{(k-1)}),$$

for all k.

Dempster et al. (1977) also showed that the likelihood function $L(\theta^{(k)})$ is bounded above. Therefore, by the Monotone Convergence Theorem, the EM algorithm always converges. However, the convergence rate of this EM algorithm could be slow. It needs more than 10,000 iterations for some cases and it sometimes converges to a saddle point rather than the MLE, as is discussed by Asmussen et al. (1996).

6.5.2.1 Fisher Information Matrix

The Fisher information matrix is very important in studying inference on parameters by measuring the information that an observation carries about the unknown parameters. The Fisher information matrix is also used in computing asymptotic variances and uncertainty of parameter estimators. Unlike the direct numerical method in section 6.5.1 where the observed Fisher Information matrix

is automatically produced by the negative estimated Hessian matrix of the log-likelihood, an EM parameter estimation method does not automatically produce an estimate of the Fisher information matrix. However, Oakes (1999) proposed a simple formula to estimate a Fisher Information matrix via the EM algorithm for general parameter estimation. Bladt et al. (2011) applied Oakes's method to provide a method to produce an estimated Fisher Information matrix for a phase type model in the case where all transition rates are freely varying and are not linked by any relations. In this study, we provide an alternative method to estimate the Fisher Information matrix, where some transition paths could share common transition rates. In this section, we derive the Fisher Information specifically for the family of phase type models described in Figure 6.8. Our method is a direct application of Oakes (1999) and the Runge-Kutta method used in the parameter estimation.

Following Oakes (1999), the Fisher information matrix is estimated by substituting $\theta_1 = \hat{\theta}$ into

$$\frac{\partial^2 \text{Loglik}(\theta_1; y)}{\partial \theta_1^2} = \left\{ \frac{\partial^2 Q(\theta_2|\theta_1)}{\partial \theta_2^2} + \frac{\partial^2 Q(\theta_2|\theta_1)}{\partial \theta_1 \partial \theta_2} \right\} |_{\theta_2=\theta_1}, \tag{6.5.22}$$

where

$$Q(\theta_2|\theta_1) = \mathbb{E}_{\theta_1}(\text{Loglik}(\theta_2\,;\,\mathbf{z}) \mid \mathbf{y}),$$

and $\mathbf{z} = (z_1, ..., z_N)$ denote the full data for the N observations. Here by substitution of conditional expectation expressions (6.5.11) - (6.5.13) and (6.5.17) - (6.5.18) into the conditional expected log-likelihood (6.5.9), we obtain

$$
\begin{aligned}
Q(\hat{\theta}|\theta) &= \mathbb{E}_\theta(\text{Loglik}(\hat{\theta}; \mathbf{z}|\mathbf{y})) \\
&= -(\hat{\alpha}_1 + \hat{\alpha}_2) \left(\sum_{n=1}^{N_1} \mathbb{E}(Z_1^{(n)}|\mathbf{y}) + \sum_{n=N_1+1}^{N_1+N_2} \mathbb{E}(Z_1^{(n)}|\mathbf{c}) \right) \\
&\quad - \hat{\lambda}_1 \left(\sum_{n=1}^{N_1} \sum_{i=2}^{k_1} \mathbb{E}(Z_i^{(n)}|\mathbf{y}) + \sum_{n=N_1+1}^{N_1+N_2} \sum_{i=2}^{k_1-1} \mathbb{E}(Z_i^{(n)}|\mathbf{c}) \right) \\
&\quad - \hat{\lambda}_2 \left(\sum_{n=1}^{N_1} \sum_{i=k_1+1}^{k_1+k_2} \mathbb{E}(Z_i^{(n)}|\mathbf{y}) + \sum_{n=N_1+1}^{N_1+N_2} \sum_{i=k_1+1}^{k_1+k_2-1} \mathbb{E}(Z_i^{(n)}|\mathbf{c}) \right) \\
&\quad + \log(\hat{\alpha}_1) \left(\sum_{n=1}^{N_1} \mathbb{E}(N_{12}^{(n)}|\mathbf{y}) + \sum_{n=N_1+1}^{N_1+N_2} \mathbb{E}(N_{12}^{(n)}|\mathbf{c}) \right) \\
&\quad + \log(\hat{\alpha}_2) \left(\sum_{n=1}^{N_1} \mathbb{E}(N_{1(k_1+1)}^{(n)}|\mathbf{y}) + \sum_{n=N_1+1}^{N_1+N_2} \mathbb{E}(N_{1(k_1+1)}^{(n)}|\mathbf{c}) \right) \\
&\quad + \log(\hat{\lambda}_1) \sum_{n=1}^{N_1} \left(\sum_{i=2}^{k_1-1} \mathbb{E}(N_{i,(i+1)}^{(n)}|\mathbf{y}) + \mathbb{E}(N_{k_1,D}^{(n)}|\mathbf{y}) \right) \\
&\quad + \log(\hat{\lambda}_1) \sum_{n=N_1+1}^{N_1+N_2} \sum_{i=2}^{k_1-1} \mathbb{E}(N_{i,(i+1)}^{(n)}|\mathbf{c}) \\
&\quad + \log(\hat{\lambda}_2) \sum_{n=1}^{N_1} \left(\sum_{i=k_1+1}^{k_1+k_2-1} \mathbb{E}(N_{i,(i+1)}^{(n)}|\mathbf{y}) + \mathbb{E}(N_{k_1+k_2,D}^{(n)}|\mathbf{y}) \right) \\
&\quad + \log(\hat{\lambda}_2) \sum_{n=N_1+1}^{N_1+N_2} \sum_{i=k_1+1}^{k_1+k_2-1} \mathbb{E}(N_{i,(i+1)}^{(n)}|\mathbf{c}) \\
&= -(\hat{\alpha}_1 + \hat{\alpha}_2) \left(\sum_{n=1}^{N_1} \frac{c_1(y_n : 1|\pi, \mathbf{T})}{f(y_n)} + \sum_{n=N_1+1}^{N_1+N_2} \frac{d_1(y_n : 1|\pi, \mathbf{T})}{S(c)} \right) \\
&\quad - \hat{\lambda}_1 \left(\sum_{n=1}^{N_1} \sum_{i=2}^{k_1} \frac{c_i(y_n : i|\pi, \mathbf{T})}{f(y_n)} + \sum_{n=N_1+1}^{N_1+N_2} \sum_{i=2}^{k_1-1} \frac{d_i(y_n : i|\pi, \mathbf{T})}{S(c)} \right) \\
&\quad - \hat{\lambda}_2 \left(\sum_{n=1}^{N_1} \sum_{i=k_1+1}^{k_1+k_2} \frac{c_i(y_n : i|\pi, \mathbf{T})}{f(y_n)} + \sum_{n=N_1+1}^{N_1+N_2} \sum_{i=k_1+1}^{k_1+k_2-1} \frac{d_i(y_n : i|\pi, \mathbf{T})}{S(c)} \right) \\
&\quad + \log(\hat{\alpha}_1) \left(\sum_{n=1}^{N_1} \alpha_1 \frac{c_2(y_n : 1|\pi, \mathbf{T})}{f(y_n)} + \sum_{n=N_1+1}^{N_1+N_2} \alpha_1 \frac{d_2(y_n : 1|\pi, \mathbf{T})}{S(c)} \right) \\
&\quad + \log(\hat{\alpha}_2) \left(\sum_{n=1}^{N_1} \alpha_1 \frac{c_{k_1+1}(y_n : 1|\pi, \mathbf{T})}{f(y_n)} + \sum_{n=N_1+1}^{N_1+N_2} \alpha_2 \frac{d_{k_1+1}(y_n : 1|\pi, \mathbf{T})}{S(c)} \right)
\end{aligned}
$$

$$
\begin{aligned}
&+\log(\hat{\lambda}_1)\sum_{n=1}^{N_1}\lambda_1\Big[\sum_{i=2}^{k_1-1}\frac{c_{i+1}(y_n:i|\pi,\mathbf{T})}{f(y_n)}+\frac{a_{k_1}(y_n|\pi,\mathbf{T})}{f(y_n)}\Big]\\
&+\log(\hat{\lambda}_1)\sum_{n=N_1+1}^{N_1+N_2}\sum_{i=2}^{k_1-1}\lambda_1\frac{d_{i+1}(y_n:i|\pi,\mathbf{T})}{S(c)}\\
&+\log(\hat{\lambda}_2)\sum_{n=1}^{N_1}\lambda_2\Big[\sum_{i=k_1+1}^{k_1+k_2-1}\frac{c_{i+1}(y_n:i|\pi,\mathbf{T})}{f(y_n)}+\frac{a_{k_1+k_2}(y_n|\pi,\mathbf{T})}{f(y_n)}\Big]\\
&+\log(\hat{\lambda}_2)\sum_{n=N_1+1}^{N_1+N_2}\sum_{i=k_1+1}^{k_1+k_2-1}\lambda_2\frac{d_{i+1}(y_n:i|\pi,\mathbf{T})}{S(c)}.
\end{aligned}
$$

We simplify using the following notations:

$$
\mathbf{ME}^{(n)}=\begin{pmatrix}
c_1(y_n:1|\pi,\mathbf{T}) & c_2(y_n:1|\pi,\mathbf{T})\\
\sum_{n=1}^{N_1}\sum_{i=2}^{k_1}c_i(y_n:i|\pi,\mathbf{T}) & \sum_{n=1}^{N_1}\sum_{i=2}^{k_1-1}c_{i+1}(y_n:i|\pi,\mathbf{T})+\frac{a_{k_1}(y_n|\pi,\mathbf{T})}{f(y_n)}\\
c_1(y_n:1|\pi,\mathbf{T}) & c_{k_1+1}(y_n:1|\pi,\mathbf{T})\\
\sum_{n=1}^{N_1}\sum_{i=k_1+1}^{k_1+k_2}c_i(y_n:i|\pi,\mathbf{T}) & \sum_{n=1}^{N_1}\sum_{i=k_1+1}^{k_1+k_2-1}c_{i+1}(y_n:i|\pi,\mathbf{T})+\frac{a_{k_1+k_2}(y_n|\pi,\mathbf{T})}{f(y_n)}
\end{pmatrix},
$$

and

$$
\mathbf{MC}^{(n)}=\begin{pmatrix}
d_1(y_n:1|\pi,\mathbf{T}) & d_2(y_n:1|\pi,\mathbf{T})\\
\sum_{n=N_1+1}^{N_1+N_2}\sum_{i=2}^{k_1-1}d_i(y_n:i|\pi,\mathbf{T}) & \sum_{n=N_1+1}^{N_1+N_2}\sum_{i=2}^{k_1-1}d_{i+1}(y_n:i|\pi,\mathbf{T})\\
d_1(y_n:1|\pi,\mathbf{T}) & d_{k_1+1}(y_n:1|\pi,\mathbf{T})\\
\sum_{n=N_1+1}^{N_1+N_2}\sum_{i=k_1+1}^{k_1+k_2-1}d_i(y_n:i|\pi,\mathbf{T}) & \sum_{n=N_1+1}^{N_1+N_2}\sum_{i=k_1+1}^{k_1+k_2-1}d_{i+1}(y_n:i|\pi,\mathbf{T})
\end{pmatrix}.
$$

Consequently, the estimated Fisher Information matrix

$\mathbf{L} = (L_{ij}) = -\Big(\frac{\partial^2 Q(\hat{\theta}|\theta)}{\partial\hat{\theta}_j\partial\hat{\theta}_i} + \frac{\partial^2 Q(\hat{\theta}|\theta)}{\partial\theta_j\partial\hat{\theta}_i}\Big)_{\hat{\theta}=\theta}$ is given by

$$\begin{aligned}
L_{ij} = &\sum_{n=1}^{N_1} \frac{1}{f(y_n)}\big[\frac{\partial}{\partial\theta_j}\mathrm{ME}_{i2}^{(n)} - \frac{\partial}{\partial\theta_j}\mathrm{ME}_{i1}^{(n)}\big] \\
&- \sum_{n=1}^{N_1} \frac{1}{(f(y_n))^2}\frac{\partial}{\partial\theta_j}f(y_n)\big[\mathrm{ME}_{i2}^{(n)} - \mathrm{ME}_{i1}^{(n)}\big] \\
&+ \sum_{n=N_1+1}^{N_1+N_2} \frac{1}{f(y_n)}\big[\frac{\partial}{\partial\theta_j}\mathrm{MC}_{i2}^{(n)} - \frac{\partial}{\partial\theta_j}\mathrm{MC}_{i1}^{(n)}\big] \\
&- \sum_{n=N_1+1}^{N_1+N_2} \frac{1}{(f(y_n))^2}\frac{\partial}{\partial\theta_j}f(y_n)\big[\mathrm{MC}_{i2}^{(n)} - \mathrm{MC}_{i1}^{(n)}\big].
\end{aligned}$$

The partial derivatives $\dfrac{\partial\mathbf{ME}^{(n)}}{\partial\theta_j}$ and $\dfrac{\partial\mathbf{MC}^{(n)}}{\partial\theta_j}$ are obtained by taking derivative of the quantities in (6.5.14)-(6.5.16) and (6.5.20)-(6.5.21) and the density function f as follows.

For $j = 1, \ldots, 4$, define $\mathbf{T}_{\theta_j}(y) = \dfrac{\partial}{\partial\theta_j}\exp(\mathbf{T}y)$ and $\mathbf{e}(y|\pi, \mathbf{T}) = \exp(\mathbf{T}y)$.

Then

$$\begin{aligned}
\frac{\partial}{\partial y}\mathbf{T}_{\theta_j}(y) &= \frac{\partial}{\partial y}\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}y) \\
&= \frac{\partial}{\partial\theta_j}\frac{\partial}{\partial y}\exp(\mathbf{T}y) \\
&= \frac{\partial}{\partial\theta_j}\big(\exp(\mathbf{T}y)\mathbf{T}\big) \\
&= \exp(\mathbf{T}y)\frac{\partial}{\partial\theta_j}\mathbf{T} + \big(\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}y)\big)\mathbf{T} \\
&= \mathbf{e}(y|\pi, \mathbf{T})\frac{\partial}{\partial\theta_j}\mathbf{T} + T_{\theta_j}(y)\mathbf{T}. \qquad (6.5.23)
\end{aligned}$$

Since $f(y) = \pi \exp(\mathbf{T}y)\mathbf{t}$,

$$\begin{aligned}\frac{\partial}{\partial\theta_j} f(y) &= \pi\big(\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}y)\big)\mathbf{t} + \pi\exp(\mathbf{T}y)\big(\frac{\partial}{\partial\theta_j}\mathbf{t}\big) \\ &= \pi\mathbf{T}_{\theta_j}(y)\mathbf{t} + \pi\mathbf{e}(y|\pi,\mathbf{T})\big(\frac{\partial}{\partial\theta_j}\mathbf{t}\big). \end{aligned} \tag{6.5.24}$$

We further define $C_{\theta_j,i}(y) = \dfrac{\partial}{\partial\theta_j}\mathbf{c}(y;i|\pi,\mathbf{T})$ and $D_{\theta_j,i}(c) = \dfrac{\partial}{\partial\theta_j}\mathbf{d}(c;i|\pi,\mathbf{T})$ for $i = 1,..,k$.

Then

$$\begin{aligned}\mathbf{C}_{\theta_j,i}(y) &= \frac{\partial}{\partial\theta_j}\mathbf{c}(y;i|\pi,\mathbf{T}) \\ &= \frac{\partial}{\partial\theta_j}\int_0^y \pi\exp(\mathbf{T}u)\mathbf{e}_i\exp(\mathbf{T}(y-u))\mathbf{t}du \\ &= \int_0^y \pi\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}u)\mathbf{e}_i\exp(\mathbf{T}(y-u))\mathbf{t}du \\ &\quad + \int_0^y \pi\exp(\mathbf{T}u)\mathbf{e}_i\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}(y-u))\mathbf{t}du \\ &\quad + \int_0^y \pi\exp(\mathbf{T}u)\mathbf{e}_i\exp(\mathbf{T}(y-u))\frac{\partial\mathbf{t}}{\partial\theta_j})du,\end{aligned}$$

and

$$\begin{aligned}\mathbf{D}_{\theta_j,i}(c) &= \frac{\partial}{\partial\theta_j}\mathbf{d}(c;i|\pi,\mathbf{T}) \\ &= \frac{\partial}{\partial\theta_j}\int_0^c \pi\exp(\mathbf{T}u)\mathbf{e}_i\exp(\mathbf{T}(c-u))\mathbf{e}du \\ &= \int_0^c \pi\exp(\mathbf{T}u)\mathbf{e}_i\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}(c-u))\mathbf{e}du \\ &\quad + \int_0^c \pi\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}u)\mathbf{e}_i\exp(\mathbf{T}(c-u))\mathbf{e}du.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{\partial}{\partial y}\mathbf{C}_{\theta_j,i}(y) &= \pi\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}y)\mathbf{e}_i\mathbf{t} \\ &\quad+\int_0^y \pi\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}u)\mathbf{e}_i\mathbf{T}\exp(\mathbf{T}(y-u))\mathbf{t}du \\ &\quad+\int_0^y \pi\exp(\mathbf{T}u)\mathbf{e}_i\frac{\partial}{\partial\theta_j}\Big(\mathbf{T}\exp(\mathbf{T}(y-u))\Big)\mathbf{t}du \\ &\quad+\pi\exp(\mathbf{T}y)\mathbf{e}_i\frac{\partial\mathbf{t}}{\partial\theta_j}+\int_0^y \pi\exp(\mathbf{T}u)\mathbf{e}_i\mathbf{T}\exp(\mathbf{T}(y-u))\frac{\partial\mathbf{t}}{\partial\theta_j})du \\ &= \pi\mathbf{T}_{\theta_j}(y)\mathbf{e}_i\mathbf{t}+\pi\mathbf{e}(y|\pi,\mathbf{T})\mathbf{e}_i\frac{\partial\mathbf{t}}{\partial\theta_j}+\mathbf{T}\mathbf{C}_{\theta_j,i}(y)+\big(\frac{\partial}{\partial\theta_j}\mathbf{T}\big)\mathbf{c}(y;i|\pi,\mathbf{T}),\end{aligned} \tag{6.5.25}$$

and

$$\frac{\partial}{\partial c}\mathbf{D}_{\theta_j,i}(c) = \pi\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}c)\mathbf{e}_i\mathbf{e} \tag{6.5.26}$$

$$+\int_0^c \pi\exp(\mathbf{T}u)\mathbf{e}_i\frac{\partial}{\partial\theta_j}\big(\mathbf{T}\exp(\mathbf{T}(c-u))\Big)\mathbf{e}du \tag{6.5.27}$$

$$+\int_0^c \pi\frac{\partial}{\partial\theta_j}\exp(\mathbf{T}u)\mathbf{e}_i\mathbf{T}\exp(\mathbf{T}(c-u))\mathbf{e}du \tag{6.5.28}$$

$$=\big[\pi\mathbf{T}_{\theta_j}(c)\mathbf{e}_i\mathbf{e}\big]+\mathbf{T}\mathbf{D}_{\theta,i}+\frac{\partial\mathbf{T}}{\partial\theta_j}\mathbf{d}(c:i|\pi,\mathbf{T}). \tag{6.5.29}$$

Therefore, the partial derivatives $\dfrac{\partial\mathbf{M}\mathbf{E}^{(n)}}{\partial\theta_j}$ and $\dfrac{\partial\mathbf{M}\mathbf{C}^{(n)}}{\partial\theta_j}$ are obtained by solving the following system of equations:

- non-censored observations

$$\begin{aligned}
\frac{d}{dy}\mathbf{e}(y|\pi,\mathbf{T}) &= \mathbf{T}\mathbf{e}(y|\pi,\mathbf{T}) \\
\frac{d}{dy}\mathbf{T}_{\theta_j}(y) &= \mathbf{e}(y|\pi,\mathbf{T})\frac{\partial \mathbf{T}}{\partial \theta_j} + \mathbf{T}_{\theta_j}(y)\mathbf{T} \\
\frac{d}{dy}\mathbf{C}(y|\pi,\mathbf{T}) &= \mathbf{T}\mathbf{C}(y|\pi,\mathbf{T}) + \mathbf{t}\otimes(\pi\mathbf{e}(y|\pi,\mathbf{T})) \\
\frac{d}{dy}\mathbf{C}_{\theta_j}(y|\pi,\mathbf{T}) &= \Big(\mathbf{t}\otimes\big(\pi\mathbf{T}_{\theta_j}(y)\big) + \big(\frac{\partial \mathbf{t}}{\partial \theta_j}\big)\otimes\big(\pi\mathbf{e}(y|\pi,\mathbf{T})\big)\Big) \\
&\quad + \mathbf{T}\mathbf{C}_{\theta_j}(y|\pi,\mathbf{T}) + \frac{\partial \mathbf{T}}{\partial \theta_j}\mathbf{C}(y|\pi,\mathbf{T})
\end{aligned}$$

where $\mathbf{e}(y|\pi,\mathbf{T}) = \exp(\mathbf{T}y)$, and $\otimes$ denotes a Kronecker product. The system of differential equations can be solved by the Runge-Kutta method with the initial value $\mathbf{e}(0|\pi,\mathbf{T}) = \mathbf{I}_p$, and $\mathbf{C}(0|\pi,\mathbf{T}) = \mathbf{C}_{\theta_j}(0|\pi,\mathbf{T}) = \mathbf{T}_{\theta_j}(0) = \mathbf{O}_p$ for all $\theta_j : j = 1,..,4$.

- right-censored observations

$$\begin{aligned}
\frac{d}{dc}\mathbf{e}(c|\pi,\mathbf{T}) &= \mathbf{T}\mathbf{e}(c|\pi,\mathbf{T}) \\
\frac{d}{dy}\mathbf{T}_{\theta_j}(c) &= \mathbf{e}(c|\pi,\mathbf{T})\frac{\partial \mathbf{T}}{\partial \theta_j} + \mathbf{T}_{\theta_j}(c)\mathbf{T} \\
\frac{d}{dc}\mathbf{D}(c|\pi,\mathbf{T}) &= \mathbf{T}\mathbf{D}^{(n)}(c|\pi,\mathbf{T}) + \mathbf{e}\otimes(\pi\mathbf{e}(c|\pi,\mathbf{T})) \\
\frac{d}{dy}\mathbf{D}_{\theta_j}(c|\pi,\mathbf{T}) &= \Big(\mathbf{e}\otimes\big(\pi\mathbf{T}_{\theta_j}(c)\big)\Big) \\
&\quad + \mathbf{T}\mathbf{D}_{\theta}(c|\pi,\mathbf{T}) + \frac{\partial \mathbf{T}}{\partial \theta_j}\mathbf{D}^{(n)}(y|\pi,\mathbf{T})
\end{aligned}$$

where $\mathbf{e}(c|\pi,\mathbf{T}) = \exp(\mathbf{T}c)$, and $\otimes$ denotes a Kronecker product. The system of differential equations can be solved by the Runge Kutta method with the initial value $\mathbf{e}(0|\pi,\mathbf{T}) = \mathbf{I}_p$, and $\mathbf{D}(0|\pi,\mathbf{T}) = \mathbf{D}_{\theta_j}(0|\pi,\mathbf{T}) = \mathbf{T}_{\theta_j}(0) = \mathbf{O}_p$ for all $\theta_j : j = 1,..,4$

6.5.2.2 Numerical Results

We consider a special case of the phase type model in Figure 6.8 where $k_1 = 4$ and $k_2 = 2$. This is the mixture of $\text{Exp}(\alpha_1) * \text{Gamma}(4, \lambda_1)$ and $\text{Exp}(\alpha_2) * \text{Gamma}(2, \lambda_2)$. We choose a sample size of 100 non-censored observations. The results suggest that the EM algorithm does not give accurate numerical results.

Parameters	True values	MLE	SD
α_1	0.15	0.06476837	0.05564376
λ_1	0.15	0.91315934	0.50344497
α_2	0.25	0.17920354	0.42455326
λ_2	0.12	0.09735084	0.01896362

6.6 Discussion of Computational Experience

In this chapter, we have considered two estimation methods which are direct quasi-Newton-Raphson optimization and an EM algorithm. The quasi-Newton-Raphson maximizes the log-likelihood reasonably precise. This method was applicable because of the relative simplicity of `Model F`, where paths do not connect except at the Origin and Death states. Tables 6.3 and 6.5 illustrated the need for large sample sizes to estimate all parameters accurately. Figures 6.2-6.7 also display histograms allowing the reader to assess the (rather slow) rate of convergence of distributions of ML estimators to normality as sample sizes get large.

In contrast, the EM method does not give parameter estimates precisely and

has very slow convergence rates. We applied the EM method to a sample discussed in Section 6.5.2.2, which is the mixture of $\text{Exp}(\alpha_1) * \text{Gamma}(4, \lambda_1)$ and $\text{Exp}(\alpha_2) * \text{Gamma}(2, \lambda_2)$, with sample size of 100. In our study, we implemented the algorithm in the `R` platform using the R-function `rk` for the Runge-Kutta equation solver. Our convergence criteria involve smallness of changes in log-likelihood of the order of accuracy 10^{-10}. As mentioned in Asmussen et al. (1996) , some drawbacks of the EM algorithm are its slow convergence rate (up to 10000 iterations often being required for reasonable convergence), and its occasional convergence to a local maximum or saddle point. Another drawback is that the E-step calculation must be performed for each observation, which is computationally burdensome in large samples. We found that very long CPU times are required to achieve convergence in the case sample sizes as large as 100, even in low-dimensional parametric examples.

Our overall conclusions are that the EM algorithm method of Asmussen et al. (1996) and Olsson (1996) for fitting phase type survival densities to right-censored survival data is primarily of theoretical interest, because the method places no restriction on the complexity of the underlying Markov chain. But in practice, even when the models are very simple, simpler than `Model F` of Figure 6.1, the computation times are prohibitively large even for moderately large datasets, and they scale roughly proportionately to sample size.

6.7 Data Analysis: Breast Cancer Mortality

In this Section, we fit the `Model F` parametric class of densities to the White Female SEER dataset on mortality in 13 US registries of breast cancer cases diagnosed between 1992 and 2001 and followed through 2002. Details concerning the data, a spline-based fitting methodology, and discussion can be found in Anderson et al. (2006) . Of the complete dataset of 243,808 cases, we analyzed only the 198,785-case subset of White females with age at diagnosis from 30 to 89, for breast-cancer mortality. Although the primary focus of the Anderson et al. study was to understand the shape of post-diagnosis hazard as a mixture of the disaggregated disease types indicated by Estrogen Receptor (ER) status, we omitted that covariate from our analysis, since our objective is to learn what a purely parametric statistical analysis using the model of Section 6.4 could have told about the likely mixture components of breast-cancer mortality in the combined population.

While Anderson et al. (2006) directly created spline-fitted hazard functions for their combined and ER-disaggregated study populations, we performed a slightly more complicated preliminary analysis designed to correct for year-of-diagnosis mortality differences, since Kaplan-Meier curves for the data stratified by diagnosis year (`DiagYr`) showed a small but clear trend of decreasing of hazards with `DiagYr`. The cumulative hazards were nearly linear for the datasets with `DiagYr` after 1996, with a slight concavity over times 6-11 years for earlier `DiagYr`'s. Since the nonparametrically fitted hazards were therefore approximately proportional across `DiagYr`, we fitted a Cox proportional hazards model with a dummy variable for `DiagYr` as the only

covariate, finding effect coefficients for `DiagYr` versus 1992 as .007, −.024, −.065, −.093, −.138, −.161, −.236, −0.285, −0.292. In all of our analyses, 0.5 was added to the raw survival times of 0:131 months. We present as our basic nonparametric mortality curve the summary survival curve for that Cox model, to which we fitted a smoothing spline using the `R` function `smooth.spline`, with smoothing parameter `spar=0.5`. Figure 6.9 shows the corresponding survival density, along with one computed the same way but with less smoothing (`spar=0.25`), along with the best fit that we were able to find to the data, a 6-parameter model which differs slightly from `Model F` in removing the direct failure paths with parameters $b_D\,\mu$, β_1, β_2, and instead inserting an extra state A between k_1 and D, with transition arcs from state k_1 to A and from A to D. In this fitted model, as in all those treated below, $k_1 = 4$ and $k_2 = 1$. (A 5-parameter variant model which looks visually identical to the 6-parameter density in Figure 6.9 is obtained by letting the μ rate-parameter in Figure 6.1 go to ∞.)

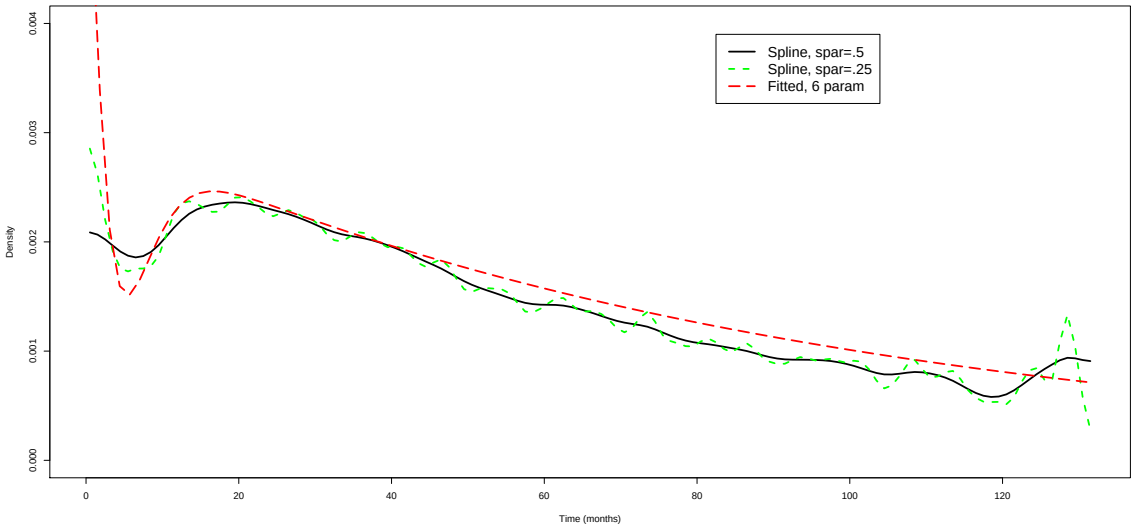


Figure 6.9: Spline and fitted density functions to the SEER 1992-2002 data on US white 30-89 female breast cancer mortality following diagnosis.

The solid spline fitted curve in Figure 6.9 closely resembles the summary all patients survival hazard pictures in Anderson et al. (2006) . The spline fit to the same Cox model summary survival, but with less smoothing (dotted curve in Figure 6.9), shows more clearly the overall features of the density which a parametric model should seek to reproduce. These features include a high initial spike in hazard, a density peak near 20 months, an approximately linear decrease of density between 20 and 120 months, and a final increase in density between 120 and 130 months. Presumably the initial hazard spike is due to immediate adverse outcomes from surgery and untreatable advanced stage cancers, and the peak and density decrease from 20 to 120 months are due to the recent successes in treating a large fraction of cancers detected at early stages. But we cannot account for the final upturn in hazard, which our models do not address at all.

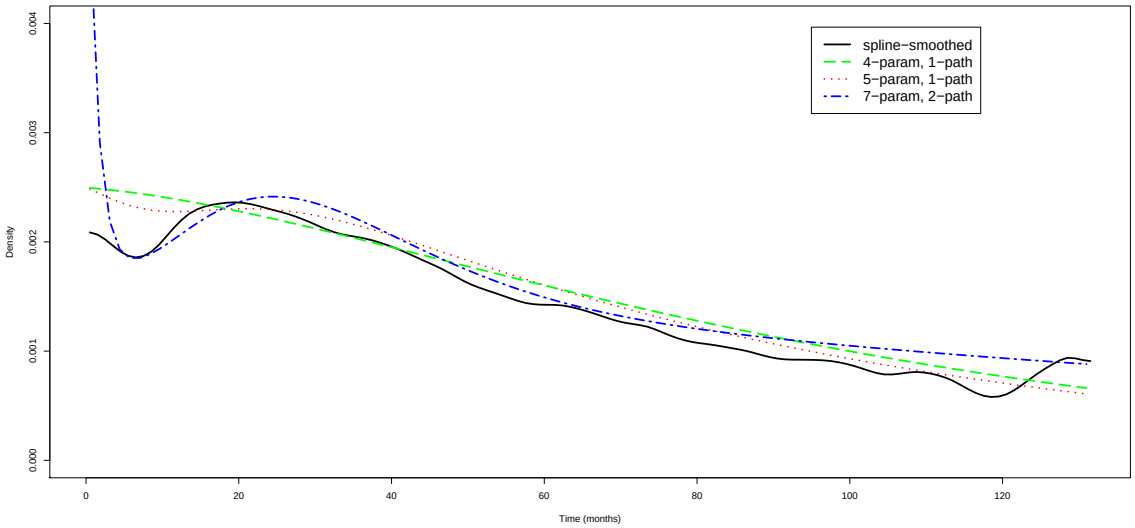


Figure 6.10: Spline and three estimated Model F densities ML fitted to the SEER 1992-2002 data as in Figure 6.9 on US female breast cancer mortality following diagnosis.

Most of the computational work done in fitting the models displayed in Figure 6.10 was done on a single set of 20,000 patient records randomly selected from the full dataset of 198,785 records. Within each model class and fixed parameter dimension, the right censored survival data log-likelihood was maximized using the `R` function `nlm`, convergence of which was very sensitive to the choice of starting parameter values. That choice often had to be guided by visual inspection of plotted density curves, a process which was sufficient for the selection of single path models within `Model F`, but adequate starting values for two-path models were found only using the mixture idea of Lemma 6.1 to combine two separate single-path models.

The models compared visually in this section can be further understood through their log-likelihood values on the SEER breast cancer data. We first clarify the re-

lationship between visual fidelity of fitted survival densities and purely statistical model comparisons via likelihood ratio tests. Table 6.9 displays ML estimated model parameters and log-likelihoods for the SEER data used in producing Figure 6.10, i.e., the SEER data on breast cancer mortality following diagnosis for white females aged 30-89. The loglikelihood differences between the models are large, because of the large sample size. For purposes of comparison, the log-likelihoods on the same data for the models whose densities are plotted in Figure 6.9 are -171112 for the spline-fitted survival density with `spar`$= 0.25$, -171699 for the spline-fitted survival density with `spar`$= 0.5$, and -172184 for the best-fitting (6-parameter, 2-path) model.

Table 6.9: Parameters and log-likelihoods for models in Figure 6.10, with $k_1 = 4, k_2 = 1, \beta_2 = 0$.

# par.	p	μ	λ_1	λ_2	β_1	b_C	b_D	logLik
4	1	0.0022	0.0190	100.	0	6.493	1.132	-172691
5	1	0.0009	0.0002	100.	0.1864	12.09	2.294	-172640
7	0.0894	0.2747	0.0001	0.0054	0.1194	1.749	0.0229	-172347

The Figures and loglikelihoods shown, and the results of other analyses not shown, demonstrate clearly that the essential features of the density curves up to 120 months can be captured only by 2-path models, in other words mixture models, within the phase type model F class. Figure 6.10 also indicates that each increase in parameter dimension allows an additional visual feature of the empirical smoothed

density – which the spline fit displays – to be captured by the parametric model: the 4-parameter one-path model captures roughly the early and late density levels and the approximate curvilinear pattern of decrease of density or hazard; the 5-parameter model begins to capture the initial hook (decrease and then increase to a local peak); and the 2-path 7-parameter model follows (and even exaggerates) the initial hook, although the less smoothed spline picture in Figure 6.9 does show a sharp initial density decrease) while closely following the local peak near 20 months.

It is well known that latent class and mixture models often have poorly identified parameters, sometimes even for strikingly large sample sizes. We have seen the same phenomenon in the information matrices for the simulated data discussed in Section 6.5.1.1 above. So we focus next on the Fisher information matrices and parameter standard errors for the fitted models, expressed for the transformed parameters, which are subvectors of $\vartheta = (\text{logit}(p), \log(\mu), \log(\lambda_1), \log(\lambda_2), \log(\beta_1), \log(b_C), \log(b_D))$. For models with respectively 4, 5, and 7 parameters, the ranges of eigenvalues of the respective observed information matrices $\hat{I}(\hat{\vartheta})$ were found to be $(1.7, 24078.0)$, $(45.6, 28907.2)$, and $(52.2, 22030.1)$. Thus, in all of the models the most accurate linear parameter combinations with unit vector coefficients have SE's of order .0065, while the least accurate have SE's of 0.14 or larger. For example, the three models give SE's for $\text{logit}(\lambda_1)$, respectively, as 0.486, 0.097, and 0.058; and the respective SE's for $\log(b_D)$ are 0.233, 0.064, and 0.028.

While the phase type models fitted to the large SEER dataset have ill-conditioned

Fisher information matrices — and therefore at least some parameters which are very badly identified — one can with some assurance achieve the qualitatively important finding, that at least two mixture components are needed for a high-quality parametric fit. The fact that in these data the ER status now represents a medically *observable* identifier of two distinct mixture components (which is essentially the point of the Anderson et al. 2006 article) corroborates this conclusion, and suggests the potential usefulness in new applications of a similar parametric statistical in detecting the presence of two separate diseases within a single diagnostic category.

6.8 Summary and Discussion

We have surveyed the broad field of parametric models for survival densities, from the vantage point of the special class of latent state stochastic transition models known as Phase type models. Our numerical illustrations and data analysis of a real breast cancer dataset show that even for relatively low dimensional models of this type, the Fisher Information matrices can be strikingly ill conditioned, and yet that certain parametric functions reflecting qualitative features of the fitted models — especially the presence or absence of extra 'paths' or mixture components — can be estimated adequately and have important interpretations. The general point we have made is that visual features of survival densities may reflect important structure about underlying mechanism of transition among minimally parameterized latent states, structure with biomedical importance for the suggestion of future research directions, such as the search for multiple diseases underlying a single diagnostic

category.

Parametric models built from mixtures are notoriously difficult to identify from moderate sample size data. The consequence of this observation for Phase type survival models is that only models with relatively simple path structure and state descriptions can have a realistic chance of being fitted stably. For this reason, it may be slightly misguided in biomedical applications to fit the complicated multistate phase type models for which the EM methods of parameter estimation were devised. As a consequence, if only models at most of the order of complexity of our `Model F` are to be fitted, then direct likelihood computation methods based on simple properties of exponential variates and mixtures of their convolutions will be applicable.

The phase type `Model F` can readily be extended to incorporate regression terms in terms of biomedical covariates for log transition rates such as $\log(\mu)$ or $\log(\lambda_1)$. Such survival regression models increase flexibility for joint models of non-homogeneous populations, in the spirit of the threshold regression models of Lee and Whitmore (2006). Analogous regressions for Coxian parameters were found to increase the model likelihood in Faddy and McClean (1999). However, the introduction of unknown coefficients for covariates might also result in ML parameter estimates with large variances. The identification of the non-intercept regression coefficients might also be strong, as we have seen for ratios of transition-rates. The empirical and numerical study of such parametric regression models is a subject of our further research.

Chapter 7

Appendix A: Preliminaries on Computational Statistics

7.1 Bootstrap

Bootstrapping, first proposed by Efron (1979), is a class of computer-intensive simulation techniques widely used for several purposes including bias removal in parameter estimation, variance estimation, and pointwise confidence interval construction. Bootstrapping could be either nonparametric (original) or parametric. In this section, we discusses only some features of bootstrap. We refer to Efron and Tibshirani (1993) for complete practical discussion of Bootstrap and Shao and Tu (1995) for more mathematical aspects.

7.1.1 Nonparametric Bootstrap

Consider the situation where we want to make inference about a real parameter $\theta = t(\mathbf{F})$ from a sample $\mathbf{x} = (x_1, ..., x_n)$ that follows the unknown distribution $\mathbf{F}$. A nonparametric bootstrap is to draw sample $x^* = (x_1^*, \ldots, x_n^*)$ with replacement from the data $\mathbf{x} = (x_1, ..., x_n)$. Therefore, for each replication b of the bootstrap, we can obtain

$$\hat{\theta}^*(b) = t(x^{*b}).$$

The bootstrap parameter estimate of θ is defined as

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*(b),$$

where B is the number of bootstrap replications. Consequently, the bootstrap estimates of bias and $\text{se}_F(\hat{\theta})$ are then defined as

$$\widehat{\text{Bias}}_B = \hat{\theta}^* - \hat{\theta}, \tag{7.1.1}$$

$$\hat{\text{se}}_B = \left(\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*(b) - \hat{\theta}^*)^2\right)^{1/2}. \tag{7.1.2}$$

7.1.2 Parametric Bootstrap

Parametric bootstrap is usually used when a parametric model is fitted to a data set. To draw a parametric bootstrap sample, instead of drawing a random sample from the data $\mathbf{x}$, we draw a sample from the parametric distribution $\hat{F}_{par}$, where $\hat{F}_{par}$ is the parametric model with model parameter estimates substituted for the true parameters. After B bootstrap samples are drawn, the bootstrap parameter estimate of θ, and the bootstrap estimates of bias and $\text{se}_F(\hat{\theta})$ are calculated in the same way as in the nonparametric bootstrap.

7.1.3 Bootstrap Confidence Interval

Let $X_1, \ldots, X_n$ be i.i.d. random variables drawn from an unknown distribution F, and let θ be the parameter of interest. A $(1-\alpha)$ confidence set for θ is a subset $\mathcal{A}_n(X_1, .., X_n)$ of $\mathbb{R}$ such that

$$P(\theta \in \mathcal{A}_n(X_1, .., X_n)) = 1 - \alpha,$$

where α is a real number in $(0, 1)$.

A two-sided $(1-\alpha)$ confidence interval of the parameter θ is defined as the interval $(\underline{\theta}(X_1, .., X_n), \overline{\theta}(X_1, .., X_n))$ such that

$$P(\theta \in (\underline{\theta}(X_1, .., X_n), \overline{\theta}(X_1, .., X_n))) = 1 - \alpha.$$

There are several bootstrap approaches to obtain a confidence interval for bootstrap estimates. In this section, two types of pointwise confidence intervals are studied: standard normal confidence interval and percentile confidence interval. Let $\hat{\theta}$ be an estimate of a parameter θ. The $(1-\alpha)100\%$ standard normal confidence interval is given by

$$[\hat{\theta} - z_{\alpha/2} \cdot \text{se}(\hat{\theta}), \hat{\theta} + z_{\alpha/2} \cdot \text{se}(\hat{\theta})],$$

where $\text{se}(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$. The $(1-\alpha)100\%$ percentile confidence interval is defined by

$$[\hat{\theta}_B^{\alpha/2}, \hat{\theta}_B^{(1-\alpha/2)}],$$

where $\hat{\theta}_B^{\alpha}$ is the $B\alpha^{th}$ value in the ordered list of the B bootstrap replications.

7.2 Spline Smoothing

Smoothing spline is a curve fitting method that is based on piecewise polynomial functions. One well known Spline smoothing method is the cubic smoothing spline, which is widely used in statistical analysis. In this section, we restrict attention to the calculation of coefficient parameters of a cubic smoothing spline.

Consider the function $f = f(x)$ on the interval with k knots with coordinates $(x_1, f_1), ..., (x_k, f_k)$. The cubic spline function $S = S_i$ is given (Pollock) for $x \in$

$[x_i, x_{i+1}]$ by

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \tag{7.2.3}$$

where x is in $[x_i, x_{i+1}]$.

In cubic splines, three conditions are required:

$$S_{i-1}(x_i) = S_i(x_i) \tag{7.2.4}$$

$$S'_{i-1}(x_i) = S'_i(x_i) \tag{7.2.5}$$

$$S''_{i-1}(x_i) = S''_i(x_i). \tag{7.2.6}$$

From (7.2.3), we have for $x \in [x_i, x_{i+1}]$

$$S'(x) = 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i \tag{7.2.7}$$

$$S''(x) = 6a_i(x - x_i) + 2b_i. \tag{7.2.8}$$

From (7.2.4)-(7.2.6) and (7.2.7)-(7.2.8),

$$b_i = 3a_{i-1}h_{i-1} + b_{i-1} \tag{7.2.9}$$

$$c_i = 3a_{i-1}h_{i-1}^2 + 2b_{i-1}h_{i-1} + c_{i-1}, \tag{7.2.10}$$

$$d_i = a_{i-1}h_{i-1}^3 + b_{i-1}h_{i-1}^2 + c_{i-1}h_{i-1} + d_{i-1}. \tag{7.2.11}$$

Therefore, given that k knots with $k - 1$ intervals are used in the cubic smoothing spline, the number of regression coefficients is $k + 2$, consisting of 4 parameters in the first interval and $k - 2$ parameters in the rest of $k - 2$ intervals.

7.3 Runge Kutta Methods

The Runge-Kutta method is a numerical method to approximate a solution of

$$\frac{dy}{dx} = f(x, y(x)), \quad y(0) = y_0.$$

The Runge-Kutta method is based on an expansion of the Taylor's series in a way that any order N has precision $O(h^N)$. Therefore the method can be very precise when the order increases, however, the computation of higher order terms is very complicated. The order most commonly used in practice is 4, which leads to computations that are easy to program yet very precise and stable.

The fourth-order Runge-Kutta formula is given as

$$\begin{aligned}
k_1 &= hf(x_n, y_n) \\
k_2 &= hf(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}) \\
k_3 &= hf(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}) \\
k_4 &= hf(x_n + \frac{h}{2}, y_n + \frac{k_3}{2}) \\
y_{n+1} &= y_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} + O(h^5).
\end{aligned}$$

7.4 Expectation-Maximization Algorithm

The Expectation-Maximum algorithm or the so called *EM* algorithm is an iterative method to find maximum likelihood estimates proposed by Dempster et al. (1977) for incomplete data problems. The algorithm consists of two main steps which are: (1) the Expectation step (E-step) where missing data are replaced by theirs expectations and (2) the Maximization step (M-step) where the maximum

likelihood function is maximized after the missing data are filled.

We begin this section by a definition of maximum-likelihood estimate. Let $\mathbf{X} = (X_1, ..., X_n)$ be a random vector with the joint density function $f(\mathbf{x}|\theta)$. Then the *likelihood function* of θ given the data $\mathbf{x}$ is defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta).$$

The maximum likelihood estimate $\hat{\theta}$ of θ is

$$\hat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{x}).$$

In the context of the *EM* algorithms, we consider the situation that the unobserved complete data X can be represented as (Y, Z), where Y is the observed data and Z is the missing data. Therefore, on the k iteration, the likelihood function is approximated by $\mathcal{Q}(\theta, \theta^{(k-1)}) = E[L(\theta, \mathbf{y}|\mathbf{Z})|\mathbf{y}, \theta^{(k-1)}]$ in the E-step. The $\mathcal{Q}(\theta, \theta^{(k-1)})$ is then maximized in the M-step by choosing $\theta^{(k)}$ such that

$$\theta^{(k)} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{(k-1)}).$$

The E- and M- steps are then repeated until a convergence is achieved. The convergent criteria generally used is a small increase of the likelihood for the successive iteration steps. Dempster et al. (1977) showed that the the likelihood function is always non-decreased after an EM iteration. That is

$$L(\theta^{(k)}) \geq L(\theta^{(k-1)}),$$

for all k.

Moreover, the authors (Dempster et al. (1977)) also showed that the likelihood function $L(\theta^{(k)})$ is bounded above. Therefore, by the Monotone convergence theorem, EM algorithm always converges.

Although the convergence of the EM algorithm is guaranteed, the method is known to have slow convergence rate and sometimes converges to a saddle point (Wu, 1983]). This leads to many alternatives and modifications for improving the rate of convergence. For example, Aitken's method (Aitkin, 1996]), Louis's method (Louis,1982) and a Conjugate gradient method (Jamshidian and Jennrich, 1993), which are the most common methods for speeding up the EM algorithm. We refer to McLachlan and Krishnan (2008) for intensive studies of the EM algorithm and its extensions in both theoretical and practical aspects.

7.5 Numerical optimization methods

In this section, we discuss a few numerical optimization methods used in our projects: Newton-Raphson method and Quasi-Newton method. We refer to Griva et al. (2009) for more details and variations of numerical optimization methods.

7.5.1 Newton-Raphson method

The Newton-Raphson method is a an iterative method to solve the equation

$$\mathbf{f}(\mathbf{x}) = \mathbf{0},$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is in $\mathbb{R}^n$ and $\mathbf{f}$ is twice differentiable.

Let $\nabla\mathbf{f}(\mathbf{x}) = (\nabla f_1(x), ..., \nabla f_n(x))$ be the Jacobian matrix of the function $\mathbf{f}$.

Therefore, by the Taylor series expansion around the point $\mathbf{x}_k$ at the k^{th} iteration step,

$$\mathbf{f}(\mathbf{x}^{(k+1)}) \approx \mathbf{f}(\mathbf{x}^{(k)}) + \nabla \mathbf{f}^T(\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = 0,$$

or

$$\mathbf{x}^{(k+1)} \approx \mathbf{x}^{(k)} - \nabla \mathbf{f}^T(\mathbf{x}^{(k)})\mathbf{f}(\mathbf{x}^{(k)}). \tag{7.5.12}$$

The iteration step is repeated until it converges, where under some assumptions of smoothness and local concavity, the Newton-Raphson method is proved to have a quadratic rate of convergence (Ortega and Rheinboldt (1970)).

7.5.2 Quasi-Newton method

In the optimization context, the Newton method is applied to the gradient function of the function to be optimized. Therefore, an evaluation of the second derivative, $\nabla^2 \mathbf{f}(\mathbf{x})$, or the the Hessian matrix is required in order to apply the Newton method which could require intensive computations. Therefore, a natural alternative to the Newton method is to approximate the Hessian matrix by another matrix $\mathbf{B}_k$ that is available by using only the first derivative. This method is so called the *quasi-Newton method*. There are many versions of the Quasi-Newton method varying by the choice of the matrix $\mathbf{B}_k$. One example is the well-known and widely used method, *BFGS*, named after its four originator: Broyden, Fletcher, Goldfarb and Shanno. The formula is given by Griva et al. (2009)

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{(\mathbf{B}_k \mathbf{s}_k)(\mathbf{B}_k \mathbf{s}_k)^T}{\mathbf{s}_k^T \mathbf{B}_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k},$$

where $\mathbf{y}_k = \nabla \mathbf{f}^T(\mathbf{x}^{(k+1)}) - \nabla \mathbf{f}^T(\mathbf{x}^{(k)})$ and $\mathbf{s}_k = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$.

The Quasi-Newton methods such as *BFGS* are proved to have a superlinear rate of convergence (Byrd et al. (1987)).

Chapter 8

Appendix B: Derivation of the first hitting time of an Ornstein-Uhlenbeck Process

In this chapter, we calculate the approximate density function of the first hitting of an Ornstein-Uhlenbeck process (2.8.19) discussed in Section 2.8.2 by applying Durbin's Approximation of the First hitting time of a Gaussian Process (Durbin, 1985) and following the algorithm studied in Lachaud (2004). We begin with two theorems of Durbin (1985) in approximating the first hitting time density of a Gaussian process.

8.1 Durbin's Approximation of the First hitting time of a Gaussian Process

Theorem 8.1 (Durbin, 1985)**.** *Let* $\big(Y(t)\big)_{t\geq 0}$ *be a centered continuous Gaussian process with covariance function* $\rho(s,t)$ *for* $0 \leq s \leq t$*. Let* p *be the density of the first hitting time of the boundary* $a = a(t)$ *"coming from below".*

We make the following hypotheses:

(H1) for all $t \geq 0$*, the boundary function* a *is continuous at point* t *and left-differentiable at point* t*;*

(H2) the covariance function ρ *is strictly positive and its first-order partial deriva-*

tives are continuous on the set $\{(s,t) \in \mathbb{R}^+; 0 \leq s \leq t\}\}$ *with the convenient left- and right-derivatives at points* $s = 0$ *and* $s = t$*;*

(H3) the variance of the increment $Y_t - Y_s$ *satisfies the condition:*

$$\lim_{s \uparrow t} \frac{Var(Y_t - Y_s)}{(t-s)} = \lambda_t,$$

with $0 < \lambda_t < \infty$*, which is equivalent to:*

$$\lim_{s \uparrow t} \big[\frac{\partial \rho(s,t)}{\partial s} - \frac{\partial \rho(s,t)}{\partial t}\big] = \lambda_t,$$

with $0 < \lambda_t < \infty$*.*

Then we get the following for p*:*

$$\forall t \geq 0, \qquad\qquad p(t) = b(t)f(t),$$

where

$$f(t) = \frac{1}{\sqrt{2\pi\rho(t,t)}} \exp\left[-\frac{a^2(t)}{2\rho(t,t)}\right],$$

and

$$b(t) = \lim_{s \uparrow t} \mathbf{E}\big[I(s,Y)\big(a(s) - Y(s)\big)|Y(t) = a(t)\big],$$

where

$$I(s,Y) = \begin{cases} 1 & \textit{if the path does not cross the boundary prior to time } s, \\ 0 & \textit{otherwise.} \end{cases}$$

Since the function b in Theorem 8.1 is not easy to calculate, Durbin provided an approximation of the function b in the following Theorem.

Theorem 8.2 (Durbin, 1985)**.** *The notations are the same as those in Theorem 8.1. Let b_1 be defined for all $t \geq 0$ by*

$$\begin{aligned} b_1(t) &= \lim_{s \uparrow t} \frac{1}{t-s} \mathbf{E}\big[I(s,Y)\big(a(s) - Y(s)\big)|Y(t) = a(t)\big] \\ &= \frac{a(t)}{\rho(t,t)} \frac{\partial \rho(s,t)}{\partial s}\Big|_{s=t} - a'(t). \end{aligned}$$

Under the three hypotheses (H1), (H2) and (H3) the function p_1 defined for all $t \geq 0$ by

$$p_1(t) = b_1(t) f(t),$$

is an approximation of p which is exact in the limit as the boundary becomes increasingly remote.

8.2 The first hitting time of an Ornstein-Uhlenbeck process

In this section, we derive the approximate density function of the first hitting of an Ornstein-Uhlenbeck process (2.8.19) discussed in Section 2.8.2 by following the algorithm studied in Lachaud (2004) .

Define a process $(Y(t))_{t \geq 0}$ by

$$Y(t) = \frac{a}{b} + \big(x_0 - \frac{a}{b}\big)e^{-bt} - X(t),$$

where $X(t)$ is the Ornstein-Uhlenbeck process.

The process $(Y(t))_{t \geq 0}$ is a centered Gaussian process, with covariance function $\rho(s,t) = \sigma^2/(2b)\big[e^{-b|t-s|} - e^{-b(t+s)}\big]$. The process starts at 0 and the first hitting time is

$$T_c^{x_0} = \inf\left\{t \geq 0 : Y(t) \geq \big(\frac{a}{b} - c\big) + \big(x_0 - \frac{a}{b}\big)e^{-bt}\right\}.$$

The boundary is $a(t) = \big(a/b - c\big) + \big(x_0 - a/b\big)e^{-bt}$.

The hypotheses (H1) and (H2) are clearly satisfied. Next, we check the hypothesis (H3). Note that, for $s \le t$,

$$\begin{aligned}\frac{\partial \rho(s,t)}{\partial s} &= \frac{\partial}{\partial s}\frac{\sigma^2}{2b}\big(e^{-b(t-s)} - e^{-b(t+s)}\big) \\ &= \frac{\sigma^2}{2}\big(e^{-b(t-s)} + e^{-b(t+s)}\big)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \rho(s,t)}{\partial t} &= \frac{\partial}{\partial t}\frac{\sigma^2}{2b}\big[e^{-b(t-s)} - e^{-b(t+s)}\big] \\ &= \frac{\sigma^2}{2}\big[-e^{-b(t-s)} + e^{-b(t+s)}\big].\end{aligned}$$

Then

$$\begin{aligned}\lim_{s\uparrow t}\big[\frac{\partial \rho(s,t)}{\partial s} - \frac{\partial \rho(s,t)}{\partial t}\big] &= \frac{\sigma^2}{2}\lim_{s\uparrow t}\Big(2e^{-b(t-s)}\Big) \\ &= \sigma^2 > 0.\end{aligned}$$

In Theorem 8.1, the approximate density function p_1 is $p_1(t) = b_1(t)f(t)$, where

$$\begin{aligned}b_1(t) &= \frac{a(t)}{\rho(t,t)}\frac{\partial \rho(s,t)}{\partial s}\Big|_{s=t} - a'(t) \\ &= \frac{\Big(\big(a/b - c\big) + \big(x_0 - a/b\big)e^{-bt}\Big)}{(\sigma^2/2b)(1 - e^{-2bt})}\frac{\sigma^2}{2}(1 + e^{-2bt}) + b\big(x_0 - a/b\big)e^{-bt} \\ &= \frac{b\Big((a/b)\big(1 - e^{-bt}\big)^2 - c\big(1 + e^{-2bt}\big) + 2x_0e^{-bt}\Big)}{1 - e^{-2bt}}.\end{aligned}$$

Therefore

$$
\begin{aligned}
p_1(t) &= b_1(t) f(t) \\
&= \frac{b\Big((a/b)\big(1-e^{-bt}\big)^2 - c\big(1+e^{-2bt}\big) + 2x_0 e^{-bt}\Big)}{1-e^{-2bt}} \\
&\quad \times \frac{1}{\sqrt{2\pi}\sqrt{(\sigma^2/2b)(1-e^{-2bt})}} \exp\left\{ \frac{b\big[\big(a/b-c\big) + \big(x_0 - a/b\big)e^{-bt}\big]^2}{\sigma^2(1-e^{-2bt})} \right\} \\
&= \frac{e^{bt}((a/b)(e^{bt}-1)^2 - c(e^{2bt}+1) + 2x_0 e^{bt})}{\sigma\sqrt{\pi}} \big(\frac{b}{e^{2bt}-1}\big)^{3/2} \\
&\quad \times \exp\big\{ \frac{-b\big[(a/b-c)e^{bt} + (x_0 - a/b)\big]^2}{\sigma^2(e^{2bt}-1)} \big\}.
\end{aligned}
$$

Chapter 9

Appendix C: Asymptotic Properties of Maximum Penalized Likelihood Estimates

In this chapter, we study asymptotic properties of penalized likelihood estimates by checking conditions of Pakes and Pollard's consistency and asymptotic Normality conditions (Pakes and Pollard, 1989). Our results are special cases of Pakes and Pollard (1989). We further study asymptotic normality properties of bootstrap estimates from a penalized likelihood model by specializing results of Chen et al. (2003).

This chapter is organized as follows. Section 9.1 reviews fundamental concepts of asymptotic statistics. Pakes and Pollard's consistency and asymptotic normality conditions are discussed in Section 9.2. Section 9.3 discusses consistency and normality conditions of penalized likelihood estimates. Section 9.4 discusses asymptotic normality conditions of bootstrap estimates from a penalized likelihood Model.

9.1 Fundamental Asymptotic Theorems

In this section, reviews of fundamental concepts and theorems used in our study are discussed.

Definition 9.1. *(Casella and Berger, 2002) A sequence of random variables, $X_1, X_2, \ldots$ converges in probability to a random variable X, denoted $X_n \rightarrow_p X$, if for every*

$\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \quad \textit{or, equivalently,} \quad \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

Definition 9.2. *(Casella and Berger, 2002) A sequence of random variables, $X_1, X_2, \ldots$ converges in distribution to a random variable X, denoted $X_n \rightsquigarrow X$, if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous.

Definition 9.3. *(Casella and Berger, 2002) A sequence of random variables, $X_1, X_2, \ldots$ converges almost surely to a random variable X if for every $\epsilon > 0$,*

$$P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1.$$

Definition 9.4. *(Van der Vaart, 1998) A class $\mathcal{F}$ of measurable functions $f : \Omega \to \mathbb{R}$ is called P-Glivenko-Cantelli if*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| \to 0 \quad a.s.$$

Definition 9.5. *(Van der Vaart, 1998)A class $\mathcal{F}$ of measurable function $f : \Omega \to \mathbb{R}$ is called P-Donsker if the sequence of processes*

$$\left\{ \sqrt{n} \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right) \quad : \quad f \in \mathcal{F} \right\}$$

converges in distribution to a tight limit process in the space $l^\infty(f)$.

Theorem 9.1. *[Weak Law of Large Numbers], (Casella and Berger, 2002) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E(X_i) = \mu_i$ and $Var(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1;$$

that is, $\bar{X}_n$ converges in probability to μ, or $\bar{X}_n \to_p \mu$.

Theorem 9.2. *[Continuous Mapping Theorem] (Van der Vaart, 1998) Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$.*

(i) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$;

(ii) If $X_n \to_p X$, then $g(X_n) \to_p g(X)$;

(iii) If $X_n \to_{a.s.} X$, then $g(X_n) \to_{a.s.} g(X)$.

Theorem 9.3. *[Central Limit Theorem] (Casella and Berger, 2002) Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $E(X_i) = \mu$ and $0 < Var(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then*

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow N(0, 1).$$

9.2 Pakes and Pollard's Consistency and Asymptotic Normality Conditions

Let $\{Z_i\}_{i=1}^n$ be i.i.d. random variable sampled from a distribution $\mathbf{P}$, and let $\mathbf{\Theta} \subseteq \mathbb{R}^k$ be a finite dimensional parameter set. Let $G : \mathbf{\Theta} \to \mathbb{R}^l$ be a deterministic vector-valued function defined on $\mathbf{\Theta}$ such that the true value θ_0 is the unique solution to $G(\theta) = 0$. We consider consistency and asymptotic normality conditions of an estimate $\hat{\theta}_n$ defined as the minimizer of the length $\|G_n(\cdot)\|$. In this section, we restate Pakes and Pollard's consistency and asymptotic normality conditions (Pakes and Pollard, 1989) which are applied to our situation in Chapter 4. We refer readers

to Chen et al. (2003) for an extension of these results to the case of semi-parametric models and an infinite dimensional parameter space.

Theorem 9.4. *(Corollary 3.2, Pakes and Pollard, 1989) Under the following conditions $\hat{\theta}_n$ converges in probability to the unique θ_0 in $\mathbf{\Theta}$ for which $G(\theta_0) = 0$:*

(9.4.1) $\|G_n(\hat{\theta}_n)\| \leq \inf_{\theta} \|G_n(\theta)\| + o_P(1)$,

(9.4.2) $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta)\| > 0$ *for each* $\delta > 0$,

(9.4.3) $\sup_{\theta} \dfrac{\|G_n(\theta) - G(\theta)\|}{1 + \|G_n(\theta)\| + \|G(\theta)\|} = o_P(1)$.

Note that

$$\frac{\|G_n(\theta) - G(\theta)\|}{1 + \|G_n(\theta)\| + \|G(\theta)\|} \leq \|G_n(\theta) - G(\theta)\|.$$

Therefore condition (9.4.3) is implied by condition (9.4.3') :

$$\sup_{\theta} \|G_n(\theta) - G(\theta)\| = o_P(1).$$

Theorem 9.5. *(Theorem 3.3, Pakes and Pollard, 1989) Let $\hat{\theta}_n$ be a consistent estimator of θ_0, the unique point of $\mathbf{\Theta}$ for which $G(\theta_0) = 0$. If*

(9.5.1) $\|G_n(\hat{\theta}_n)\| \leq \inf_{\theta} \|G_n(\theta)\| + o_P(n^{-1/2})$;

(9.5.2) $G(\cdot)$ is differentiable at θ_0 with an $l \times k$ derivative matrix Γ of full rank;

(9.5.3) for every sequence $\{\delta_n\}$ of positive numbers that converges to zero,

$$\sup_{\|\theta-\theta_0\|<\delta_n} \frac{\|G_n(\theta) - G(\theta) - G_n(\theta_0)\|}{n^{-1/2} + \|G_n(\theta)\| + \|G(\theta)\|} = o_P(1);$$

(9.5.4) $\sqrt{n}G_n(\theta_0) \rightsquigarrow N(0, V)$;

(9.5.5) θ_0 is an interior point of $\mathbf{\Theta}$;

then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\big(0, (\Gamma^T\Gamma)^{-1}\Gamma^T V \Gamma (\Gamma^T\Gamma)^{-1}\big).$$

Note that,

$$\frac{\|G_n(\theta) - G(\theta) - G_n(\theta_0)\|}{n^{-1/2} + \|G_n(\theta)\| + \|G(\theta)\|} \leq \sqrt{n}\|G_n(\theta) - G(\theta) - G_n(\theta_0)\|.$$

Therefore condition (9.5.3) is implied by condition (9.5.3'):

$$\sup_{\|\theta-\theta_0\|<\delta_n} \|G_n(\theta) - G(\theta) - G_n(\theta_0)\| = o_P(n^{-1/2}).$$

9.3 Consistency and Asymptotic Normality Conditions for Penalized Likelihood Estimates

In this section, we specializing Theorems 9.4 and 9.5 to prove consistency and asymptotic normality conditions of maximum penalized likelihood parameters. Let $X_1, \ldots, X_n$ be a random sample drawn from a distribution P, where the log-likelihood function $l(\theta|\cdot)$ satisfies the conditions:

(A1) The parameter space $\mathbf{\Theta}$ is compact,

(A2) $E(l(\theta|X))$ is continuous,

(A3) All third-order derivatives $\dfrac{\partial^3}{\partial\theta_i\theta_j\theta_k} l(\theta|X)$ exist for all fixed X,

(A4) $E(\nabla l(\theta|X)^2) < \infty$ where ∇ is the gradient operator with respect to θ,

(A5) $E(\nabla^{\otimes 2} l(\theta|X)^2) < \infty$ where $\nabla^{\otimes 2}$ is the Hessian operator with respect to θ.

The penalized likelihood function is defined as

$$\mathbf{l}(\theta|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} l(\theta|X_i) - \frac{1}{\sqrt{n}}\big[p(\theta)\big],$$

where $p(\theta)$ is convex and continuous.

In this section, we study consistency and asymptotic normality properties of a penalized likelihood estimate, $\hat{\theta}$, by checking conditions of Pakes and Pollard (1989). To apply Theorems 9.4 - 9.5, we define the functions G_n and G as

$$G_n(\theta) = -\left[\frac{1}{n}\sum_{i=1}^{n} l(\theta|X_i) - \mathbb{E}(l(\theta_0|X_1))\right] + \frac{1}{\sqrt{n}}\left[p(\theta) - p(\theta_0)\right], \tag{9.3.1}$$

and

$$G(\theta) = -\mathbb{E}\big[l(\theta|X_1) - l(\theta_0|X_1)\big]. \tag{9.3.2}$$

(I) Claim : $\sqrt{n}\left(\|G_n(\hat{\theta})\| - \inf_\theta \|G_n(\theta)\|\right) \to 0$ a.s.

Note that,

$$\| G_n(\hat{\theta}) \| = \inf_\theta \| G_n(\theta) \| . \tag{9.3.3}$$

This verifies conditions (9.4.1) and (9.5.1).

(II) Claim : $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta)\| > 0$ for each $\delta > 0$.

Suppose instead that there is a $\delta > 0$ such that $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta)\| = 0$. Since $\mathbf{\Theta}$ is compact, $C_\delta = \{\theta \in \mathcal{B} : \|\theta - \theta_0\| \geq \delta\}$ is closed. Therefore there is a $\theta' \in C_\delta$ such that $G(\theta') = 0$. This contradicts the assumption that θ_0 is the unique point such that $G(\theta) = 0$. Therefore $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta)\| > 0$ for each $\delta > 0$. The condition (9.4.2) is then verified.

(III) Claim: $\mathcal{L} = \{l(\theta|\cdot) : \theta \in \mathbf{\Theta}\}$ is *P-Glivenko-Cantelli* and *P-Donsker*.

Since $l(\theta|X)$ is differentiable as a function of θ, by the Mean Value Theorem, for all $\theta_1, \theta_2 \in \mathbf{\Theta}$,

$$l(\theta_1|X) - l(\theta_2|X) = \nabla l(\theta'|X) \cdot (\theta_2 - \theta_1), \tag{9.3.4}$$

where $\cdot$ is the dot product and $\theta' = \theta_1 + \lambda(\theta_2 - \theta_1)$, $0 < \lambda < 1$.

By the Cauchy-Schwarz inequality,

$$|l(\theta_1|X) - l(\theta_2|X)| \leq \left\| \sup_\theta \nabla l(\theta|X) \right\| \|\theta_2 - \theta_1\|. \tag{9.3.5}$$

Since $\mathbf{\Theta}$ is compact, $\mathcal{L}$ is compact and then $\mathbb{E}|(\sup_\theta \nabla l(\theta|X))| < \infty$. Then (Example 19.7, p.271, Van der Vaart, 1998) $\mathcal{L}$ is *P-Glivenko-Cantelli* and *P-Donsker*. That is

$$\sup_\theta \| \frac{1}{n} \sum_{i=1}^n l(\theta|X_i) - \mathbb{E}(l(\theta|X_1)) \| = o(1) \quad a.s., \tag{9.3.6}$$

and

$$\sqrt{n} \sup_\theta \| \frac{1}{n} \sum_{i=1}^n l(\theta|X_i) - \mathbb{E}(l(\theta|X_1)) \| = O_P(1). \tag{9.3.7}$$

(IV) Claim : $\sup_\theta \|G_n(\theta) - G(\theta)\| \to 0$ a.s.

$$\begin{aligned} G_n(\theta) - G(\theta) &= -\big[\frac{1}{n} \sum_{i=1}^n l(\theta|X_i) - \mathbb{E}(l(\theta_0|X_1))\big] + \frac{1}{\sqrt{n}}\big[p(\theta) - p(\theta_0)\big] \\ &\quad - \Big(- \mathbb{E}\big[l(\theta|X_1) - l(\theta_0|X_1)\big]\Big) \\ &= -\Big[\frac{1}{n} \sum_{i=1}^n l(\theta|X_i) - \mathbb{E}\big[l(\theta|X_1)\Big] + \frac{1}{\sqrt{n}}\big[p(\theta) - p(\theta_0)\big]. \end{aligned}$$

By (9.3.6),

$$\sup_\theta \| \frac{1}{n} \sum_{i=1}^n l(\theta|X_i) - \mathbb{E}\big[l(\theta|X_1) \| \to 0 \quad \text{a.s.} \tag{9.3.8}$$

By (9.3.8),

$$\begin{aligned}\sup_\theta \parallel G_n(\theta) - G(\theta) \parallel &\le \sup_\theta \left\| -\Big[\frac{1}{n}\sum_{i=1}^n l(\theta|X_i) - \mathbb{E}\big[l(\theta|X_1)\big]\Big] \right\| \\ &\quad + \sup_\theta \parallel \frac{1}{\sqrt{n}}\big[p(\theta) - p(\theta_0)\big] \parallel \\ &\le o_{\{a.s.\}}(1) + o(1) \\ &\le o_{\{a.s.\}}(1).\end{aligned}$$

Hence $\sup_\theta \|G_n(\theta) - G(\theta)\| = o_{\{a.s.\}}(1)$, and so is $o_P(1)$.

Therefore condition (9.4.3) is verified.

(V) Claim: for every sequence $\{\delta_n\}$ of positive numbers that converges to zero,

$$\sup_{\|\theta-\theta_0\|<\delta_n} \|G_n(\theta) - G(\theta) - G_n(\theta_0)\| = o_P(n^{-1/2}).$$

Note that

$$\begin{aligned}G_n(\theta) - G(\theta) - G_n(\theta_0) &= -\big[\frac{1}{n}\sum_{i=1}^n l(\theta|X_i) - \mathbb{E}(l(\theta_0|X_1))\big] \\ &\quad + \frac{1}{\sqrt{n}}\big[p(\theta) - p(\theta_0)\big] + \mathbb{E}\big[l(\theta|X_1) - l(\theta_0|X_1)\big] \\ &\quad + \left[\frac{1}{n}\sum_{i=1}^n l(\theta_0|X_i) - \mathbb{E}(l(\theta_0|X_1))\right] \\ &= -\left[\frac{1}{n}\sum_{i=1}^n l(\theta|X_i) - E(l(\theta|X_1))\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^n l(\theta_0|X_i) - E(l(\theta_0|X_1)) + \frac{1}{\sqrt{n}}\left[p(\theta) - p(\theta_0)\right].\end{aligned} \tag{9.3.9}$$

Let $Y_n(\theta) = \frac{1}{n}\sum_{i=1}^n l(\theta|X_i) - E(l(\theta|X_1))$. Then, by Taylor series expansion,

$$Y_n(\theta) = Y_n(\theta_0) + (\theta - \theta_0)^T \nabla Y_n(\theta'), \tag{9.3.10}$$

where $\theta' = \theta_0 + \lambda(\theta - \theta_0)$ and $\lambda \in (0,1)$.

Therefore

$$\begin{aligned}\sqrt{n}\big(Y_n(\theta) - Y_n(\theta_0)\big) &= (\theta - \theta_0)^T \sqrt{n}\Big[\nabla Y_n(\theta')\Big] \\ &= (\theta - \theta_0)^T \sqrt{n}\Big[\frac{1}{n}\sum_{i=1}^{n} \nabla l(\theta'|X_i) - \mathbb{E}(\nabla l(\theta'|X_1))\Big]\end{aligned} \tag{9.3.11}$$

Substitute $\nabla l(\theta|X_i)$ for $l(\theta|X_i)$ into (9.3.7) under the condition (A5),

$$\sqrt{n}\sup_{\theta} \parallel \Big[\frac{1}{n}\sum_{i=1}^{n} \nabla l(\theta'|X_i) - \mathbb{E}(\nabla\theta'|X_1)\Big] \parallel = O_P(1). \tag{9.3.12}$$

Since $\parallel (\theta - \theta_0) \parallel \leq \delta_n = o(1)$ and (9.3.12) holds,

$$\sqrt{n} \sup_{\|\theta - \theta_0\| < \delta_n} \parallel \big(Y_n(\theta) - Y_n(\theta_0)\big) \parallel = o_P(1). \tag{9.3.13}$$

Since p is continuous and $\parallel (\theta - \theta_0) \parallel \leq \delta_n = o(1)$,

$$\sup_{\|\theta - \theta_0\| < \delta_n} \parallel p(\theta) - p(\theta_0) \parallel = o(1). \tag{9.3.14}$$

Hence, by (9.3.9)- (9.3.14),

$$\sup_{\|\theta - \theta_0\| < \delta_n} \|G_n(\theta) - G(\theta) - G_n(\theta_0)\| = o_P(n^{-1/2}).$$

The condition (9.5.3') is then verified.

(VI) Claim : $\sqrt{n}G_n(\theta_0) \rightsquigarrow N(0, V)$.

Since $\big\{X_i\big\}_{i=1}^{n}$ are i.i.d., then by the Central Limit Theorem,

$$\sqrt{n}G_n(\theta_0) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} l(\theta_0|X_i) - E(l(\theta|X_1))\right] \rightsquigarrow N(0, V),$$

where $V = \mathrm{Var}_\theta\big(l(\theta_0|X_1)\big)$. Therefore the condition (9.5.4) is verified.

From (I)-(VI), and Theorems 9.4 and 9.5, $\hat{\theta}_n \to_p \theta_0$ and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically Normally distributed.

9.4 The Bootstrap

In this section, we study asymptotic normality properties of bootstrap estimates from a penalized likelihood model. We begin our section with two theorems of Chen, Linton and Keilegom [Chen et al., 2003]. The original versions are stated for Semiparametric models, but we restrict the theorems to a fully parametric case in this section.

Let $\{X_i^*\}_{i=1}^n$ be drawn randomly with replacement from $\{X_i\}_{i=1}^n$.
Let $G_n^*(\theta) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} g(X_i^*, \theta)$ for each θ where $g : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}$ is a measurable function such that $G(\theta) = \mathbb{E}(g(X_i, \theta)) = 0$ if and only if $\theta = \theta_0$.

Theorem 9.6. *Suppose that $\{X_i\}_{i=1}^n$ are i.i.d. and $\theta_0 \in int\,\boldsymbol{\Theta}$ satisfies $\mathbb{E}(G(X_i, \theta_0)) = 0$; that is $\hat{\theta}_n - \theta_0 = o_{a.s.}(1)$; that conditions (9.5.1)- (9.5.6) hold with 'in probability' replaced by 'almost surely' in the conditions (9.5.1) and (9.5.3'). Suppose*

(9.6.3B) $\displaystyle\sup_{\|\theta-\theta_0\|<\delta_n} \| \, G_n^*(\theta) - G_n(\theta) - \big\{G_n^*(\theta_0) - G_n(\theta_0)\big\} \, \| = o_{p^*}(n^{-1/2})$ *for all sufficiently small positive constants δ_n.*

(9.6.4B) $\sqrt{n}\big\{G_n^*(\hat{\theta}_n) - G_n(\hat{\theta}_n)\big\} = \mathcal{N}(0, V_1) + o_{p^*}(1)$, *for some covariance matrix V_1.*

Then, $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ converges in distribution to a $\mathcal{N}(0, \Omega)$ distribution in P^*-probability.

Theorem 9.7. *Let $\{X_i\}_{i=1}^n$ be i.i.d. with $\mathbb{E}(g(X_i, \theta_0)) = 0$. Suppose that each component of g take the form $g(x, \theta) = g_c(x, \theta) + g_{lc}(x, \theta)$ and satisfies:*

(9.7.1) g_c *is Hölder continuous with respect to* θ*, in the sense that*

$$|g_c(x,\theta_1) - g_c(x,\theta_2)| \le b(x)\|\theta_1 - \theta_2\|^{s_1},$$

for some constant $s_1 \in (0,1]$ *and some measurable functions* $b(\cdot)$ *with* $E[b_j(X)]^r < \infty$ *for some* $r \ge 2$*.*

(9.7.2) g_{lc} *is for some* $r \ge 2$ *locally uniformly* $L_r(P)$ *continuous with respect to* θ*:*

$$\Big(\mathbb{E}\big[\sup_{\{\theta':\|\theta-\theta'\|<\delta\}} |g_{lc}(X,\theta') - g_{lc}(X,\theta)|^r\big]\Big)^{1/r} \le K\delta^s,$$

for all $\theta_0 \in \mathbf{\Theta}$*, for all sufficiently small positive values* $\delta = o(1)$*, and for some constants* $s \in (0,1]$*,* $K > 0$*.*

(9.7.3) $\mathbf{\Theta}$ *is a compact subset of* $\mathbb{R}^p$*.*

Then conditions (9.5.3') and (9.6.3B) hold.

9.4.1 Asymptotic Normality Properties of Bootstrap Estimates from a Penalized Likelihood Model

In this section, we study the asymptotic normality of bootstrap estimates from a penalized likelihood model where the distribution is from the exponential family by checking assumptions of Theorem 9.6.

(C1) Claim: $\hat{\theta}_n \to \theta_0$ a.s.

To prove that $\|\hat{\theta}_n - \theta_0\| = o_{\{a.s.\}}(1)$, we can follow the proof of Theorem 9.4 as follows.

From (II) For all $\delta > 0$, there is $\epsilon(\delta) > 0$, such that $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta)\| = \epsilon(\delta) > 0$.

Therefore, for all $n \in \mathbb{N}$, for all ω in the probability space Ω , $\|\hat{\theta}_n(\omega) - \theta_0\| > \delta$ implies $\|G(\hat{\theta}_n(\omega))\| \geq \epsilon(\delta)$. Hence,

$$P(\lim_{n\to\infty} \|\hat{\theta}_n - \theta_0\| > \delta) \leq P(\lim_{n\to\infty} \|G(\hat{\theta}_n)\| \geq \epsilon(\delta)).$$

So, it suffices to show that $\|G(\hat{\theta}_n)\| = o_{\{a.s.\}}(1)$.

$$\begin{aligned}
\|G(\hat{\theta}_n)\| &= \|G(\hat{\theta}_n) + G_n(\hat{\theta}_n) - G_n(\hat{\theta}_n)\| \\
&\leq \|G_n(\hat{\theta}_n)\| + \|G(\hat{\theta}_n) - G_n(\hat{\theta}_n)\| \\
&\leq \|G_n(\hat{\theta}_n)\| + o_{\{a.s.\}}(1)\big(1 + \|G(\hat{\theta}_n)\| + \|G_n(\hat{\theta}_n)\|\big) \quad \text{(by (IV))}.
\end{aligned} \tag{9.4.15}$$

By rearranging (9.4.15), we have

$$\|G(\hat{\theta}_n)\| \left(1 - o_{\{a.s.\}}(1)\right) \leq o_{\{a.s.\}}(1) + \|G_n(\hat{\theta}_n)\|(1 + o_{\{a.s.\}}(1)). \tag{9.4.16}$$

From (IV), $\|G_n(\theta_0)\| = o_{\{a.s.\}}(1)$. Then, from (I),

$$\|G_n(\hat{\theta})\| \leq o_{\{a.s.\}}(1) + \|G_n(\theta_0)\| = o_{\{a.s.\}}(1). \tag{9.4.17}$$

Therefore, by (9.4.16) and (9.4.17),

$$\|G(\hat{\theta}_n)\| = o_{\{a.s.\}}(1). \tag{9.4.18}$$

The condition (C1) is then verified.

(C2) Claim: $\sup_{\|\theta-\theta_0\|<\delta_n} \| \ G_n^*(\theta) - G_n(\theta) - \left\{G_n^*(\theta_0) - G_n(\theta_0)\right\} \| = o_{p^*}(n^{-1/2})$ for all sequences $\delta_n = o(1)$.

Since the function $g(x, \theta)$ is convex as a function of θ and $\mathbf{\Theta}$ is compact, $g(x, \theta)$ is Lipschitz continuous. Therefore, by Theorem 9.7 the condition (9.6.3B) is satisfied.

(C3) Claim: $\sqrt{n}\{G_n^*(\hat{\theta}_n) - G_n(\hat{\theta}_n)\} = \mathcal{N}(0, V_1) + o_{p^*}(1)$, for a covariance matrix V_1.

Following Chen et al. (2003), and by (VI),

$$\sqrt{n}G_n(\theta_0) \rightsquigarrow N(0, V).$$

By (V) and (C1),

$$\sqrt{n}\|G_n(\hat{\theta}_n) - G(\hat{\theta}_n) - G_n(\theta_0)\| = o_P(1). \tag{9.4.19}$$

Therefore,

$$\sqrt{n}\{G_n(\hat{\theta}_n) - G(\hat{\theta}_n)\} \rightsquigarrow N(0, V). \tag{9.4.20}$$

By Giné and Zinn (1990),

$$\sqrt{n}\{G_n^*(\hat{\theta}_n) - G_n(\hat{\theta}_n)\} \rightsquigarrow N(0, V).$$

This verifies condition (9.6.4B).

From conditions (C1)-(C3), $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$ is asymptotically normally distributed.

Chapter 10

Appendix D: Graphical Results for Bootstrap studies in Chapter 3

This chapter is intended to be a graphical supplement for the bootstrap confidence intervals mentioned in Chapter 3. Two types of pointwise confidence intervals are studied in this chapter: standard normal confidence interval and percentile confidence interval. The parameters used in this study are derived from the SSLC model. For each parameter θ , the model estimate of the parameter is denoted by $\hat{\theta}$ and the bootstrap estimate is defined by

$$\hat{\theta}^{(*)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}.$$

The $(1-\alpha)100\%$ bootstrap standard normal confidence interval is given by

$$[\hat{\theta}^{(*)} - z_{\frac{\alpha}{2}} \cdot se(\hat{\theta}^{(*)}), \hat{\theta}^{(*)} + z_{\frac{\alpha}{2}} \cdot se(\hat{\theta}^{(*)})],$$

where $se(\hat{\theta}^{(*)})$ is the estimated standard error of $\hat{\theta}^{(*)}$.

The $(1-\alpha)100\%$ bootstrap percentile confidence interval is defined by

$$[\hat{\theta}_B^{\frac{\alpha}{2}}, \hat{\theta}_B^{(1-\frac{\alpha}{2})}],$$

where $\hat{\theta}_B^{\alpha}$ is the $B\alpha^{th}$ value in the ordered list of the B replications.

In many applications including our cases, there are some extreme values present. These extreme values make the estimated standard deviation too large compared to the true value and consequently the corresponding normal curve is too flat. To avoid

such a problem each overlaid normal curve in the histogram of each parameter θ in this section will be drawn from a normal distribution with mean $\hat{\theta}^{(*)}$ and standard deviation $\hat{\sigma}_\theta$. The standard deviation $\hat{\sigma}_\theta$ is estimated by

$$\hat{\sigma}_\theta = \frac{Q_3(\theta^*) - Q_1(\theta^*)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)},$$

where Φ is the standard normal distribution function, and $Q_1(\theta)$ and $Q_3(\theta)$ are the 1^{st} and 3^{rd} quartiles of bootstrap values of the parameter θ, respectively.

Heart Diseases

Figures 10.1-10.31 show histograms of bootstrapped samples and graphical comparisons of 95% bootstrap pointwise confidence intervals and 95% bootstrap pointwise confidence interval widths between the two types of confidence intervals for parameter estimates. Figures 10.1-10.5 suggest the comparability between the two types of intervals of α_a's. Figures 10.6-10.9 show symmetric shapes of histograms for bootstrapped samples for all α_a's. The 95% bootstrap pointwise confidence interval widths for α_a's are smaller at old ages than at young ages. Figures 10.10-10.11 demonstrate some differences between the two intervals of β_a at young ages, ages 1-12 years. The differences between the two intervals are also noticed for $\gamma_{p,1}$'s and $\gamma_{p,2}$'s as shown in Figures 10.16- 10.17 and 10.20-10.21, respectively. These differences occur because of non-normality of histograms for the bootstrap samples as we can see non-symmetric shapes of histograms of $\hat{\beta}_a$'s , $\hat{\gamma}_{p,1}$'s, and $\hat{\gamma}_{p,2}$'s in Figures 10.12-10.15, 10.18-10.19, and 10.22-10.23, respectively. Figures 10.32-10.35 show the corresponding comparisons for log mortality rate estimates at some selected ages.

The figures show agreements of the two types of 95% bootstrap pointwise confidence intervals for estimated log mortality rates. Figures 10.36-10.37 show histograms of bootstrapped samples of log mortality rates at a selected age, 14 years. The figures confirm the normality assumption of log mortality rate distribution.

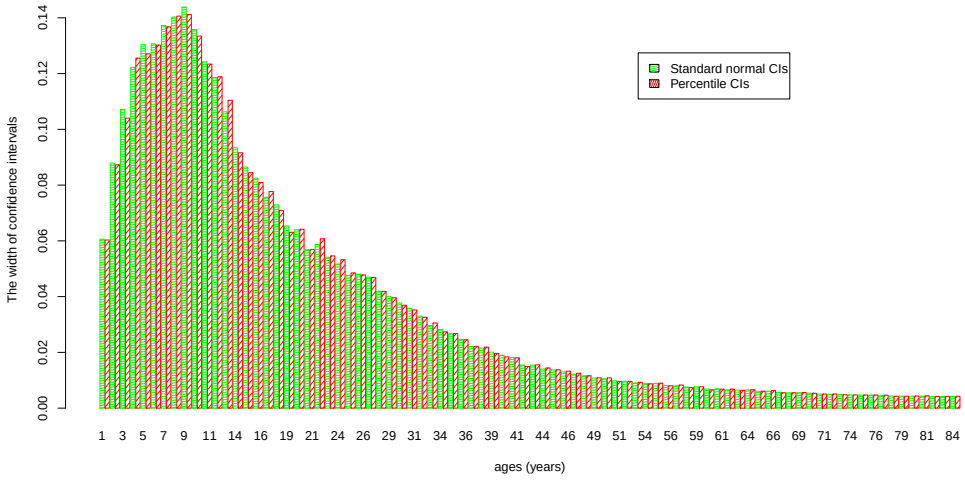


Figure 10.1: Heart diseases: 95% bootstrap pointwise confidence interval widths of $\alpha_a : a = 1, \ldots, 84$ obtained from percentile and standard normal intervals.

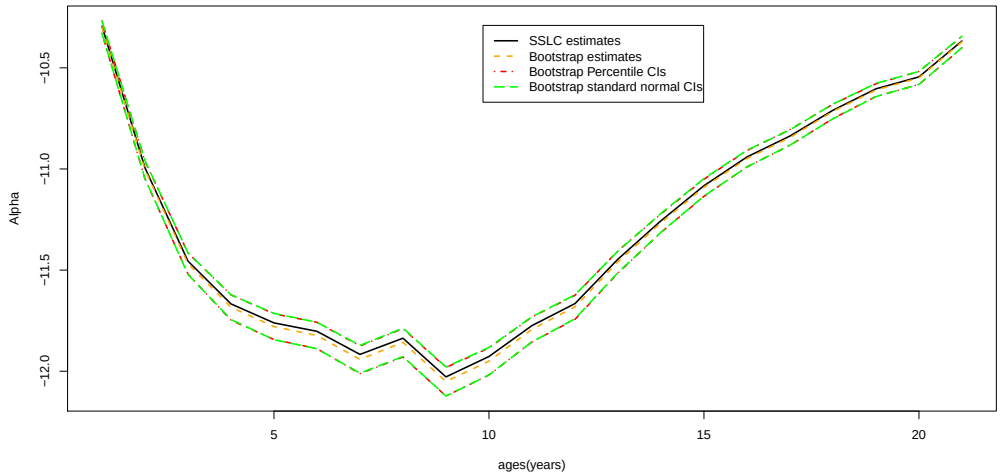


Figure 10.2: Heart diseases : $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 1, \ldots, 21$ and corresponding 95% bootstrap pointwise confidence intervals.

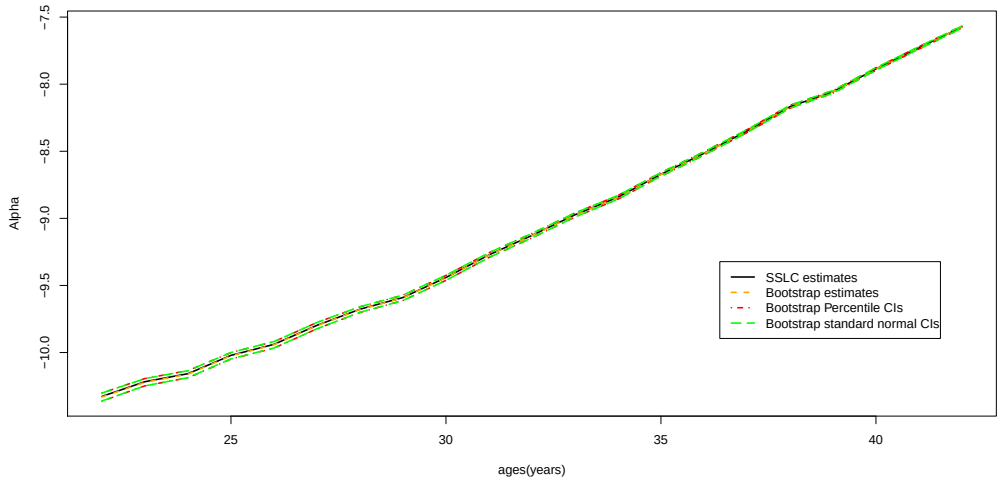


Figure 10.3: Heart diseases: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 22, \ldots, 42$ and corresponding 95% bootstrap pointwise confidence intervals.

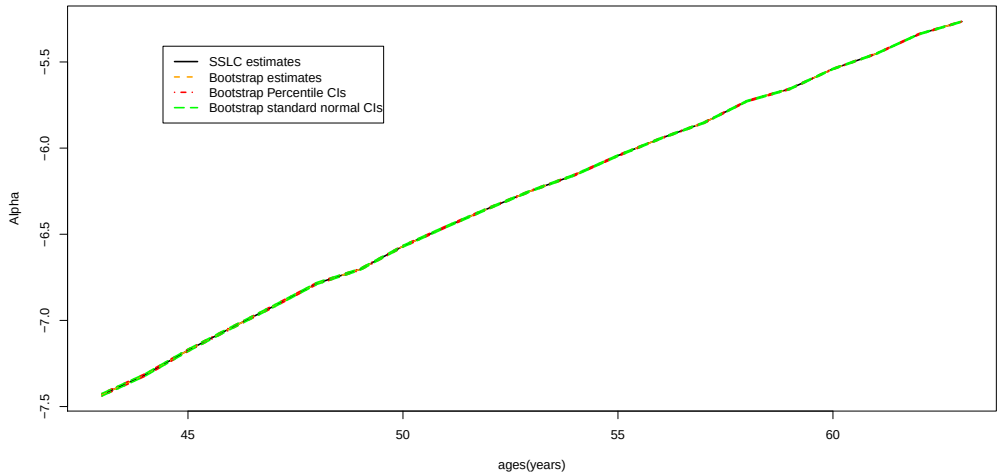


Figure 10.4: Heart diseases: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 43, \ldots, 63$ and corresponding 95% bootstrap pointwise confidence intervals.

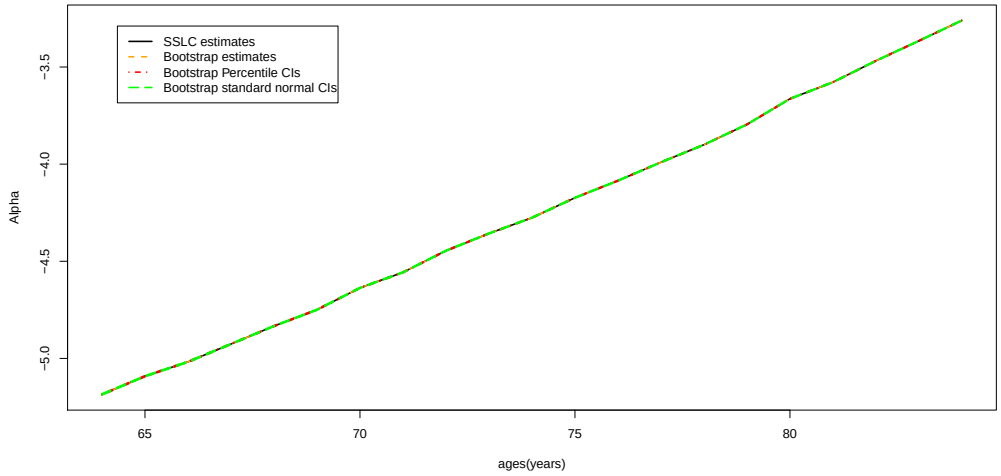


Figure 10.5: Heart diseases: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 64, \ldots, 84$ and corresponding 95% bootstrap pointwise confidence intervals.

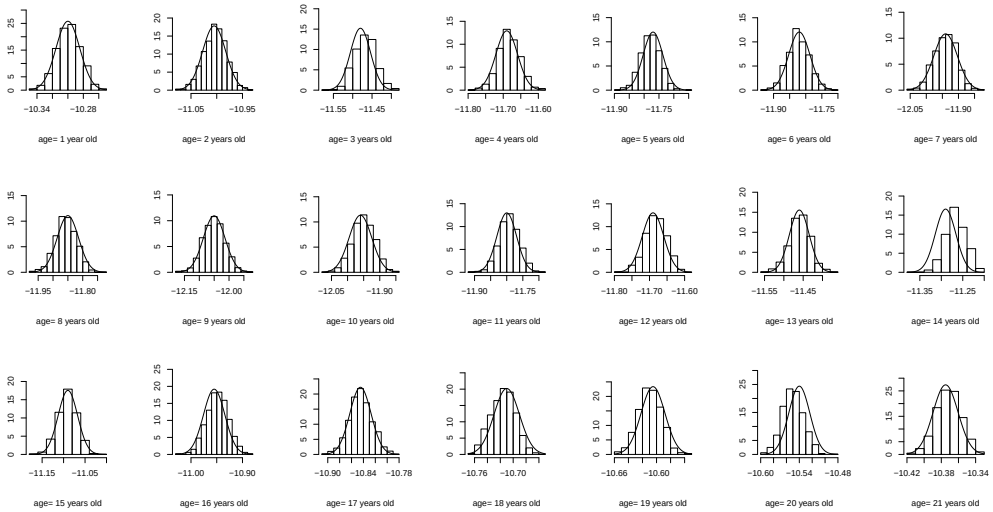


Figure 10.6: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 1, ..., 21$.

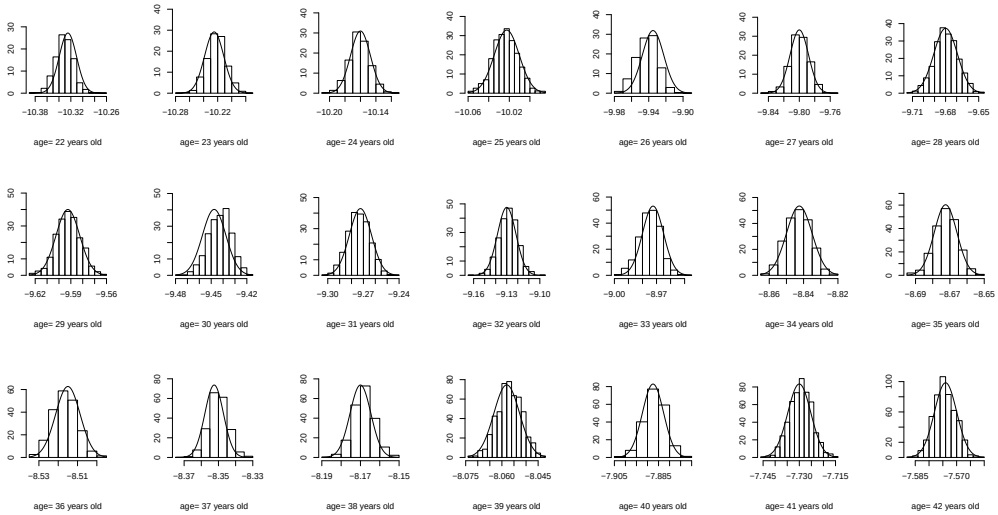


Figure 10.7: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 22, ..., 42$.

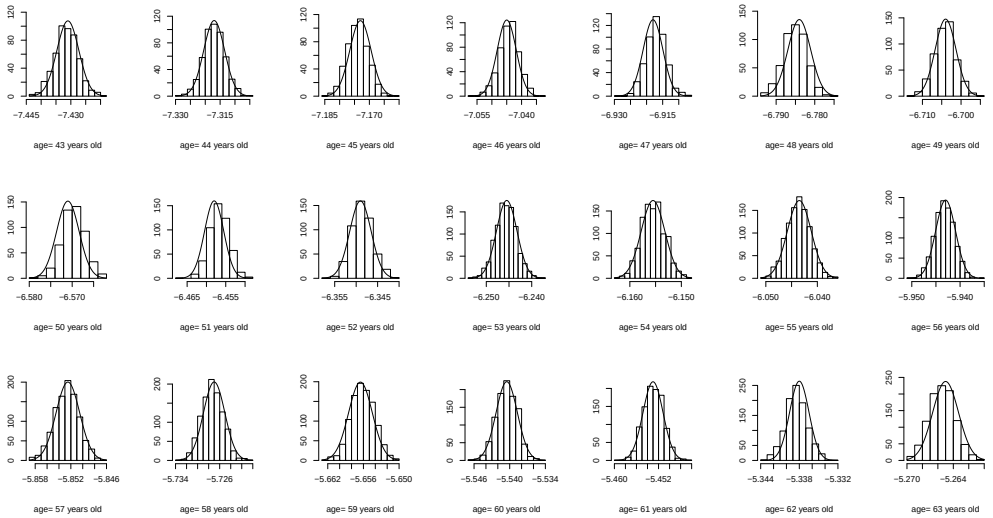


Figure 10.8: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 43, ..., 63$.

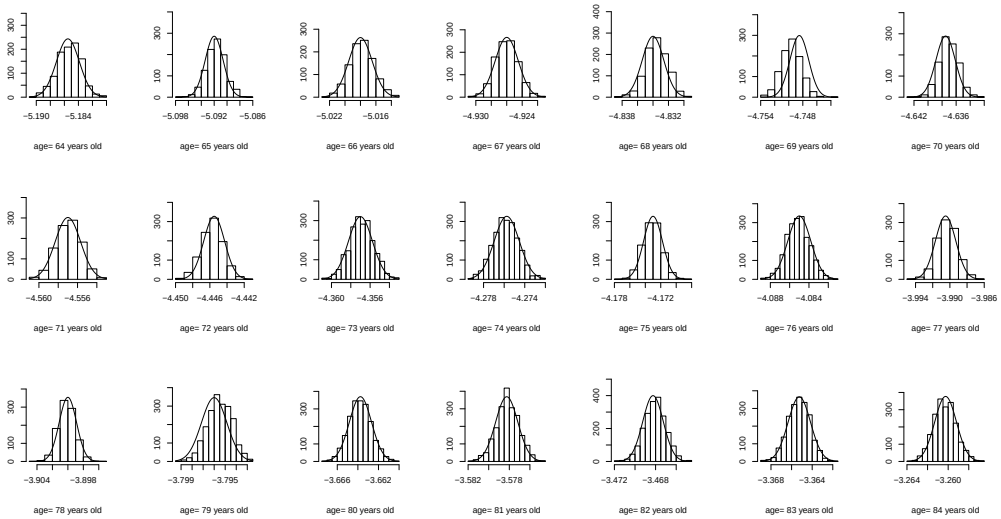


Figure 10.9: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 64, ..., 84$.

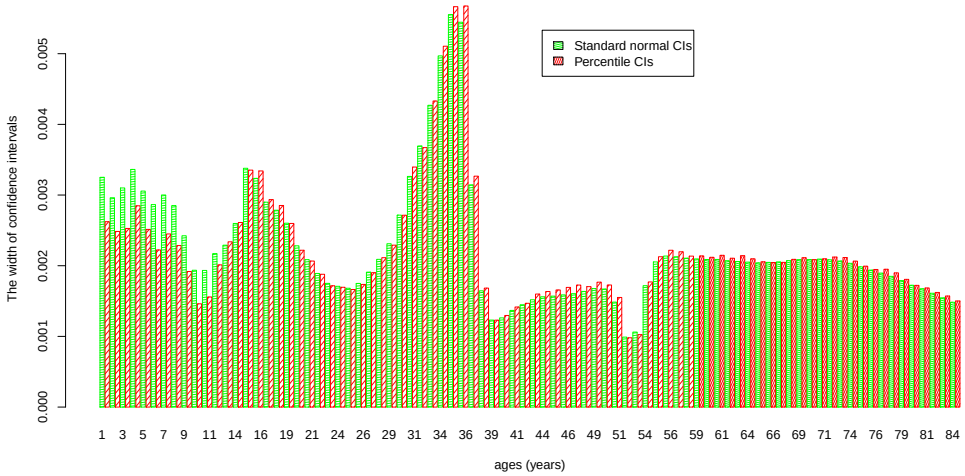


Figure 10.10: Heart diseases: 95% bootstrap pointwise confidence interval widths of $\beta_a : a = 1, \ldots, 84$ obtained from percentile and standard normal intervals.

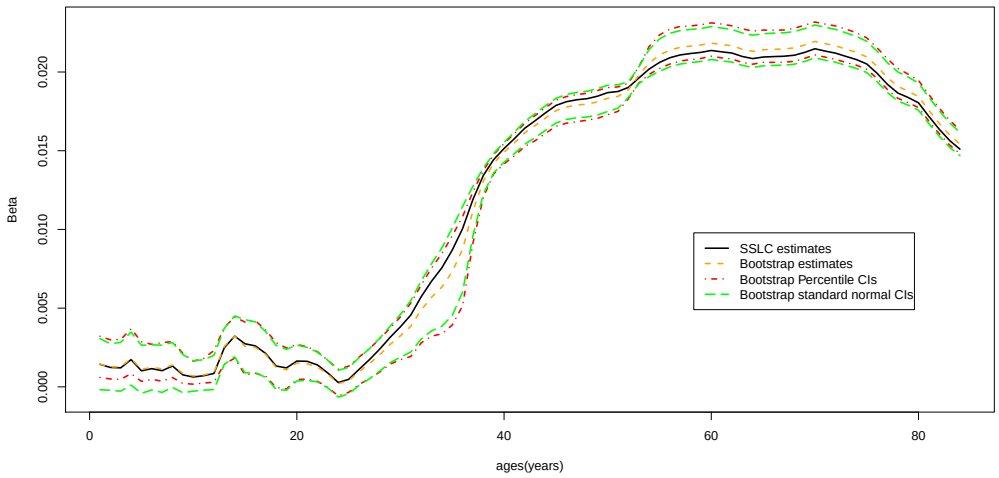


Figure 10.11: Heart diseases: $\hat{\beta}_a, \hat{\beta}_a^{(*)} : a = 1, \ldots, 84$ and corresponding 95% bootstrap pointwise confidence intervals.

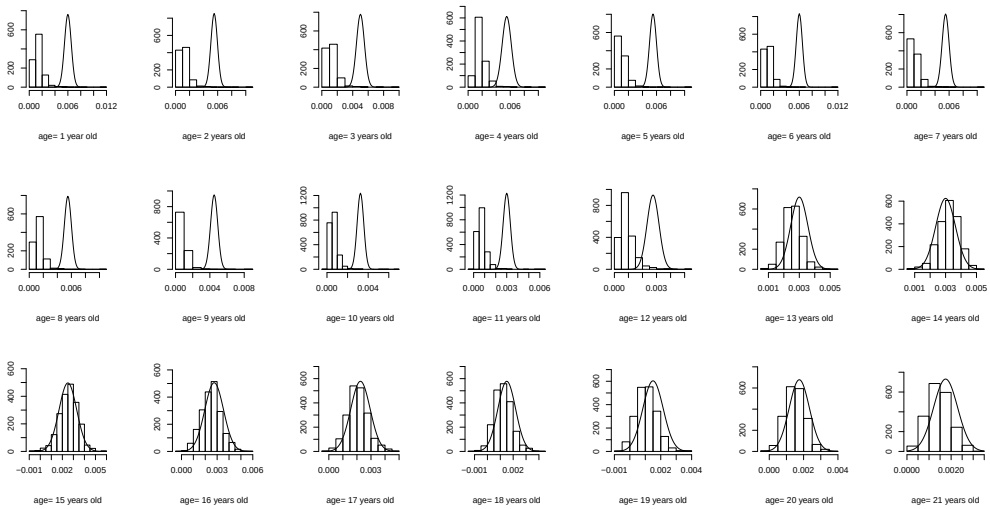


Figure 10.12: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 1, ..., 21$.

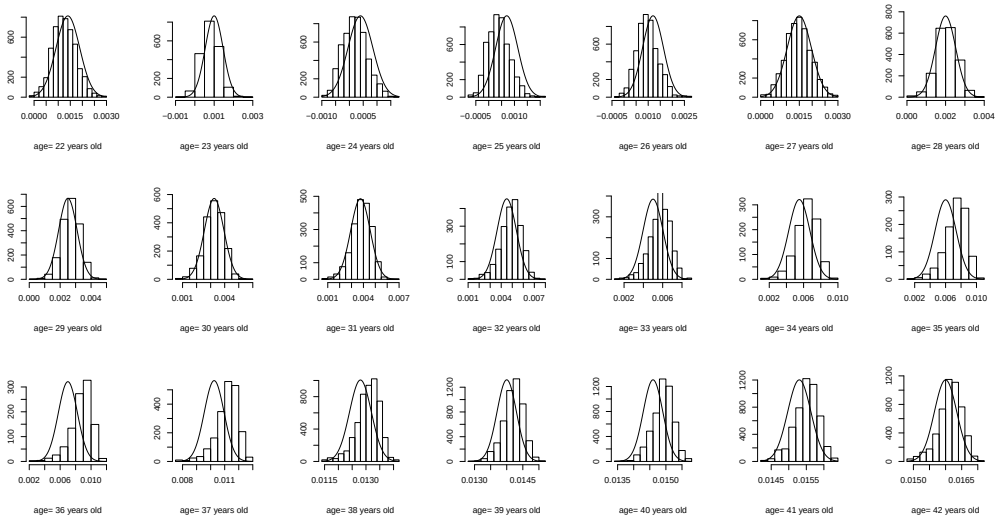


Figure 10.13: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 22, ..., 42$.

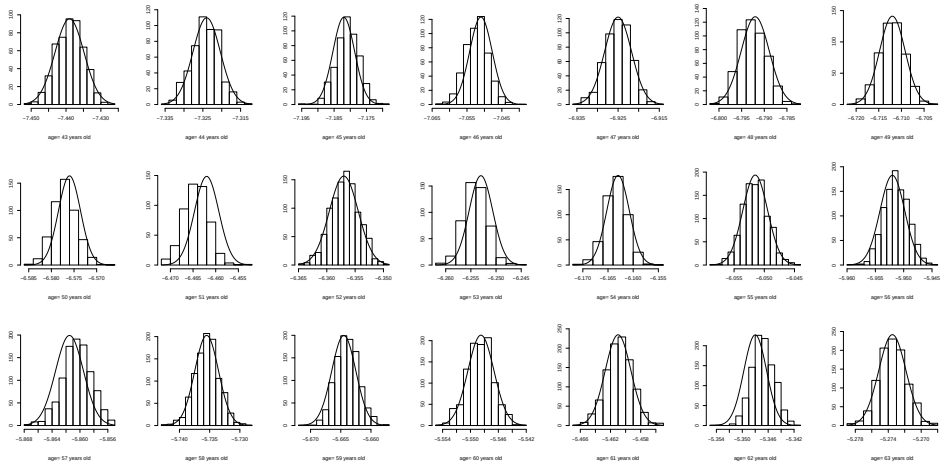


Figure 10.14: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 43, ..., 63$.

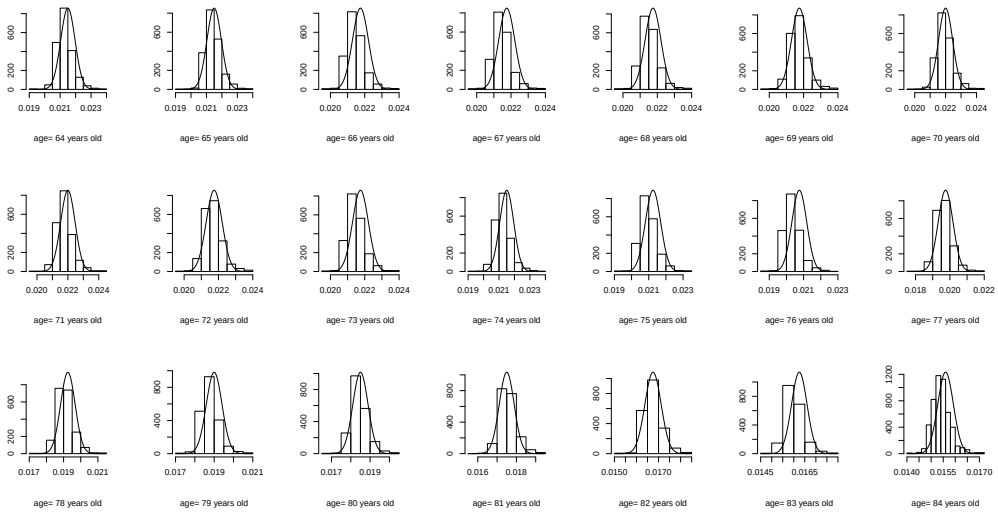


Figure 10.15: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 64, ..., 84$.

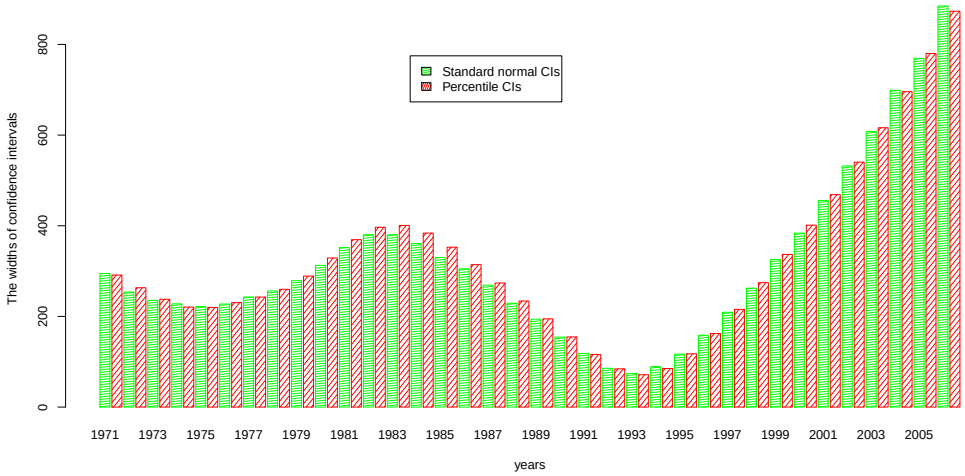


Figure 10.16: Heart diseases: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,1} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

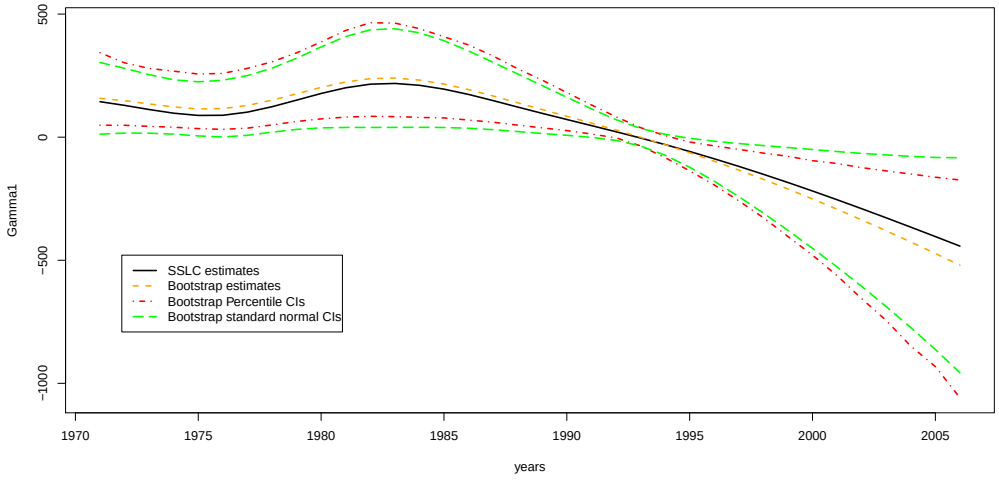


Figure 10.17: Heart diseases: $\hat{\gamma}_{p,1}, \hat{\gamma}_{p,1}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

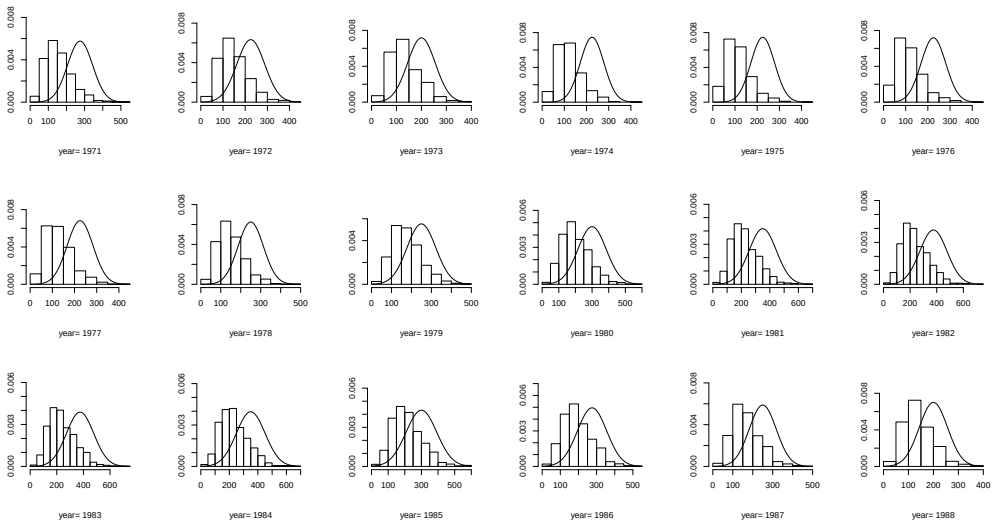


Figure 10.18: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,1} : p = 1971, ..., 1988$.

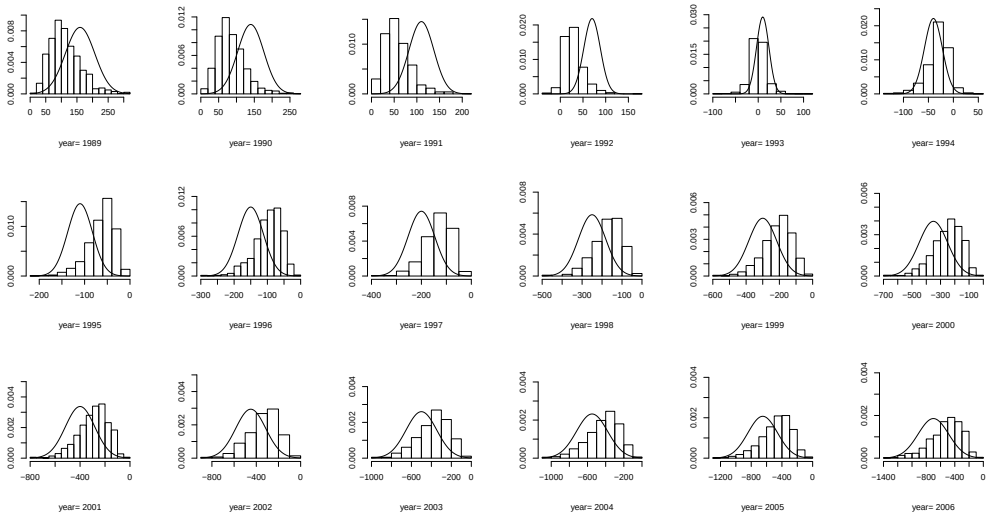


Figure 10.19: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,1} : p = 1989, ..., 2006$.

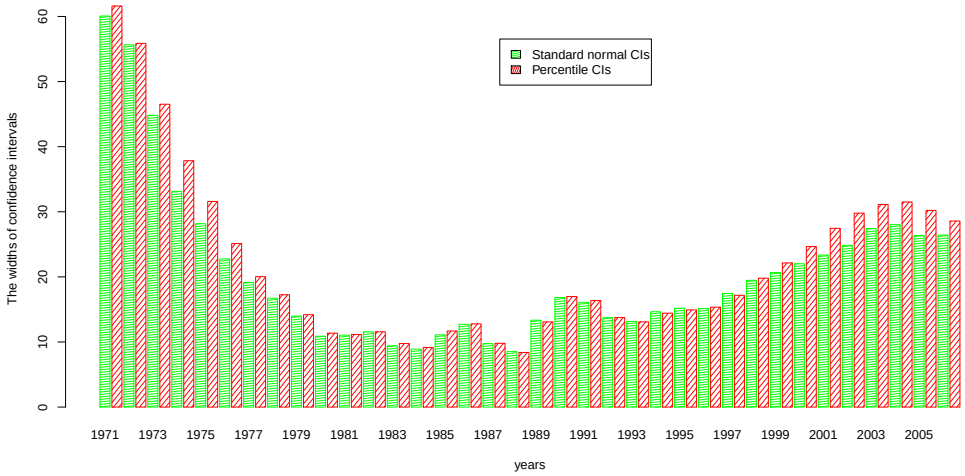


Figure 10.20: Heart diseases: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,2} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

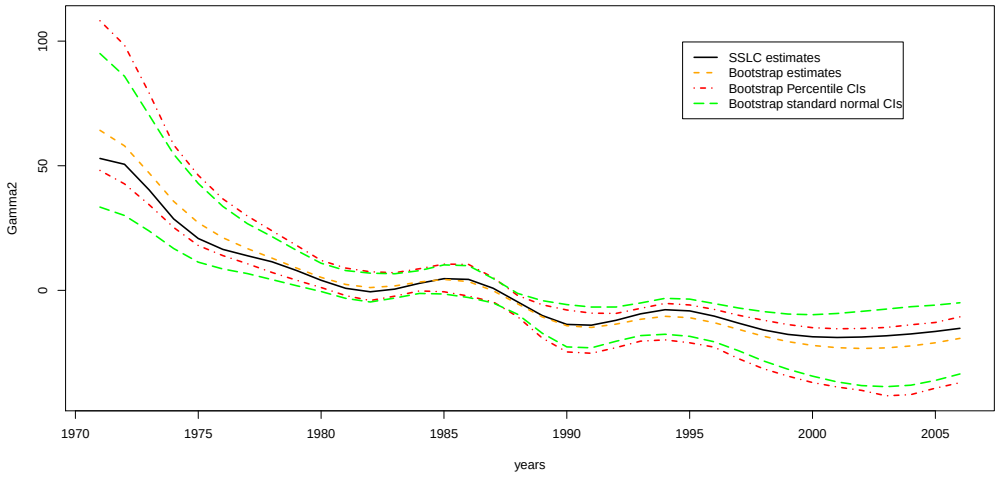


Figure 10.21: Heart diseases: $\hat{\gamma}_{p,2}, \hat{\gamma}_{p,2}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

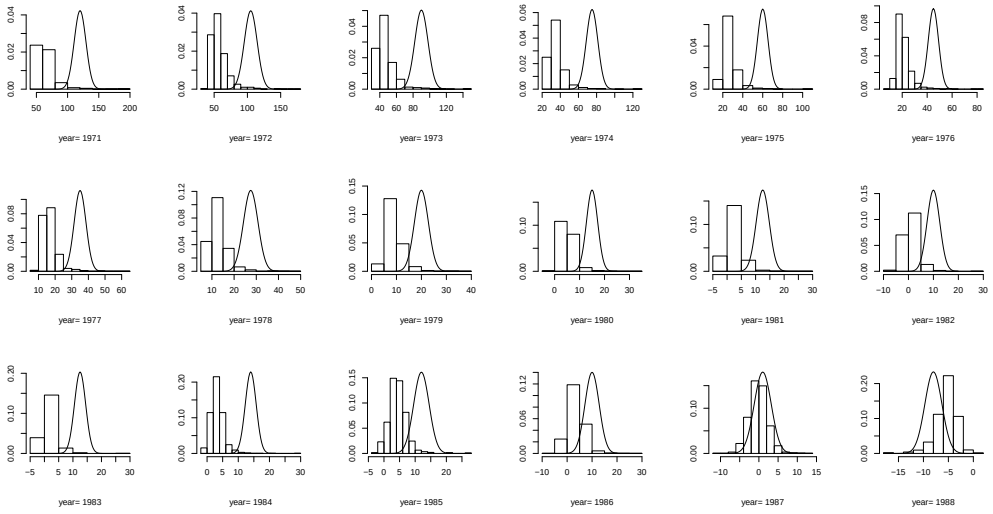


Figure 10.22: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,2} : p = 1971, ..., 1988$.

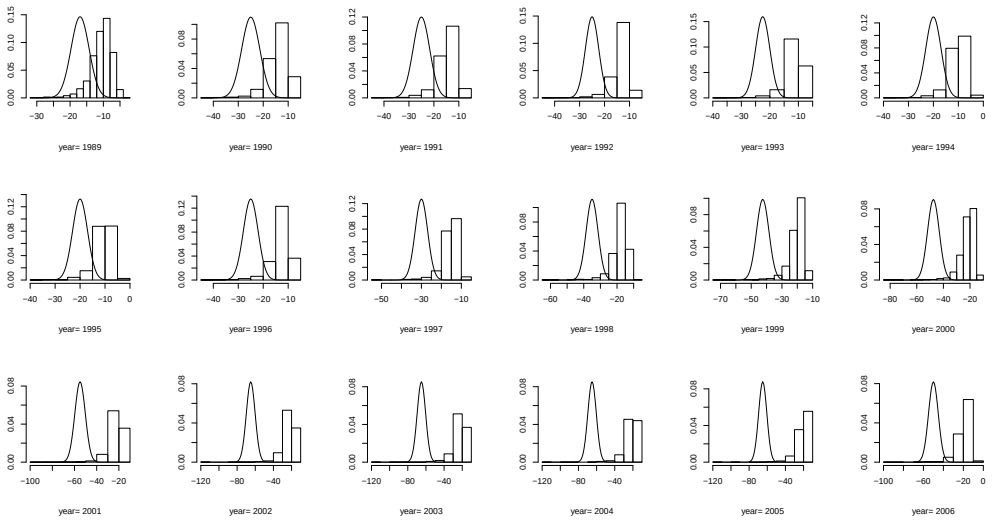


Figure 10.23: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,2} : p = 1989, ..., 2006$.

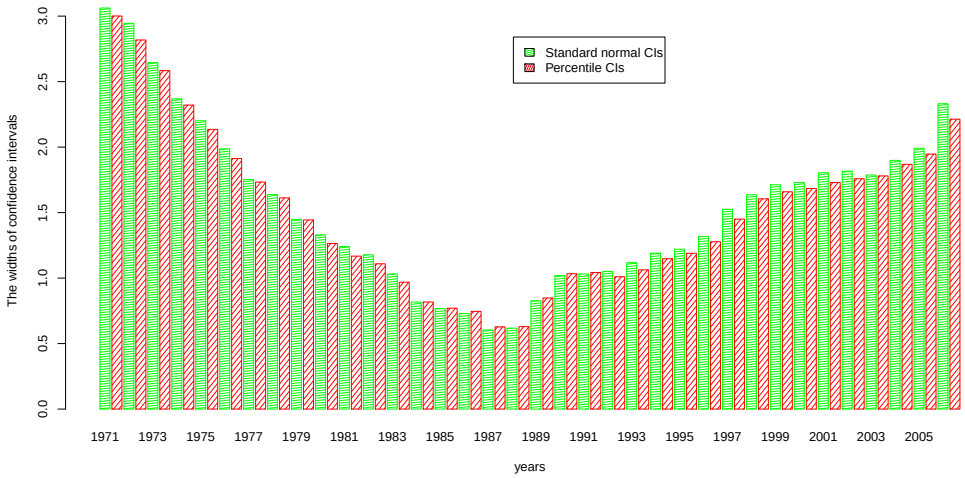


Figure 10.24: Heart diseases: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,3} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

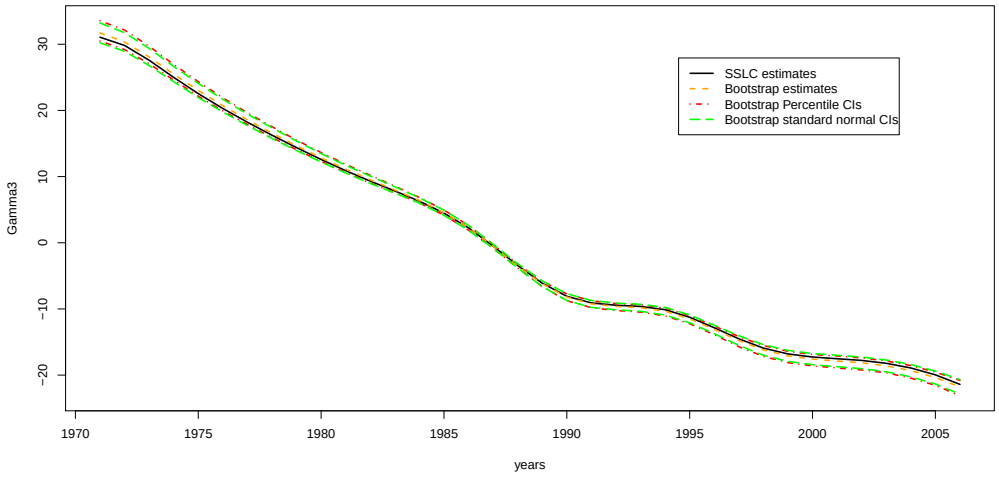


Figure 10.25: Heart diseases: $\hat{\gamma}_{p,3}, \hat{\gamma}_{p,3}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

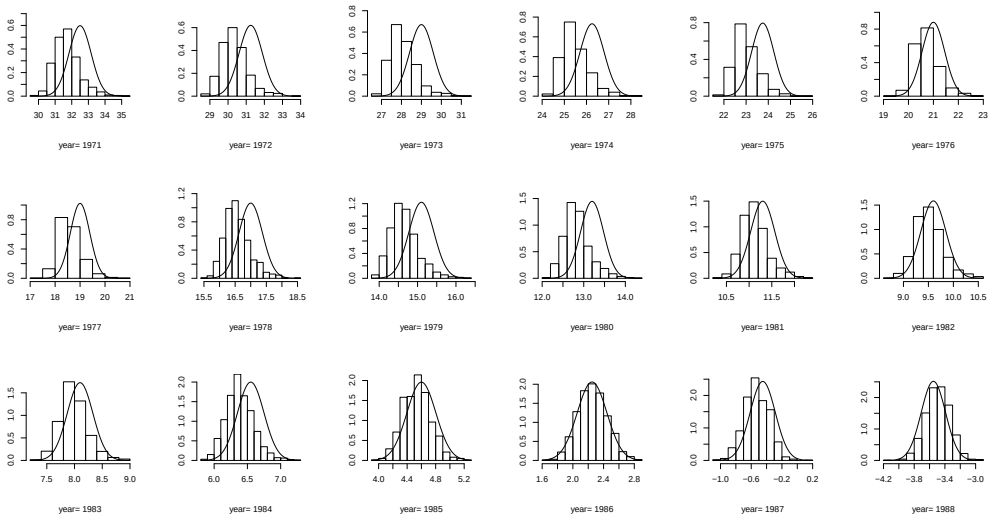


Figure 10.26: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,3} : p = 1971, ..., 1988$.

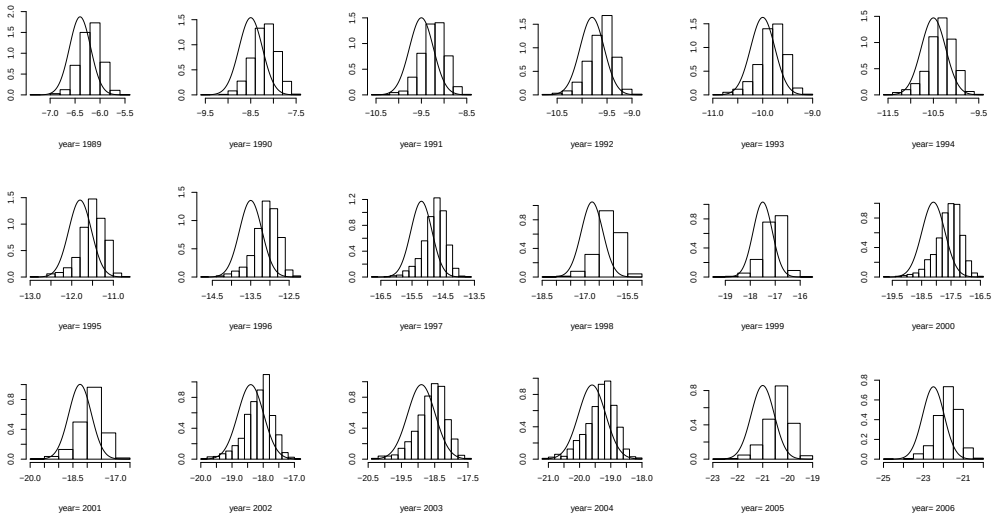


Figure 10.27: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,3} : p = 1989, ..., 2006$.

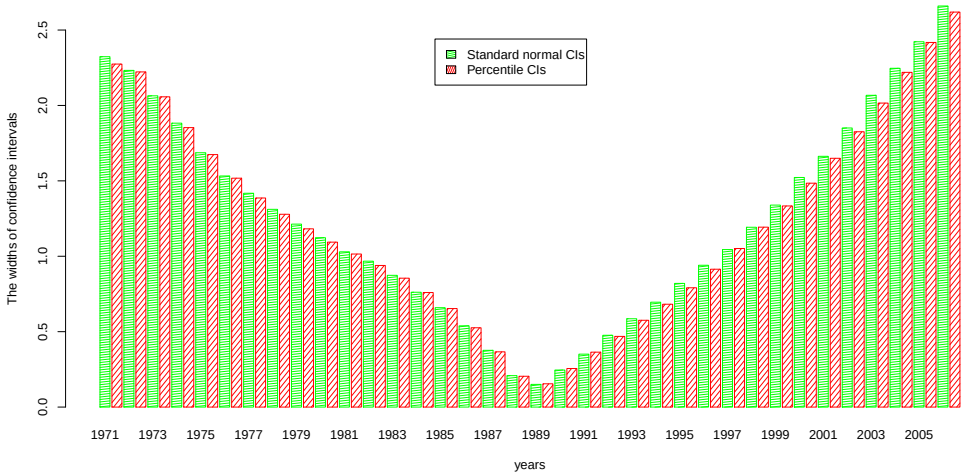


Figure 10.28: Heart diseases: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,4} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

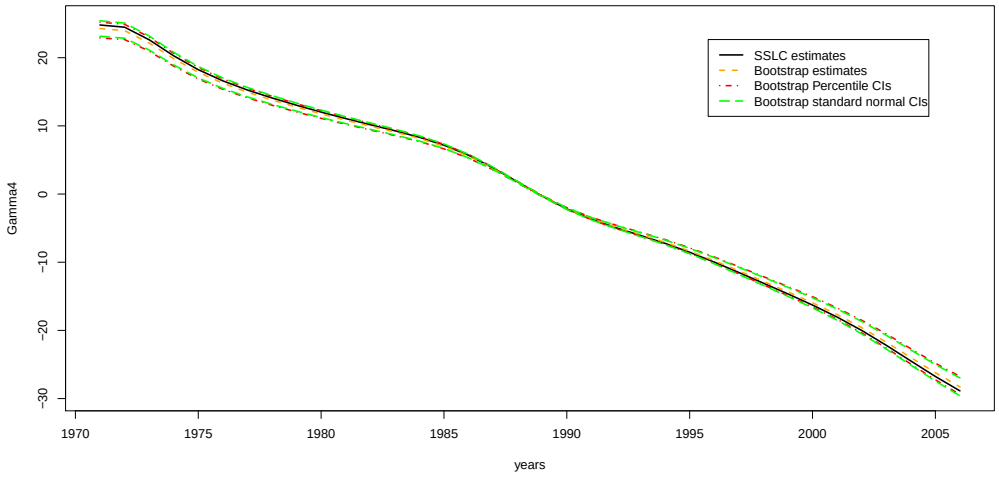


Figure 10.29: Heart diseases: $\hat{\gamma}_{p,4}, \hat{\gamma}_{p,4}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

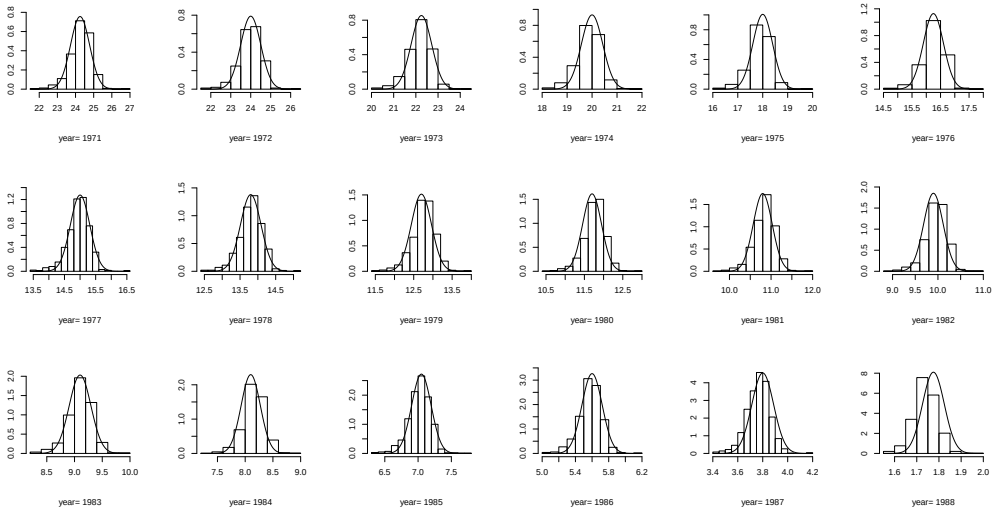


Figure 10.30: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,4} : p = 1971, ..., 1988$.

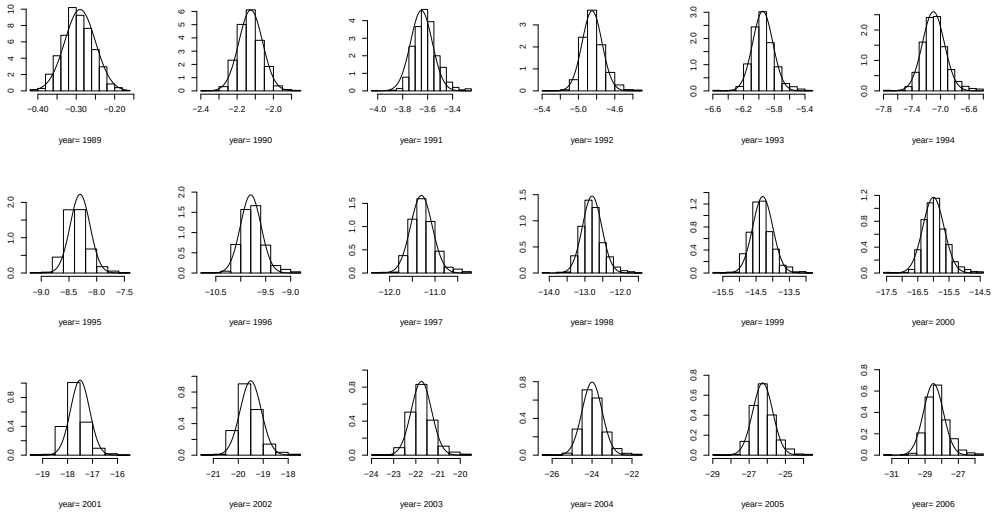


Figure 10.31: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,4} : p = 1989, ..., 2006$.

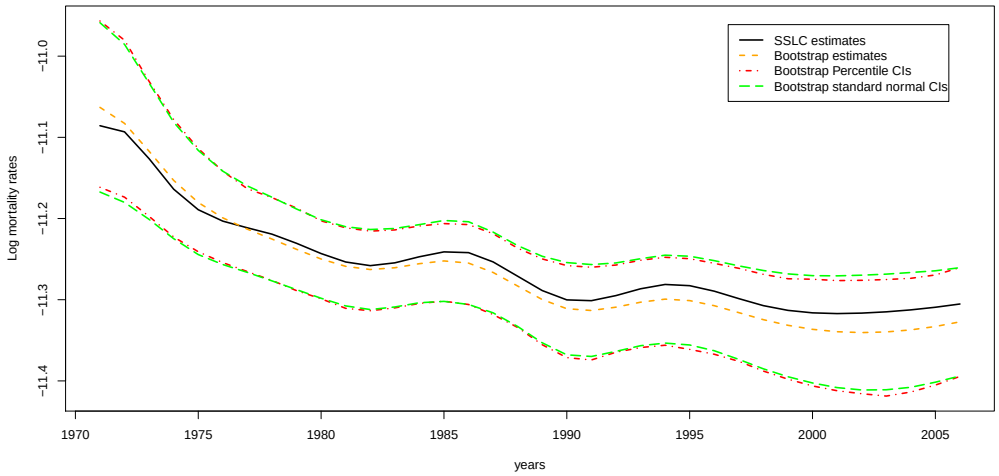


Figure 10.32: Heart diseases: Log mortality rate estimates at age 14 years and 95% bootstrap pointwise confidence intervals.

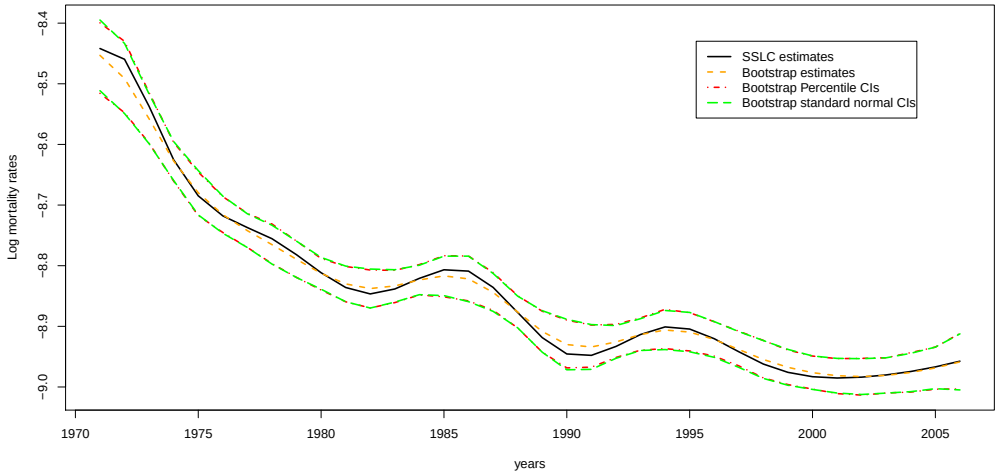


Figure 10.33: Heart diseases: Log mortality rate estimates at age 34 years and 95% bootstrap pointwise confidence intervals.

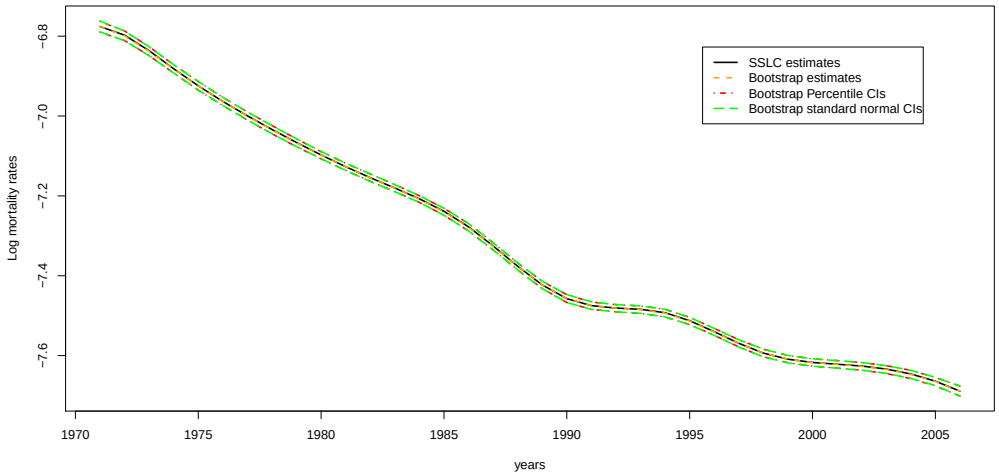


Figure 10.34: Heart diseases: Log mortality rate estimates at age 44 years and 95% bootstrap pointwise confidence intervals.

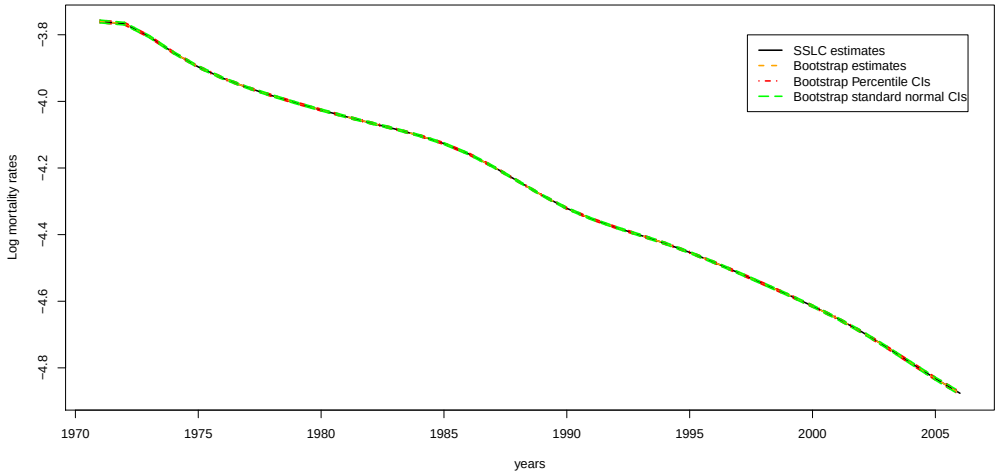


Figure 10.35: Heart diseases: Log mortality rate estimates at age 74 years and 95% bootstrap pointwise confidence intervals.

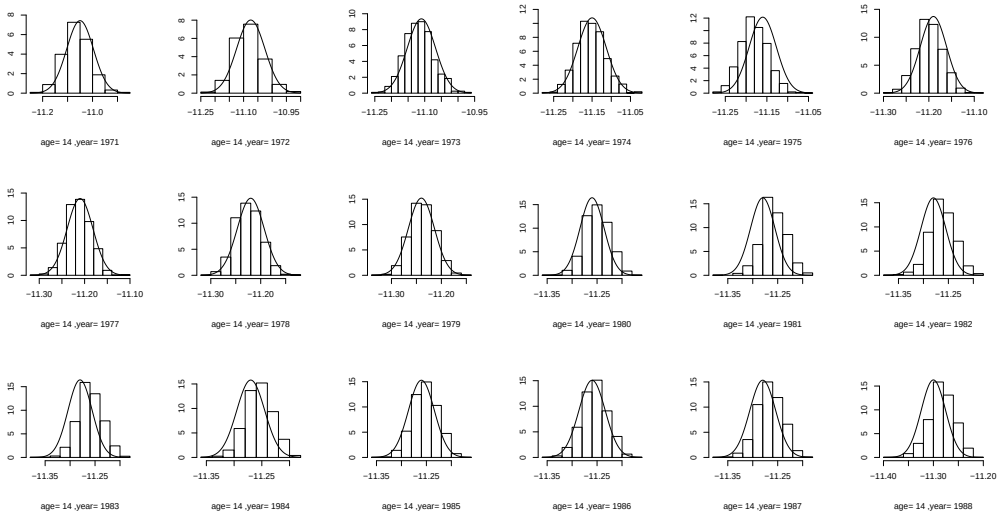


Figure 10.36: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\log(\hat{\lambda}_{14,p}) : p = 1971, ..., 1988$.

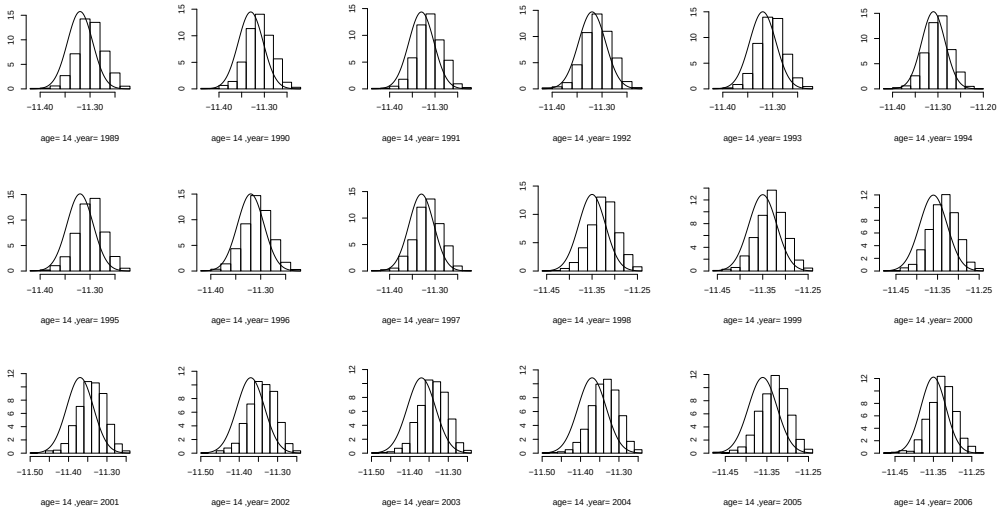


Figure 10.37: Heart diseases: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\log(\hat{\lambda}_{14,p}) : p = 1989, ..., 2006$.

Cancer

Figures 10.38-10.64 show graphical comparisons of 95% bootstrap pointwise confidence intervals and 95% bootstrap pointwise confidence interval widths between the two types of confidence intervals for parameter estimates. Figures 10.38-10.42 and 10.47-10.48 show that the two types of intervals for $\alpha_a, \beta_a : a = 1, \ldots, 84$ coincide. Histograms of the bootstrapped samples also agree to the corresponding normal curves, as shown in Figures 10.43-10.46 and 10.49-10.52 for α_a and β_a, respectively. Figures 10.53-10.64 also show agreements between the two types of bootstrap pointwise confidence intervals for period-effect terms $\gamma_{p,i} : p = 1971, \ldots, 2006 \ ; \ i = 1, 2, 3$. Figures 10.65-10.68 show the corresponding comparisons for log mortality rate estimates at some selected ages. Histograms with corresponding normal curves of bootstrapped samples of log mortality rate estimates at age 14 years are presented in Figures 10.69-10.70.

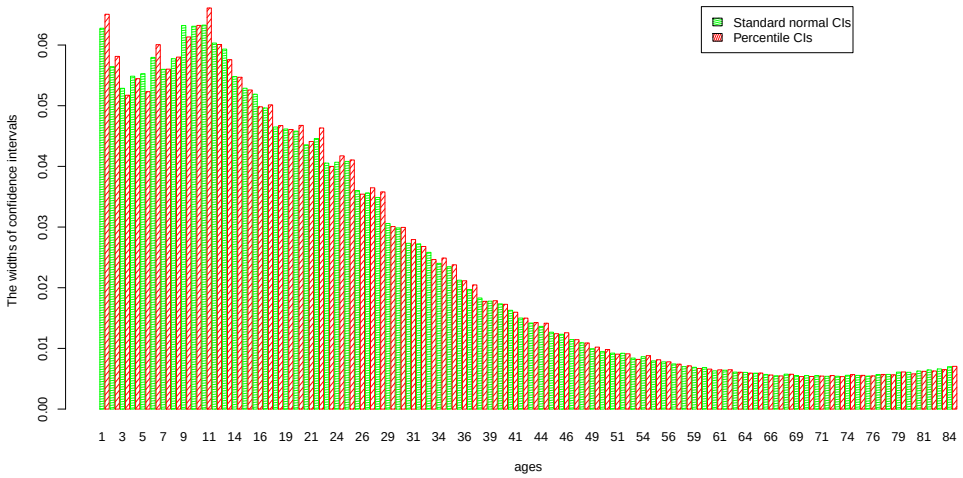


Figure 10.38: Cancer: 95% bootstrap pointwise confidence interval widths of α_a : $a = 1, ..., 84$ obtained from percentile and standard normal intervals.

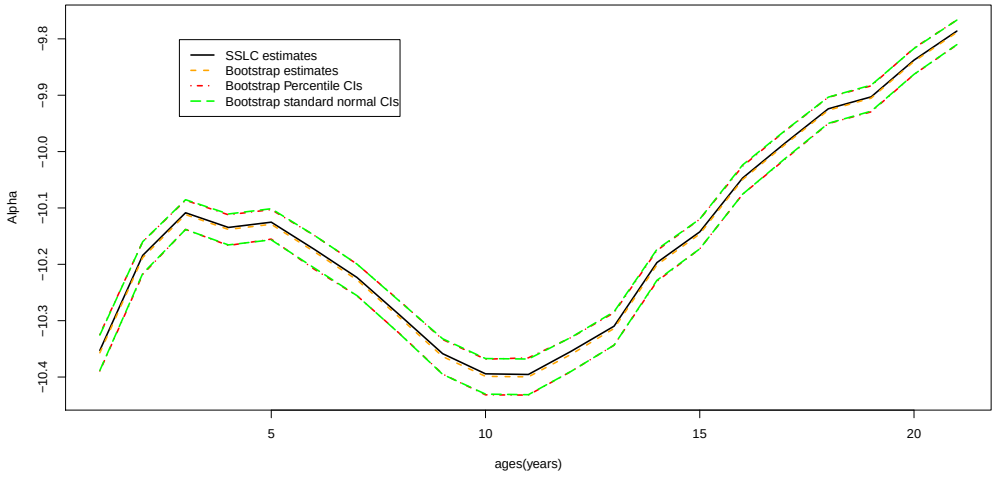


Figure 10.39: Cancer: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)}$: $a = 1, ..., 21$ and corresponding 95% bootstrap pointwise confidence intervals.

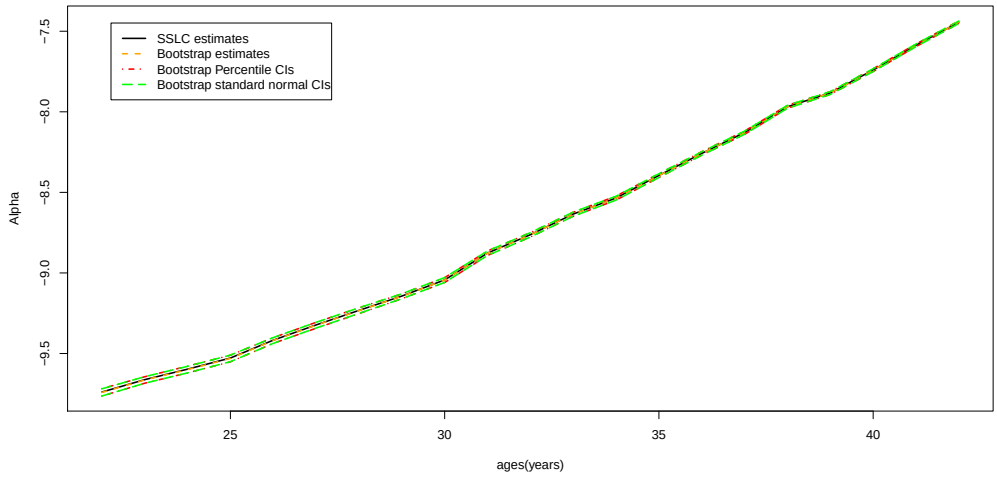


Figure 10.40: Cancer: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 22, ..., 42$ and corresponding 95% bootstrap pointwise confidence intervals.

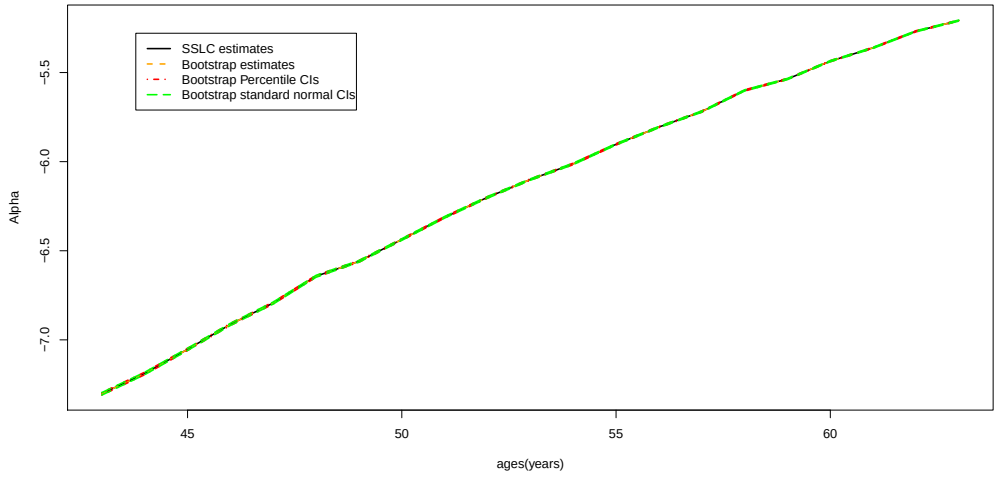


Figure 10.41: Cancer: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 43, ..., 63$ and corresponding 95% bootstrap pointwise confidence intervals.

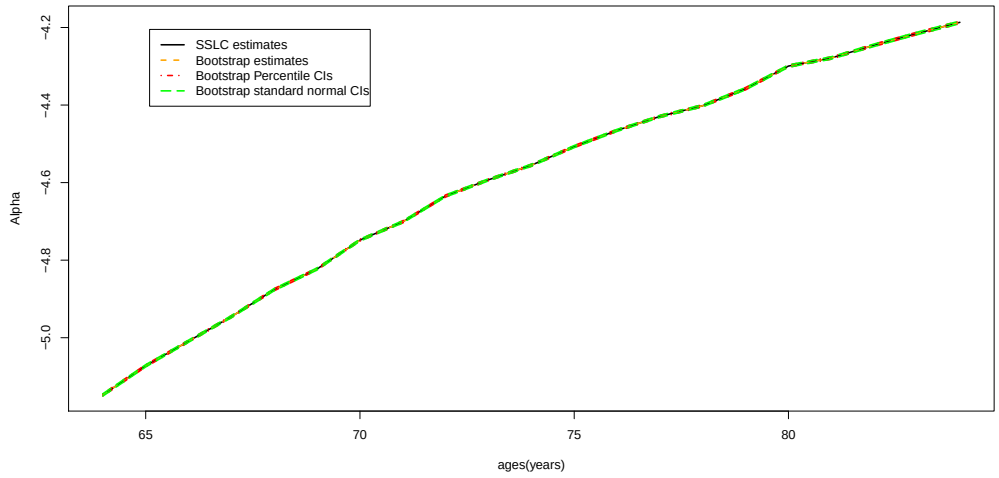


Figure 10.42: Cancer: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 64, ..., 84$ and corresponding 95% bootstrap pointwise confidence intervals.

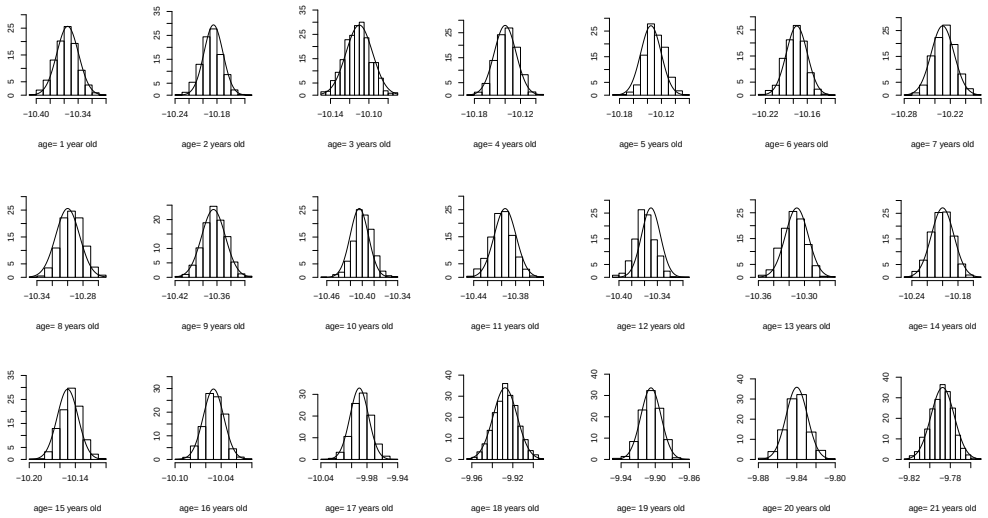


Figure 10.43: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 1, ..., 21$.

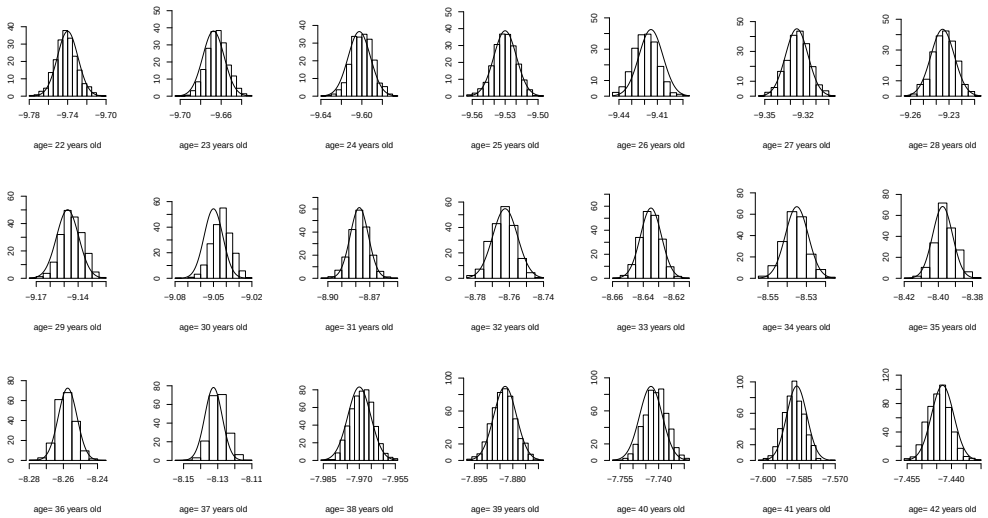


Figure 10.44: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 22, ..., 42$.

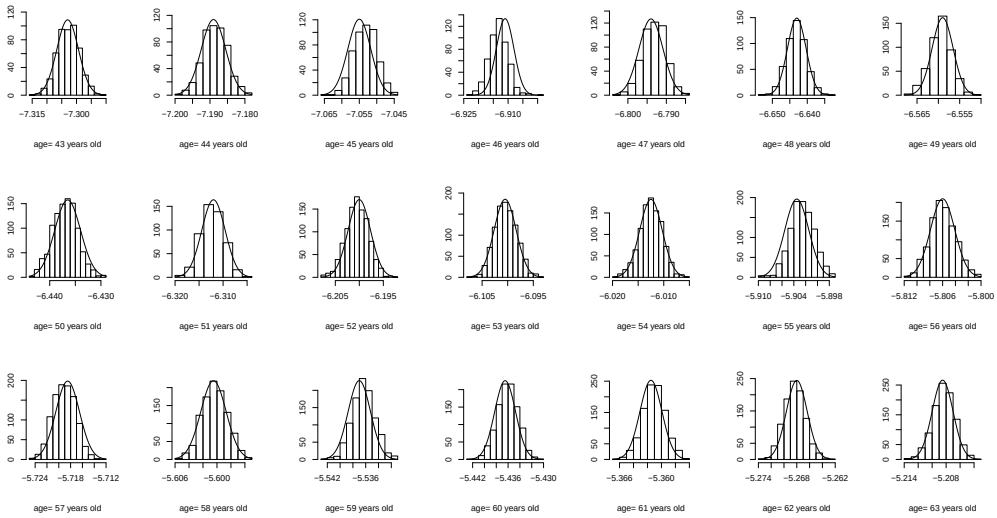


Figure 10.45: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 43, ..., 63$.

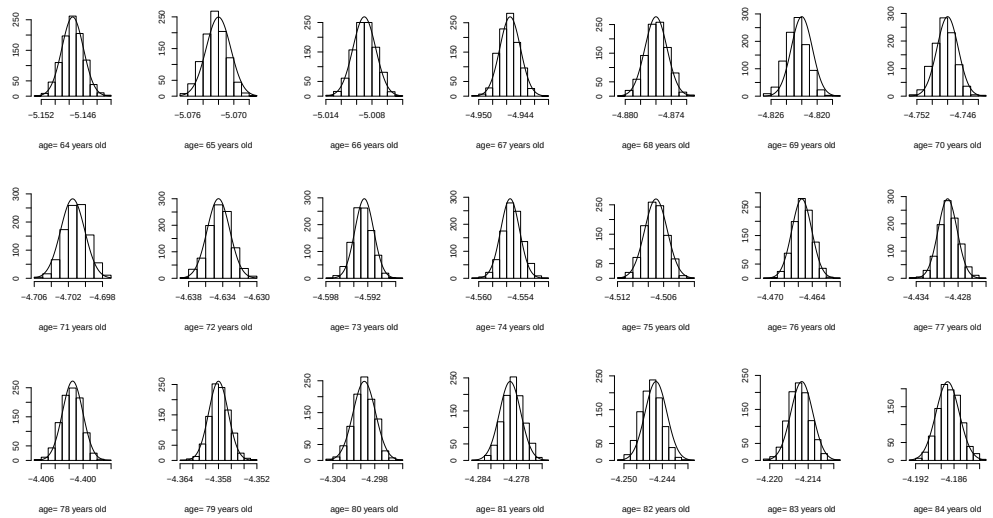


Figure 10.46: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 64, ..., 84$.

Figure 10.47: Cancer: 95% bootstrap pointwise confidence interval widths of β_a : $a = 1, ..., 84$ obtained from percentile and standard normal intervals.

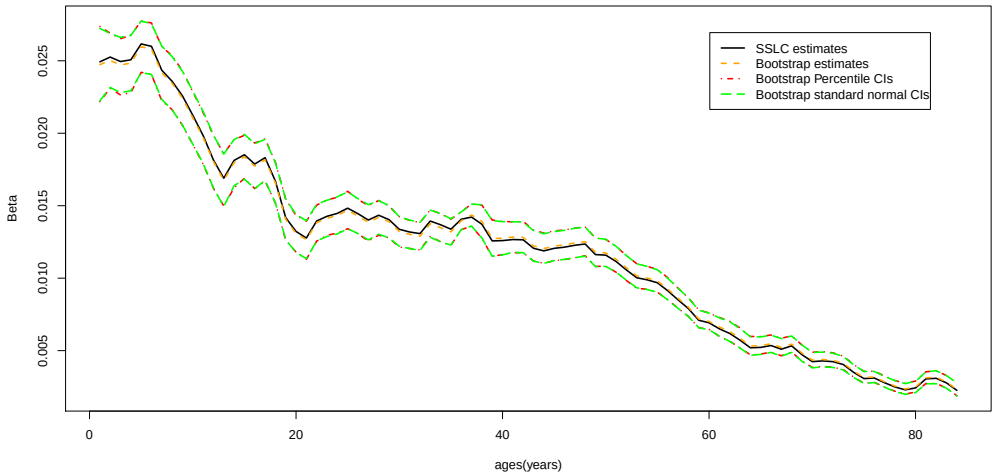


Figure 10.48: Cancer: $\hat{\beta}_a, \hat{\beta}_a^{(*)} : a = 1, ..., 84$ and corresponding 95% bootstrap pointwise confidence intervals.

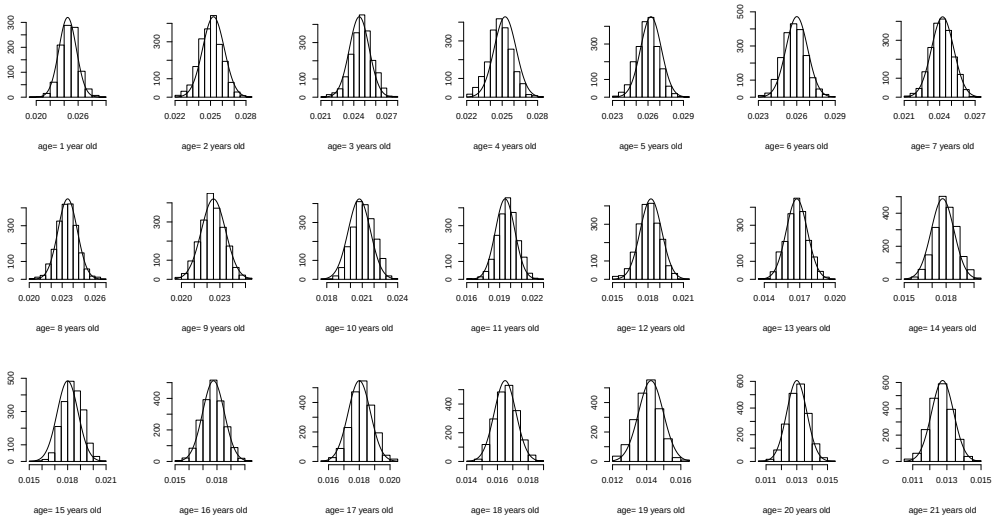


Figure 10.49: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 1, ..., 21$.

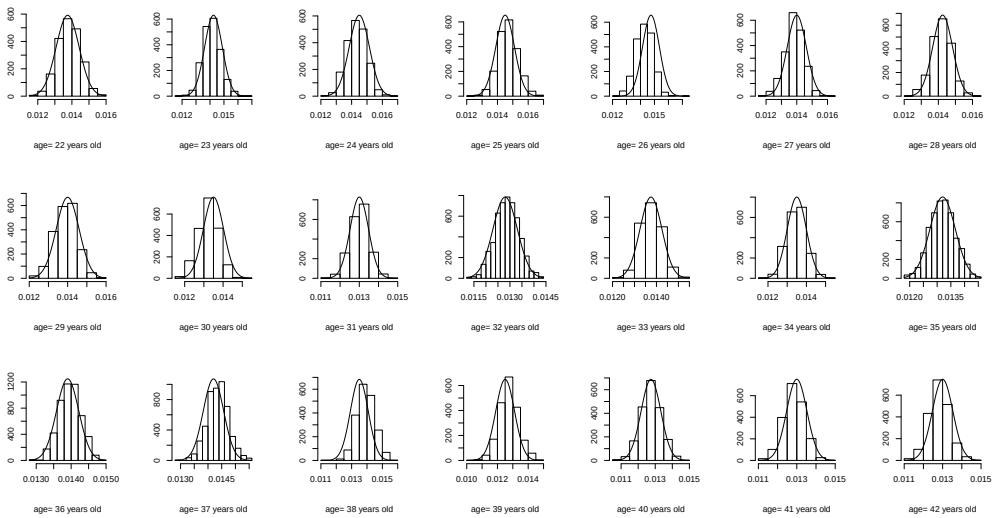


Figure 10.50: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 22, ..., 42$.

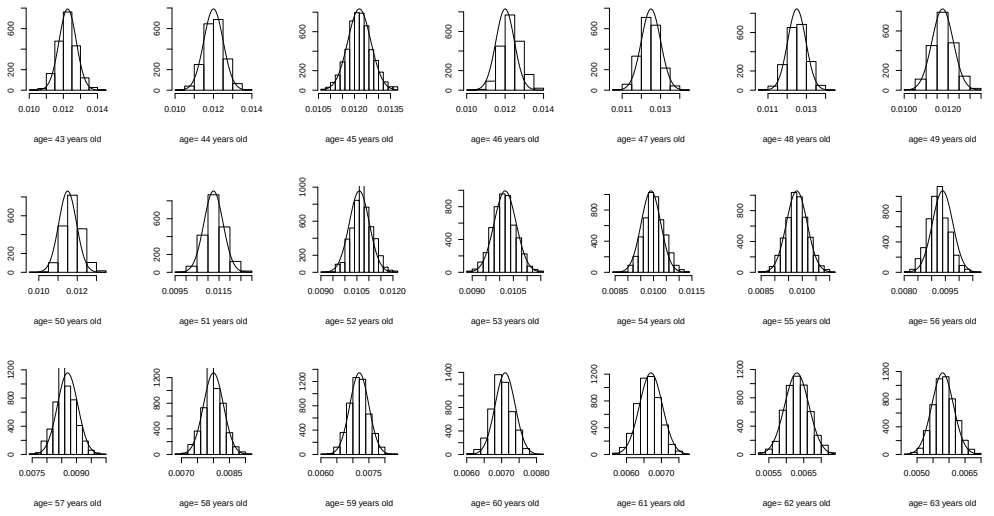


Figure 10.51: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 43, ..., 63$.

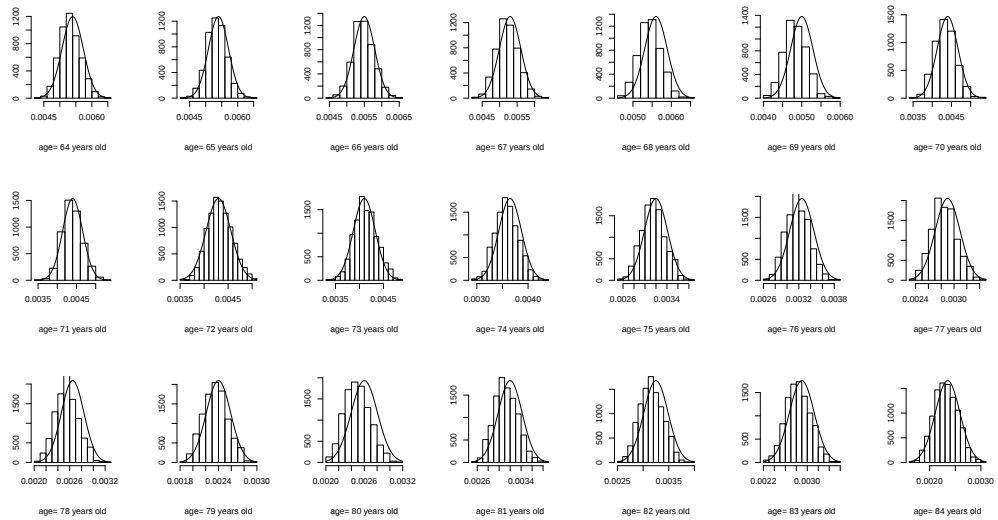


Figure 10.52: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 64, ..., 84$.

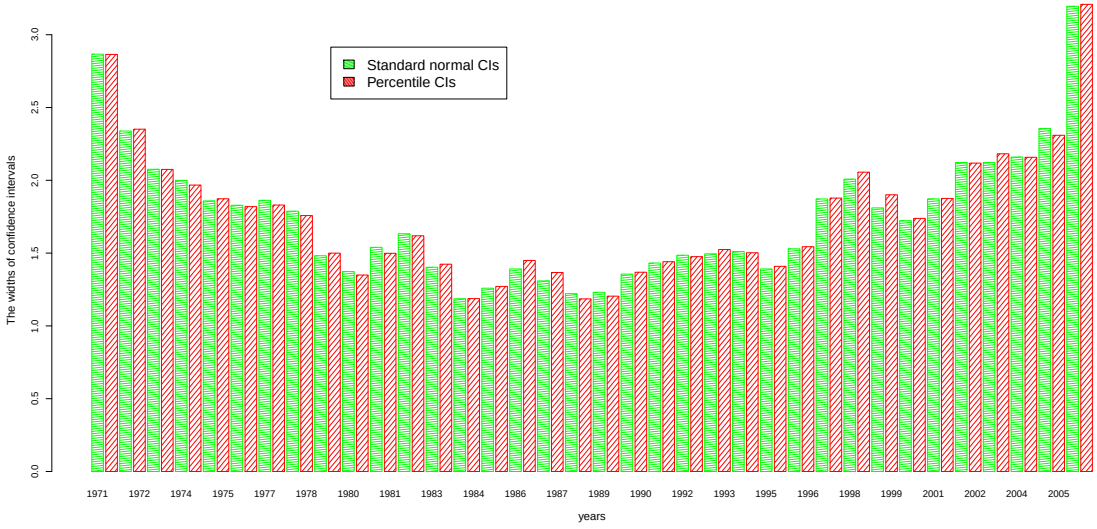


Figure 10.53: Cancer: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,1}$: $p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

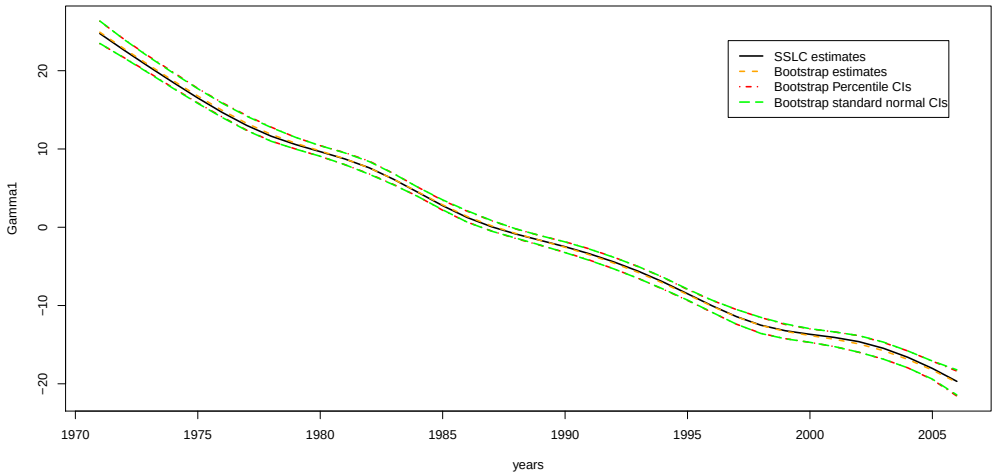


Figure 10.54: Cancer: $\hat{\gamma}_{p,1}, \hat{\gamma}_{p,1}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

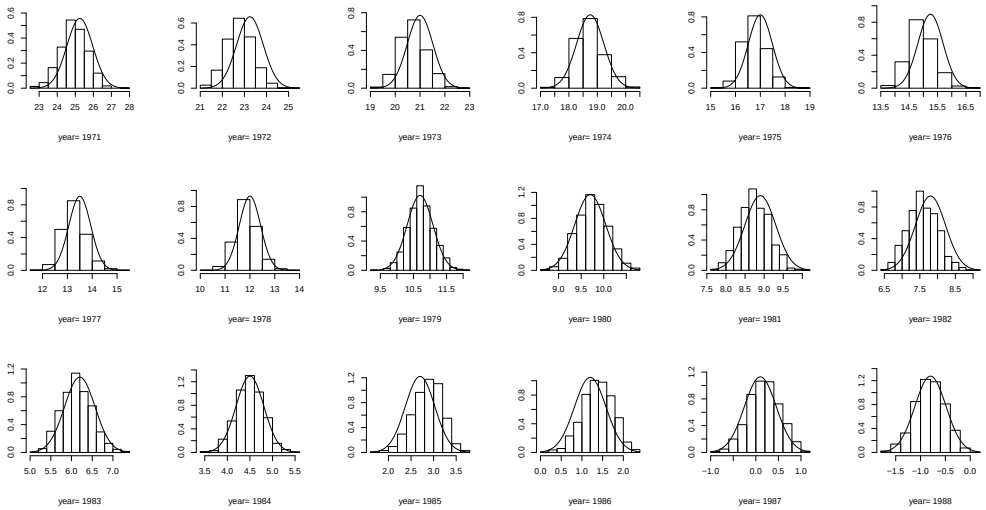


Figure 10.55: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,1} : p = 1971, ..., 1988$.

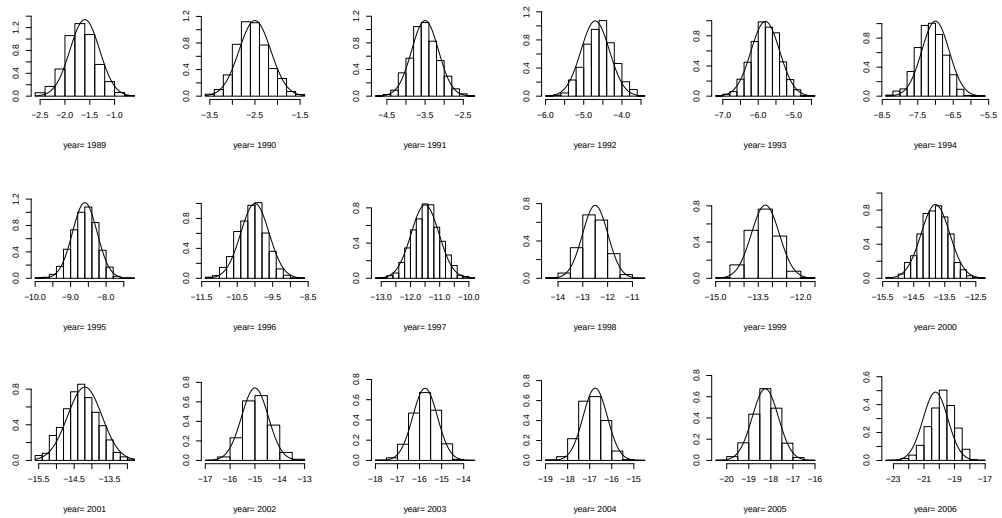


Figure 10.56: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,1} : p = 1989, ..., 2006$.

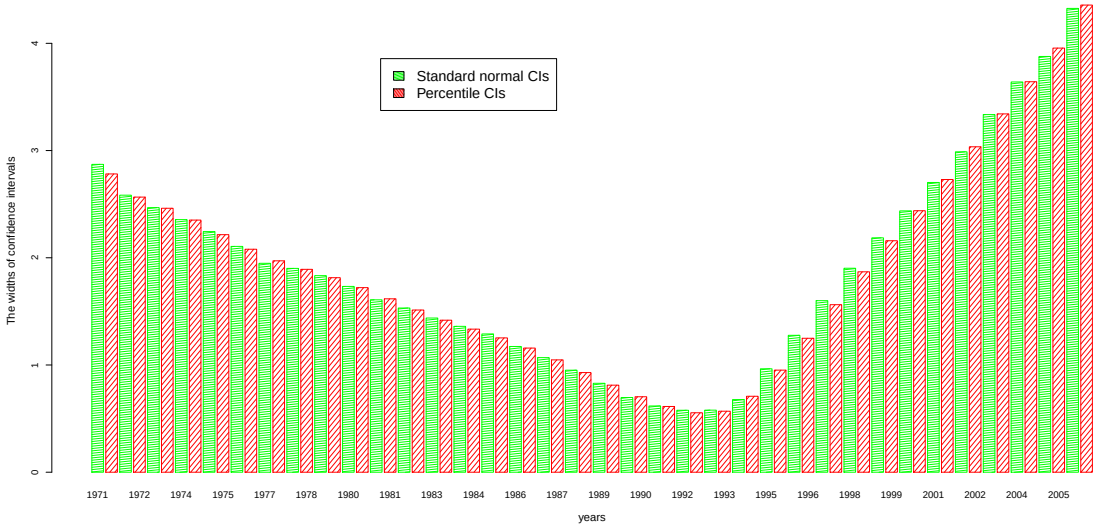


Figure 10.57: Cancer: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,2}$: $p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

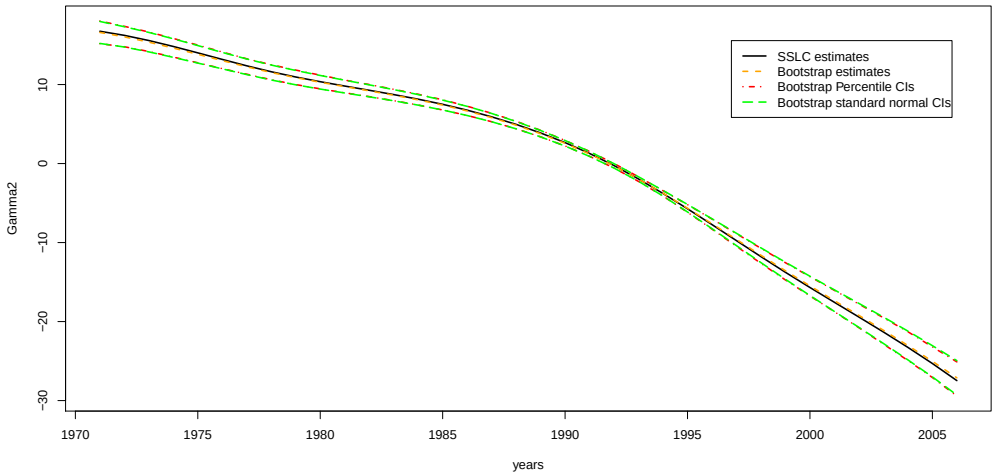


Figure 10.58: Cancer: $\hat{\gamma}_{p,2}, \hat{\gamma}_{p,2}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

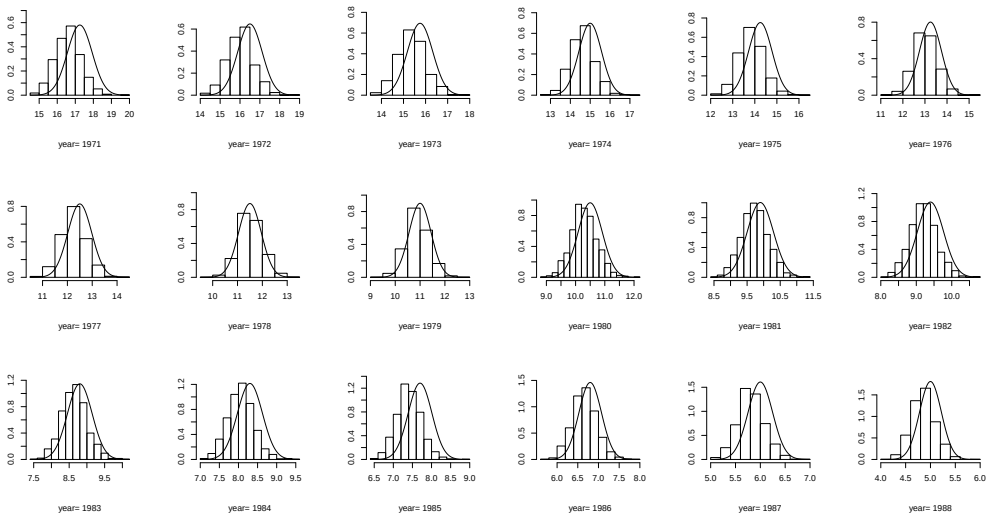


Figure 10.59: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,2} : p = 1971, ..., 1988$.

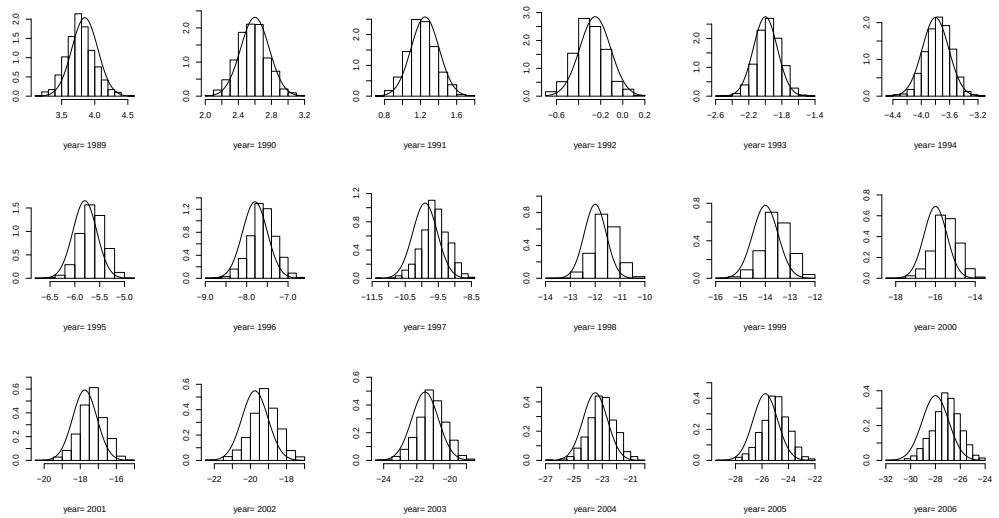


Figure 10.60: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,2} : p = 1989, ..., 2006$.

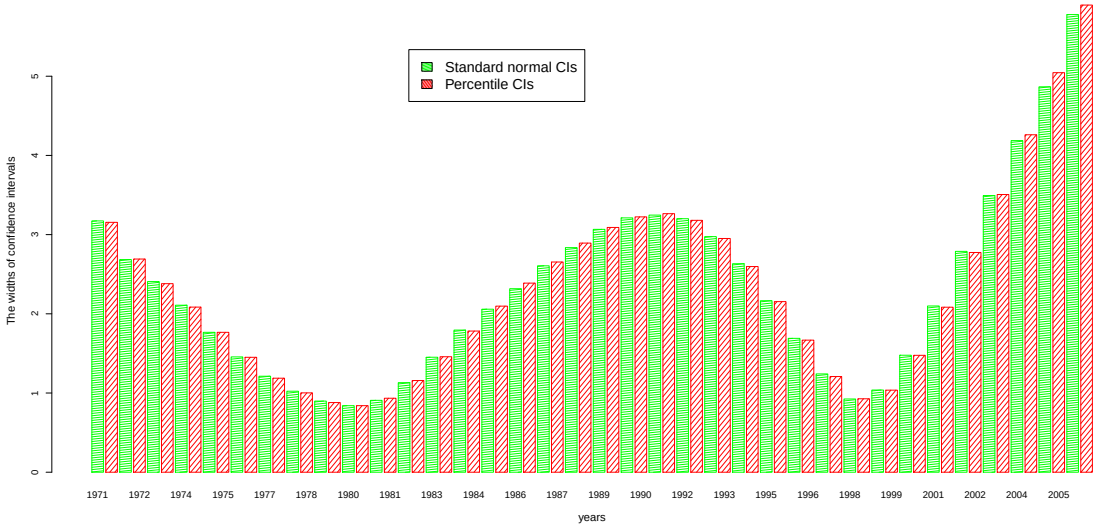


Figure 10.61: Cancer: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,3}$: $p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

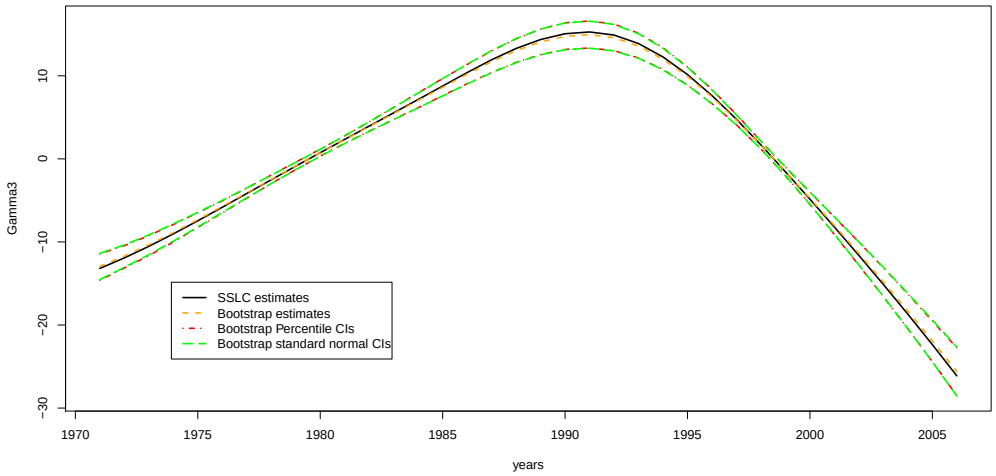


Figure 10.62: Cancer: $\hat{\gamma}_{p,3}, \hat{\gamma}_{p,3}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

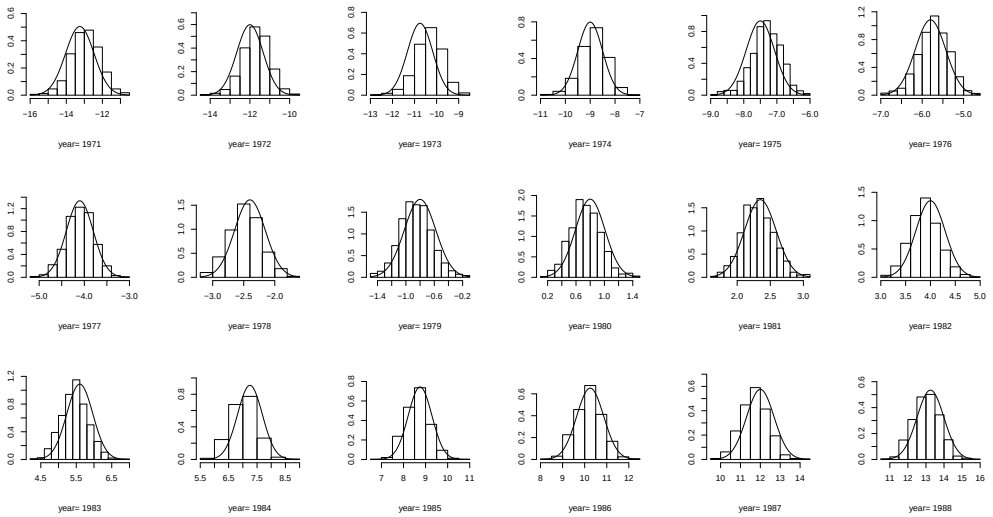


Figure 10.63: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,3} : p = 1971, ..., 1988$.

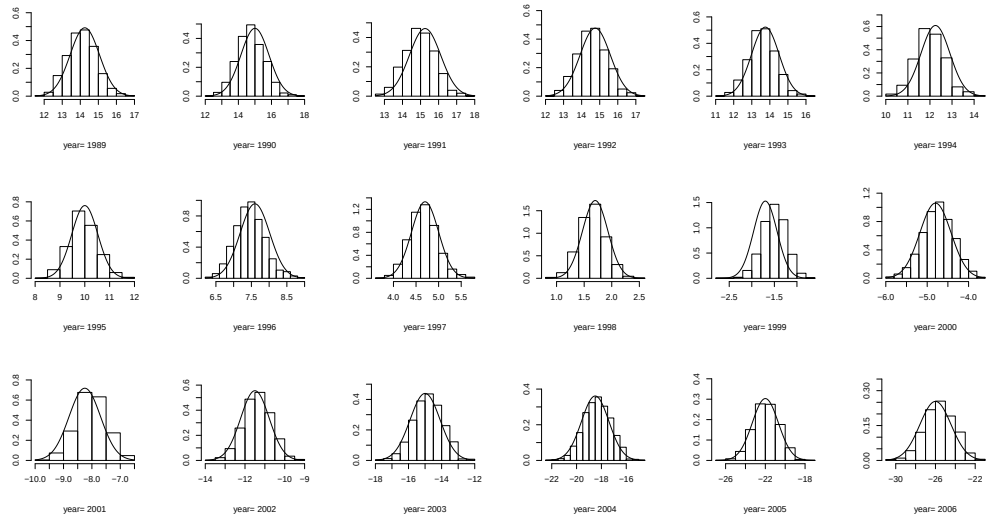


Figure 10.64: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,3} : p = 1989, ..., 2006$.

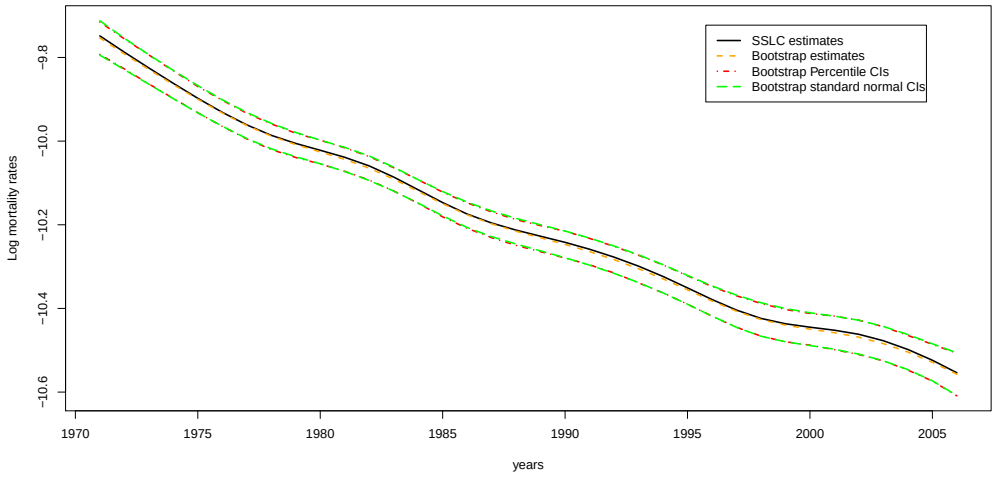


Figure 10.65: Cancer: Log mortality rate estimates at age 14 years and 95% bootstrap pointwise confidence intervals.

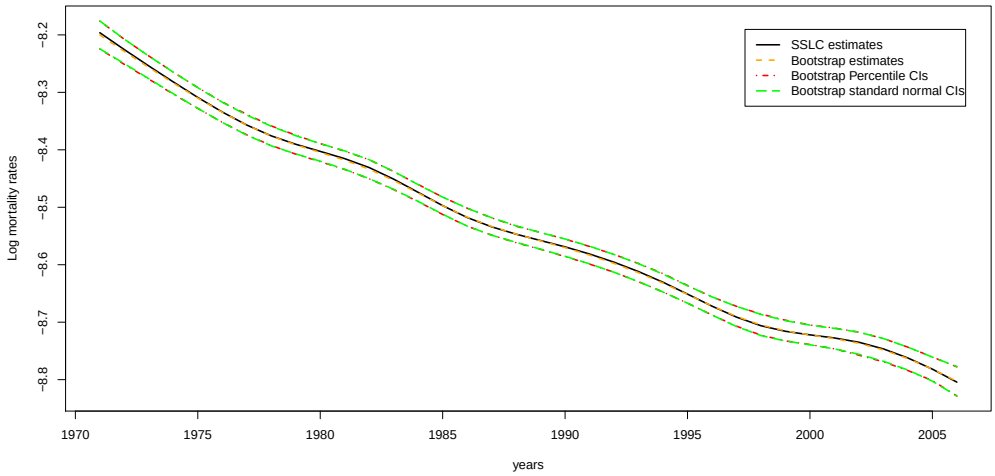


Figure 10.66: Cancer: Log mortality rate estimates at age 34 years and 95% bootstrap pointwise confidence intervals.

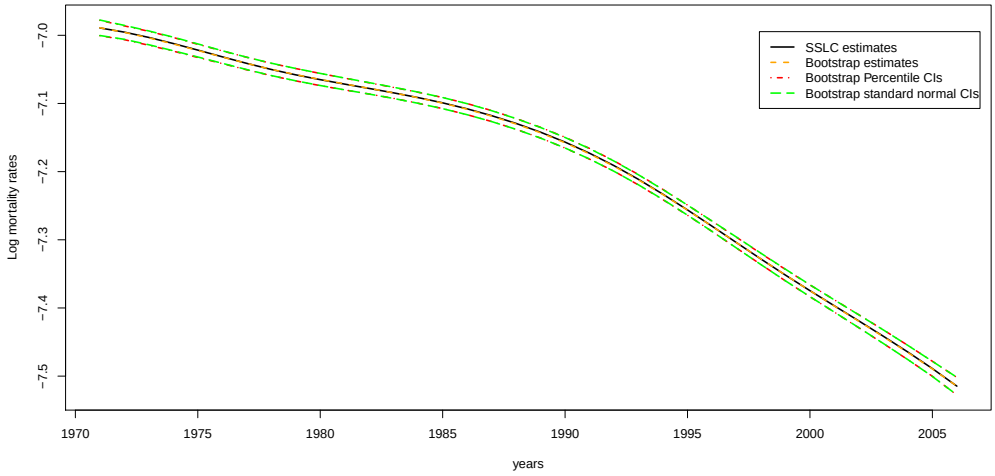


Figure 10.67: Cancer: Log mortality rate estimates at age 44 years and 95% bootstrap pointwise confidence intervals.

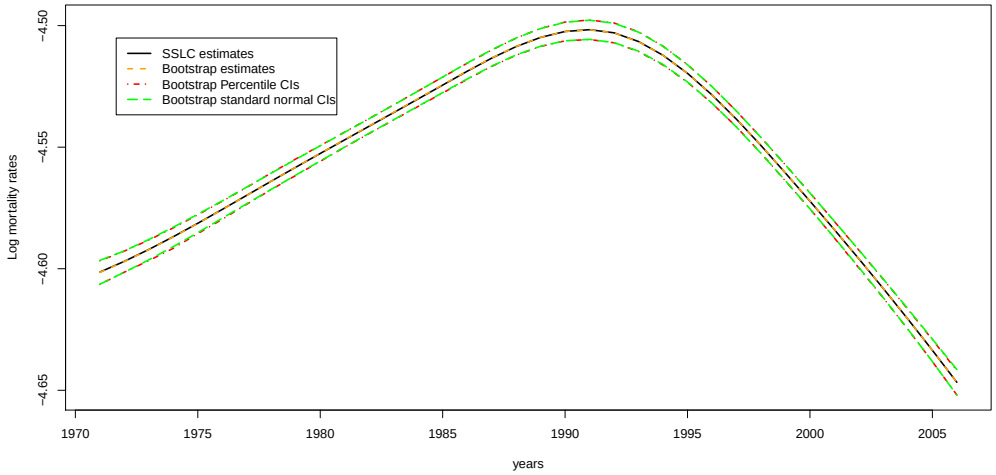


Figure 10.68: Cancer: Log mortality rate estimates at age 74 years and 95% bootstrap pointwise confidence intervals.

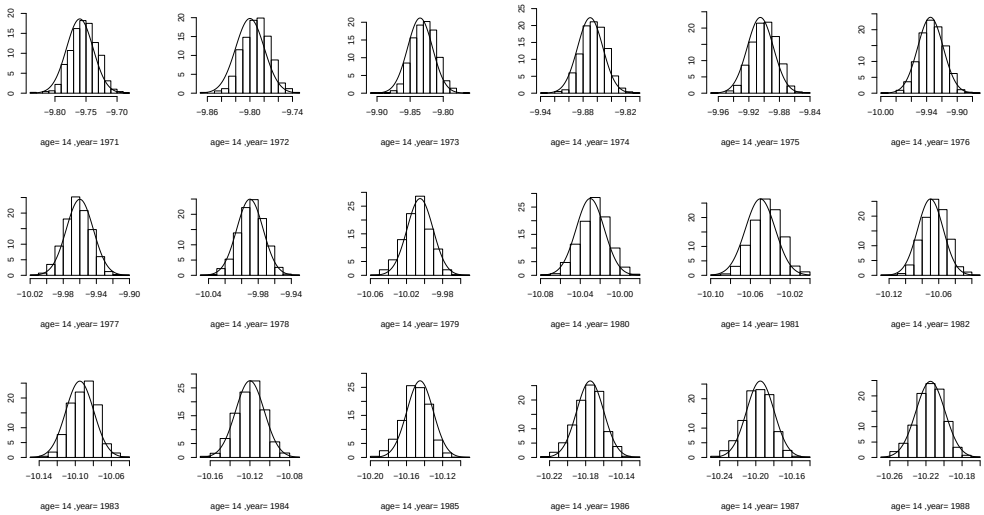


Figure 10.69: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\log(\hat{\lambda}_{14,p}) : p = 1971, ..., 1988$.

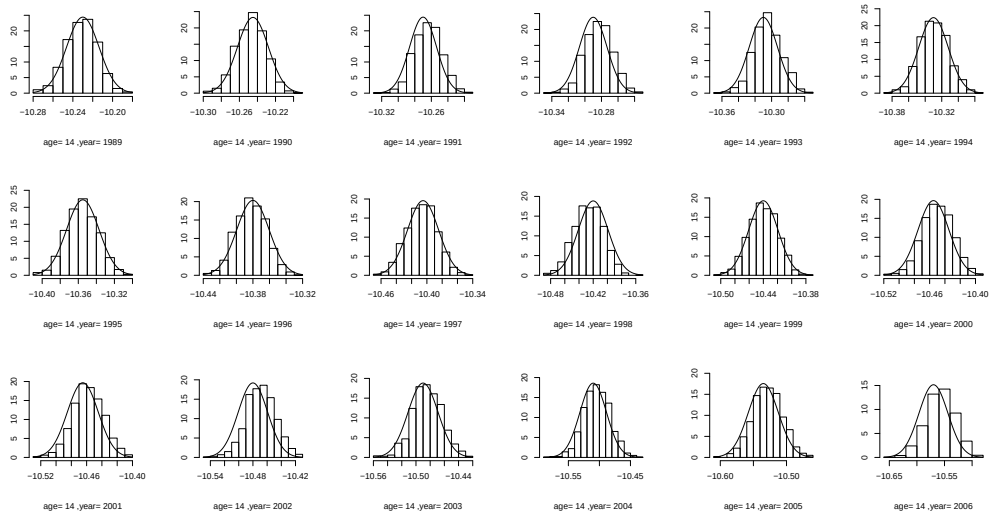


Figure 10.70: Cancer: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\log(\hat{\lambda}_{14,p}) : p = 1989, ..., 2006$.

Accidents

Figures 10.71-10.101 show histograms of bootstrapped samples and graphical comparisons of 95% bootstrap pointwise confidence intervals and 95% bootstrap pointwise confidence interval widths between the two types of confidence intervals for parameter estimates. The figures suggest that the two types of intervals coincide in most cases. Figure 10.97 shows small deviations of histograms from normal curves of $\hat{\gamma}_{p,3}$ at middle periods (years 1991-1995). Figures 10.102-10.105 show the corresponding comparisons for log mortality rate estimates at some selected ages. Figures 10.106-10.107 demonstrate histograms of bootstrapped samples of log mortality rates at a selected age, 14 years. The figures suggest that the bootstrapped samples of log mortality rate estimates follow normal distributions.

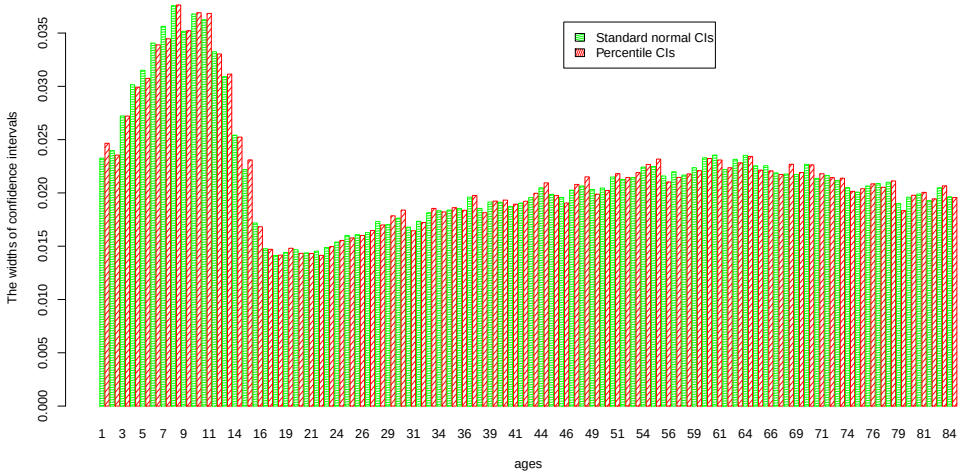


Figure 10.71: Accidents: 95% bootstrap pointwise confidence interval widths of $\alpha_a : a = 1, \ldots, 84$ obtained from percentile and standard normal intervals.

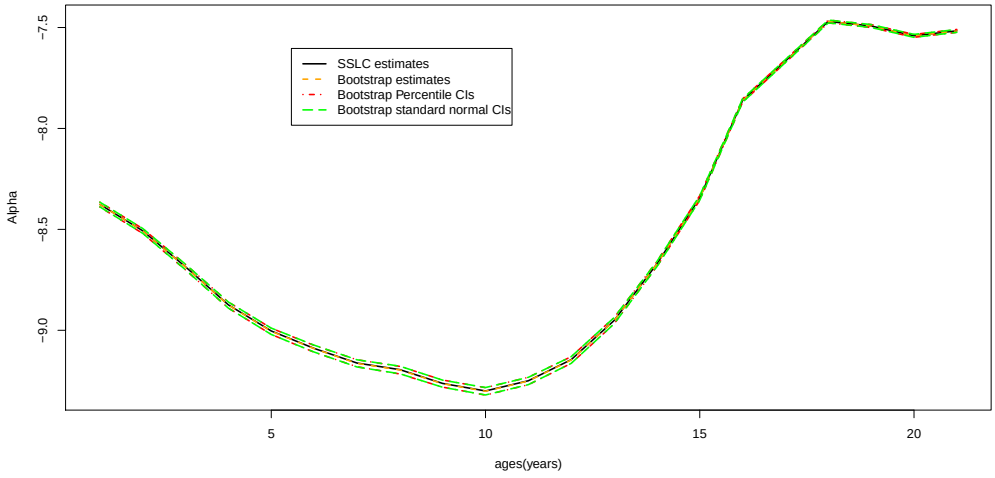


Figure 10.72: Accidents: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 1, \ldots, 21$ and corresponding 95% bootstrap pointwise confidence intervals.

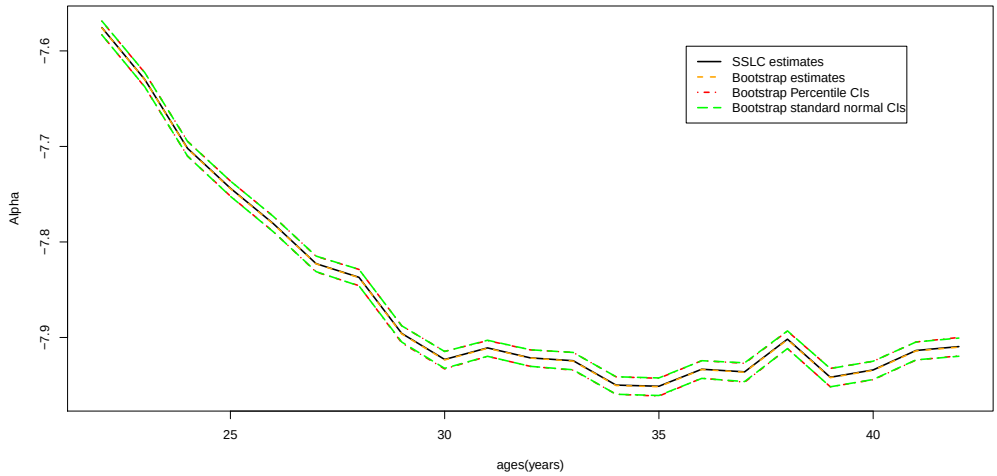


Figure 10.73: Accidents: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 22, \ldots, 42$ and corresponding 95% bootstrap pointwise confidence intervals.

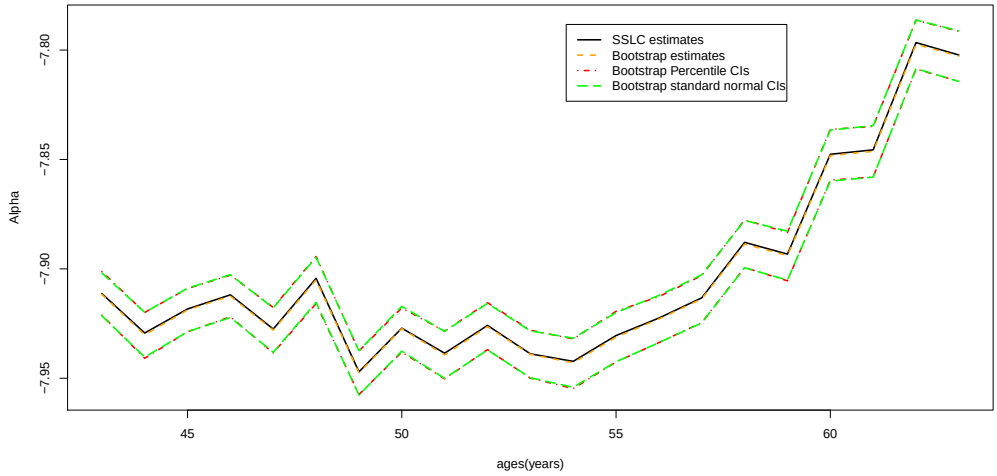


Figure 10.74: Accidents: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 43, \ldots, 63$ and corresponding 95% bootstrap pointwise confidence intervals.

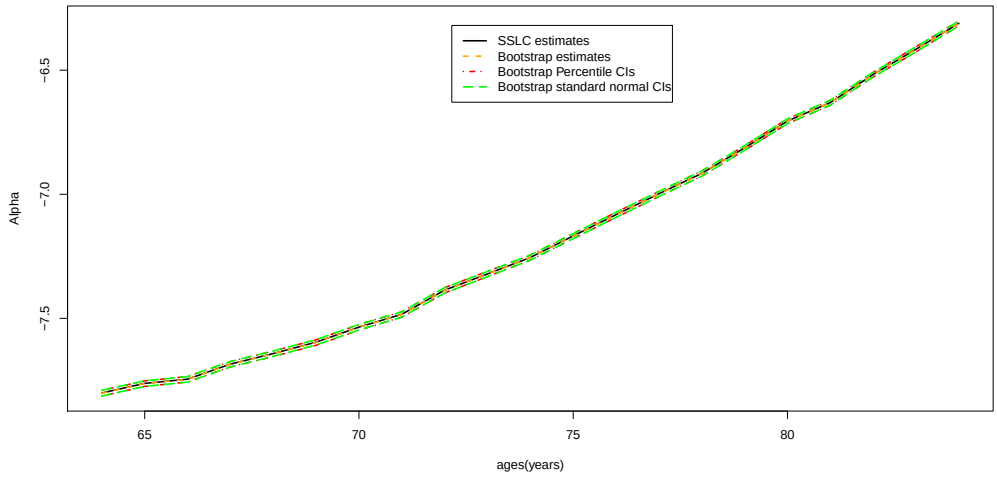


Figure 10.75: Accidents: $\hat{\alpha}_a, \hat{\alpha}_a^{(*)} : a = 64, \ldots, 84$ and corresponding 95% bootstrap pointwise confidence intervals.

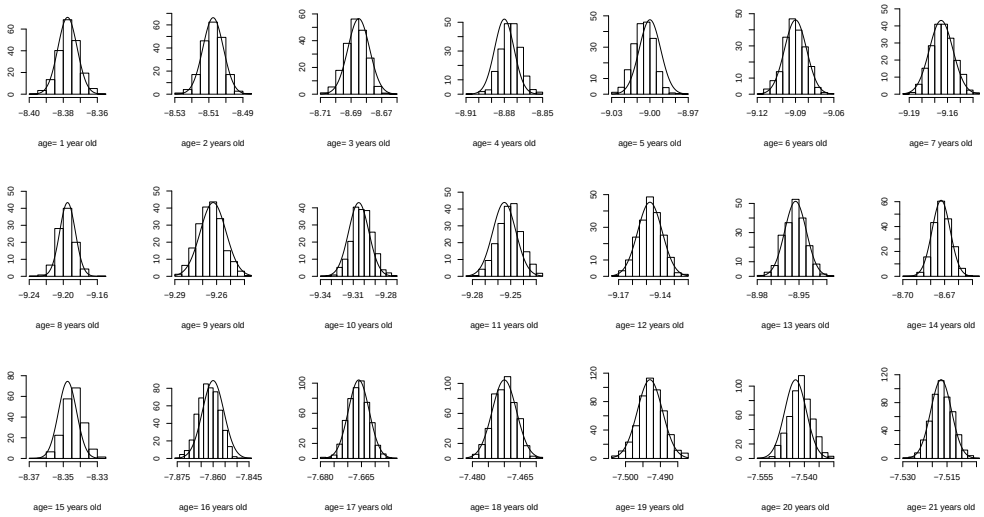


Figure 10.76: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 1, ..., 21$.

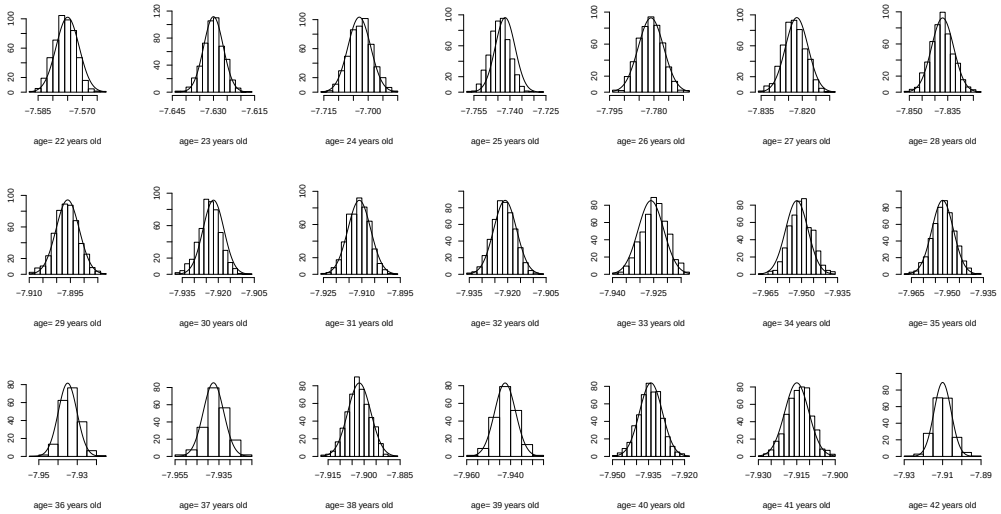


Figure 10.77: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 22, ..., 42$.

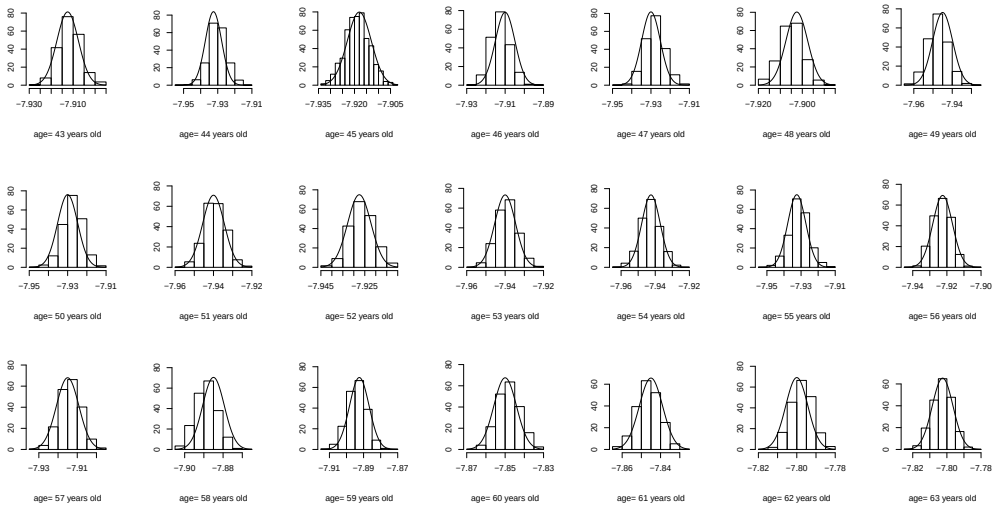


Figure 10.78: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 43, ..., 63$.

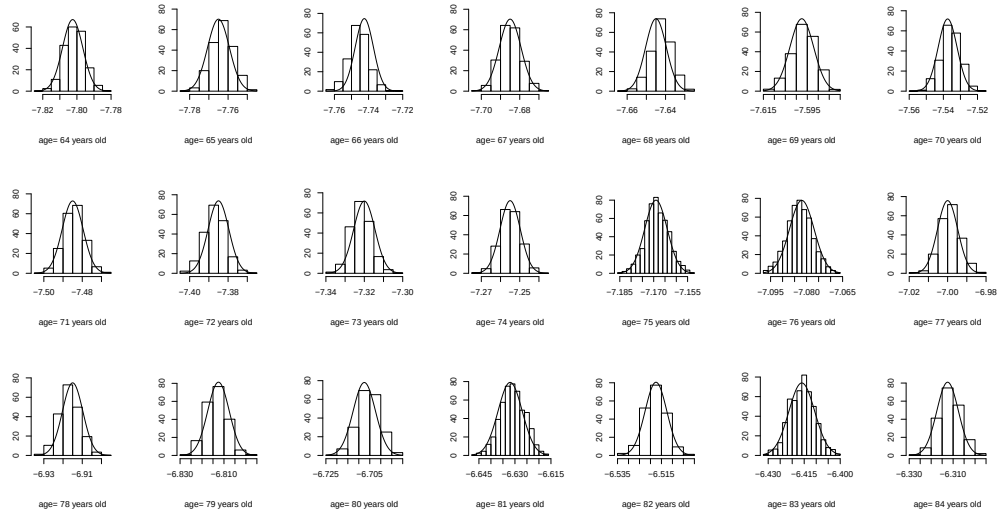


Figure 10.79: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\alpha}_a : a = 64, ..., 84$.

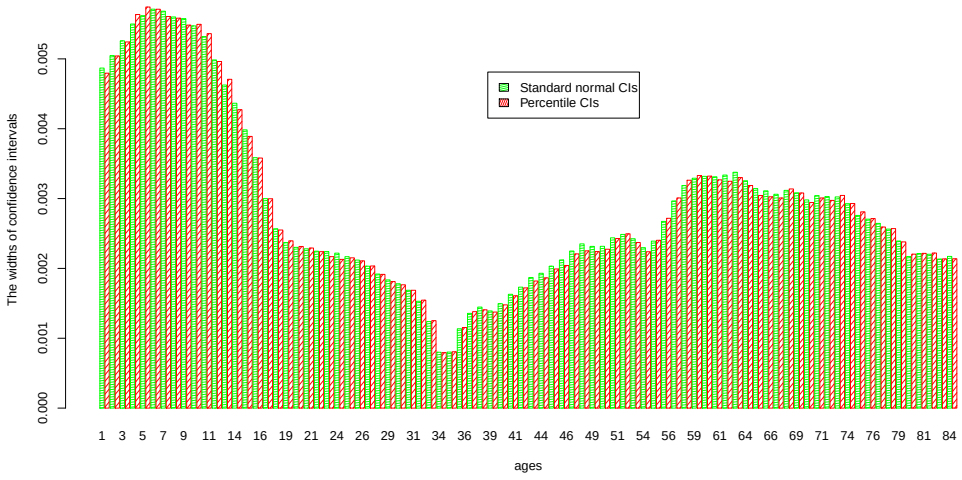


Figure 10.80: Accidents: 95% bootstrap pointwise confidence interval widths of $\beta_a : a = 1, \ldots, 84$ obtained from percentile and standard normal intervals.

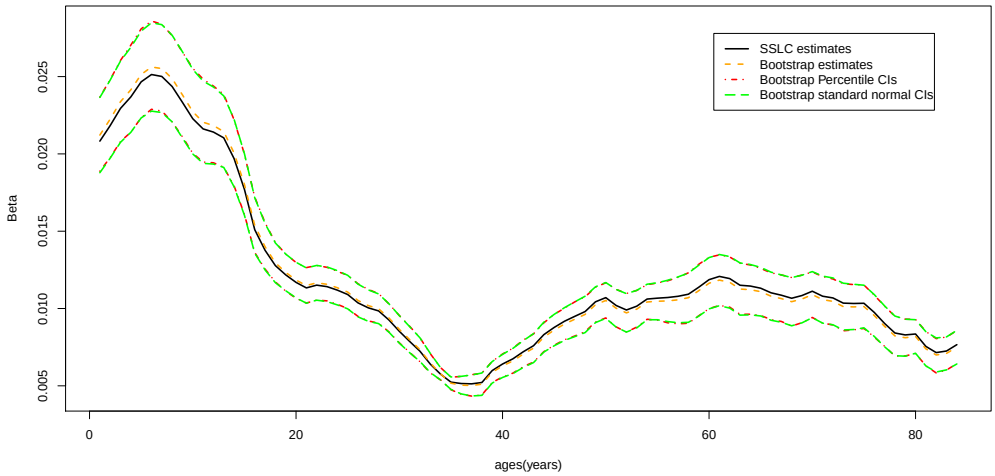


Figure 10.81: Accidents: $\hat{\beta}_a, \hat{\beta}_a^{(*)} : a = 1, \ldots, 84$ and corresponding 95% bootstrap pointwise confidence intervals.

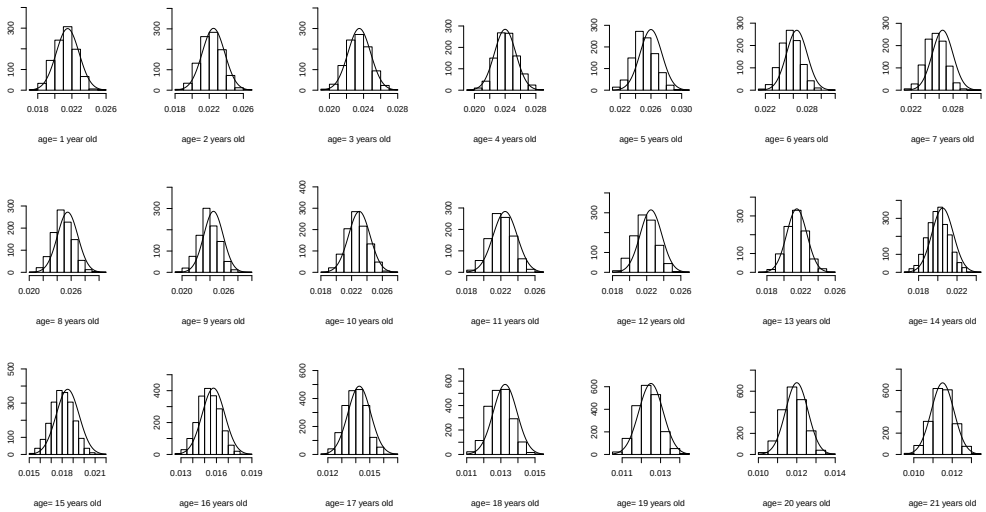


Figure 10.82: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 1, ..., 21$.

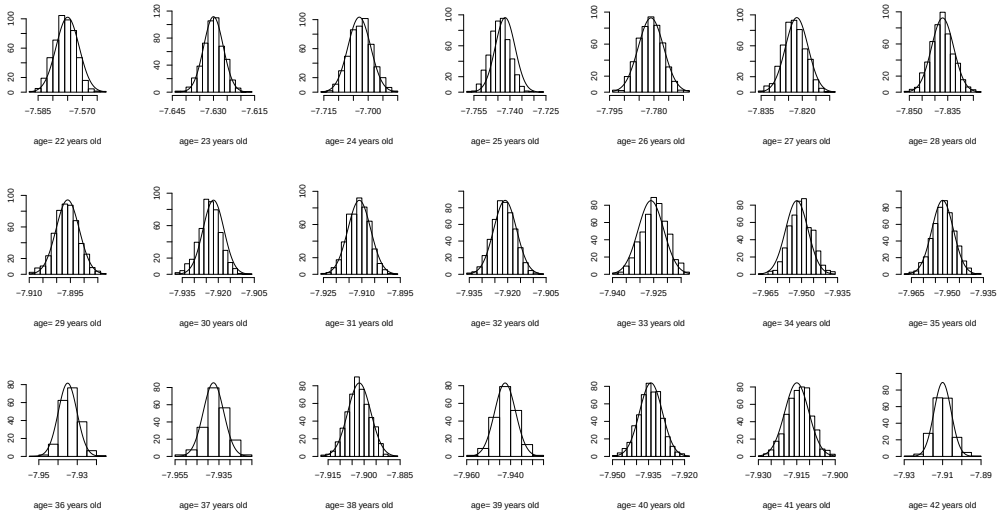


Figure 10.83: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 22, ..., 42$.

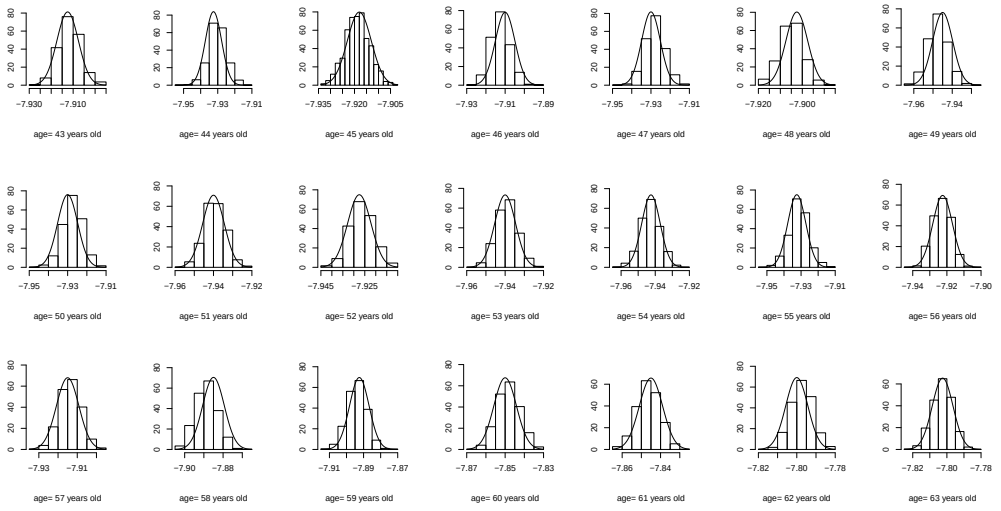


Figure 10.84: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 43, ..., 63$.

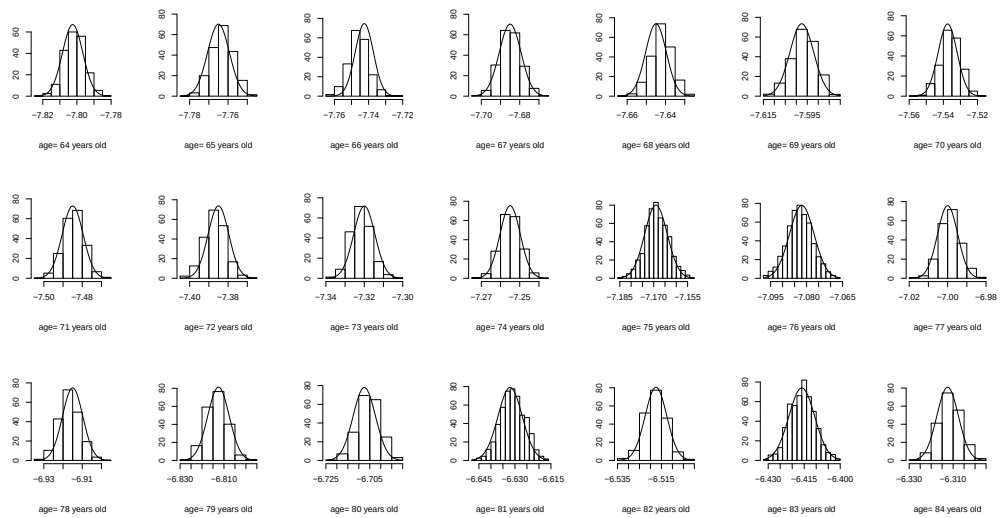


Figure 10.85: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\beta}_a : a = 64, ..., 84$.

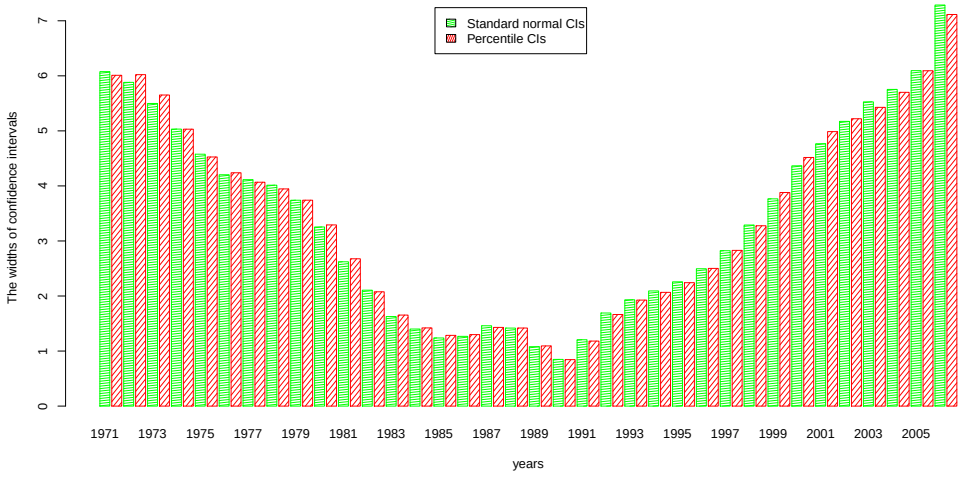


Figure 10.86: Accidents: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,1} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

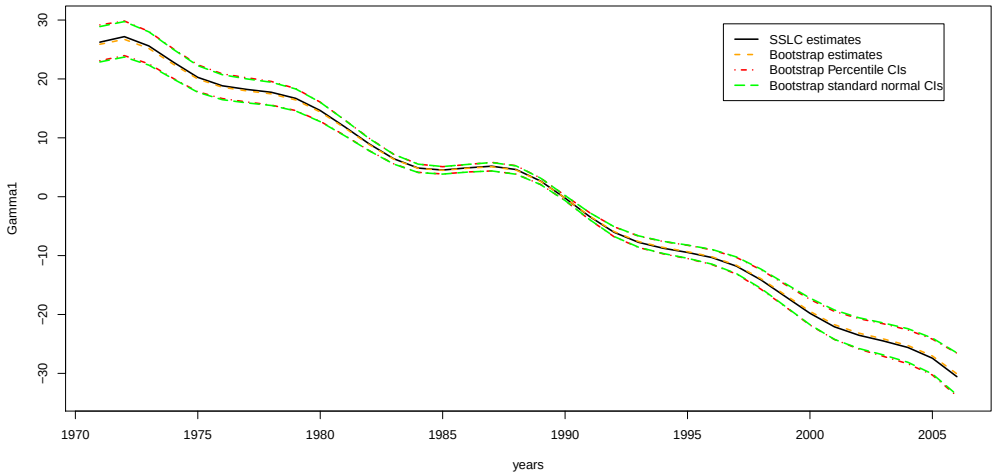


Figure 10.87: Accidents: $\hat{\gamma}_{p,1}, \hat{\gamma}_{p,1}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

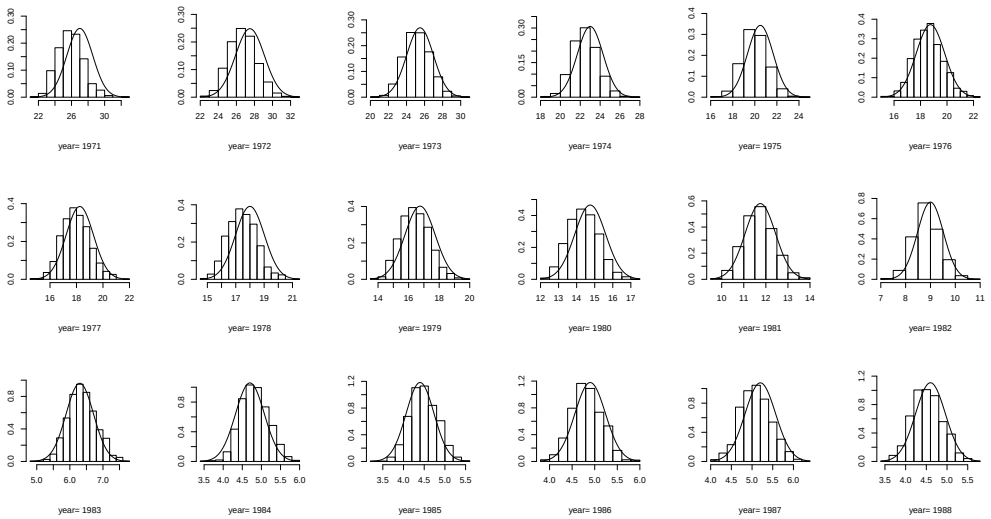


Figure 10.88: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,1} : p = 1971, ..., 1988$.

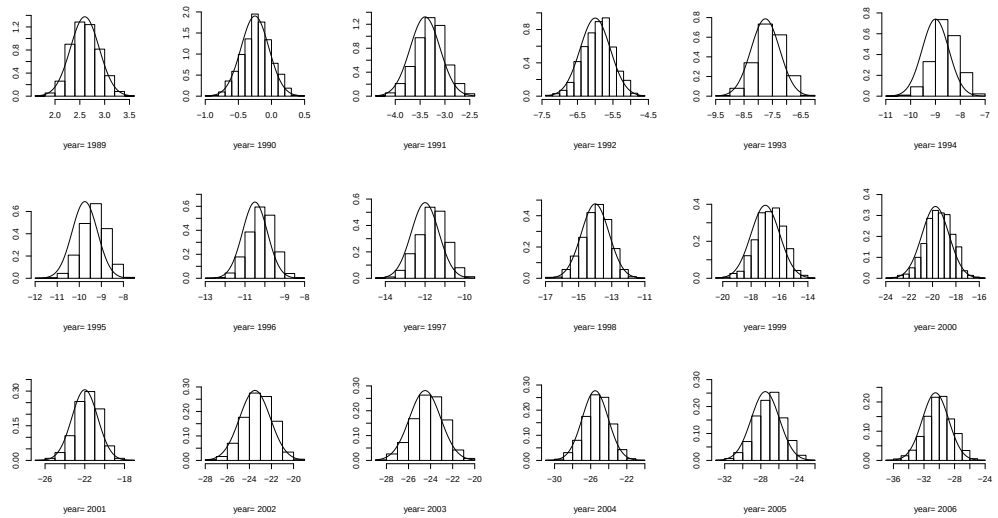


Figure 10.89: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,1} : p = 1989, ..., 2006$.

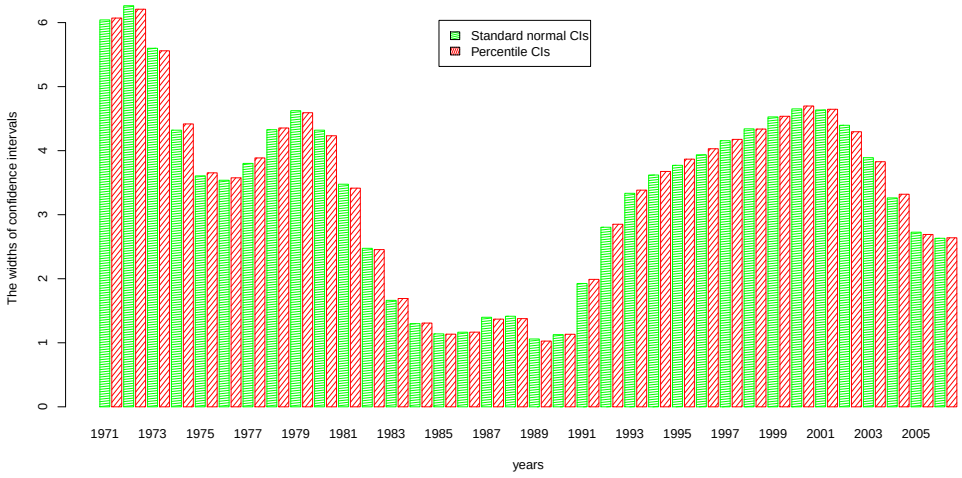


Figure 10.90: Accidents: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,2} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

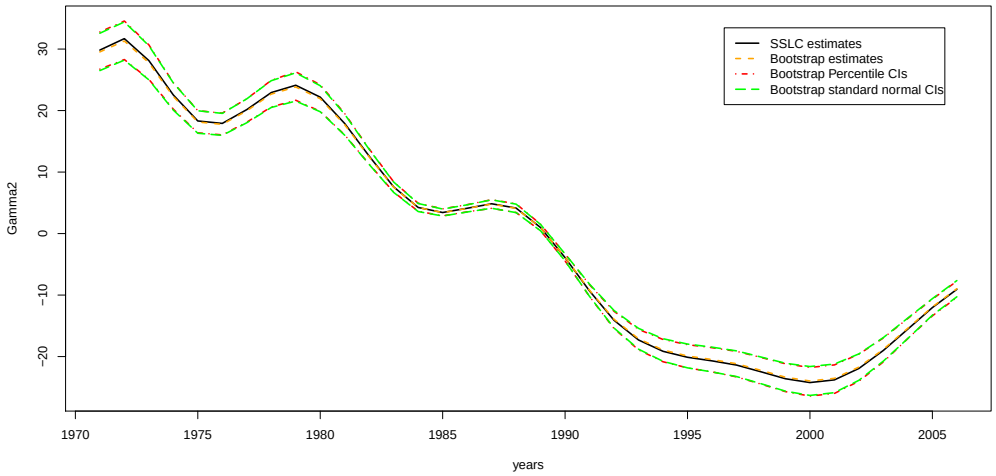


Figure 10.91: Accidents: $\hat{\gamma}_{p,2}, \hat{\gamma}_{p,2}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

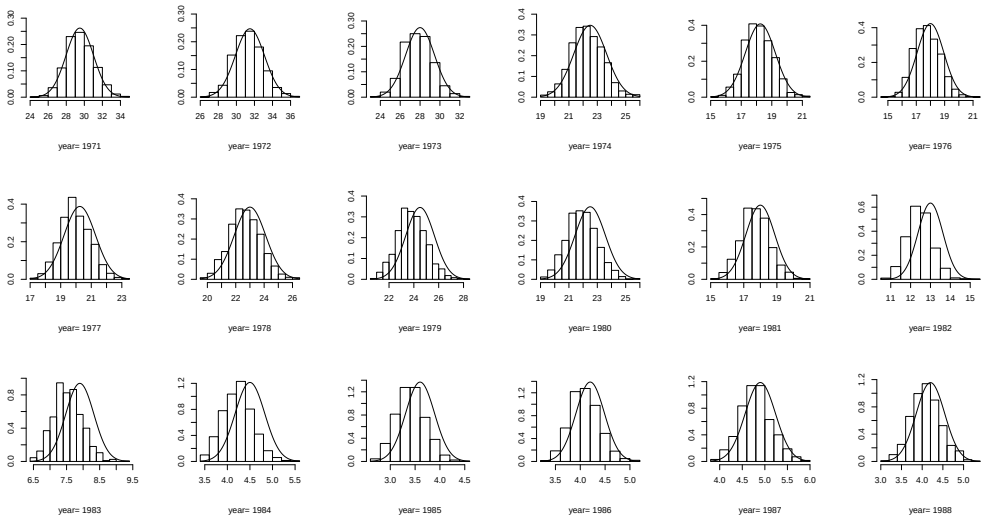


Figure 10.92: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,2} : p = 1971, ..., 1988$.

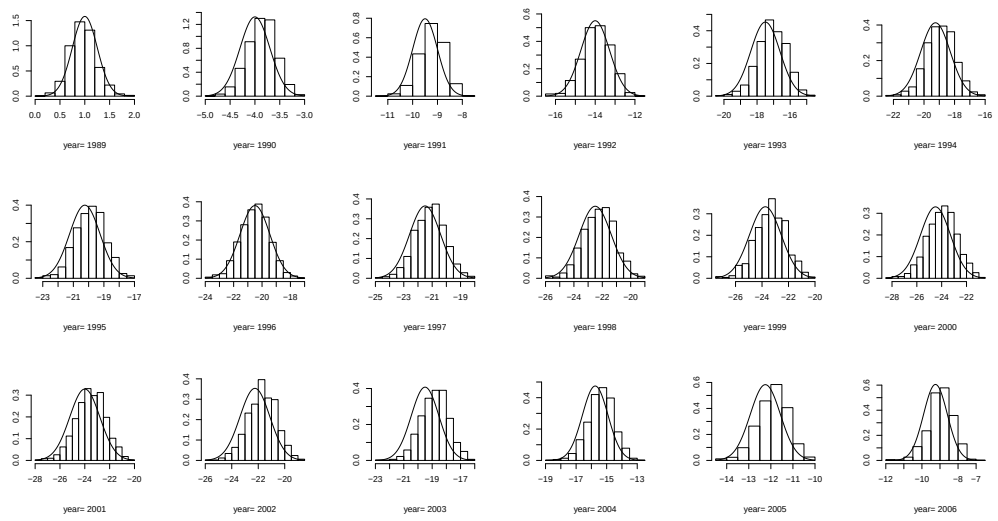


Figure 10.93: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,2} : p = 1989, ..., 2006$.

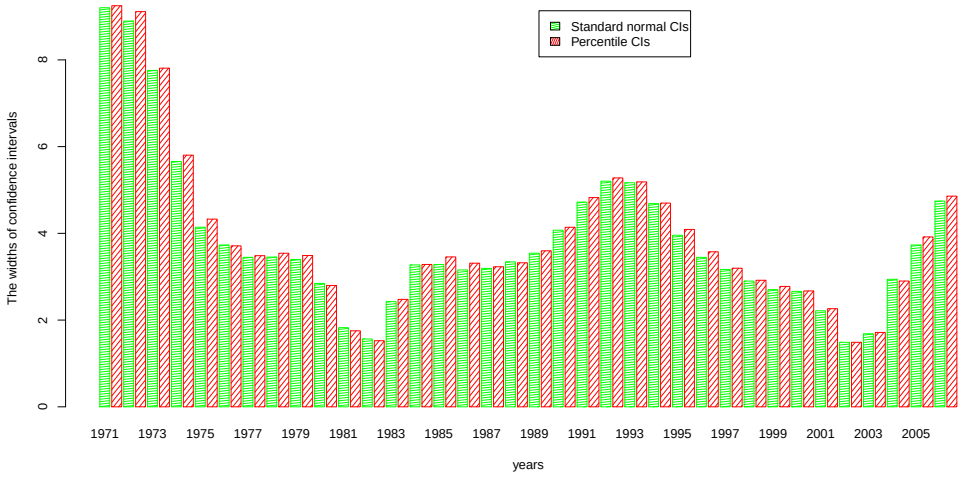


Figure 10.94: Accidents: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,3} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

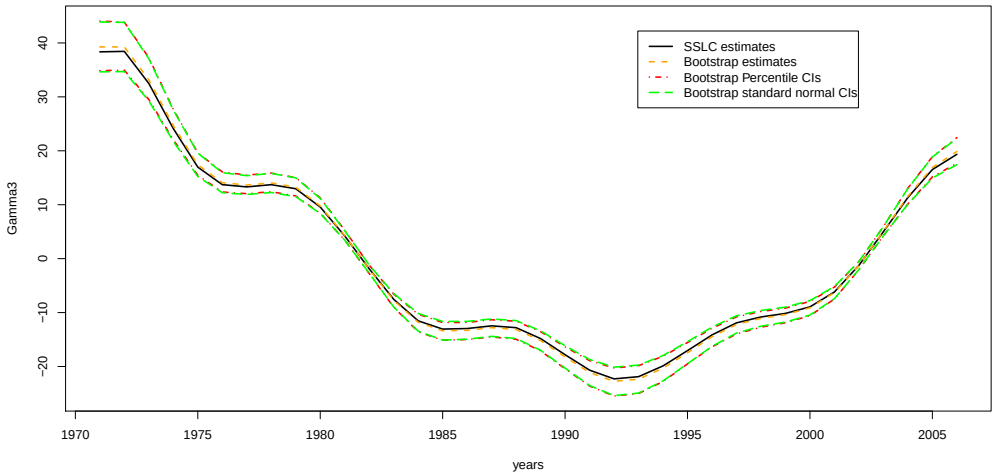


Figure 10.95: Accidents: $\hat{\gamma}_{p,3}, \hat{\gamma}_{p,3}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

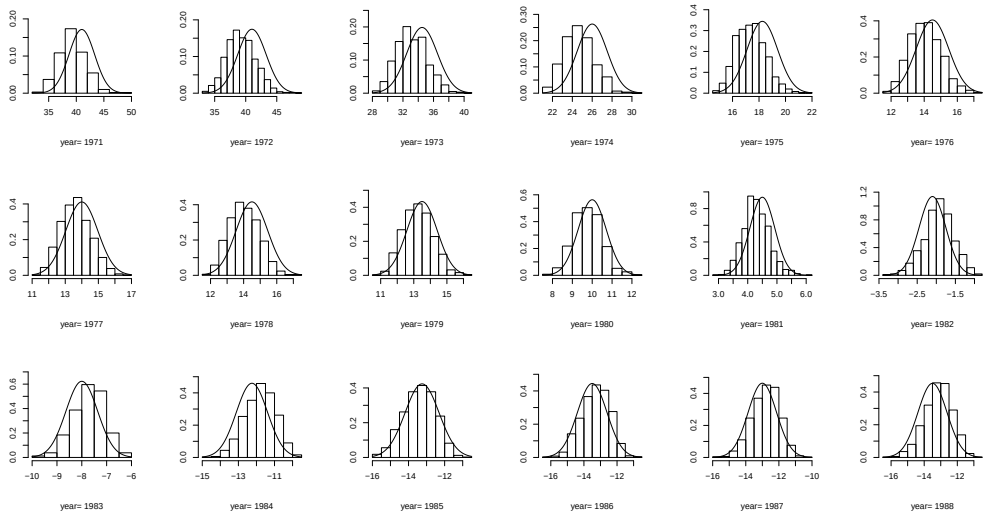


Figure 10.96: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,3} : p = 1971, ..., 1988$.

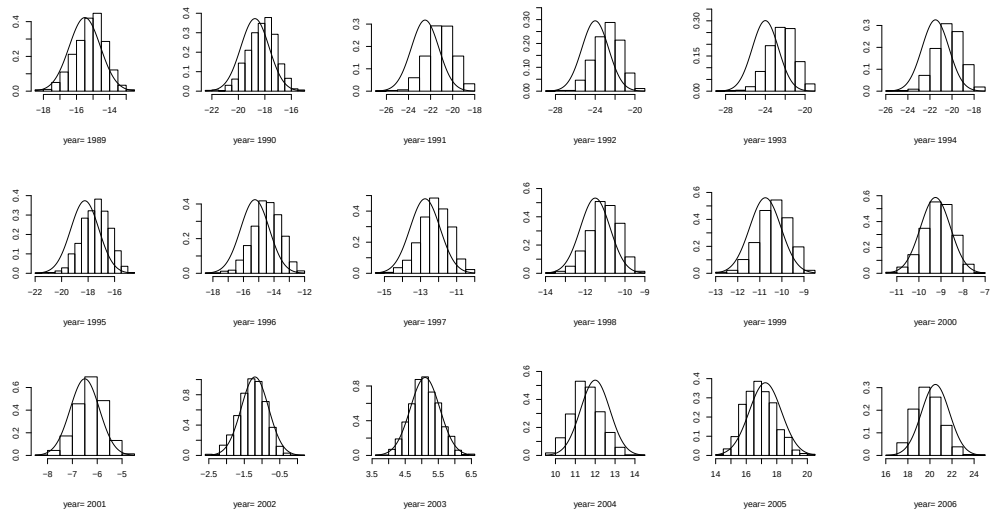


Figure 10.97: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,3} : p = 1989, ..., 2006$.

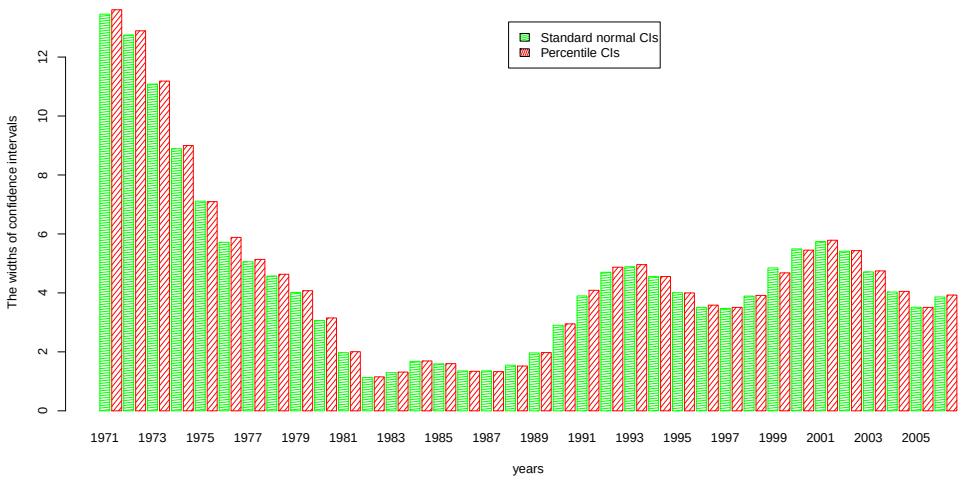


Figure 10.98: Accidents: 95% bootstrap pointwise confidence interval widths of $\gamma_{p,4} : p = 1971, ..., 2006$ obtained from percentile and standard normal intervals.

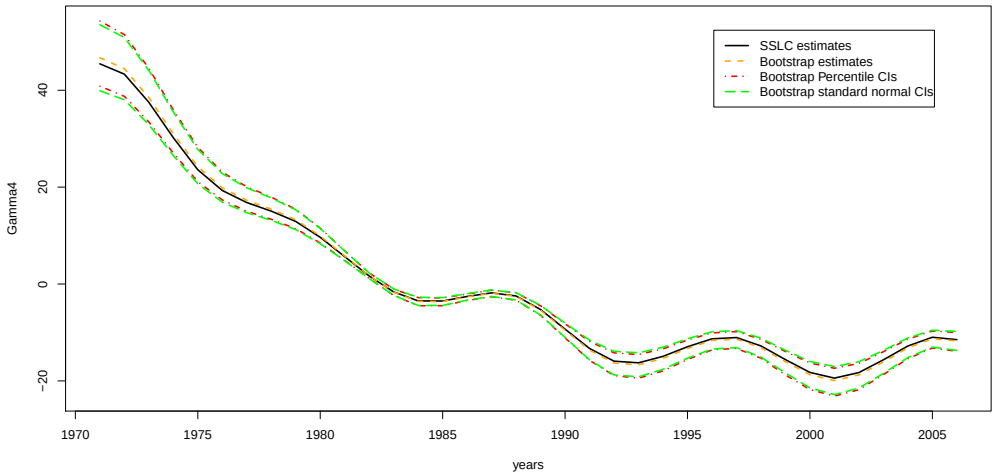


Figure 10.99: Accidents: $\hat{\gamma}_{p,4}, \hat{\gamma}_{p,4}^{(*)} : p = 1971, ..., 2006$ and corresponding 95% bootstrap pointwise confidence intervals.

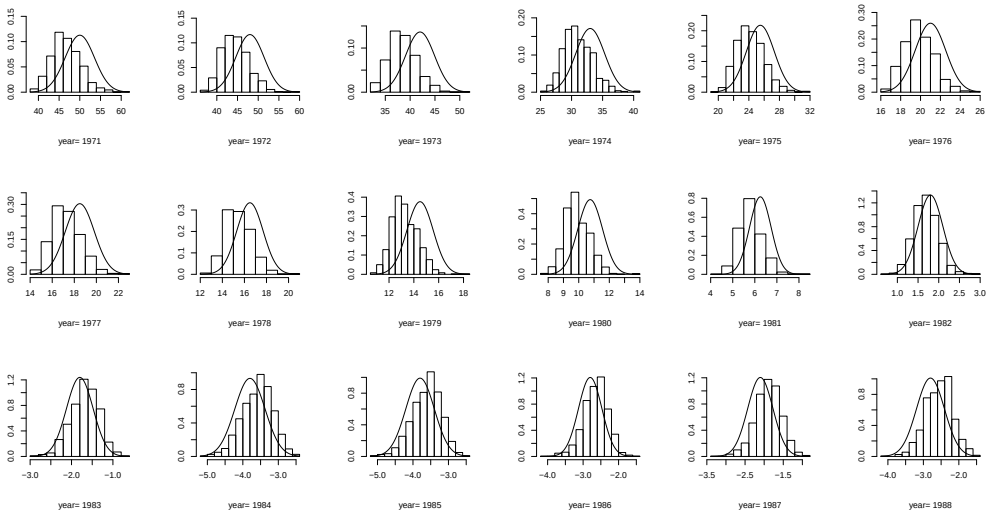


Figure 10.100: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,4} : p = 1971, ..., 1988$.

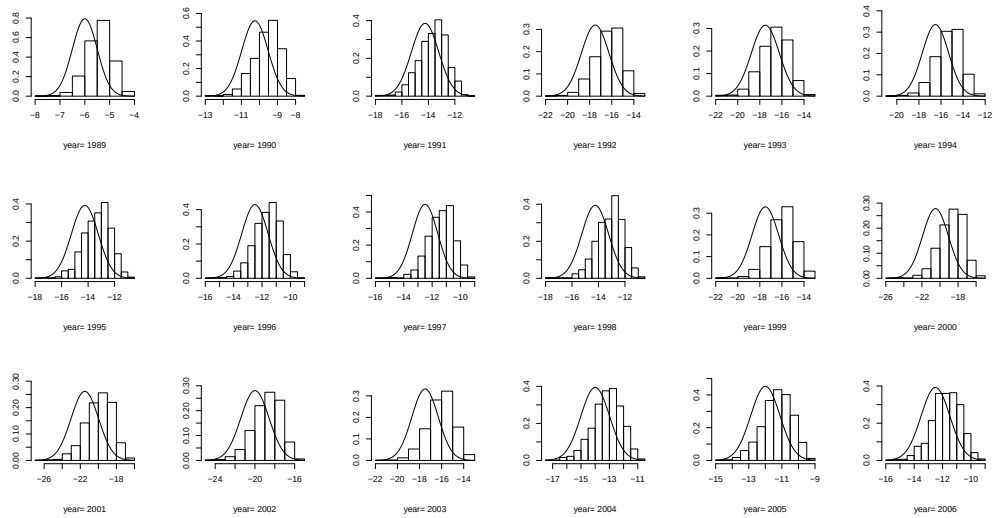


Figure 10.101: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\hat{\gamma}_{p,4} : p = 1989, ..., 2006$.

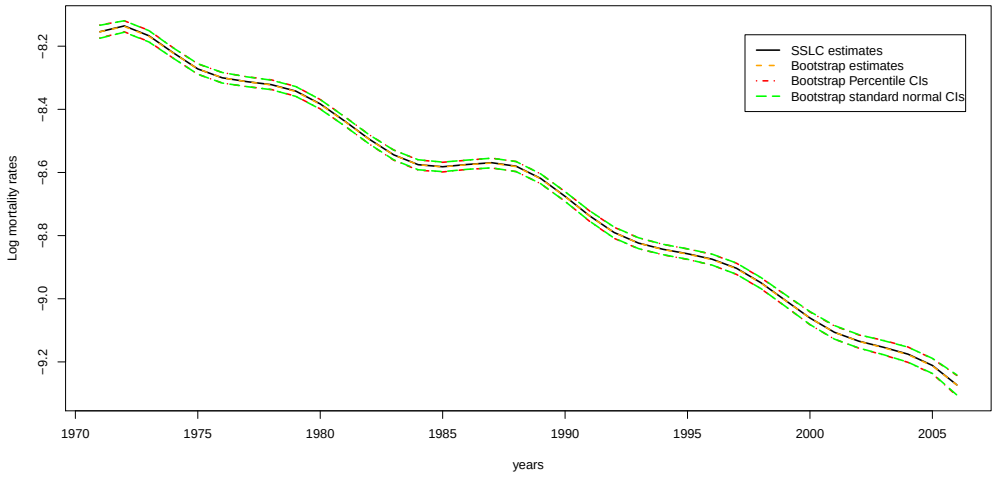


Figure 10.102: Accidents: Log mortality rate estimates at age 14 years and 95% bootstrap pointwise confidence intervals.

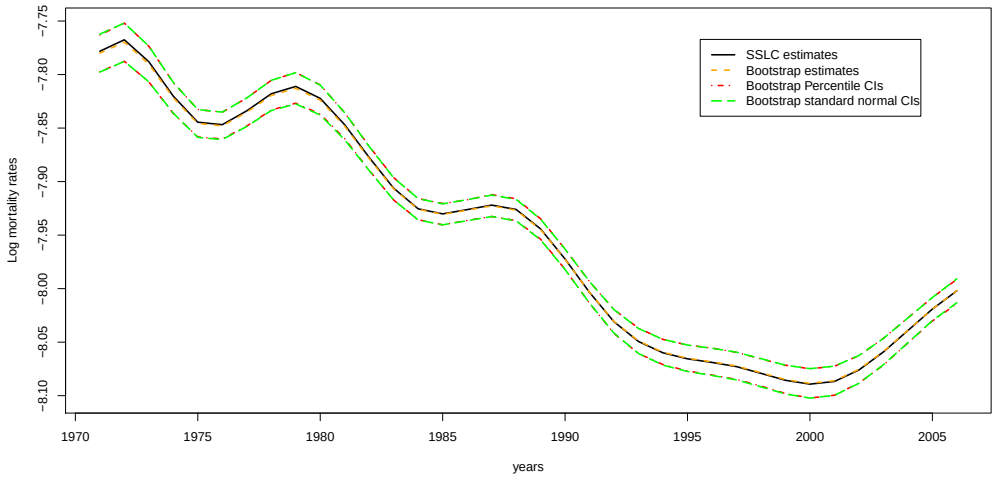


Figure 10.103: Accidents: Log mortality rate estimates at age 34 years and 95% bootstrap pointwise confidence intervals.

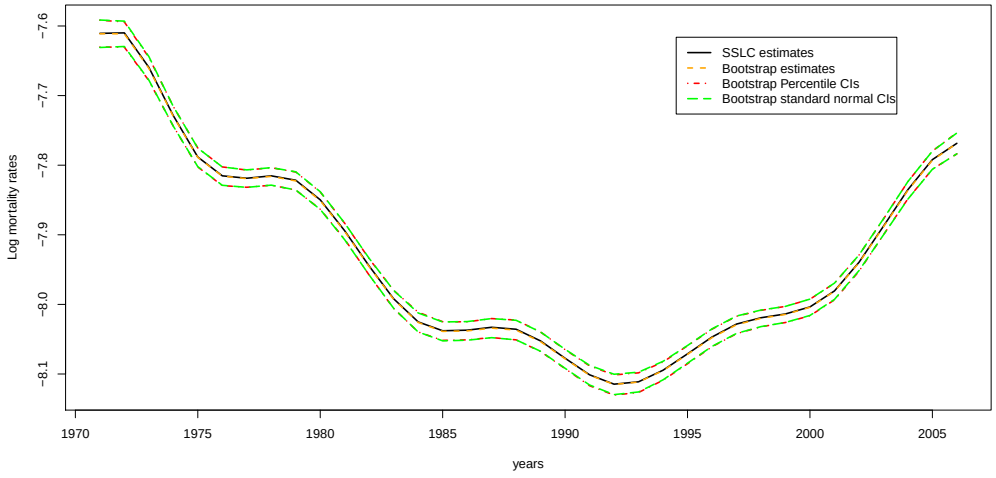


Figure 10.104: Accidents: Log mortality rate estimates at age 44 years and 95% bootstrap pointwise confidence intervals.

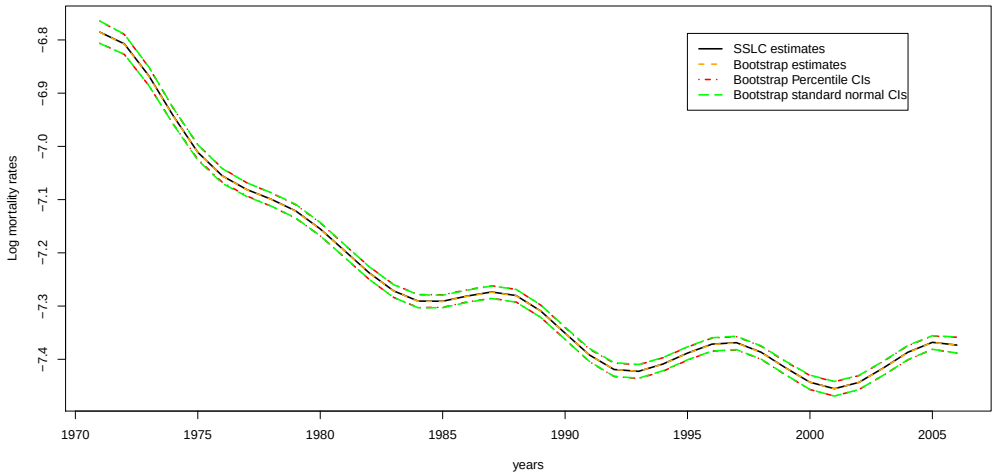


Figure 10.105: Accidents: Log mortality rate estimates at age 74 years and 95% bootstrap pointwise confidence intervals.

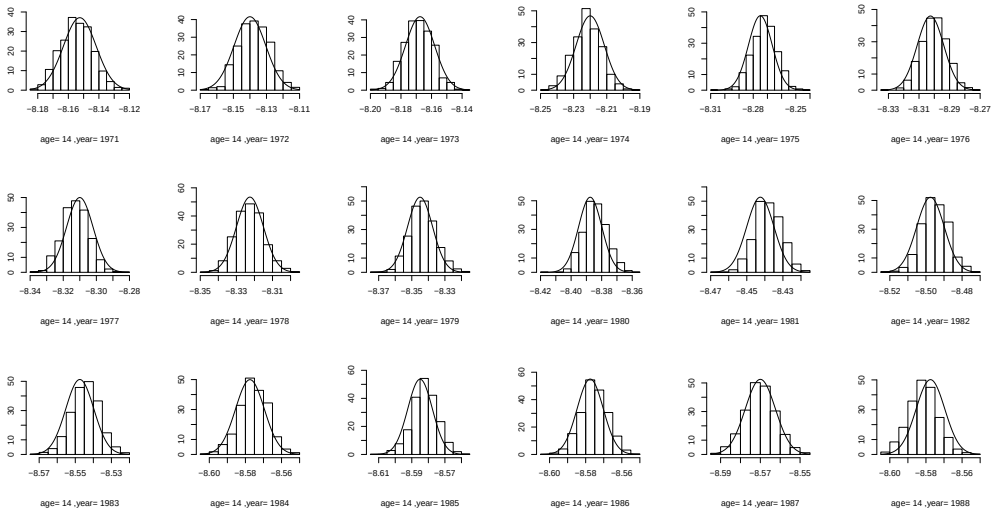


Figure 10.106: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\log(\hat{\lambda}_{14,p}) : p = 1971, ..., 1988$.

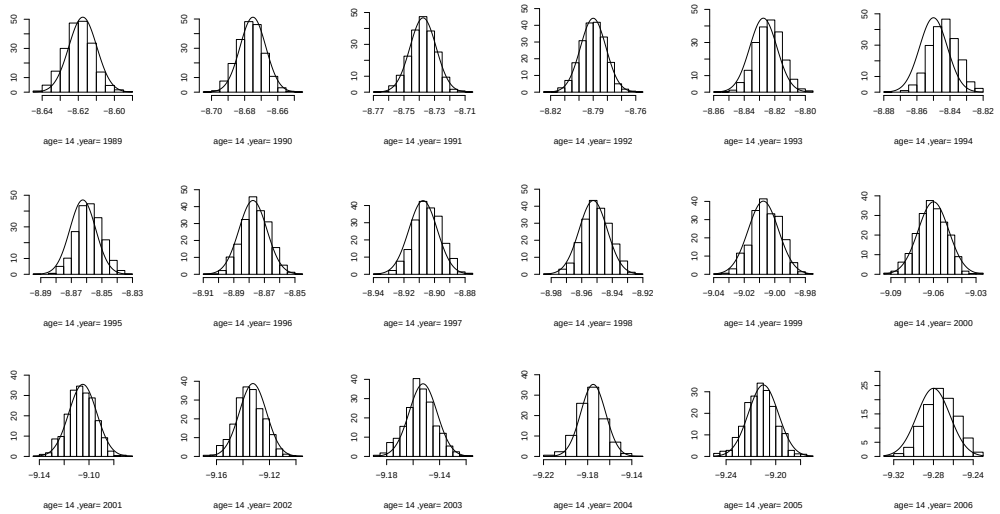


Figure 10.107: Accidents: Histograms of 1000 bootstrap replications with corresponding overlaid normal curve of $\log(\hat{\lambda}_{14,p}) : p = 1989, ..., 2006$.

In conclusion, the 95% percentile pointwise confidence intervals and the 95% standard normal pointwise confidence intervals coincide in most cases in particular in cancer and accidents cases. Some differences between the two intervals for parameter estimates are found in heart diseases. Even through they are mostly compatible, the percentile intervals are recommended in general cases because it has transformation-respecting property (Efron and Tibshirani, 1993).

The LC and SSLC estimates of log mortality rates and percentile confidence intervals

This section shows an exhibition of LC and SSLC estimates of log mortality rates at some selected ages with their corresponding 95% bootstrap percentile pointwise confidence intervals for heart diseases, cancer and accidents. Crude estimates

mentioned in the figures are the logarithms of the proportions $\tilde{\lambda}_{a,p}$'s.

Heart diseases

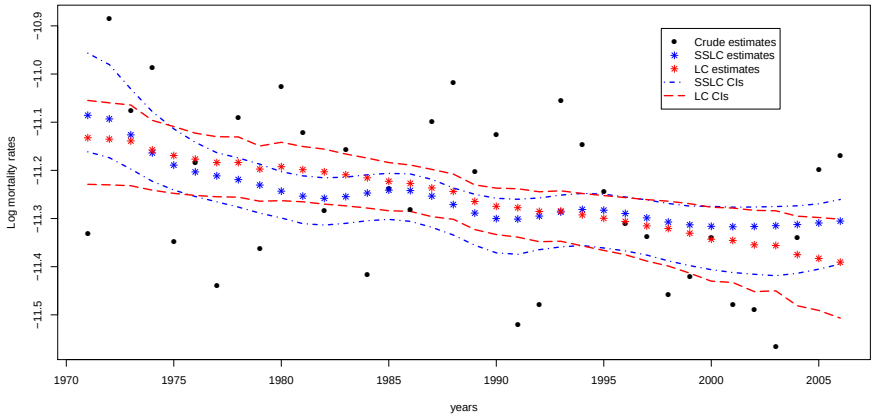


Figure 10.108: Heart diseases: Log mortality rate estimates at age 14 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

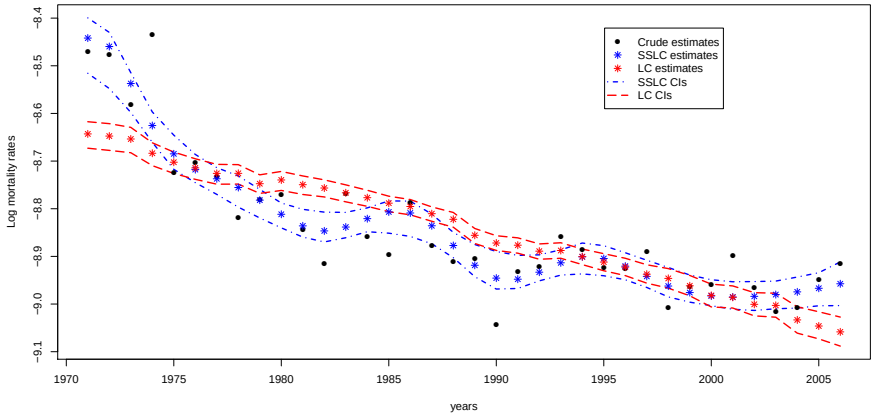


Figure 10.109: Heart diseases: Log mortality rate estimates at age 34 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

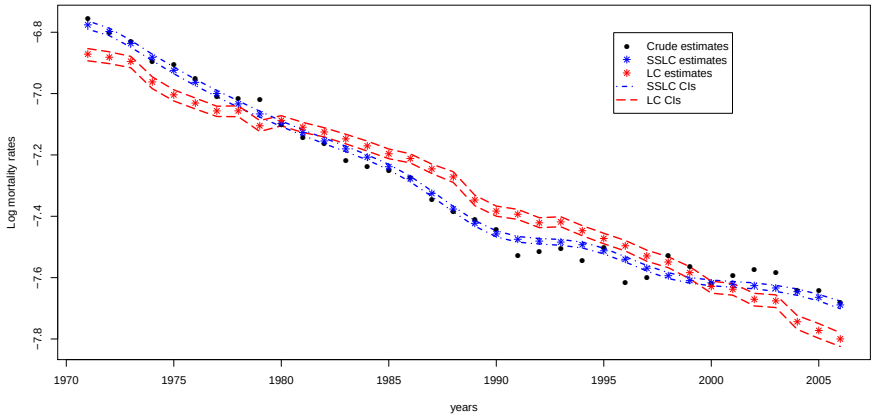


Figure 10.110: Heart diseases: Log mortality rate estimates at age 44 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

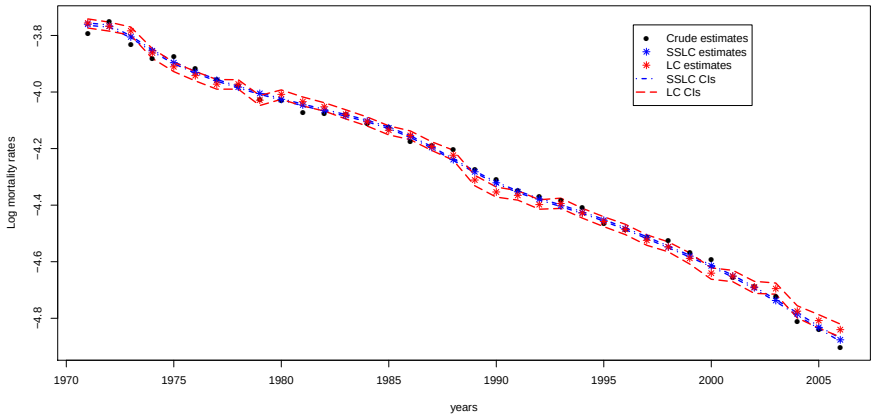


Figure 10.111: Heart diseases: Log mortality rate estimates at age 74 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

Cancer

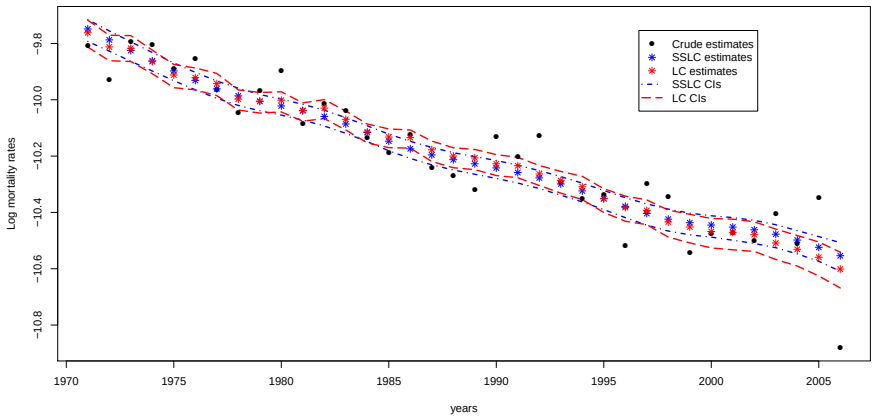


Figure 10.112: Cancer: Log mortality rate estimates at age 14 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

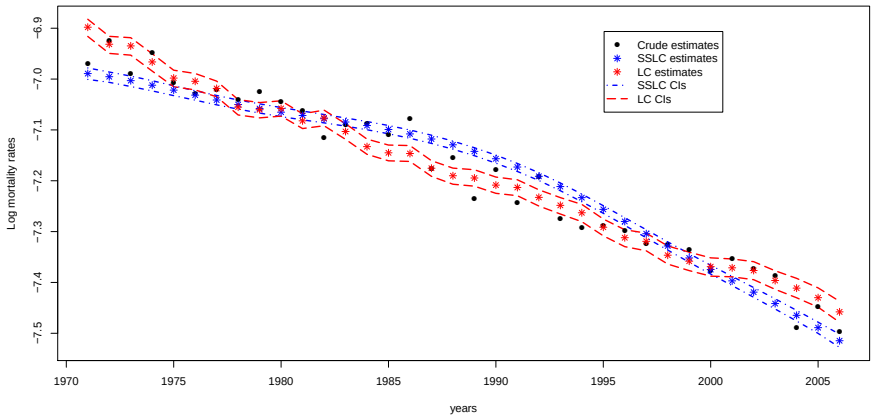


Figure 10.113: Cancer: Log mortality rate estimates at age 44 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

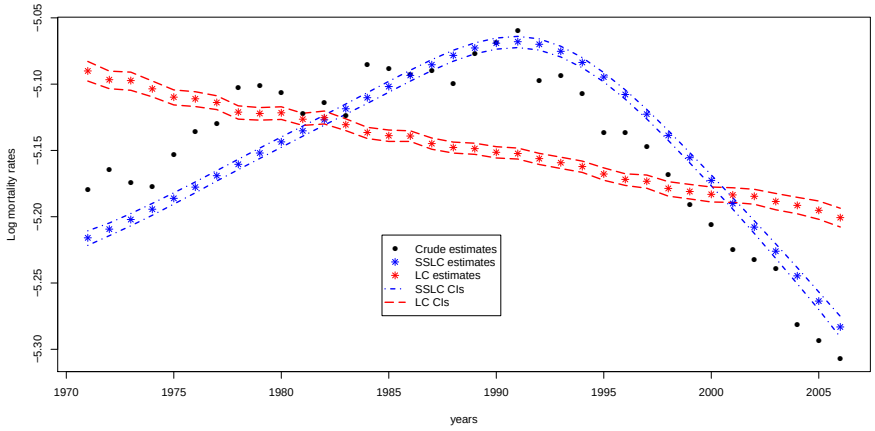


Figure 10.114: Cancer: Log mortality rate estimates at age 64 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

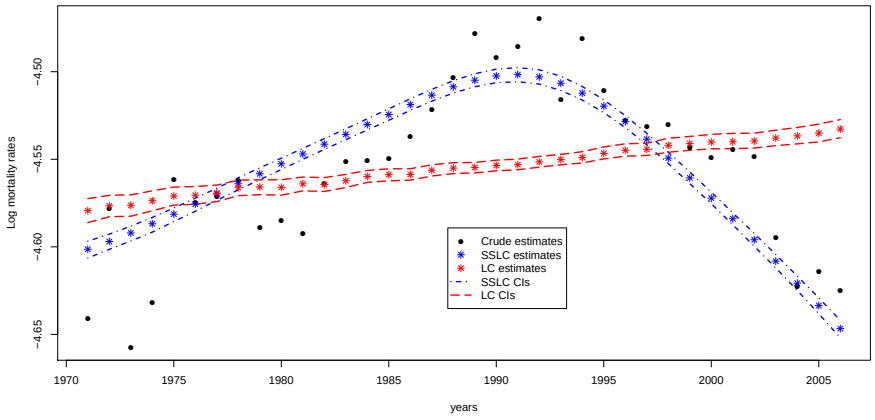


Figure 10.115: Cancer: Log mortality rate estimates at age 74 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

Accidents

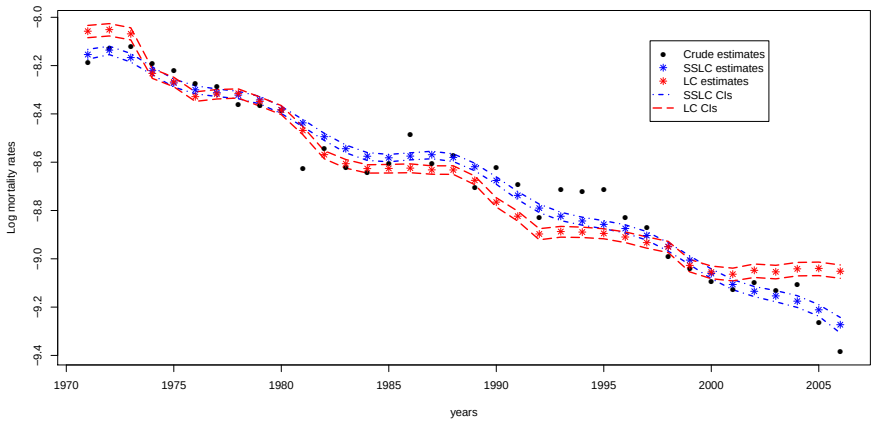


Figure 10.116: Accidents: Log mortality rate estimates at age 14 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

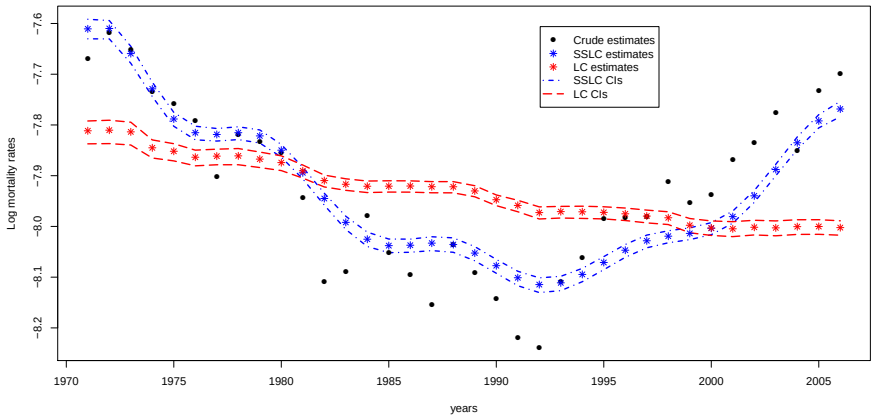


Figure 10.117: Accidents: Log mortality rate estimates at age 44 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

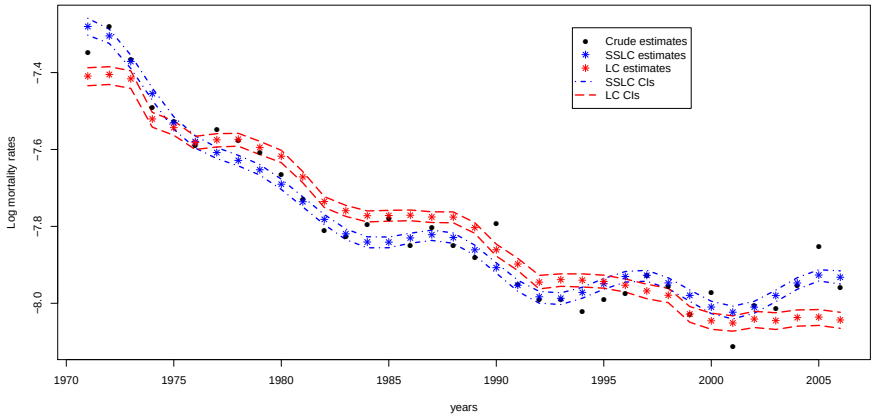


Figure 10.118: Accidents: Log mortality rate estimates at age 64 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

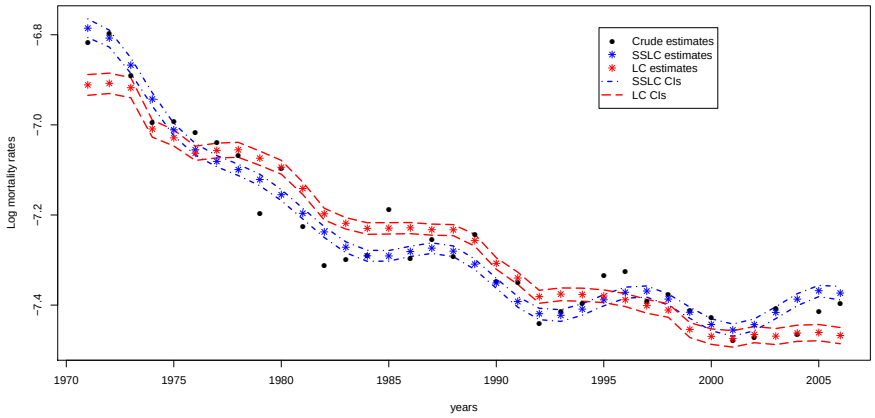


Figure 10.119: Accidents: Log mortality rate estimates at age 74 years obtained from LC and SSLC models and corresponding 95% bootstrap percentile pointwise confidence intervals.

Bibliography

[1] O.O. Aalen. Phase type distributions in survival analysis. *Scand. Journal of Statistics*, 22:447–463, 1995.

[2] O.O. Aalen, Ø. Borgan, and H.K. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Springer, 2008.

[3] O.O. Aalen and H. Gjessing. Understanding the shape of the hazard rate: A process point of view. *Statistical Science*, 16:1–22, 2001.

[4] O.O. Aalen and H.K. Gjessing. Survival models based on the Ornstein-Uhlenbeck Process. *Lifetime Data Analysis*, 10:407–423, 2004.

[5] M. Aitkin and I. Aitkin. A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*, 6:127–130, 1996.

[6] J. M. Alho. Discussion of Lee (2000). *North American Actuarial Journal*, 4:91–93, 2000.

[7] J.M. Alho and B.D. Spencer. *Statistical Demography and Forecasting*. Springer, 2005.

[8] W. Anderson, B. Chen, I. Jatoi, and P. Rosenberg. Effects of estrogen receptor expression and histopathology on annual hazard rates of death from breast cancer. *Breast Cancer Res. Treat.*, 100:121–126, 2006.

[9] E. Arias. United States Life Tables, 2006. *National Vital Statistics Reports*, 54, 2006.

[10] P. Armitage and R. Doll. The age distribution of cancer and a multistage theory of carcinogeneisis. *British Jour. Cancer*, 8:1–12, 1954.

[11] S. Asmussen. Phase-type representations in random walk and queueing problems. *Annals of Probability*, 20:772–789, 1992.

[12] S. Asmussen, O.Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorithm. *Scand. Journal of Statistics*, 23:419–441, 1996.

[13] M. Ausin, M. Wiper, and R. Lillo. Bayesian estimation for the M/G/1 queue using a phase-type approximation. *Jour. Statist. Planning & Inference*, 118:83–101, 2004.

[14] J. Balka, A. F. Desmond, and P.D. McNicholas. Review and Implementation of cure models based on first hitting times for Wiener processes. *Lifetime Data Analysis*, 15:147–176, 2009.

[15] M. Bladt, A. Gonzalez, and S.L. Lauritzen. The estimation of phase-type related functionals using Markov Chain Monte Carlo methods. *Scand. Actuarial Journal*, 4:280–300, 2003.

[16] A. Bobbio, A. Horváth, M. Scarpa, and M. Telek. Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. *Performance Evaluation*, 54:1–32, 2003.

[17] A. Bobbio, A. Horváth, and M. Telek. Matching three moments with minimal acyclic phase-type distributions. *Stochastic Models*, 21:303–326, 2005.

[18] J. Bongaarts. Population aging and the rising cost of public pensions. *Population and Development Review*, 30:1–23, 2004.

[19] H. Booth, R.J. Hyndman, L. Tickle, and P.D. Jong. Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research*, 15:289–310, 2006.

[20] H. Booth, J. Maindonald, and L. Smith. Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56:325–336, 2002.

[21] H. Booth and L. Tickle. The future aged: new projections of Australia's elderly population. *Australasian Journal on Aging*, 22:196–202, 2003.

[22] H. Booth and L. Tickle. Mortality modelling and forecasting: a review of methods. *Annals of Actuarial Science*, 3:3–43, 2008.

[23] N. Bowers, H. Gerber, J. Hickman, D. Jones, and C. Nesbitt. *Actuarial Mathematics, 2nd ed.* Society of Actuaries, Schaumburg, IL, 1997.

[24] D. R. Brillinger. A justification of some common laws of mortality. *Transac. Soc. Actuaries*, XIII:116–119, 1961.

[25] D. R. Brillinger. The natural variability of vital rates and associated statistics. *Biometrics*, 42:693–734, 1986.

[26] N. Brouhns, M. Denuit, and I.V. Keilegon. Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, 3:212–224, 2005.

[27] N. Brouhns, M. Denuit, and J.K. Vermunt. Measuring the longevity risk in mortality projections. *Bulletin of the Swiss Associatiopn of Actuaries*, pages 105–130, 2002.

[28] N. Brouhns, M. Denuit, and J.K. Vermunt. A Poisson log-bilinear regression approach to the construction of projected life tables. *Insurance: Mathematics and Economics*, 31:373–393, 2002.

[29] A. Buonocore, A. G. Nobile, and L.M. Ricciardi. A new integral equation for the evaluation of first-passage-time probability density. *Advance Applied Probbaility*, 19:784–800, 1987.

[30] R.H. Byrd, J. Nocedal, and Y.-X. Yuan. Global convergence of a class of quasi-newton methods on convec problems. *SIAM Journal on Numerical Analysis*, 24:1171–1190, 1987.

[31] B. Carstensen. Age-period-cohort models for the Lexis diagram. *Statisics in Medicine*, 26:3018–3045, 2007.

[32] A. Case and C. Paxson. Sex differences in Morbidity and Mortality. *Demography*, 42:189–214, 2005.

[33] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury, 2002.

[34] X. Chen, O. Linton, and I.V. Keilegom. Estimation of Semiparametric Models when the Criterion Function is not smooth. *Econometrica*, 71:1591–1608, 2003.

[35] H. Chernoff and E.L. Lehmann. The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit. *The Annals of Mathematical Statistics*, 25, 1954.

[36] D. Clayton and E. Schifflers. Models for temporal variation in cancer rates I: Age-Period and Age-Cohort models. *Statistics in Medicine*, 6:449–467, 1987.

[37] A. Delwarde, M. Denuit, and P. Eilers. Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: a penalized log-likelihood approach. *Statistical Modelling*, 7:29–48, 2007.

[38] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Jour. Royal Statist. Society, Series B (Methodologcal)*, 39:1–38, 1977.

[39] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge, 1998.

[40] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051- 07-0, URL http://www.R-project.org.

[41] J. Durbin. The first passage density of a continuous Gaussian process to a general boundary. *Journal of Applied Probability*, 22:99–122, 1985.

[42] B. Efron. Bootstrap Methods: another look at jackknife. *The Annals of Statistics*, 7:1–26, 1979.

[43] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[44] M. Fackrell. Modelling healthcare systems with phase-type distributions. *Health Care Manag. Sci.*, 12:11–26, 2009.

[45] M. Faddy and S. McClean. Analysing data on lengths of stay of hospital patients using phase-type distributions. *Appl. Stoch. Models in Bus. and Indust.*, 15:311–317, 1999.

[46] W. Feller. *Introduction to Probability Theory and its Applications, 2nd ed.* Wiley, 1972.

[47] S.E. Fienberg and W.M. Mason. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology*, 1979.

[48] L. Garg, S. McClean, B.J. Meenan, and P. Millard. Phase-type survival trees and mixed distribution survival trees for clustering patients's hospital length of stay. *Informatica*, 22:57–72, 2011.

[49] F. Girosi and G. King. Understanding the Lee-Carter mortality forecasting method. Available at `http://gking.harvard.edu/files/le.pdf`.

[50] F. Girosi and G. King. *Demorgraphic Forecasting*. Princeton University Press, New Jersey, 2008.

[51] B. Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Phil. Trans. Royal Society of London*, 115:513–585, 1825.

[52] I. Griva, S.G. Nash, and A. Sofer. *Linear and Nonlinear Optimization*. SIAM, 2009.

[53] L. Heligman and J. Pollard. The age pattern of mortality. *Jour. Institute of Actuaries*, 107:49–175, 1980.

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

[55] R.J. Hyndman and M.S. Ullah. Robust forecasting of mortality and fertility rates: a functinal data approach. *Computational Statistics & Data Analysis*, 51:4942–4956, 2007.

[56] E. Ishay. Fitting phase-type distributions to data from a telephone call center. *Thesis*, 2002.

[57] M. Jamshidian and R.I. Jennrich. Conjugate gradiant acceleration of the EM algorithm. *Journal of the American Statistical Association*, 88:221–228, 1993.

[58] M.A. Johnson. Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and Erlang distributions. *ORSA Jour. Computing*, 5:69–83, 1993.

[59] M.C. Koissi, A.F. Shapiro, and G. Höngäs. Evaluating and extending the Lee-Carter model for mortality forecasting: bootstrap confidence interval. *Insurance: Mathematics and Economics*, 38:1–20, 2006.

[60] A. Kostaki. A nine-parameter version of the Heiligman-Pollard formula. *Mathematical Population Study*, 3:277–288, 1992.

[61] B. Lachaud. Hitting times of Ornstein-Uhlenbeck Processes, 2004.

[62] T. Lancaster. A stochastic model for duration of a strike. *Journal of Royal Statistical Society: Series A*, 135:257–271, 1972.

[63] D. H. Lawlor, S. Ebrahim, and G. D. Smith. Sex Matters: Secular and Geographical trends in Sex Differences in Coronary Heart diseas Mortality. *British Medical Joournal*, 323:541–545, 2001.

[64] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. SIAM, 1974.

[65] M.-L. Lee and G. Whitmore. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, 21:501–513, 2006.

[66] R.D. Lee and L. Carter. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87:659–671, 1992.

[67] R.D. Lee and L. Carter. Modeling and Forecasting US sex differentials in mortality. *International Journal of Forecasting*, 8:393–411, 1992.

[68] S. Lee and X.S. Lin. Modelling Insurance Losses and Calculating Risk Measures via a Mixture of Erlangs. *North American Actuarial Journal*, 17:107–130, 2010.

[69] X.S. Lin. The moments of the time of ruin, the surplus before ruin, and the deficit at ruin. *Insurance: Mathematics and Economics*, 27:19–44, 2000.

[70] X.S. Lin and X. Liu. Markov aging process and phase-type law of mortality. *North Amer. Actuarial Journal*, 11:92–109, 2007.

[71] X.S. Lin and G.E. Willmot. Analysis of a defective renewal equation arising in ruin theory. *Insurance: Mathematics and Economics*, 25:63–84, 1999.

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

[73] H. Lundstrom and J. Qvist. Mortality forecasting and trend shifts: an application of the Lee-Carter model to Swedish mortality data. *International Statistical Review*, 72:37–50, 2004.

[74] K. Manton and E. Stallard. A two disease model of female breast cancer: mortality in 1969 among white females in the United States. *Journal of National Cancer Institute*, 64:9–16, 1980.

[75] C. McGrory, A. Pettitt, and M. Faddy. A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Jour. Computational Statist. & Data Anal.*, 53:4311–4321, 2009.

[76] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 2008.

[77] A. Molarius and TITLE=Self-rated Health, Chronic Diseases, and Symptoms Among Middle-Ages and Elderly Mean and Women S. Janson. *Journal of Clinical Epidemiology*, 55:364–370, 2002.

[78] S.H. Moolgavkar. Fifty years of the multistage model: Remarks on a landmark paper. *Int. J. Epidemiology*, 33:1182–183, 2004.

[79] M.F. Neuts. Probability of phase type. *Liber Amicorum Prof. Emiritus H. Florin: Department of Mathematics. Belgium: University of Louvain*, pages 173–206, 1975.

[80] M.F. Neuts. *Matrix Geometric Solutions in Stochastic Models: an Algorithmic Approach*. The Johns Hopkins University Press, 1981.

[81] A. G. Nobile, L.M. Ricciardi, and L. Secerdote. Exponential trends of Ornstein-Uhlenbeck first-passage-time densities. *Journal of Applied Probability*, 22:360–369, 1985.

[82] D. Oakes. Direct calculation of the information matrix via the EM algorithm. *J. R. Statist.Soc.B*, 61:479–482, 1999.

[83] C.A. O'Cinneide. On non-uniqueness of representations of phase-type distributions. *Stochastic Models*, 5:247–259, 1989.

[84] M. Olsson. The `EMpht` programme. Available at `home.imf.au.dk/smus/dl/EMusersguide.ps`.

[85] M. Olsson. Estimation of phase-type distributions from censored data. *Scand. Jour. Statistics*, 23:443–460, 1996.

[86] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press,NewYork, 1970.

[87] A. Pakes and D. Pollard. Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57:1027–1057, 1989.

[88] F. C. Pampel. Cigarette Use and The narrowing Sex Differential in Mortality. *Population and Development Review*, 28:77–104, 2002.

[89] J. Pitman and M. Yor. Bessel processes and infinitely divisible laws. *Stochastic integrals. Proc. Sympos., Univ. Durham, Durham, 1980*, 1981.

[90] D.S.G. Pollock. Smoothing with cubic splines. http://r.789695.n4.nabble.com/file/n905996/SPLINES.PDF.

[91] S. H. Preston and H. Wang. Sex mortality diferences in the united states: The role of cohort smoking patterns. *Demography*, 43:631–646, 2006.

[92] A. Renshaw and S. Haberman. Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections. *Journal of the Royal Statistical Society*, 52:119–137, 2003a.

[93] A. Renshaw and S. Haberman. Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33:255–272, 2003b.

[94] L. M. Ricciardi and S. Sato. First-passage-time density and moments of the Ornstein-Uhlenbeck process. *Journal of Applied Probability*, 25:43–57, 1988.

[95] C. Robertson, S. Gandini, and P. Boyle. Age-Period-Cohort Models: A Comparative Study of Available Methodologies. *Journal of Clinical Epidemiology*, 52:569–583, 1999.

[96] A. Rogers and K. Gard. Applications of the Heiligman-Pollard model mortality schedule. *Population Bulletin of the United nations*, 30:79–105, 1991.

[97] P. S. Rosenberg and W. F. Anderson. Proportional Hazard models and age-period-cohort analysis of cancer rate. *Statitics in Medicine*, 29:1228–1238, 2010.

[98] P. S. Rosenberg and W. F. Anderson. Age-period-cohort models in cancer surveillance research: Ready for prime time ? *Cancer Epidemiology, Biomarkers and Prevention*, 20:1263–1268, 2011.

[99] R. Schoen. *Dynamic Population Models*. Springer, 2006.

[100] B. Sengupta. Markov processes whose steady-state distribution is matrix-exponential with an application to the GI/G/1 queue. *Adv. Appl. Probability*, 21:159–180, 1989.

[101] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer, 1995.

[102] D.J. Sharrow, S.J. Clark, M.A. Collinson, K. Kahn, and S.M. Tollman. The age-pattern of increases in mortality affected by HIV: Bayesian fit of the Heiligman-Pollard model to data from the Agincourt HDSS field site in rural Northeast South Africa, 2010.

[103] M. Sherris and C. Njenga. Modeling Mortality with a Bayesian Vector Autoregression. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id =1776532.

[104] P-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, Boston, 2005.

[105] J. M. G. Taylor, W. G. Cumberland, and J. P. Sy. A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, 89:727–736, 1994.

[106] A. Thümmler, P. Buchholz, and M. Telek. A novel approach for phase-type fitting with the EM algorithm. *IEEE Trans. Depend. Secure Computing*, 3:245–258, 2006.

[107] D. C. Trost, E.A. Overman, J.H. Ostroff, W. Xiong, and P. Marchc. A model for liver homeostasis using modified mean-reverting OrnsteinUhlenbeck process. *Computational and Mathematical Methods in Medicine*, 11:27–47, 2010.

[108] L. M. Verbrugge. The Twain Meet: Empirical Explanations of Sex Differences in Health and Mortality. *Journal of Health and Social Behavior*, 30:282–304, 1989.

[109] J. K. Vermunt. LEM: A General Program for the Analysis of Categorical Data. Users' Manual. Tilburg University, Tilburg, The Natherlands., 1979.

[110] J. K. Vermunt. *Log-linear Models for Event Histories*. Sage,Thousand oakes, 1979.

[111] A. Voulgaraki, B. Kedem, and R. Wei. Estimation of death rates in U.S. states with small subpopulations. Preprint, 2008.

[112] H. Wang and S. H. Preston. Forecasting United States Mortality using cohort smoking histories. *Proceedings of the National Academy of Sciences*, 106:393–398, 2009.

[113] R. Wei, B. Nandram, and D. Bhatta. A Bayesian Analysis of US Mortality Curves for Race-Sex Domains by Small Area. Preprint 2011, submitted.

[114] J.S. Weitz and H.B. Fraser. Explaining mortality rate plateaus. *PNAS*, 98:15383–15386, 2001.

[115] J. R. Wilmoth. Is the pace of Japanease mortality decline converging toward international trends? *Population and Development Review*, 1998.

[116] J.R. Wilmoth. Computational methods for fitting and extrapolating the Lee-Carter model for mortality change. Techical Report, 1993.

[117] C.F.J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.