UiT

THE ARCTIC UNIVERSITY OF NORWAY

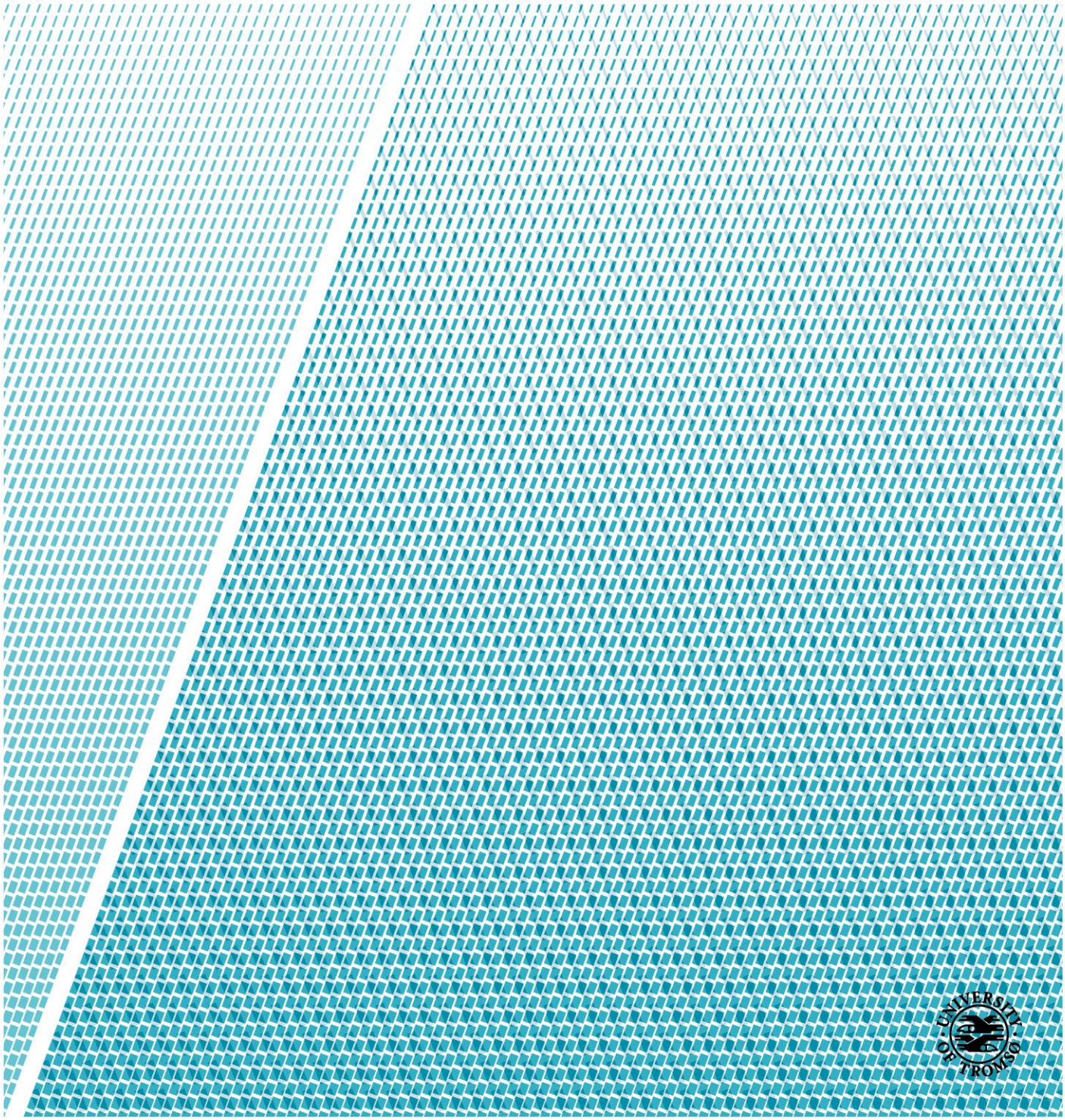
Faculty of science and technology

Bayesian analysis of the occurrence of myocardial infarction using the LGM framework

Yohannes Tesfay

STA-3900 Master's thesis in Statistics

May 2019



Acknowledgement

I would like to express the outmost appreciation to my supervisor Sigrunn Holbek Sørbye for the guidance, patience and support throughout my thesis. I am deeply grateful for the time and effort dedicated in steering me in the right direction during this learning process. Without her this dissertation would not have been possible.

I would like to thank my dear Lamort for the support. I also want to thank all of my friends that were supportive during this process.

Abstract

The goal of this thesis is analyse the incidence rate of the first ever myocardial infarction (MI) and the survival time after the MI. The data used for this purpose is from the Tromsø study surveys collected in the period from 1974 to 2008. This thesis provides a general introduction to latent Gaussian models and the methodology of integrated nested Laplace approximation. Specifically, the data is analysed using Bayesian age period cohort models and cox proportional hazards models.

Contents

Chapter 1

Introduction

1.1 Myocardial Infarction

Cardiovascular diseases (CVD) such as Heart diseases are often reposted as the leading causes of death world wild (Murray et al., 2010). According to the Norwegian institute of public health (NIPH), heart diseases are also one of the leading causes of death in Norway, and about one fifth of the population in Norway is diagnosed with at least one type of CVD or in risk of developing one.

Myocardial infarction (MI) is a common type of coronary heart disease and its incidence has been in a decline from 2001 to 2014 in all age groups (Sulo et al., 2018, 2014) A considerable amount of the reduction is attributed to changes in cardiovascular risk factors such as smoking, high levels of cholesterol and levels of physical activity (Mannsverk et al., 2016). In addition, the Norwegian Prescription database reports that there is an increased number of medical drug use for both prevention and treatment of cardiovascular diseases in all age groups.

The mortality due to MI has been declining in Norway since 1976 according to the Norwegian cause of death register. There is a higher proportion of men above the age of 75 among those who die as a result of MI. This over representation of men is also seen in the incidence of MI (Albrektsen et al., 2016; Jortveit et al., 2014). In addition, the inhabitants of the two northern most counties are affected

by these issues in a higher proportion than the rest of the country.
A large portion of the general population is affected by MI. Even though the incidence rate and mortality is decreasing, there is still a lots of work that remains to be done in this field. One of them is to find the groups that exposed to the risk of MI. Therefore, in this thesis we are interested in looking at the incidence rates of MI and survival time after MI. To do that, data from the Tromsø study surveys will be used. From here on, we refer to the first ever myocardial infarction as MI.

1.2 Overview of the Tromsø Study surveys

Since the data analysis in this thesis will be based on data from the Tromsø study surveys, an introduction to the population based study is provided. Most of the information about the Tromsø Study surveys is from the website of the Tromsø study `https://uit.no/forskning/forskningsgrupper/gruppe?p_document_id=367276`.
The Tromsø Study, with 45000 participants and seven surveys, is the highest visited and most comprehensive health survey in Norway. Since its beginning in 1974, the population study has evolved in complexity and depth. Each of those surveys have taken on challenging public heath issues and revealed interesting medical findings that have helped improve the health condition of the community. The most notable results based on the Tromsø Study are that consumption of boiled coffee raises the level of serum cholesterol (Thelle et al., 1983) and the importance of HDL cholesterol (Miller et al., 1977) .
In this thesis, various statistical analysis will be conducted using the Tromsø study dataset consisting of 39 870 subjects that participated in at least one of the first six surveys. Figure 1.1 and table 1.1 show the age and gender distribution of all the enrolled participants by survey. It started off with just male participants between the ages of 20 and 49. However, the second survey was expanded to include female participants as well. The entire data consists of roughly equal number of male

and female participants. Figure 1.2 displays a histogram with the horizontal axis showing the number of surveys and the vertical axis showing total number of participants with the given number of participation. The histogram shows that a large portion of the participants has return for the subsequent surveys.

The dataset contains the age at each medical examination, gender, date of each medical examination, date of their first myocardial infarction, date of death and date of emigration from Tromsø (if the participant has moved from Tromsø, had a heart attack and/or died) of all the 39 870 subjects. The date of each case of MI was extracted by looking for the MI through the discharge summaries from the University Hospital of northern Norway (UNN). In addition to the discharge summaries, the Norwegian cause of death registry and death certificates were also extensively reviewed for deaths that did not take place at UNN, in case the patients died of MI. The registration process was overseen by experienced medical doctors (Hopstock et al., 2011).

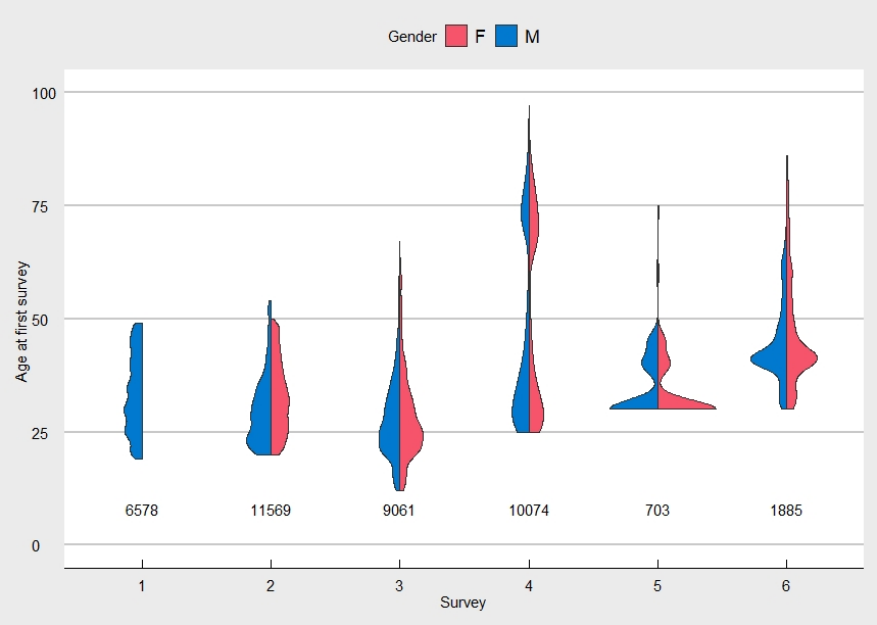


Figure 1.1: *The distribution of participants of Tromsø study by age, gender and survey. In addition, the number of first time participants is given.*

Cardiovascular diseases were widespread among middle aged Northern Norwegian men in the 1970s. The search for the cause of the pandemic cardiovascular disease lead to the initiation of the Tromsø study in 1974. The first survey, Tromsø 1 was therefore started in 1974 with the invitation of 8866 men between 20 and 49 to take part in the survey. About 74% of those invited chose to participate. These participants answered a questionnaire, gave a short interview followed by measurements of height, weight, blood pressure and giving a blood sample. The Tromsø study has since broaden its focus from just cardiovascular diseases to problems like mental health, dental health and also cancer.

Table 1.1: *The distribution of participants of Tromsø study by survey and gender.*

Survey	**Compliance rate**	**Gender**	**Age/ Avg age**	**No. total Part.**	**No. 1st time Part.**
Tromsø 1	74%	Male	20-49/	6578	6578
		Female	-	-	-
Tromsø 2	74%	Male	20-54	8457	3456
		Female	20-50	8121	8121
Tromsø 3	75%	Male	12-64	10937	4314
		Female	12-67	10823	4747
Tromsø 4	77%	Male	25-97	12805	4356
		Female	25-94	14187	5718
Tromsø 5	57%	Male	30-75	3483	310
		Female	30-75	4579	392
Tromsø6	74%	Male	30-86	6053	884
		Female	30-86	6928	1001

Tromsø 2 took place in 1979/80 and is the first Tromsø survey to include women attendees. The participants of Tromsø 1, women between 20 and 50 years old and men between the ages of 20 and 25 were invited to participate in the Tromsø 2 survey. All of the participants went through a similar process to the subjects

of the Tromsø 1 survey.

Tromsø 3 expanded on the number of participants by inviting the family members of the participants of Tromsø 1 and 2 and randomly selected 10% of those between the ages of 12 and 19. Almost 22000 of Tromsøs inhabitants between 12 and 67 participated in the survey in 1986/87.

All the citizens of Tromsø Municipality who were above the age of 25 at the time of the survey was invited to take part in Tromsø 4 in 1994/95. This was the largest invited group, and made Tromsø 4 the largest survey among the Tromsø Study surveys. There were more than 27 000 participants between the ages of 25 and 97 years. The questionnaires and medical examinations have evolved in complexity since the ones in the Tromsø 1 survey. The fifth Tromsø study took place in 2001 and included a group that attended the Tromsø 4 study and another group as part of a nationwide health survey. There were 8130 participants in the 2001 Tromsø 5 survey, among which 703 were first time participants.

Tromsø 6 was carried out in 2007-08 with almost 13000 participants. The participants went through a thorough questionnaire about their mental and physical health. 1885 of the participants were first time participants of the Tromsø study. Data from the most recent survey, Tromsø 7, were not available for this thesis.

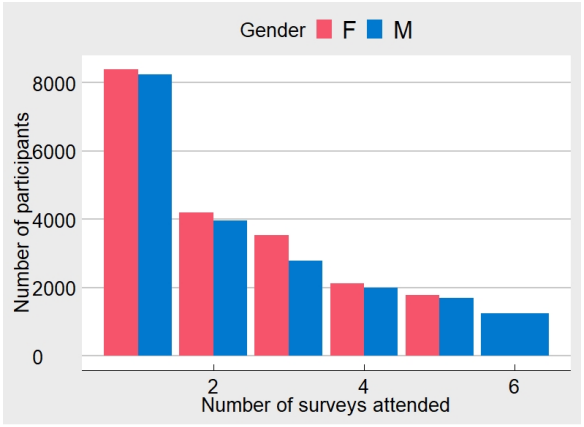


Figure 1.2: *The number of surveys attended by participants of the Tromsø Study.*

A large proprtion of the invited population has participated in the Tromsø studies. The majority of the surveys have about 75% participation rate. This rate is higher for women than men. The participation rate tends to increase with age except for the oldest

participants.

1.3 Objectives and outline of the thesis

The objective of this thesis is to give a temporal analysis of the incidence rate of MI in the municipality of Tromsø. Also the survival time after an MI among the population is studied. In particular, this thesis focuses on gender differences and characteristics in different age groups. All analysis is performed using the programming language R.

In chapter 2, the data is presented and preliminary analysis is performed using basic statistical methods. To address the issue of increased age among the participants during the study period more advanced statistical methods are used. Chapter 3 presents a Bayesian framework for analysing a wide class of statistical models known as latent Gaussian models (LGM). These models are analysed using the methodology of integrated nested Laplacae approximation (INLA). The models analysed in chapters 4 and 5 are subclasses of LGM. These include an age period cohort model to study the incidence rate of MI and Cox proportional hazards model to analyse survival time. A discussion and the final conclusions are provided in chapter 6.

Chapter 2

Data and preliminary analysis

This chapter presents the datasets that will be analysed in the upcoming chapters. A preliminary analysis on the occurrence of MI will be also be provided in this chapter. In section 2.1, data that will be used in the analysis of incidence rate of MI will be presented. Next, survival data will be introduced in section 2.2. The chapter will then be concluded in section 2.3 by the discussion of the preliminary analysis based on age and gender of the participants who suffered from MI, and monthly and seasonal variation of the incidence of MI.

2.1 Description of incidence of MI data

In the datasets, there are registered dates of occurrence of MI prior to the enrolment time of the participants. This is due to the comprehensive retrospective process of MI registration. As a result, the enrolment time of the participants does not play any role in the analysis of incidence rates of MI. The first registered MI took place in August 1962, about 12 years before the first Tromsø study survey took place. This was used at the start of registration. While, the last registered MI, in November of 2014 denotes the end of the registration.

As a consequence of that, every participant in the Tromsø study can be thought of as eligible for MI before their first attendance in the study. In figure 2.1, the

number of MI per year, average age, total number of participants and proportion of male participants is displayed, and the counting started in 1962. Total number of participants denotes the number of participants that were alive and yet to have MI at the given time. This was decided by extracting their birth year using their age at the time of participation for each participant. At time of death or MI, the participants are then removed from the list of the risk. The average age and the proportion of male participants was then calculated using the risk set at the given time.

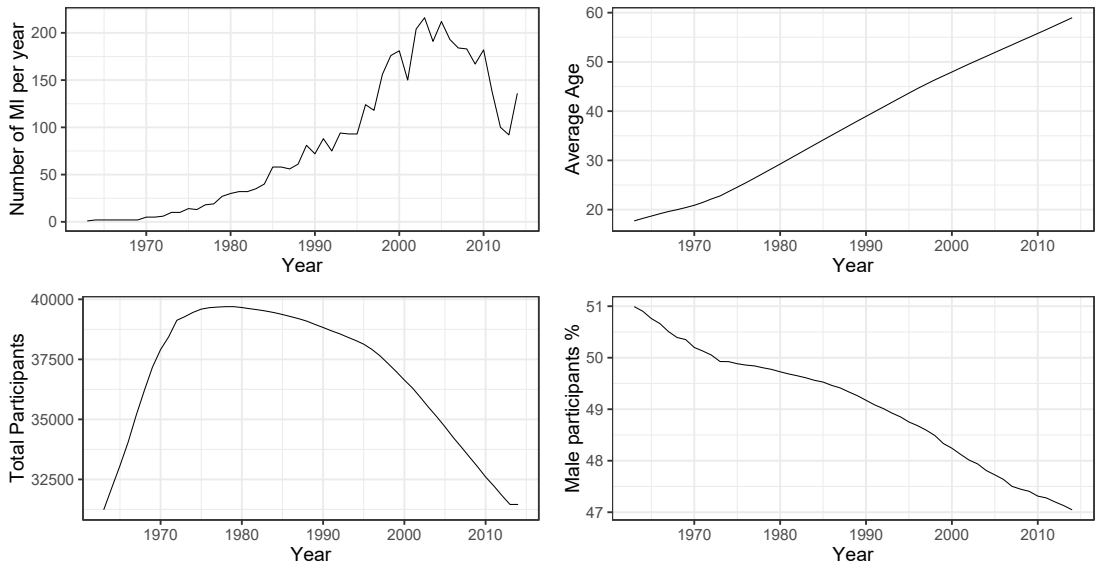


Figure 2.1: *The total number of MI per year (top left). The annual average age (top right). The total number of participants (bottom left) and the proportion of male participants in percent (bottom left).*

In the beginning, the annual number of MI remained constant near zero up until 1970. Next, the total number increased steadily for the following 20 years. It then accelerated faster in the 90s before an abrupt drop beyond 2000. Simultaneously, the proportion of the female population and average age increase linearly from the time of the first to the last registered MI. The average age was just below 18 in 1962 and increased to 59 in 2014, while the proportion of women increased

form 49 % to 52% in the same time period. Meanwhile, the total number of participants increases sharply from 31000 to 39600 in the first 10 years, and it drops back to 31458 in 2014.

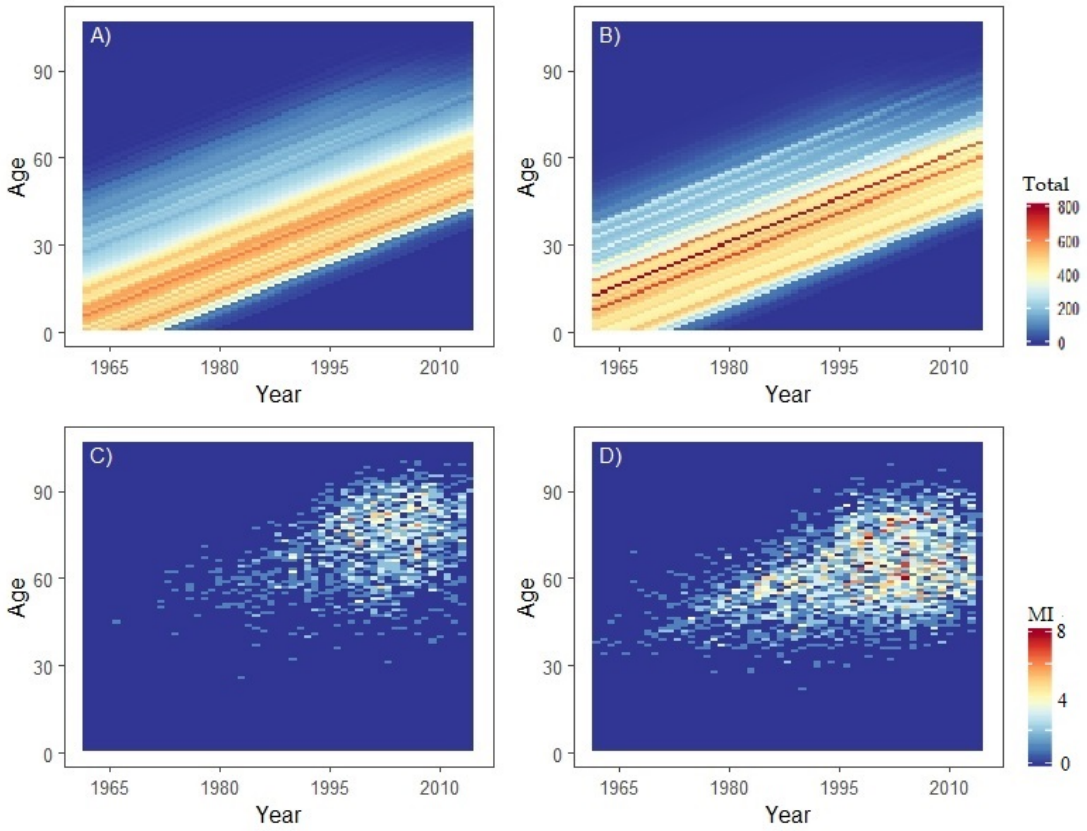


Figure 2.2: *A) Total female participants distributed by year and age B) Total male participants distributed by year and age C) Total MI in the female population distributed by year and age D) Total MI among the male participants distributed by year and age*

Figure 2.2 has age in the vertical axis and years from the first to the last registered MI in the horizontal axis. The total number of females and males by age and year are color coded into figure 2.2 A) and B) receptively. For instance, the total number of 40 years old females in 1980 are given according the legend on the right hand side of the figure. Similarly, the total number of MIs for females and males

are displayed in figure 2.2 C) and D) receptively. The individual birth cohorts can be followed along the diagonal lines. Such figures are known as Lexis diagram (Lexis, 1875)

Nr of MI	Gender	Age	Year	Birth Cohort	Total
0	1	40	1982	1942	361
0	2	40	1992	1942	318
0	1	41	1982	1941	218
0	2	41	1982	1941	287
0	1	42	1982	1940	298
0	2	42	1982	1940	277
⋮	⋮	⋮	⋮	⋮	⋮
1	1	60	1995	1935	190
1	2	60	1995	1935	194
⋮	⋮	⋮	⋮	⋮	⋮
0	1	80	2014	1934	133
0	2	80	2014	1934	112

Table 2.1: *A list of all the ages between the 40 and 80 and years between 1982 and 2014 with total participant in a particular row above 100. Nr of MI denotes the total number of MI in a given gender, age and year. Females are represented by 1 and males by 2 under the gender column. Total denotes the total number of participants that are eligible for MI in the specified gender, age and year. All rows are given a specific sets of age, period in years and the birth Cohort.*

The count data in table 2.1 consist of the number of MI for a given age, year and gender. All the ages between 40 and 80 in the time between 1982 and 2014 with total number of participants in the age group that exceeds 100 are represented. Figure 2.2 C) and D) show that males and female below the age of 40 rarely suffer from MI and that there are few registered MI prior to 1982. Those groups are omitted from the table because of the small number of MIs.

2.2 Survival data

In addition to the dates of the MI of the participants, the date of death of the participants is part of the Tromsø study dataset which is available for this thesis. The first registered death was registered in august of 1974, while the last registered death took place in march of 2017. Between august 1974 and march 2017, a total of 8441 have died, among them 2753 have also had a MI. Table 2.2 presents the total number of deaths and the number of dead participants who had suffered MI, divided by gender. Table 2.3 summarizes the age at the time of death of the participants by gender.

	Males (%)	**Female (%)**	**Total (%)**
Dead	4830 (24%)	3611 (18%)	8441 (21%)
Dead & MI	1790 (63%)	963 (69%)	2753 (65%)

Table 2.2: *The number of participants that have died*

As expected, table 2.2 shows a higher proportion of the participants who have had MI have died than the overall participants. In the overall population, the proportion of male participants who have died is higher than that of females. However, higher proportion of females who have had encountered MI have died than their male counterparts. Not only do those with MI die at higher proportion, their median age at the time of death is 3 year higher than the overall population as can be seen in table 2.3. And there is in addition about a 10 difference in the median age at the time of death between sexes.

In figure 2.3, the percentage of participants who have died after having had MI is plotted against the number of weeks they lived after MI. Among the total male participants who have had MI followed by death, 579 (32%) of them died within the first week. Similarly, 329 (34 %) of females died within the first week. Following the first week after the MI, the percentage of death declines markedly for both sexes. The longest time a male has lived after MI is 50 years, and the longest time a female lived after MI is 38 years.

	Age at the time of death					
	Min	**Median**	**Max**	**Min**	**Median**	**Max**
	Dead			**Dead & MI**		
Overall	15	75	106	26	78	104
Males	16	71	101	29	74	104
Females	15	80	106	26	83	99

Table 2.3: *The minimum, median and maximum age at the time of death grouped by gender and MI*

Table 2.4 shows that the proportion of the participants with MI that live past the first week decreases with age for both sexes. In addition, the median and maximum number of weeks the participants of Tromsø study with MI decreases progressively with age for males. While there is an increase for females in the first two age groups, the trend reverses for the last four age groups. The proportion of participants that had MI and who are still alive after march 2017 (last registered death in the dataset) decreases from 51% for the youngest age group to just under 7% in the oldest age group for males. The corresponding decrease for females is from 40% to 8%.

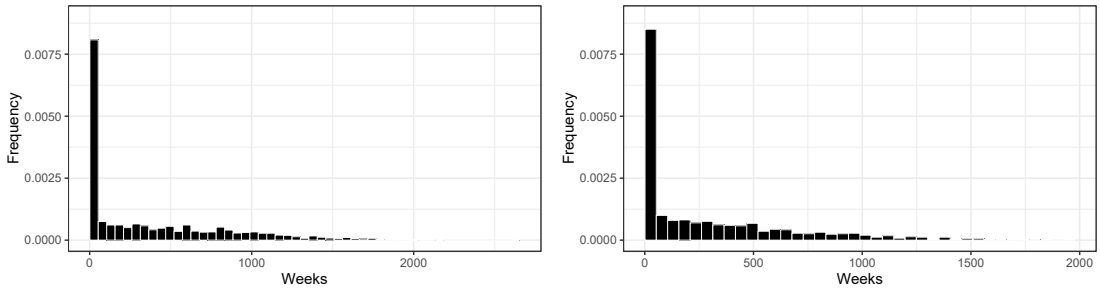


Figure 2.3: *Survival time of males (left) and females (right) (each bin represents 52 weeks starting at week 1).*

Based on the tables in this section,the life course of the participants after the incidence of MI depends on age and gender. The tables so far do not include

	Males				Females			
Age	**MI(Dead)**	**Wk 1(%)**	**Med**	**Max**	**MI(Dead)**	**Wk 1(%)**	**Med**	**Max**
(0-40]	101 (44)	14 (13.9)	678	2612	10 (4)	2 (20)	157	793
(40-50]	423 (206)	64 (15.1)	477	2123	60 (17)	5 (8.3)	419	1948
(50-60]	822 (443)	140 (17)	336	2117	207 (96)	24 (11.6)	629	1966
(60-70]	731 (430)	132 (18.1)	306	1606	315 (178)	41 (13)	451	1431
(70-80]	530 (433)	134 (25.5)	116	1113	392 (294)	87 (22.2)	168	1102
(80+)	251 (234)	95 (37.9)	6	782	406 (374)	170 (41.9)	3	950
Total	2858 (1790)	579 (20.3)	174	2612	1390 (963)	329 (23.7)	115	1966

Table 2.4: *The total number of participants with MI (the number of part. with MI that have died), Total death in week 1 (in %), Median number of weeks participants lived before death, and maximum number of weeks before death grouped by age and gender*

the participants that have had MI and are still alive. Including the participants that are alive after the last registered death gives a more complete data for the analysis of life span of the participants who had suffered from MI. In order to utilize the dataset that includes the participants that are alive and have suffered from MI, survival analysis models have to be applied to the dataset. Next, the structuring of the data that will be used in chapter 5 will be described.

All of the participants that have suffered form MI are included in the data. Each participant is given an ID number from 1 to 4248. Time from the incidence of the MI to death for the dead participants is given in days and weeks in parenthesis. The number of days from the incidence MI to the last registered death is given for the participants that were alive at the end of death registration. Counting the days/weeks starts at the time of MI. Therefore, the smallest value the time to column can have is 1. Furthermore, death of participants is marked by 1 in the censor column, while the participants who are alive are assigned the number 0 in the censor column. Gender of the participants is also included as part of the dataset. Number 1 is given to males and 2 to females. In addition, Age of

ID	Time to (weeks)	Censor	Gender	Age MI	Year MI	Month MI
1	4685 (670)	0	2	63	2004	5 (Spr)
2	1 (1)	1	2	60	2005	1(win)
3	2609(373)	0	2	64	2010	2(win)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4247	1247(179)	0	1	51	2013	10(win)
4248	5083(727)	0	2	59	2003	4(win)

Table 2.5: *Structuring of the data which will be used in the survival analysis models in chapter 5*

the participants, year and season at the time of the incidence of the MI included. The data in table 2.5 is the data behind the survival analysis in this thesis.

2.3 Preliminary analysis of occurrence of MI

In this section, the difference in the occurrence of MI with age and gender will be investigated. In addition, the monthly and seasonal variations of the occurrences of MI and death will be examined. This involves standard hypothesis tests described in Walpole et al. (2013).

2.3.1 Age and gender

Of the total participants in the first six Tromsø study surveys, 49.9 % are males and 50.1 % females. Even though the percentage distribution of males and females is similar, more than twice as many males have had MI compared to the female participants. In table 2.6, the total number of males and females who have had MI are displayed.

A hypothesis test with the null and alternative hypothesis in (2.1) is used to determine if the observed difference is significant. The null hypothesis states that

	Males	Females	Total
Total Participants	19896	19974	39870
Participants with MI	2858	1390	4248

Table 2.6: *Number of males and females in the Tromsø study survey who have had MI*

the proportion of men p_m that have had MI is equal to the proportion among women.

$$\begin{aligned} H_0 : &\quad p_m = p_f \\ H_1 : &\quad p_m \neq p_f. \end{aligned} \tag{2.1}$$

Due to the large sample sizes of both genders, the difference between the point estimates are approximately normal distributed with $\mu_{p_m - p_f}$ and $\sigma^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$. The hypothesis test in (2.1) is done using the test statistic,

$$Z = \frac{\hat{p}_m - \hat{p}_f}{\sqrt{\hat{p}(1-\hat{p})(1/n_m + 1/n_f)}}, \tag{2.2}$$

where Z is a standard normal distribution, n_m is the number of male participants, n_f is the number of female participants and $\hat{p}$ is the pooled estimate of both male and female participants. With $Z \approx 24$, the null hypothesis can be rejected in favour of the alternative hypothesis with $p - value \ll 0.01$.

The earliest a participant has had MI in the Tromsø study dataset is at the age of 22, while the oldest was 100 years. Figure 2.4A shows that women tend to have MI at a higher age than men.

The median age at the first MI is 65 years, 74 years for the female participants and 61 for males. The mean age is 72 years for the female and 62.3 years for the male participants. The dotted lines in figure 2.4*A* mark the average ages of the occurrence of the first heart attacks of the subjects.

The difference between mean ages at the time of MI males and femalesis tested by using (2.3). The null hypothesis states that the expected mean age of the male

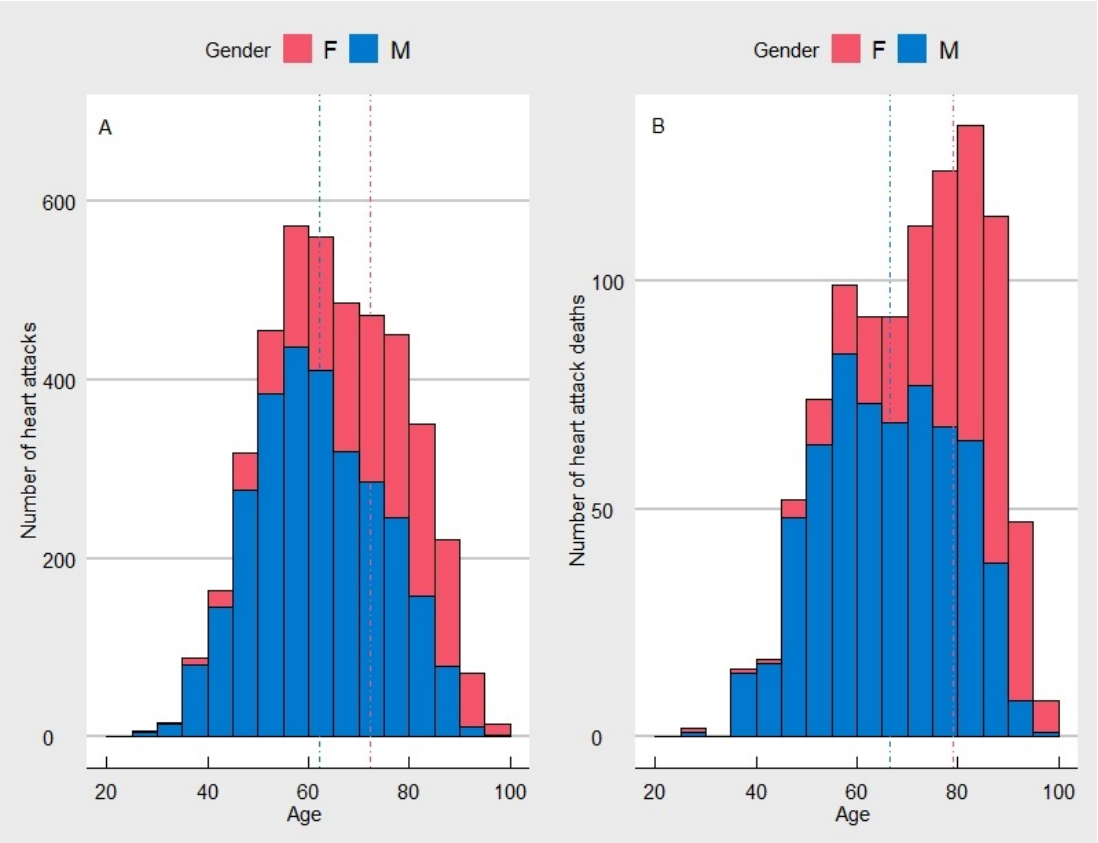


Figure 2.4: *A) The distribution of ages at the first heart attacks by gender. The dotted lines indicate the average age of the subjects at their first heart attack B) The distributions of ages at the time death for participants who died within three weeks of their first heart attack by gender.*

μ_m and female μ_f participants at the time of MI are equal, while the alternative hypothesis states the opposite, i.e.

$$\begin{aligned} H_0 : &\quad \mu_m = \mu_f \\ H_1 : &\quad \mu_m \neq \mu_f. \end{aligned} \tag{2.3}$$

The hypothesis test is based on the test statistic ,

$$T = \frac{\hat{\mu}_m - \hat{\mu}_f}{\sqrt{(s_m^2/n_m + s_f^2/n_f)}}. \tag{2.4}$$

with unknown and different sample variances s_m^2 and s_f^2 for males and females respectively.
And T is approximately t-distributed with ν degree of freedom defined as follows:

$$\nu = \frac{(s_m^2/n_m + s_f^2/n_f)}{(s_m^2/n_m)^2/(n_m - 1) + (s_f^2/n_f)^2/(n_f - 1)},$$

where $\hat{\mu}_m$ and $\hat{\mu}_f$ are the average age at the time of MI for males and females respectively. With $t = 19.035$,and $\nu \approx 38575$, the null hypothesis is rejected with $p - value \ll 0.01$.

2.3.2 Monthly and seasonal variation

AVOVA will be performed using random complete block (RCB) design to test for monthly or seasonal differences in incidence of MI. In addition, when the ANOVA reveals significant differences between population means, Tukeys procedure for comparison of mean will be utilized to test which populations exhibit the difference. Therefore, a short introduction to the method will also be given. ANOVA tests the difference between more than three population means. This is done by performing a hypothesis test. The null and alternative hypotheses are in the case of the monthly death rate variations is given by

$$\begin{aligned} H_0 &: \quad \mu_{1.} = \mu_{2.} = \ldots = \mu_{12.} \\ H_1 &: \quad \text{The } \mu_{m.} \text{ are not equal.} \end{aligned} \tag{2.5}$$

The RCB design is used when the effects of the factor of interest is affected by a second factor that is not of interest. The primary factors are then grouped by each secondary factor. Such groups are known as blocks. The RBC design is used in order to get a clearer vision of how the monthly death rates change with minimal impact from the annual fluctuations. RCB design is used with the monthly variation in the incidence of MI as the primary effect and the annual increase in the incidence of MI as the secondary factor. The annual effects are

then blocked off to control their effects on the between month variations. This can be done by grouping all the monthly rates in a year in table 2.7.

Month	**Block:**	**1975**	**1976**	...	**2016**	**Total**	**Mean**
1		$y_{1\,1975}$	$y_{1\,1976}$	$\cdots$	$y_{1\,2016}$	$T_{1\,.}$	$\bar{y}_{1\,.}$
2		$y_{2\,1975}$	$y_{2\,1976}$	$\cdots$	$y_{2\,2016}$	$T_{2\,.}$	$\bar{y}_{2\,.}$
3		$y_{3\,1975}$	$y_{3\,1976}$	$\cdots$	$y_{3\,2016}$	$T_{3\,.}$	$\bar{y}_{3\,.}$
$\vdots$		$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
12		$y_{12\,1975}$	$y_{12\,1976}$	$\cdots$	$y_{12\,2016}$	$T_{12\,.}$	$\bar{y}_{12\,.}$
Total		$T_{.\,1975}$	$T_{.\,1976}$	$\cdots$	$T_{.\,2016}$	$T_{..}$	
Mean		$\bar{y}_{.\,1975}$	$\bar{y}_{.\,1976}$	$\cdots$	$\bar{y}_{.\,2016}$		$\bar{y}_{..}$

Table 2.7: *RCB design*

The values of $y_{m\,j}$ are the sums of the daily number of MI within a month. Thus, due to the central limit theorem they can be assumed to be normal distributed with mean $\mu_{m\,j}$ and variance σ^2. The monthly averages denoted by $\mu_{m\,.}$ for $m = 1, 2, ..., 12$ are

$$\mu_{m,.} = \frac{1}{42} \sum_{j=1975}^{2016} \mu_{m,j}\,,$$

for $j = 1975, 1976, ..., 2016$. The annual averages $\mu_{.\,j}$ also known as the average of the j^{th} block are

$$\mu_{.\,j} = \frac{1}{12} \sum_{m=1}^{12} \mu_{m\,j}\,.$$

Whereas, the total mean μ is

$$\mu_{..} = \frac{1}{12 \times 42} \sum_{m=1}^{12} \sum_{j=1975}^{2016} \mu_{m\,j}\,.$$

As part of the hypothesis test the various sums-of-squares are defined as follows:

$$
\begin{aligned}
SST &= \sum_{m=1}^{12}\sum_{j=1976}^{2016}(y_{m\,j}-\bar{y}_{..})^2 \qquad \text{total sum of squares}\\
SSM &= 12\sum_{m=1}^{12}(y_{m\,.}-\bar{y}_{..})^2 \qquad \text{monthly sum of squares}\\
SSB &= 42\sum_{j=1976}^{2016}(y_{.\,j}-\bar{y}_{..})^2 \qquad \text{block sum of squares}\\
SSE &= \sum_{m=1}^{12}\sum_{j=1976}^{2016}(y_{m\,j}-\bar{y}_{m\,.}-\bar{y}_{.\,j}+\bar{y}_{..})^2 \quad \text{error sum of squares}
\end{aligned}
$$

and

$$
\begin{aligned}
SST &= SSM + SSB + SSE\\
s_m^2 &= \frac{SSM}{dof_m} \quad s_j^2 = \frac{SSB}{dof_y} \quad f = \frac{s_m^2}{s_j^2}
\end{aligned} \tag{2.6}
$$

The null hypothesis is rejected in favour of the alternative hypothesis at α level of significance if $f = \frac{s_m^2}{s^2} > f_\alpha[dof_m, dof_m \cdot dof_y]$.

Figure 2.5 and figure 2.6 show the monthly variation in the incidence rate of MI. The black line in figure 2.5 shows the total monthly MI from first registered MI to the last one. The largest number of MIs were registered in January with 411 cases. June had the smallest number of MIs with 315 cases. The incidence rates of MI have a local maximum in July.

The red line in figure 2.5 shows the total number MIs registered in each month multiplied by $\frac{30}{\text{days per month}}$, and the MIs from 1963 to 2013 are added to make sure each month has equal number of observations.

The within and between month variation are graphically displayed in the box plot in figure 2.6. It looks like the between month variation become less prominent when presented with the within month variations. The hypothesis test introduced in (2.5) is used to test whether there is monthly difference in the incidence rate of MI. Using RCB design, the sum-of-squares and test statistic defined in (2.6), the test is carried out. The results of the test is summarized in table 2.8.

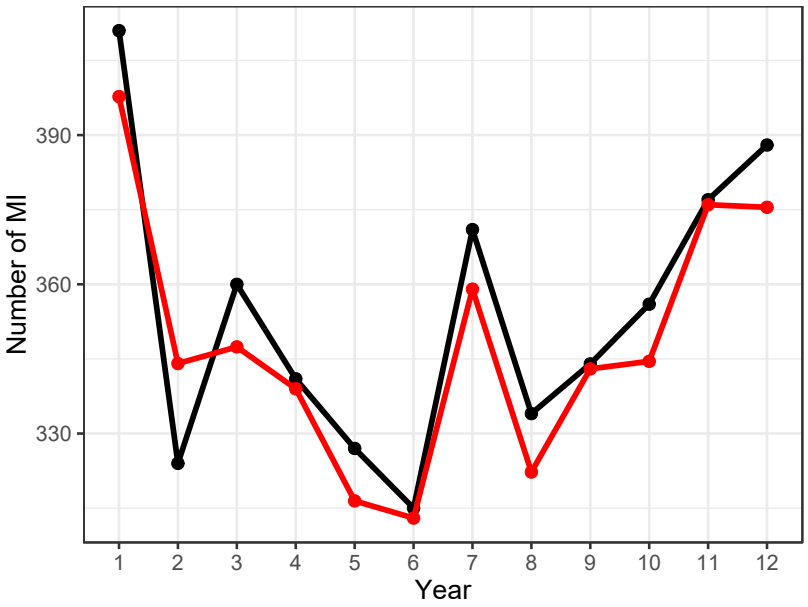


Figure 2.5: *The number of MI in each month summed up is shown in black, while the red line shows the number total number of death adjusted for the number of days and taking equal number of months(Jan-1963 to Des-2013)*

Variation	**Sum of sq**	**Dof**	**Mean sq**	**f** $(P(>f))$
Month	140.2	11	12.75	2 (0.03)
Block(Year)	20363.4	50	407	
Error	3558	550	6.47	

Table 2.8: *ANOVA table for the monthly variation in incidence rate of MI using the RCB design*

The results presented in table 2.8 reveal that $f = 2$, and that the null hypothesis in (2.5) can be rejected in favour of the alternative hypothesis with the $P-value = 0.03$.

To further investigation on which months are significantly different the Tukey's procedure is used. Tukey's procedure is used to test paired comparisons by making a simultaneous confidence interval for all pairs. This ensures the preservation

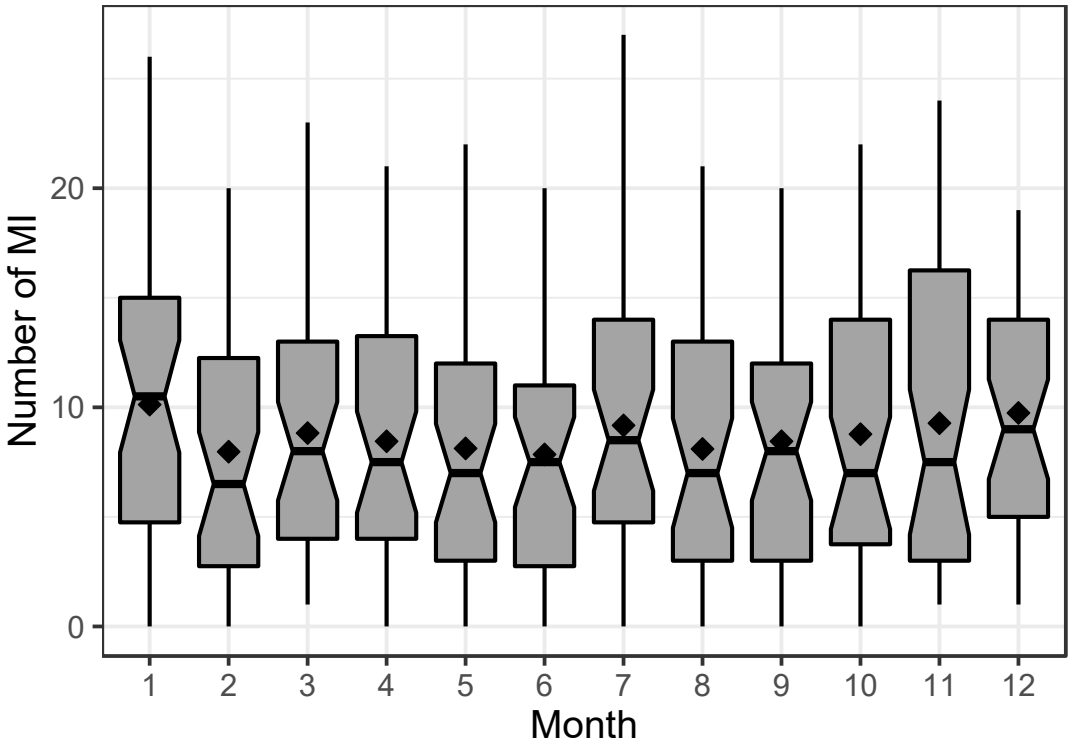


Figure 2.6: *The monthly variation in the number of MI displayed in a boxplot. The black dot localizes the mean*

of the rate of type I error. Meaning, the preservation of the type I error rate means that there is a probability α that at least one of the pairs will be falsely found to be different. The simultaneous confidence interval is constructed based on the number of means to be compared (k), the significance level (α) and the degrees of freedom(v) (Tukey, 1949).

Using Tukey's procedure, the difference between means is found significant if $|y_{i,.} - y_{j,.}| > q(\alpha, k, v)\sqrt{\frac{s^2}{n}}$, where $q(\alpha, k, v)$ is the upper quantile for α based on the studentized range distribution (Tukey, 1949).

Since at least two of the monthly MI rates are unequal, Tukey's procedure was applied to find which months were significantly different. There test revealed that significant differences were found only between the months of January and June with P-adj=0.047.

An alternative way to test for seasonal differences is to study the incidence rates of MI during the winter, spring, summer and autumn. The four season in Tromsø according to the Norwegian weather forecasting site `https://yr.no/` are divided as; winter between November 6 and April 10, spring between April 11 and June 22, summer between June 23 and August 24 and autumn between August 25 and November 5. Following the seasonal division, the number of MI in each seasons were summed and presented by the black line in the right panel of figure 2.7. The number of MI during the winter is far are larger than any of the other seasons. However, there are many more days in the winter in Tromsø than in the other seasons. To adjust for those differences the total number of MI in the season were divided by the number of days in the seasons and multiplied by 30. The result is presented by the red line in the same figure.

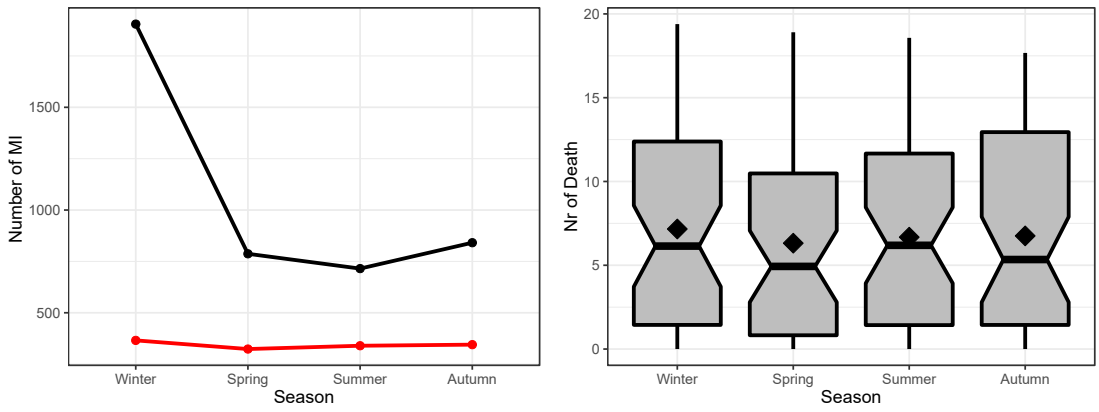


Figure 2.7: *The number of MI in each season summed up is shown in black and the adjusted monthly incidence rates of MI of each season in red is presented (left). The seasonal variations and within season variations diplayed in a boxplot. The black dot denotes the mean seasonal MI rates per 30 days (right).*

Based on the seasonal division presented earlier, the total numbers of MIs in each season are added together and presented in figure 2.7. The majority of the MI occurred in the winter season, while the smallest number of MIs occurred during the summer. However, after adjusting for the number of days and the total

number of seasons (starting in Jan. 1963 and ending in Des. 2013), the least number of MI occurre in the spring and the most in the winter. Moreover, the between variations seem smaller when presented with the large within variations. The following hypothesis test was conducted to verify if the seasonal differences are significant by blocking for the year effects:

$$\begin{aligned} H_0 &: \quad \mu_{Wi.} = \mu_{Sp.} = \mu_{Su.} = \mu_{Au.} \\ H_1 &: \quad \text{The } \mu_{seasonal,.} \text{ are not equal.} \end{aligned} \tag{2.7}$$

The ANOVA method was applied to investigate the significance of the seasonal differences. This was done using the hypothesis test with the null and alternative hypothesis presented in (2.7). The annual differences are blocked as in table 2.7 and the sum of squares and test statistic defined in (2.6) are used in the testing the hypothesis with. The results are presented in table 2.9

Variation	**Sum of sq**	**Dof**	**Mean sq**	**f** $(P(> f))$
Season	18.5	3	6.156	2.88 (0.038)
Block(Year)	6593.2	50	61.71	
Error	320.5	150	2.137	

Table 2.9: *ANOVA table for the Seasonal variation in incidence rate of MI using the RCB design*

The result of the ANOVA method presented in 2.9 reveals the null hypothesis can be rejected in favour of the alternative hypothesis that states that least to of them are different with P-value= 0.038.
Since at least two of the seasonal means were different, the Tukey's method for comparing means was applied. The results are presented in table 2.10. It shows that there is significant difference between the occurrence rates of MI in winter and spring seasons with $P_{adj} = 0.021$.

	- Spring		**- Summer**		**- Autumn**	
	Diff	P_{adj}	Diff	P_{adj}	Diff	P_{adj}
Winter	0.847	**0.021**	0.488	0.334	0.412	0.488
Spring	-	-	-0.359	0.603	-0.435	0.438
Summer	-	-	-	-	-0.077	0.993

Table 2.10: *Results of Tukey's comparison of mean monthly incidence rates of MI in the four seasons*

Chapter 3

Theory

In this chapter, the Bayesian inference techniques are presented. The presentations starts with a brief introduction in section 3.1. Following the introduction, in section 3.2 Bayesian modelling framework is described. Then, the latent Gaussian models are presented in section 3.12. Thereafter, the INLA methodology is explained in section 3.4. Section 3.5 presents the summary statistics. Finally the chapter is closed by a discussion about prior choises in section 3.6 and the presentation of the R-INLA package in sec 3.7

3.1 Introduction

In the reminder of this thesis the statistical analysis will be mainly conducted using Bayesian methods. The main goal of statistical inference in the Bayesian setting is to compute probability densities for the unknown parameters and the quantities yet to be observed. The prior beliefs about the parameters and unobserved quantities are updated using the available data. Bayes Theorem (3.1) plays a central role in incorporating the prior knowledge and the information in the data. In Bayesian statistics, probability is viewed as the measure of uncertainty and thus is subjective. Hence, the unknown parameters are viewed as random variables rather than fixed values. Probabilistic statements can therefore

be made about model parameters (Gelman et al., 2013)

3.1.1 Bayes Theorem

Bayes theorem was first discovered by Thomas Bayes and Pierre Simone Laplace and states that

$$\pi(B|A) = \frac{\pi(A \cap B)}{\pi(A)}, \tag{3.1}$$

where $\pi(A) = \sum_i \pi(A|B_i)\pi(B_i)$ for $B_i \cap B_j = \emptyset$ when $i \neq j$.
The application of Bayes theorem can be expanded beyond the estimation of probability of events. It can be used to infer the probability distribution of unknown parameters. First, the full joint probability distribution of the data and parameters has to be specified as follows:

$$\pi(\boldsymbol{\theta}, \boldsymbol{y}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta}),$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution of the parameter $\boldsymbol{\theta}$ and $\pi(\mathbf{y}|\boldsymbol{\theta})$ is the sampling distribution. The posterior distribution of the parameter $\boldsymbol{\theta}$ given the observed data $\mathbf{y}$ and the prior distribution $\pi(\boldsymbol{\theta})$ is:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})}{\pi(\mathbf{y})} \propto \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta}), \tag{3.2}$$

where $\pi(\boldsymbol{y})$ is the marginal or prior predictive distribution of the observed data.The marginal posterior distribution $\pi(\boldsymbol{y}) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})$ for discrete $\boldsymbol{\theta}$, whereas $\pi(\boldsymbol{y}) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ for the continuous case. The marginal distribution is independent of the model parameters, and works as a normalizing constant. Exchangeability of random variables and the structuring of the data is an important part of Bayesian data analysis.

3.2 Bayesian Modelling

3.2.1 Exchangeability

Exchangeability of random variables plays the same role in Bayesian statistics as independence of random variables do in classical statistics. Independence is a stronger mathematical condition than exchangeability and therefore implies exchangeability (Cordani and Wechsler, 2006). A sequence of random variables or model parameters are said to be exchangeable in their joint distribution if

$$\pi(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \pi(\boldsymbol{y}_{\Omega(1)}, ..., \boldsymbol{y}_{\Omega(n)})$$

for all combinations of Ω and for all subsets of the random variables or model parameters (Bernardo, 1996). Exchangeability in the joint distributions of random variables and model parameters can be assumed if there is no meaningful information in the indexes of the aforementioned random variables and model parameters (Gelman et al., 2013). If observations $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are exchangeable random variables, then

$$\pi(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \int \left(\prod_{i=1}^{n} \pi(\boldsymbol{y}_i|\boldsymbol{\theta}) \right) \pi(\boldsymbol{\theta}) d(\boldsymbol{\theta}) \tag{3.3}$$

where $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is a parametric model from which the $\boldsymbol{y}$'s are drawn from and $p(\boldsymbol{\theta})$ represents the prior probability distribution of $\boldsymbol{\theta}$. The concept of exchangability of variables and parameters plays a vital role in the Bayesian approach to data analysis and justifies the presence of prior distribution (Bernardo, 1996).

In many practical applications, modelling the data as a single sequence of exchangeable random variables is an over simplification (Bernardo and Smith, 1994). If a sequence of random quantities can be subdivided into exchangeable groups, then each group can be assigned a group specific prior distribution. In other words, given a group identity, all observations in the given group are exchangeable. Data structured in such a manner is referred to as partially exchangeable. Moreover, a hierarchy can be introduced if the parameters of the group specific

prior distribution are also exchangeable. Hyperprior distributions are assigned to the parameters of the prior distributions (Gelman et al., 2013).
A common application of hierarchical model is when random variables $\boldsymbol{y}_1, ..., \boldsymbol{y}_J$ depend on some observed values $\boldsymbol{z}_1, ..., \boldsymbol{z}_J$ for j=1,...,J . The observed values are often referred to as covariates. In such cases, the random quantities $\boldsymbol{y}_j$'s are not exchangeable, however the pairs $(\boldsymbol{y}_j, \boldsymbol{z}_j)$ are exchangeable. Hence, the observations are conditionally exchangeable (Gelman et al., 2013; Bernardo and Smith, 1994).

3.2.2 Structured additive regression models

Regression models are one of the most useful and practically applicable ways of studying the relationship between variables. Ordinary linear regression models are the simplest regression models where the conditional distribution $\pi(y_j|\boldsymbol{z}_j, \boldsymbol{\theta}_j)$ of the response variables $\boldsymbol{y}$ given the explanatory variable $\boldsymbol{z}$ are normally distributed and has a constant variance. The response variables given the explanatory variables have to be exchangeable given all the relevant information that distinguishes the response variables from each other are included in the explanatory variables. In other words, the pairs $(\boldsymbol{z}, y)_j$ are exchangeable (Gelman et al., 2013).
Ordinary linear regression models cannot be applied to response variables that are not normally distributed. Generalized linear models (GLM) widen the practical applicability of ordinary linear models beyond response variables with normal distributed error terms. This is achieved by connecting the linear predictor to the response variable through a link function $g(\cdot)$ (Gelman et al., 2013). The distribution of a conditionally independent response variable y_i given the covariates belongs to the exponential distribution if its density function can be presented in the following form:

$$\pi(y_i|\theta, \phi) = exp\left(\frac{y\theta_i - b(\theta_i)}{a(\phi)} w_i + c(y_i, w_i, \phi)\right). \tag{3.4}$$

For response variables that are members of the exponential family, the expected values and variances are:

$$E[y_i] = \mu_i = b'(\theta_i) \qquad var[y_i] = \sigma_i^2 = \phi b''[\theta_i],$$

where $\theta_i = g(\mu_i)$ is known as a canonical parameter, $\phi > 0$ is a dispersion parameter which is independent of the link function and $g(\cdot)$ is a link function commonly referred to as a canonical link function (Fahrmeir et al., 2013).
Generalized linear mixed models (GLMM) are a further expansion on the GLM by incorporating independent random effects component in the linear predictor (Fahrmeir et al., 2013).
A more general formulation is structured additive regression models which also includes non-linear random effects and temporal trends. Therefore, structured additive regression models can be viewed as a collection of the most commonly applied parametric and semi-parametric regression models. They can be fitted to conditionally exchangeable response variables y_i with a probability density function that belongs to the exponential family (Fahrmeir et al., 2013). The regression model for the response variables $\boldsymbol{y}$ with the predictor $\boldsymbol{\eta}$ is

$$g(\mu_i) = \eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f_k(c_{ki}) + \epsilon, \text{ for i} = 1, ..., n \tag{3.5}$$

where α is the intercept, $\{\beta_j\}$ are the linear effects of covariates $\boldsymbol{z}_j$ and $\{f_k(\cdot)\}$ are non-linear trends or non-linear functions of covariates $\boldsymbol{c}_k$. The non-linear trends can for example represent autoregressive time series models, models for smoothing and spatial effects. They can also include independent and identical normal distributed random effects (Fahrmeir et al., 2013).
In the Bayesian framework a prior distribution is given to all the parameters $\{ \eta_1, .., \eta_n \ , \mu, \beta_1, ..., \beta_{n_\beta}, f_1(\cdot), ..., f_{n_f}(\cdot) \ \}$ and the hyperparameters $\boldsymbol{\theta}$. The prior distribution is thus:

$$\pi(\eta_1, .., \eta_n, \alpha, \beta_1, ...\beta_{n_\beta}, f_1(\cdot), ..., f_{n_f}(\cdot), \boldsymbol{\theta}) = \pi(\eta_1, .., \eta_n, \alpha, \beta_1, ...\beta_{n_\beta}, f_1(\cdot), ..., f_{n_f}(\cdot)|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

The posterior distribution of the parameters is

$$\pi(\eta_1, .., \eta_n, \alpha, \{\beta\}, \{f_k(\cdot)\}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \prod_{i=1}^{n} \pi(y_i|\eta_i)\pi(\eta_1, .., \eta_n, \alpha, \{\beta\}, \{f_k(\cdot)\}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{3.6}$$

3.3 The Computational framework: Latent Gaussian Models

The goal of Bayesian inference is to obtain the marginal posterior distribution of the unknown parameters and hyperparameters in (3.6). Analytical derivation of the posterior distributions for complicated and multi-dimensional Bayesian models such as the structured additive regression models are usually inconvenient and computationally infeasible. The progress made in computational power of computers over the last half a century has however made fast simulation of random processes and numerical integration possible. These advances has made the Bayesian modelling a viable method for increasingly complicated probability models (Gelman et al., 2013).

Simulation-based Markov chain Monte Carlo (MCMC) method is a popular choice for the computation of the marginal posterior distribution of the unknown parameters and hyperparameters in (3.6). However, the development of integrated nested Laplace approximation (INLA) over the last decade has presented the users with a quicker, more accurate and more user friendly alternative for a subgroup of the structured additive regression models known as latent Gaussian models (Rue et al., 2009).

3.3.1 Latent Gaussian Models

Latent Gaussian Models (LGM) are a subclass of Bayesian structured additive regression models with a Gaussian prior distribution assigned to the intercept α, the fixed effects $\{\beta\}$ and non-linear effects $\{f_k(\cdot)\}$ in (3.6). The predictors and

the parameters of additive effects together are known as the latent field (Rue et al., 2009).

LGMs are three stage hierarchical models. The first stage specifies the conditionally exchangeable observations y_i through the likelihood function $\prod_i \pi(y_i|\boldsymbol{x_i}, \boldsymbol{\theta_1})$, and the second stage defines the latent Gaussian field $\boldsymbol{x}$. Finally, the third stage specifies priors of the hyperparameters $\pi(\boldsymbol{\theta})$. Mathematically, these stages are summarized as follows:

$$\text{Stage 1. } L(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta_1}) = \prod_{i=1}^{n} \pi(y_i|\boldsymbol{x_i}, \boldsymbol{\theta_1}) \tag{3.7}$$

$$\text{Stage 2. } \boldsymbol{x}|\boldsymbol{\theta_2} \sim N(0, \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)) \tag{3.8}$$

$$\text{Stage 3. } \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \tag{3.9}$$

where $\boldsymbol{x}$ represents the unobserved components i.e. the latent field α, $\{\beta\}$,$\{f_k(\cdot)\}$ and η_i for $i = 1, ..., n$ with a Gaussian prior distributions when $\boldsymbol{Q}^{-1}(\boldsymbol{\theta_2})$ denotes the precision matrix (inverse of the covariance matrix) of the unobserved components and $\pi(\boldsymbol{\theta})$ is the hyperprior distribution (Rue et al., 2009).

3.3.2 Gaussian Markov Random Fields

The task of estimating large number of marginal posterior distribution has enormous computational cost. INLA produces fast and accurate results if the latent Gaussian fields has a Markov property. A Gaussian latent field $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is said to be Gaussian Markov random field (GMRF) if it has a multivariate Gaussian distribution in the following form:

$$\pi(x) = (2\pi)^{\frac{-n}{2}} |\boldsymbol{Q}|^{\frac{1}{2}} exp\left[\frac{-1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})\right], \tag{3.10}$$

with the mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$. The Markov property in (3.10) is embedded in the precision matrix $\boldsymbol{Q}$ as conditional independence between individual components x_i and x_j of the Gaussian latent field given all the other components $\boldsymbol{x}_{-ij} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$. If that is the case,

then

$$x_i \perp x_j \,|\, \boldsymbol{x}_{-ij} \quad \Leftrightarrow \quad Q_{ij} = 0$$

where Q_{ij} is the ij^{th} entry of the precision matrix $\boldsymbol{Q}$. Consequently, the precision matrices of GMRFs are often sparse (Rue and Held, 2005). Numerical operations such as factorization of the sparse precision matrices of GMRFs have high computational speed compared to their dense counterparts. The equations that involve the precision matrix such as the computation of the marginal variances depend on factorization of $\boldsymbol{Q}$. Therefore the Markov properties of the latent fields is vital in enhancing the computational speed of INLA (Rue et al., 2009).
The full Bayesian model for LGMs is thus:

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \prod_{i=1}^{n} \pi(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_1)|\mathbf{Q}(\boldsymbol{\theta}_2)|^{\frac{1}{2}} exp(\boldsymbol{x}^T \mathbf{Q}(\boldsymbol{\theta}_2)\boldsymbol{x})\pi(\boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|\mathbf{Q}(\boldsymbol{\theta}_2)|^{\frac{1}{2}} exp\left(\boldsymbol{x^T} Q(\boldsymbol{\theta}_2)^{-1}\boldsymbol{x} + \sum_{i=1}^{n} log\{\pi(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_1)\}\right) \quad (3.11)\end{aligned}$$

3.3.3 Laplace approximation

Laplace approximation of integrals is a major part of the INLA algorithm as the name suggests. The integral of a function is approximated by the density function of normal distribution. The process has two steps. The first step involves the Taylor series expansion of the log of the function to be integrated as follows:

$$f(x)dx = exp\{log[f(x)]\}dx$$

The second order Taylor series expansiongives :

$$log[f(x)] \approx log[f(x_0)] + (x - x_0)\frac{\partial log[f(x)]}{\partial x}|_{x=x_0} + \frac{(x}$$

where $x = x_0$ is the mode of log[f(x)]. Thus, $(x - x_0)\frac{\partial log[f(x)]}{\partial x}|_{x=x_0} = 0$

$$log[f(x)] \approx log[f(x_0)] + \frac{(x - x_0)^2}{2}\frac{\partial^2 log[f(x)]}{\partial^2 x}|_{x=x_0}$$

The second step is the approximation by normal distribution.

$$f(x)dx \approx exp\left(log[f(x_0)] + \frac{(x-x_0)^2}{2}\frac{\partial^2 log[f(x)]}{\partial^2 x}|_{x=x_0}\right)dx$$
$$= f(x_0)exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right)dx \tag{3.12}$$

Here, the term under the integration sign is Gaussian with $\mu = x_0$ and $\sigma = 1/\frac{\partial^2 log[f(x)]}{\partial^2 x}|_{x=x_0}$. Therefore, the function $f(x)$ can be approximated using the Gaussian distribution. Thus, any non negative and integrable function can be approximated by a Gaussian distribution using the result in (3.12). In this chapter, this Gaussian approximation given by the Laplace method will be referred to as Gaussian approximation.

3.4 The INLA Methodology

In Bayesian statistics, the goal is computation of the marginal posterior distribution of all the unknown quantities. To achieve this, INLA uses the combinations of numerical integration techniques and Gaussian approximation. At first, the joint posterior distribution of hyperparameters ($\pi(\boldsymbol{\theta}|\boldsymbol{y})$) has to be computed. Then, the marginal posterior distributions of the latent field $\boldsymbol{x}$ have to be computed. Each component of the latent field is given by:

$$\pi(x_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\, \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \tag{3.13}$$

Finally, the marginal posterior distribution of each hyperparameter θ_i, which is given by:

$$\pi(\theta_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}_{-j}} \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j}, \tag{3.14}$$

have to calculated (Rue et al., 2009, 2017)

Gaussian approximation in (3.12) is used in the approximation of the joint posterior distribution of the hyperparameters. The joint posterior distribution of the hyperparameters is given by:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \tag{3.15}$$

The joint distribution of the latent field given the observation and hyperparameters is:

$$\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}) \propto exp\left(\boldsymbol{x^T}Q(\boldsymbol{\theta}_2)\boldsymbol{x} + \sum_{i=1}^{n} log\{\pi(y_i|\boldsymbol{x_i},\boldsymbol{\theta}_1)\}\right),$$

The joint posterior of the latent field conditioned on the observation and hyperparameters is close to GMRF since the effect of conditioning on the observation only affects the diagonal elements of the precision matrix and the GMRF structure of the precision $\boldsymbol{Q}$ remains. Therefore, the Gaussian approximation is likely to perform well (Rue et al., 2017).

As a result, the joint posterior of the hyperparametes can be approximated by changing the denominator in (3.15) to its Gaussian approximation as follows:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \left.\frac{\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{x=x^*(\theta)} = \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$$

where $\boldsymbol{x^*(\theta)}$ is the mode of latent field which is calculated by an iterative Newton-Raphson method and $\tilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ is the Gaussian approximation of $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ at the mode (Rue et al., 2009).

Next, the marginal posterior distribution of each component of the latent field can be computed. The posterior distribution of the latent field is presented in (3.13). In order to get to the marginal posterior distributions of the latent field, $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$ has to be obtained from $\pi(\boldsymbol{x},|\boldsymbol{\theta},\boldsymbol{y})$. There are three alternatives to accomplish that task. The first and simplest method involves Gaussian approximation of $\pi(\boldsymbol{x},|\boldsymbol{\theta},\boldsymbol{y})$. Then obtaining the marginals distributions from $\pi(\boldsymbol{x},|\boldsymbol{\theta},\boldsymbol{y})$ and the marginal variance using the Cholesky decomposition. Choosing this approach prioritizes computational speed at the expense of accuracy, since the Gaussian approximation strategy is the quickest and least accurate of the three (Rue et al., 2009).

The second method is termed Laplace approximation and is the most accurate and with the highest computational cost of the three options. In the Laplace

approximation method each $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ is approximated by:

$$\begin{aligned}\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y}) &\propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \\ &\approx \left.\frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\tilde{\pi}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}_{-i}=x_i^*(x_i,\boldsymbol{\theta})} = \tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})\end{aligned}$$

where x_{-i} contains all the components of $\boldsymbol{x}$ except the i^{th}, $x_i^*(x_i, \boldsymbol{\theta})$ is the mode of $\pi(\boldsymbol{x}_{-i}|x_{,i}\boldsymbol{\theta}, \boldsymbol{y})$ and $\tilde{\pi}(\boldsymbol{x}_{-i}|x_i\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation of $\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$. The computational cost of the Laplace approximation method piles up since the approximation needs to be computed for each component of $\boldsymbol{x}$ and that the dimension of $\boldsymbol{x}$ exceeds the number of observations (Rue et al., 2009).

The third method is called the simplified Laplace approximation and it represents a compromise between the numerically fast Gaussian and the accurate Laplace approximation. Taylor series expansion of the Laplace approximation $\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ is used to correct skewness and the position of the mean of the Gaussian approximation. This is the default due to its accuracy and computational cost improvements on the above methods (Rue et al., 2009).

The computation of the marginal posterior distributions of the latent field is completed by integrating out the hyperparameters as in (3.13). This is done numerically by approximating $\pi(x_i|\boldsymbol{y})$ by taking the sum of θ with weights Δ_k as in (3.16) (Rue et al., 2009)

$$\pi(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{y}, \boldsymbol{\theta_{(k)}})\pi(\boldsymbol{\theta_{(k)}}|\boldsymbol{y})\Delta_k \tag{3.16}$$

There are two approaches to finding the evaluation points for the integration $\boldsymbol{\theta}_{(k)}$ and summation weights Δ_k. The first method is known as grid strategy, and it involves reparametrization of $\boldsymbol{\theta}$ using the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and negative Hessian matrix (the second partial derivatives) at the mode. The process leads to a set of points that locate where the high density of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is through reparametrized and mutually orthogonal variables. This method has high computational cost, and should only be used when the number of hyperparameters is low (Rue et al., 2009).

When the number of hyperparameters exceeds 6, the computational cost of the grid strategy becomes too high and an alternative is needed (Rue et al., 2017). The second strategy is termed Central Composite Design strategy (CCD) and speeds up the integration process with minor loss of accuracy to the approximation. In this approach, the negative Hessian and the mode of $p(\boldsymbol{\theta}|\boldsymbol{y})$ is used to locate integration points (fewer integration points than in the grid strategy) around the center to be evaluated using eq. (3.16) (Rue et al., 2009).

The joint posterior distribution of the hyperparameters is sufficient if the aim is to compute the marginal posterior distribution of the latent field. However, the marginal posterior distributions of the hyperparameters can be of interest. The grid and CCD strategy can be used to integrate out $\boldsymbol{\theta}_{-i}$ from $\boldsymbol{\theta}$ if the number of hyperparameters is small, since

$$\pi(\theta_i|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-i} \tag{3.17}$$

In both the grid and CCD strategy, the mode and negative Hessian matrix of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ has to be computed. Therefore, these values can be used to approximate $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ with a multivariate Gaussian distribution. Then, using a weighted mixture of Gaussian distributions with varying weights across different axes, skewness can be added to the multivariate Gaussian distribution of the joint posterior of the hyperparameters. Different weight can also be used for the values below and above zero in the same axis. In the end, the marginal distributions can be calculated using (3.17). This method is known as asymmetric Gaussian interpolation and adds minimal computational cost since some of the values already had been computed, but as the number of hyperparameters grow, the results become unstable (Martins et al., 2013).

The default method for approximating the marginal posterior distribution is a numerical integration free algorithm. In this method, the already computed mode and negative Hessian matrix of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ are used to approximate the joint posterior of the hyperparameters. This is then used to to compute the conditional expectation of $E(\boldsymbol{\theta}_{-j}|\theta_j)$ which can be used to explore θ_j. The last step involves

approximation of the marginal posterior of the hyper parameters by a mixture of Gaussian distribution (Martins et al., 2013).

3.5 Summary Statistics

Once the marginal posterior distributions of the unknown parameters and hyperparameters are computed, the results are summarized and presented using a point estimate along with a credible interval. The point estimates of the parameters are usually given by the posterior mean, MAP-estimator or median of the marginal posterior distributions.

3.5.1 Point estimate

The posterior mean is one of the most commonly used point estimators in the Bayesian setting (Cowles, 2013) and is defined as:

$$\hat{\theta} = E(\theta|y) = \int_{-\infty}^{\infty} \theta\pi(\theta|y)d\theta.$$

The maximum a posteriori (MAP) estimate can also be used as a point estimator for an unknown parameter and is defined as:

$$\hat{\theta}_{MAP} = \underset{\theta}{argmax}\ \pi(\theta|y).$$

The posterior median is another option to finding a point estimate for the unknown parameter. The posteror median is the value of the parameters that fulfils the following condition:

$$\int_{-\infty}^{\hat{\theta}_{med}} \pi(\theta|y)d\theta - \int_{\hat{\theta}_{med}}^{\infty} \pi(\theta|y)d\theta = 0.$$

If the posterior distribution is unimodal and symmetric the posterior mean, MAP estimator and the posterior median produce equal point estimates (Cowles, 2013).

3.5.2 Bayesian credible intervals

The point estimates are usually accompanied by a credible interval (CI). CIs are the Bayesian equivalent to the frequentist confidence interval. However, the Bayesian CIs are computed from the marginal posterior distribution of the unknown parameters and they are interpreted differently from their frequentist counterparts. The Bayesian CI is interpreted as: The parameter has $100(1-\alpha)\%$ probability of being located in the given interval. Since there is a probability distribution of the parameters, there are countless ways of constructing a CI. The two most common types of CI are described below.

Equal-tailed credible interval

The $100(1-\alpha)\%$ equal-tailed Bayesian credible interval for the population parameter θ is $a < \theta < b$, where a and b are given by the following expression (Walpole et al., 2013).

$$\int_{-\infty}^{a} \pi(\theta|y)d\theta = \int_{b}^{\infty} \pi(\theta|y)d\theta = \frac{\alpha}{2}$$

The advantage to the equal-tailed credible interval is that it is easy to calculate. However, if the posterior distribution is not symmetric and unimodal, some points outside the credible interval will have higher density values than some of the points inside the interval. That leads to the equal tailed intervals not being the shortest possible intervals (Cowles, 2013).

Highest Posterior Density interval

Highest Posterior Density(HPD) intervals are an alternative to the equal tailed intervals. The points inside an HPD interval has higher pdf values than all the points outside the interval. Therefore, the HDP intervals are guaranteed to be the shortest intervals (Cowles, 2013). The HPD interval is defined by the region R

$$R = \{\boldsymbol{\theta}\,; \pi(\boldsymbol{\theta}) \geq c\}$$

where c is the largest number that:

$$\int_{\boldsymbol{\theta} \in R} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 - \alpha.$$

The equal tailed inteval and HPD region cover the same region if the posterior is symmetric and unimodal(Cowles, 2013).

3.6 Prior distributions

Bayesian statistical inference depends on the posterior distribution which is obtained by updating the prior beliefs by new evidence. In this section, the assignment of prior distribution to the unknown parameters will be discussed. Therefore, the choice of prior distributions and the inclusion of information to the model through these prior distribution plays a major role in Bayesian framework. The issue related to choosing a prior distribution for a parameter has been approached from different angels and the topic is vast. However, the various prior distributions can be broadly subdivided into non-informative, weakly informative and informative prior distributions. Non-informative prior distributions, also known as objective prior distributions, are designed to have minimal impact on the posterior distribution so that the data alone can be the source of inference. They often produce the same results as maximum likelihood estimates. Jeffreys' priors (Jeffreys, 1945) and reference priors (Berger et al., 2009) are examples of the non-informative prior distributions. In contrast,the informative prior distributions that aim to construct a prior distribution that reflect the current knowledge on the values of the parameters and the uncertainties that surround the knowledge about the parameters in question (Gelman et al., 2013). Alternatively, the weakly-informative prior distributions can be used. They constrain the parameters to a reasonable range of values without incorporating strong prior information (Gelman et al., 2013; Simpson et al., 2017).

3.6.1 The latent field

The prior distribution to the latent Gaussian fields (3.8) in the three stage hierarchical models described above is straight forward since they are defined to be Gaussian (Rue et al., 2017). In the following chapters vague (large variance) Gaussian prior distribution (non-informative) centred at zero will be used for the fixed effects and the intercept. In the structured additive regression models in (3.5), a prior has to be assigned to the model components $\{f_k\}$. Random walk models are often assigned as prior distributions to model temporal dependence (Besag et al., 1995; Knorr-Held and Rainer, 2001; Riebler and Held, 2010a)

Random walk models are a type of GMRF with precision matrices that are not of full rank, which makes them improper. The order of such GMRF, also known as intrinsic GMRF (IGMRF), are defined based on their rank deficiency. In one dimension, they are constructed using the forward difference of the order given by their rank deficiency (Rue and Held, 2005). Here, random walk of order 1 and 2 will be of interest.

A prior based on the first order random walk (RW1) is constructed using first order differences of $\boldsymbol{x} = (x_1, \ldots, x_n)$. The first order difference Δx_i is

$$\text{RW1: } \Delta x_i = x_i - x_{i-1} \sim N(0, \tau^2), \quad i = 2, \ldots, n$$

The joint distribution of $\boldsymbol{x}$ is also normal distribution and is given by

$$\begin{aligned} \pi(\boldsymbol{x}|\tau) &\propto \tau^{-(n-2)/2} exp\bigg(-\frac{1}{2\tau^2} \sum_i (x_i - x_{i-1})^2 \bigg) \\ &= \tau^{-(n-2)/2} exp\bigg(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q}(\tau) \boldsymbol{x} \bigg). \end{aligned} \tag{3.18}$$

The precision matrix $\boldsymbol{Q}(\tau)$ is

$$\boldsymbol{Q}(\tau) = \tau \boldsymbol{R}^{(rw1)}_{(dim\,\boldsymbol{x} \times dim\,\boldsymbol{x})} = \tau \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix},$$

AS a result, the RW1 gives $x_i|x_{i-1}, \ldots, x_1 \sim N(x_{i-1}, \tau^2)$, And a vague normal distributed prior is assigned to x_1. When the RW1 is assigned as a prior, deviation from a constant level is penalized. Therefore, the resulting estimated effect is smoothed. In addition, an increase in the parameter τ leads to a even smoother effects vector (Rue and Held, 2005).

Similarly, the second order random walk (RW2) is constructed using the second order difference of $\boldsymbol{x} = (x_1, \ldots, x_n)$. The second order difference is then given by

$$\text{RW2: } \Delta^2 x_i = (x_i - x_{i-1}) - (x_{i-1} + x_{i-2}) \sim N(0, \tau^2), \quad i = 3, \ldots, n$$

The joint probability distribution of $\boldsymbol{x}$ is:

$$\begin{aligned} \pi(\boldsymbol{x}|\tau) &\propto \tau^{-(dim\,\boldsymbol{x}-2)/2} exp\bigg(-\frac{1}{2\tau^2} \sum_i [(x_i - x_{i-1}) - (x_{i-1} - x_{i-2})]^2 \bigg) \\ &= \tau^{-(n-2)/2} exp\bigg(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q}(\tau) \boldsymbol{x} \bigg). \end{aligned} \tag{3.19}$$

And $\boldsymbol{Q}(\tau)$ is

$$\boldsymbol{Q}(\tau) = \tau \boldsymbol{R}^{(rw2)}_{(n\times n)} = \tau \begin{bmatrix} 1 & -2 & 1 & & & & & \\ -2 & -5 & -4 & 1 & & & & \\ 1 & -4 & 6 & -4 & 1 & & & \\ & 1 & -4 & 6 & -4 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & 1 & -4 & 6 & -4 & 1 \\ & & & & 1 & -4 & 5 & -2 \\ & & & & & 1 & -2 & 1 \end{bmatrix},$$

Here, $x_i|x_{x-1}, \ldots, x_1 \sim N(2x_{i-1} - x_{i-1}, \tau^2)$, and vague prior normal distributions are then assigned to x_1 and x_2. In this case, deviations from a linear trend is penalize. RW2 model leads to more smoothing than the RW1 model (Rue and Held, 2005).

3.6.2 Assigning priors to hyperparameters

Penalized complexity prior

In the following chapters penalised complexity (PC) prior distributions introduced by Simpson et al. (2017) will be used as prior distributions for the hyperparameters. PC prior distributions are weakly informative prior distributions Simpson et al. (2017) boiled down the desired properties of a prior distributions to four principles. The following four principles are used to design PC prior distribution for the hyperparameters.

∗ **Principle 1:** Occams Razor

The idea of a base and flexible models are crucial aspects of the first principle. A base model can be defined as the simplest class of models among models with the density $\pi(\boldsymbol{x}|\xi)$ and a flexibility parameter ξ. Simplicity for such models is defined by the flexibility parameter ξ and the base model of a class has $\xi = 0$. For instance, the base model for an iid model $N(0, \xi)$ would be that there is

no random effect ($\xi = 0$). As the value of the flexibility parameter increases, the flexible model increases in complexity. The first principle which is based on **Occam's razor** prefers the simpler base model $\pi(\boldsymbol{x}|\xi = 0)$ to the more flexible model $\pi(\boldsymbol{x}|\xi)$ (Simpson et al., 2017).

$*$ **Principle 2:** Measure of complexity

Building on the first principle, the second principle presents the **measure of complexity** between the flexible model and the base models. The information lost when a flexible model f is approximated with the simpler base model g is measured by the Kullback-Liebler divergence(KLD) which is defined as:

$$KLD(f(\boldsymbol{x}) \parallel g(\boldsymbol{x})) = \int f(\boldsymbol{x}) log\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right) d\boldsymbol{x}$$

where $g(\boldsymbol{x})$ represents the base model $\pi(\boldsymbol{x}|\xi = 0)$ and $f(\boldsymbol{x})$ represents the flexible model $\pi(\boldsymbol{x}|\xi)$ (Kullback and Leibler, 1951). The complexity of the flexible model compared to the base model is then measured by

$$d(f(\boldsymbol{x}) \parallel g(\boldsymbol{x})) = \sqrt{2\, KLD(f \parallel g)}$$

The distance $d(f(\boldsymbol{x}) \parallel g(\boldsymbol{x}))$ is used as the measure of when comparing the flexible model and the base model in that order. In other words, it is unidirectional (Simpson et al., 2017).

$*$ **Principle 3:** Constant rate penalisation

The third principle introduces a memoryless **constant rate penalization** that only depends on the relative deviation from a simpler model and has a constant decay rate. The prior distribution on the flexibility parameter is then given by:

$$\pi(\xi) = \lambda e^{\lambda d(\xi)} \left| \frac{\partial\, d(\xi)}{\partial\, \xi} \right|$$

where $d(\xi) = d(\pi(\boldsymbol{x}|\xi) \parallel \pi(\boldsymbol{x}|\xi = 0))$. Exponential distribution is therefore assigned to the probability distribution of the distance $d(\xi)$, and the mode is located at $d(\xi) = 0$ (Simpson et al., 2017).

∗ **Principle 4:** User-defined scaling

The forth and final principle allows the **user to define the scaling** of the flexibility parameter. This is done by allowing the user to control the value of U which is given by

$$Prob(Q(\xi) > U) = \alpha,$$

where $Q(\xi)$ is a transformation of the flexibility parameter and U reflects the value of the upper limit and α is the probability assigned to values exceeding the U (Simpson et al., 2017).
PC prior distributions helps the user design prior distributions for the hyperparameters that avoid overfitting by giving appropriate weight to the base model. In general, these prior distributions are weakly informative, and the informaiton supplied depends on the value assigned to U. However, they can also be used to construct informative prior distributions.

PC prior distribution for precision parameter

In both (3.18) and (3.19), the precision parameter τ is added to the model as a hyperparameter. Such precision matrices can be assigned PC prior distributions. Take for instance the RW2 model with the joint distribution in (3.19), has a base model that puts all the mass at the center with $\tau_0 = \infty$.
PC priors for the precision parameter of the random effect, RW1 and RW2 prior distribution can be constructed using the Kullback-Leibler divergence between two multivariate Gaussian distribution $\pi_1(\boldsymbol{x})$ and $\pi_2(\boldsymbol{x})$ with mean $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. It is is given by:

$$\begin{aligned} KLD\Big(\pi_1(\boldsymbol{x}) \parallel \pi_2(\boldsymbol{x})\Big) &= \int log\left(\frac{\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right)\pi_1(\boldsymbol{x})\,d\boldsymbol{x} \\ &= \frac{1}{2}\Bigg(tr(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - n - log\Big(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\Big) \\ &\qquad + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\Bigg), \end{aligned}$$

where n is the dimension of $\boldsymbol{x}$. The $KLD(\cdot \parallel \cdot)$ for the random effects, RW1 and RW2 are thus,

$$\begin{aligned} KLD\Big(\pi(\boldsymbol{x}|\xi) \parallel \pi(\boldsymbol{x}|\xi=0)\Big) &= \frac{n\tau_0}{2\tau} + \frac{n}{2} log\Big(\frac{\tau}{\tau_0}\Big) + \frac{n}{2} \\ &= \frac{n\tau_0}{2\tau}\left(1 + \frac{\tau}{\tau_0} log\Big(\frac{\tau}{\tau_0}\Big) - \frac{\tau}{\tau_0}\right), \end{aligned}$$

In this case, n is $dim\,\boldsymbol{x} - k$, where k is the rank deficiency of precision matrix $\boldsymbol{Q}$. Since the base model has $\tau \ll \tau_0$,

$$KLD\Big(\pi(\boldsymbol{x}|\xi) \parallel \pi(\boldsymbol{x}|\xi=0)\Big) = \frac{n\tau_0}{2\tau}$$

Consequently, the increase in complexity of the flexible model compared to the base model can be measured as:

$$d\Big(\pi(\boldsymbol{x}|\xi) \parallel \pi(\boldsymbol{x}|\xi=0)\Big) = \sqrt{2\,KLD\Big(\pi(\boldsymbol{x}|\xi) \parallel \pi(\boldsymbol{x}|\xi=0)\Big)} = \sqrt{n\frac{\tau_0}{\tau}}$$

Based on the third principle, the hyperprior for τ is then designed using an exponential prior on $d\Big(\pi(\boldsymbol{x}|\xi) \parallel \pi(\boldsymbol{x}|\xi=0)\Big)$. Hence,

$$\begin{aligned} \pi(d(\tau)) &= \lambda_d\, exp(\lambda_d\, d(\tau)) \\ \pi(\tau) &= \frac{1}{2}\lambda_d \sqrt{n\,\tau_0}\,\tau^{-3/2} exp(\lambda_d \sqrt{n\,\tau_0}\,\tau^{-1/2}) \\ \pi(\tau) &= \frac{\lambda}{2}\,\tau^{-3/2} exp(\lambda\,\tau^{-1/2}), \qquad \tau > 0, \lambda > 0. \end{aligned} \tag{3.20}$$

Evidently, the PC hyperprior disrtibution for the precision parameter is type-2 Gumbel distribution. Moreover, the PC prior for standard deviation is exponential distribution with the parameter λ.

In the end, the forth principle allows the user to define the parameters based on $\pi(1/\sqrt{\tau} > U) = \alpha$.

$$\pi(1/\sqrt{\tau} > U) = \alpha \quad \Leftrightarrow \quad \lambda = \frac{-ln(\alpha)}{U}$$

When the PC prior is being used, the parameters U and α specified.

Scaling the prior distributions

IGMRF's such as the RW1 and RW2 models are used as prior distributions to model dependency structure. These different types of prior distributions penalize different deviances. The RW1 penalizes deviance from a fixed level, while RW2 model penalizes deviance from a linear trend (Rue and Held, 2005). Thus, they have incomparable range of deviance. Furthermore, the size and shape of the precision matrices of the RW1 and RW2 models, affect their marginal variances (Sørbye and Rue, 2014). The marginal variances for an IGMRF with a random precision parameter τ are defined as

$$\sigma_\tau(x_i) = \frac{\sigma_{\{\tau=1\}}(x_i)}{\sqrt{\tau}} \approx \frac{\sigma_{ref}(\boldsymbol{x})}{\sqrt{\tau}}, \qquad \text{for } i = 1, \ldots, n,$$

where

$$\sigma_{ref}(\boldsymbol{x}) = exp\left(\frac{1}{n}\sum_{i=1}^{n} log\Big(\boldsymbol{\Sigma}^*_{ii}\Big)\right)$$

Here, $\boldsymbol{\Sigma}^*$ is the generalized inverse of the precision matrix $\boldsymbol{Q}$ and $\sigma_{ref}(\boldsymbol{x})$ is the geometric mean of the marginal variances with a fixed precision parameter $\tau = 1$ (Sørbye and Rue, 2014).

As a result, setting and interpreting a hyperprior distributions for different models or the same model with different size and/or shape precision matrices becomes very challenging for the unscaled IGMRF models (Sørbye and Rue, 2014).

Scaling of the RW models leads to the generalized variance $\sigma_{ref}(\boldsymbol{x})$ being equal to 1. This can be done by rescling $\boldsymbol{R}^{RW}$ using a parameter κ (Sørbye and Rue, 2014). Once the RW models are scale, the PC prior distribtution can be assigned to the scaled τ. The parameter τ can then be interpreted as the deviation from the null space of $\boldsymbol{R}^{RW}$ (Simpson et al., 2017). Hence, the scaled RW models are invariant to changes in the shape and size of the precision matrix $\boldsymbol{Q}$. In addition, the precision parameter τ is comparable across different models (Sørbye and Rue, 2014).

The PC prior for the scaled precision parameter is then set on :

$$\pi(\frac{1}{\sqrt{\tau}} > \frac{U}{\sigma_{ref}}) = \alpha \tag{3.21}$$

3.7 R-INLA

R-INLA is the R package through which the Bayesian inference with INLA methodology was implemented in this thesis. The test version of the package is often updated and it is available at `http://www.r-inla.org/` . Therefore, the test version was preferred to compute the various parameters in this thesis.

Chapter 4

Temporal analysis of the incidence rate of MI

The analysis performed in chapter 2 reveals that incidence rates of MI varies with time. In addition to the temporal variation, there is a gender related variation in the incidence rate MI. The incidence rate of MI is higher among males than females. Furthermore, the analysis revealed that MI strikes males at a younger age than females. This chapter provides further analysis of MI incidence rates using Bayesian Age-Period-Cohort(APC) models.

In section 4.1, a general introduction to the APC models is given. Then, a Bayesian Age-period (AP) model is introduced and the results of the analysis are presented in 4.2. Following the section about bayesian AP model, in section 4.3 multivariate Bayesian APC models are provided with the results of the analysis. This chapter is then concluded by a discussion of the results.

4.1 Age-period-cohort models

Temporal variations in disease and mortality rates can be attributed to changes in age, time of event (Period) and time of birth of the group of people encountering the event (birth cohort). The difference in the incidence rates of a given event

across age is referred to as age effect. The period effect describes the changes to the rate related to observation time of the event. Finally, the cohort effect explains variations in the rates due to differences in birth time. The age-birth-cohort(APC) models are often used to understand how these three effects affect the observed patterns and make predictions about future outcomes (Clayton and Schifflers, 1987b).

Univariate APC models are applied to disease or mortality data without stratifications such as gender and geographic locations (Riebler and Held, 2010a). In such cases, the total number of events y_{ij} at age i and time period j are Poisson distributed with the rate $n_{ij} \times \lambda_{ij}$,

$$y_{ij} \sim Pios(n_{ij} \times \lambda_{ij}),$$

where n_{ij} is the total number of people at risk in the given age and time period (Clayton and Schifflers, 1987b). Poisson distribution is a member of the exponential family of distributions with $E[y_{ij}] = \mu_{ij} = exp(\eta_{ij}) = \lambda_{ij}\, n_{ij}$, $var[y_{ij}] = \sigma^2_{ij} = exp(\eta_{ij}) = \lambda_{ij}\, n_{ij}$ and the conanical link function is $log(\cdot)$ (Fahrmeir et al., 2013). The univariate APC model is defined by the predictor η_{ij}.

$$\eta_{ij} = log(\lambda_{ij}) = \alpha + \varphi_i + \gamma_j + \psi_k, \tag{4.1}$$

where α is the intercept, φ_i is the age effect of i^{th} age for $i = 1, \ldots, I$, γ_j is the period effect of time j for $j = 1, \ldots, J$ and ψ_k is the cohort effect of the k^{th} birth cohort for $k = 1, \ldots, K$. The index of the cohort effect is $k = (I - i) + j$ and $K = (I - 1) + J$ when the period and age are equally spaced (Riebler and Held, 2010a).

Unique identification of the intercept term α in (4.1) requires additional constraints on the additive time effects. A sum-to-zero constraint is often imposed on the additive time effects to achieve that. The constraint imposes that $\sum_i \varphi_i = 0$, $\sum_j \gamma_j = 0$ and $\sum_j \psi_k = 0$ (Holford, 2005).

A second identifiability problem plagues the APC models due to the linear dependency between the time effects. If two of the three time effects are known,

the third effect is automatically given (since birth cohort $k = (I - i) + j$). The linear dependency of the age, period and cohort effects causes the presence of a higher number of parameters to be estimated, than the data can uniquely estimate (Clayton and Schifflers, 1987b). Consequently, by any transformations

$$\varphi_i^* \to \varphi_i + a\left(i - \frac{I+1}{2}\right),\ \gamma_j^* \to \gamma_j - a\left(j - \frac{J+1}{2}\right),\ \psi_k^* \to \psi_k + a\left(k - \frac{K+1}{2}\right) \tag{4.2}$$

with any a produce time effects that comply with the sum-to-zero constraints. Also, the transformed time effects result in the same predictor η_{ij}. In other words, a transformation of one of the time effects, can be countered by a weighted transformations of the two other effects to produce the same predictor η_{ij} (Riebler and Held, 2010a). This implies that, the individual time effects can not be distinctly identified (Holford, 2005). However, deviations from the linear trends can be identified (Clayton and Schifflers, 1987b).

4.2 Bayesian Age period model

By limiting the analysis to only two out of the three time effects, the linear dependency is avoided and the effects can be identified. In the quest to find how the incidence rate of MI vary with age and time, univariate Bayesian age-period (AP) model is used separately for males and females. To analyse the temporal patterns of incidence rates of MI, the AP model is a good option since the age composition of the Tromsø study participants is progressively increasing as illustrated by figure 2.2 A) and B) and age standardizing of the MI rates by the standard methods do not give a a satisfactory result.

Let y_{ijg} define the number of MI of age i, year j and gender g are assumed to be Poisson distributed with the rate $n_{ijg} \times \lambda_{ijg}$, where n_{ijg} is the number of participants that are eligible for their MI for the given age, year and gender.

To apply the AP model, groups with too few participants have to be omitted. Here, this includes males and females below the age of 40 and above the age of

80, time period up to 1980 and in addition the age/period groups with less than 100 total participants are omitted from this study due to the limited number of MI observed in those groups. Thus $i = 40, ..., 80$, $j = 1982, ..., 2014$ and $g = 1, 2$ for females and males respectively.

The Poisson distribution is equidispersed, meaning that the variance and mean are equal. For some practical applications, the equidispersion property of the poisson model can be too strong (Fahrmeir et al., 2013). However, the equidispersion assumption of the Poisson distribution can be relaxed using normal iid random effect ($N(0, \tau)$) (Besag et al., 1995).

The structured additive regression model described in(3.5) in chapter 3 for the response variable y_{ijg} with the predictor η_{ijg} is thus,

$$log(\lambda_{ijg}) = \eta_{ijg} = \alpha_g + \varphi_{i,g} + \gamma_{j,g} + \kappa_{ij} \tag{4.3}$$

where α_g is the gender specific intercept, $\varphi_{i,g}$ are the gender specific age effects, $\gamma_{j,g}$ are the gender specific period effects, κ_{ijg} are the iid normal random effects and the offset is given by the total number of eligible people for their MI, n_{ijg}.

The model is fitted to data in table 2.1 using the LGM framework described in chapter 3. The three stage hierarchical Poisson model in (4.3) is as follows:

$$\begin{aligned} &\text{Stage 1:} \quad y_{ijg}|\eta_{ijg} \sim Poisson\Big(n_{ijg} \times exp(\eta_{ijg})\Big) \\ &\text{Stage 2:} \qquad \boldsymbol{x}_g = (\boldsymbol{\eta}_g, \alpha_g, \boldsymbol{\varphi}_g, \boldsymbol{\gamma}_g, \boldsymbol{\kappa}) \\ &\text{Stage 3:} \qquad \boldsymbol{\tau}_g \sim \pi(\boldsymbol{\tau}_g), \end{aligned}$$

INLA is used to estimate the various parameters in the model. Since the parameter estimation is done in the Bayesian framework, prior distribution are assigned to the unknown parameters. As mentioned in chapter 3, independent Gaussian vague prior distributions are assigned to the fixed effects such as the gender specific intercepts and, while scaled random walk of second order in (3.19) is assigned to the gender specific time effects as similarity between time adjacent observation is assumed. PC prior in (3.21) are then assigned to the hyperparameters of the scaled time effects and the random effects. In addition, the Bayesian AP model is also fitted to different age groups (40-50,50-60,60-70 and 70-80) separately.

4.2.1 Results

After fitting the Bayesian AP model to the data, the marginal posterior distribution of the intercepts for females and males is computed(see figure 4.1). The posterior distribution of intercept is normal distributed. the intercept of the females is smaller than that of males with the posterior mean $\alpha_1 = -6.2$ with $sd_{\alpha_1} = 0.05$. The posterior mean for their male peers is $\alpha_2 = -4.9$ with $sd_{\alpha_2} = 0.03$ for males.

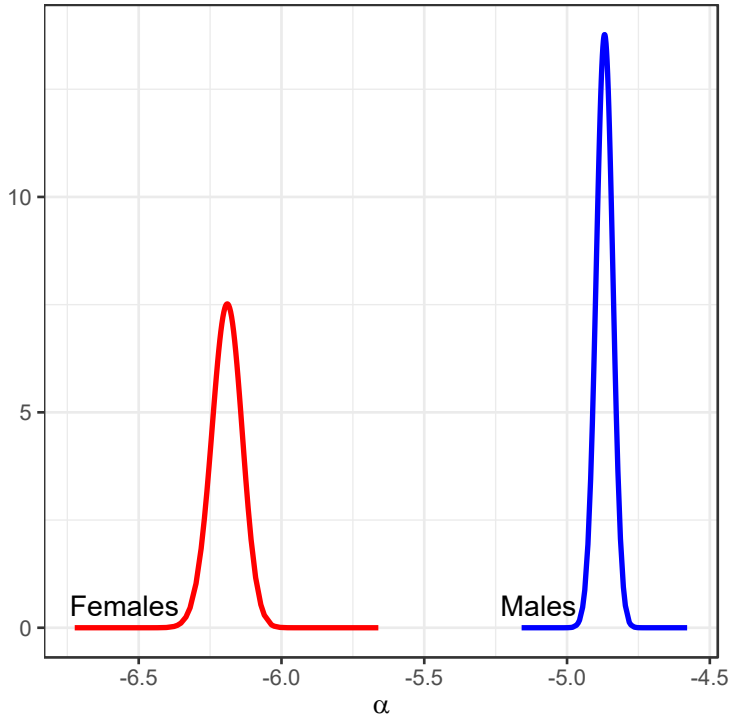


Figure 4.1: *The marginal posterior distribution intercepts of the Bayesian AP model for males and females*

Gender specific scaled age effects are presented in figure 4.2 with the red area marking the 95% credible interval of the age effects for females and the blue area for males between the ages of 40 and 80. Both males and females have a monotonically increasing age effects. An increase in age effect implies increase in the incidence rate of MI with age.

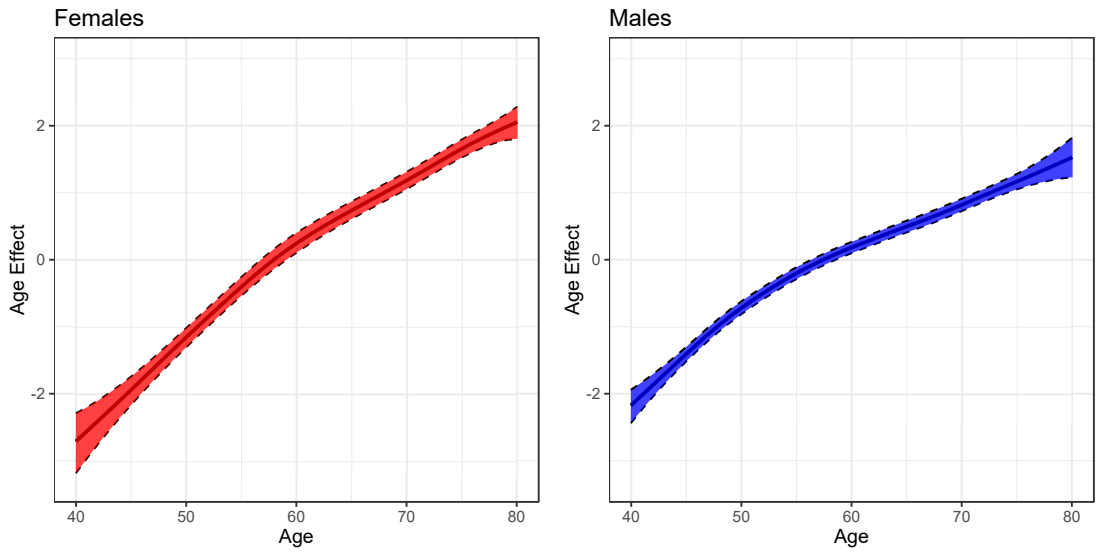


Figure 4.2: *The marginal posterior mean of age effects with a 95% credible interval of the Bayesian AP model for female (left) and male (right)*

Figure 4.3 displayes the scaled period effect of females (left) and males (right) with the shaded area representing 95% credible interval. The period effect of females has wider interval between 1980 and 1990. The number of MI are smaller in that period than the time after 1990 as can be seen in figure 2.2C). The period effect remains more or less constant until the year 2000 (a slight decrease for males). Around the year 2000, a major drop in the period effect can be observed for both sexes. Similar to the age effect, a decrease in the period effect is associated with a decrease in the incidence rate of MI. For both males and females, there is a decreasing incidence rate of MI after the year 2000.

Figures 4.4 in the left panel show the marginal posterior distribution of the variances of the age effect, and the right panel shows the marginal posterior distribution of standard deviation of the period effect for males (blue) and females(red). The standard deviations of both the age and period effects for both sexes are similar.

The RW2 model that was assigned as prior to the age effects are scaled to have

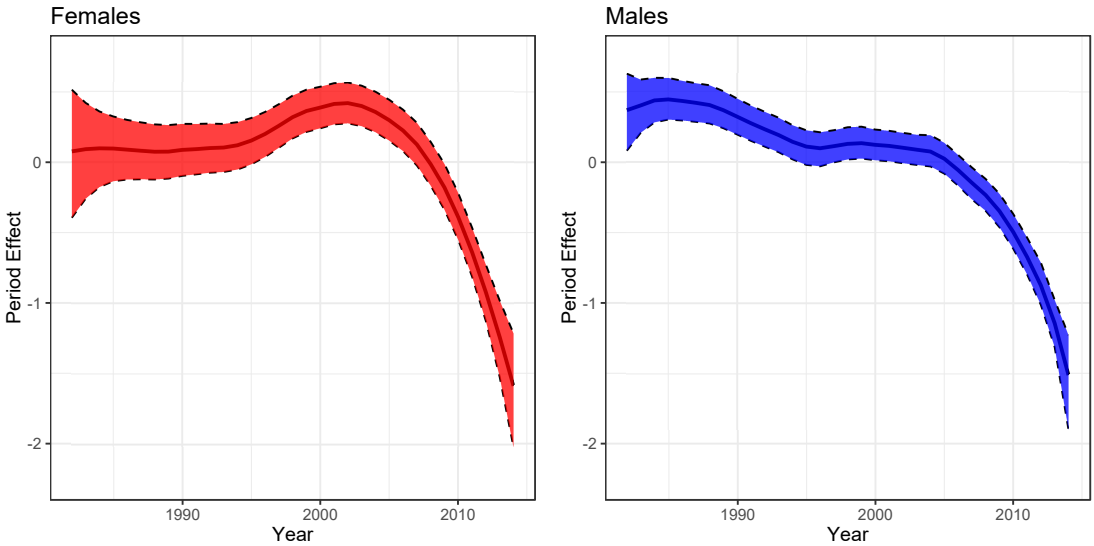


Figure 4.3: *The posterior mean of period effects with a 95% credible interval of the Bayesian AP model for female (left) and male (right)*

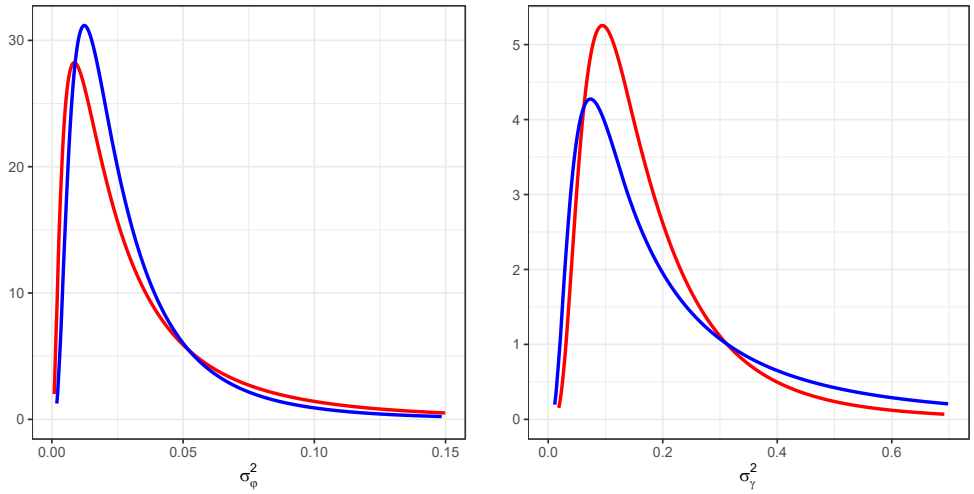


Figure 4.4: *The standard deviations of the age effects of the Bayesian AP model (left), and the standard deviation of the period effect(right) for females in red and males in blue*

a generalised variance that equals to 1. This implies that, the hyperprior distributions that are set on the precision parameters are set on the deviation of the model from the null space of the random walk model. Therefore, the marginal posterior distributions of the standard deviations of both the gender specific ef-

fects are comparable. Their resemblance indicates that the RW2 models have applied similar degrees of smoothing for both males and females.

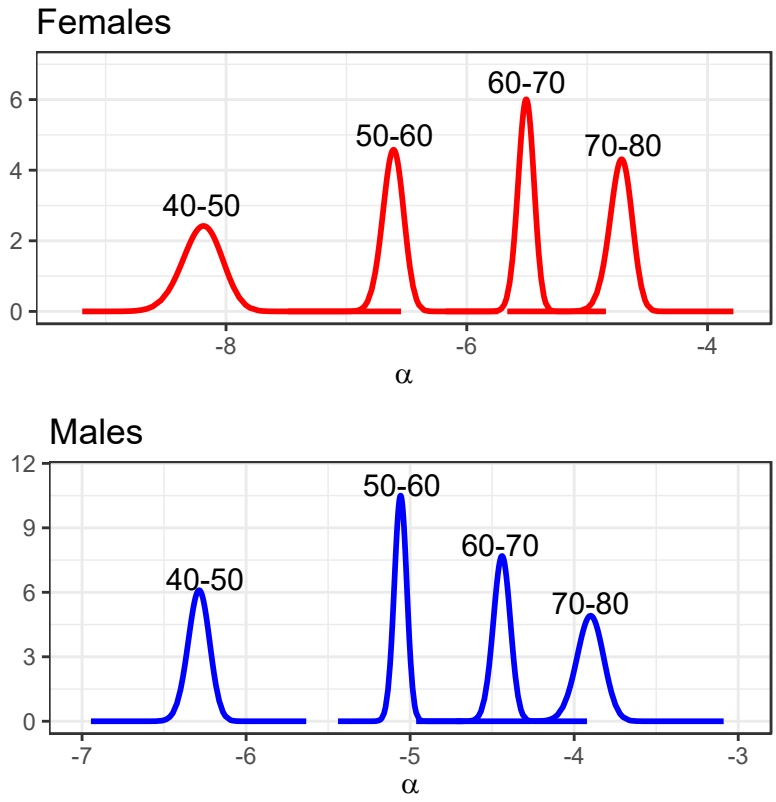


Figure 4.5: *The posterior distributions of age group specific intercepts for females (top) and for males (bottom)*

In figure 4.5, the marginal posterior distribution of age group specific intercepts are presented for females (top) and males (bottom). For both males and females, the mean value of the intercepts increase for the older age groups. Moreover, females have smaller mean intercepts than their male peers in the same age group. Age group specific scaled age and period effects for each sex were computed using the Bayesian AP models for the 40-50, 50-60, 60-70, and 70 - 80 age groups. In figure 4.6, the age group specific age effects are displayed. Similar results are observed for both males and females. There is an increase in the age effect for 40-50, 50-60 and 60-70 age groups. While, the 70-80 age group has a more constant

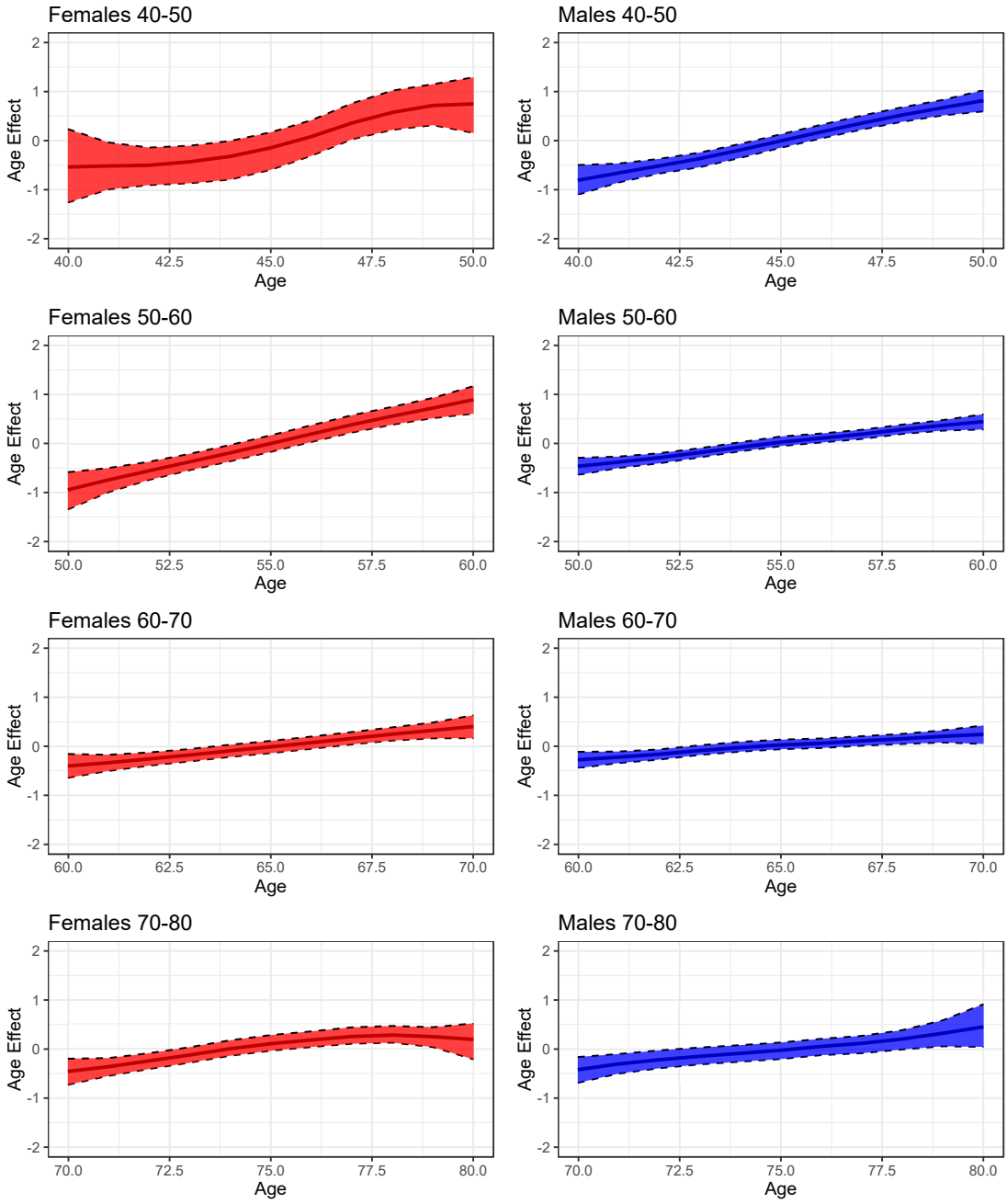


Figure 4.6: *Posterior mean of the age group specific age effects for females (left) and males (right) with the shaded region representing the 95% credible interval*

age effect.

In figure 4.7, the age group specific period effects are displayed. The first three female age groups have a constant period effect until between 2000 and 2005. The constant period effect is then followed by a decreasing trend. Whereas, the 70-80 female age group has a wide credible interval between 1987 to 1995 followed by a constant effect until 2007. After 2007, the age group specific period effect is decreasing.

The 40-50,50-60,70-80 male age groups have a constant period effect until 2000-2005, and this period is followed by a decreasing trend until 2014. However, the 50-60 male age group has a slightly decreasing trend from the beginning in 1982 to 2014.

The incidence rate of MI is thus increasing with age for both genders from the age of 40 to 80. And incidence rate of MI decreases after the year 2000.

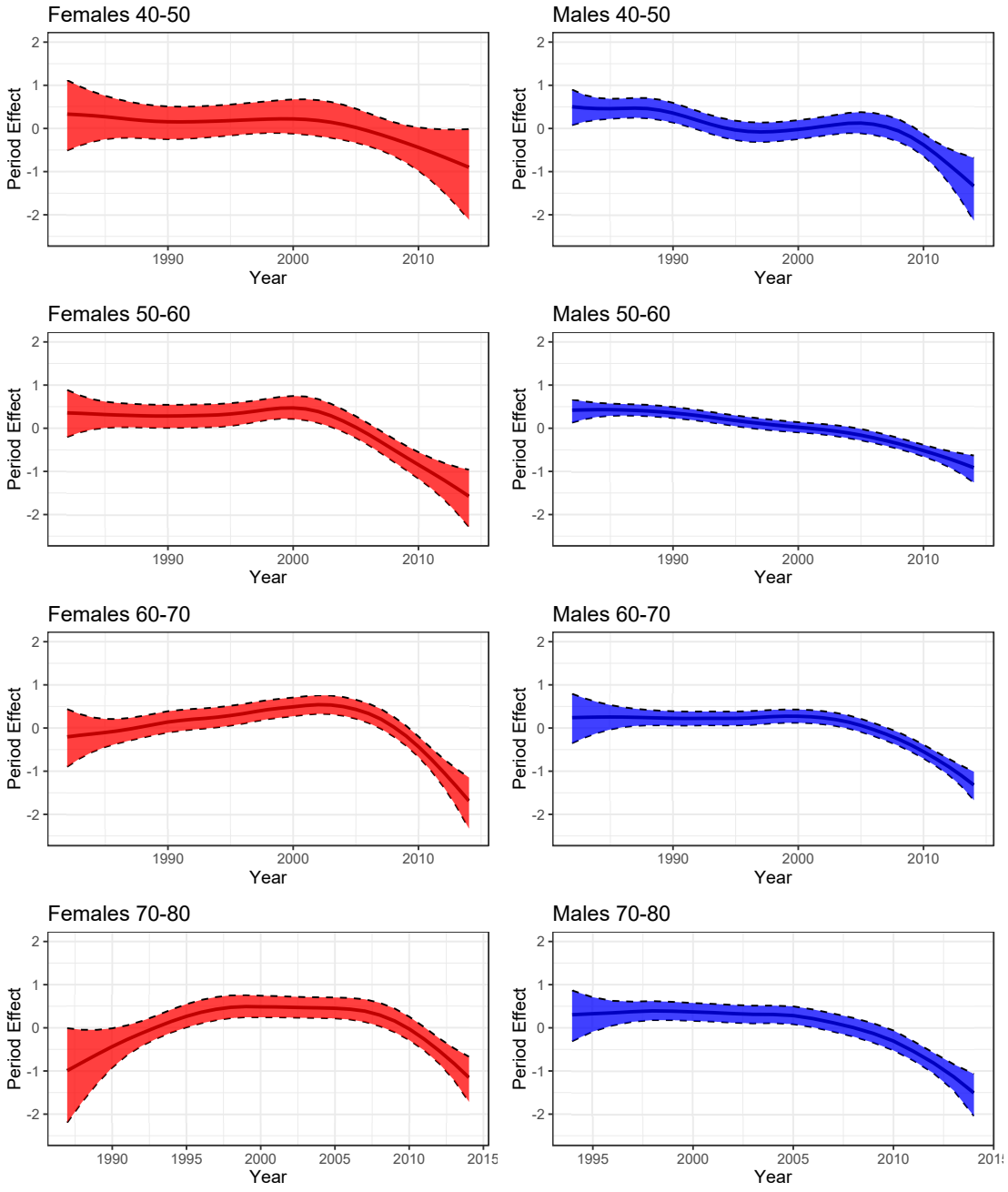


Figure 4.7: *The posterior mean of the age group specific period effects for females (left) and males (right) with the shaded region representing the 95% credible interval*

4.3 Multivariate Bayesian age-period-cohort models

In this section, multivariate Bayesian APC models will be used to study how the incidence rate of MI varies across gender. Such models are useful when jointly analysing stratified mortality or disease rate data (Riebler and Held, 2010a). When using the multivariate APC models, the incidence of MI y_{ijg} of age i, in time period j and gender g is poisson distributed with the rate $n_{ijg} \times \lambda_{ijg}$ where n_{ijg} represents the total number of participants that are alive and yet to have their first MI at time j, age i and gender g. An additional iid random effect is added to the multivariate APC model as was the case in the AP model to account for possible overdispersion. The predictor η_{ijg} is thus:

$$\eta_{ijg} = \alpha_g + \varphi_{ig} + \gamma_{jg} + \psi_{kg} + \kappa_{ijg}, \tag{4.4}$$

where φ_{ig} is the gender dependent age effect, γ_{jg} is the period effect of gender g, ψ_{kg} is the cohort effect of the gender g and κ_{ijg} is the random effect. When jointly analysing the incidence rate using the multivariate APC models, one or more of the time effects can be assumed to be common for both the sexes, while the others can vary across gender (Riebler and Held, 2010a). If the age effect in (4.4) is assumed to be common for both genders, then the gender specific age effects φ_{ig} will be replaced by φ_i. This can be done for the other time effects as well.

As with the univariate case, the sum-to-zero constraint has to be imposed on the multivariate APC model if the gender specific intercepts α_g are to be uniquely identified. In addition, the linear effects of the three time effects can not be uniquely identified (Holford, 2005) Although unique identification of the time effects of the multivariate APC models are not possible without additional constraints, the identifiable non-linear trends can be utilised (Clayton and Schifflers, 1987b; Riebler and Held, 2010a). Among the identifiable aspects of the multivariate APC is the time effect differences between the stratum such as the difference

between the age/period/cohort effects of males and females, if at least one of the time effects is common for both genders (Riebler and Held, 2010a). Following the convention used in Riebler and Held (2010a) capital A,P and C will be used to emphasise the common time effects, whereas the lowercase letters a,p and c will be used to specify the gender specific effects. Take for instance APc model, the uppercase A and P imply that the age and period effects are common for both males and females, while separate cohort effects are assumed.

To begin with, let two of the three time effects be common for both genders. Such multivariate APC model can be the APc where the age and period effects are the common effects, with the cohort effect varying across gender. In this case the difference in between the cohort effects of males and females is identifiable and given by:

$$\begin{aligned} \Delta_k &= \psi_{k,1} - \psi_{k,0} \\ \Delta_\alpha &= \alpha_1 - \alpha_0 \\ \tilde{\Delta}_j &= \Delta_\alpha + \Delta_k, \end{aligned} \tag{4.5}$$

where Δ_k is the difference between the k^{th} cohort effects of males and females, Δ_α is the difference between the intercepts and $\tilde{\Delta}_k$ is the adjusted difference between the cohort effects and is interpreted as log relative risk. All the differences in (4.5) are identifiable since they do not depend on the transformation term a given in (4.2) (Riebler and Held, 2010a). The same analogy in (4.5) is applicable to the aPC and ApC models by changing the log relative risk of cohort effects $\tilde{\Delta}_k$ to log relative risk of the age effects $\tilde{\Delta}_i$ and the log relative risk of the period effects $\tilde{\Delta}_j$ respectively.

Furthermore, only one of the time effects can be allowed to be common for both genders. Suppose the age effect is set to be common for both genders and there are gender specific period and cohort effects so that the resulting multivariate APC model becomes Apc. The set up in (4.5) changes slightly to

$$\Delta_k = \psi_{k,1} - \psi_{k,0}$$
$$\Delta_j = \gamma_{j,1} - \gamma_{j,0}$$
$$\Delta_\alpha = \alpha_1 - \alpha_0$$

Since both the gender specific period and cohort effects are allowed to vary, the log relative risk becomes $\tilde{\Delta}_{jk}$ and is defined as

$$\tilde{\Delta}_{jk} = \Delta_\alpha + \Delta_k + \Delta_j$$
$$\tilde{\Delta}_j = \frac{1}{K}\sum_k \tilde{\Delta}_{jk} \text{ and } \tilde{\Delta}_k = \frac{1}{J}\sum_j \tilde{\Delta}_{jk}, \tag{4.6}$$

where $\tilde{\Delta}_j$ is the average log relative risk of period j and $\tilde{\Delta}_k$ is the average log relative risk of cohort k (Riebler and Held, 2010a). Similar average log relative risk for the time effects can be computed for the aPc and apC models using (4.6).

Moreover, any presumable correlation between the gender specific time effects can be added to the model. Addition of the correlations can narrow the credible interval of the log and average log relative risk of the time effects in (4.5) and (4.6). For the uncorrelated time effects, second order random walk prior distributions are used. In the case of correlated time effects, random walk prior distributions with added correlation component can be used (Riebler et al., 2010b)

A uniform correlation matrix plays a vital role in driving correlated multivariate random walk prior distributions. The uniform correlation matrix is a square matrix with ones in the diagonal and the unknown correlation parameter ρ elsewhere

as follows:

$$C = (1-\rho)I + \rho J = \begin{bmatrix} 1 & \rho & \dots & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \dots & \dots & \rho & 1 \end{bmatrix}_{R\times R}$$

$$C^{-1} = \begin{bmatrix} a & b & \dots & \dots & b \\ b & a & b & \dots & b \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ b & \dots & \dots & b & a \end{bmatrix}_{R\times R},$$

$$\text{where } a = \frac{-(R-2)\rho + 1}{(\rho-1)[(R-1)\rho+1]}$$
$$b = \frac{\rho}{(\rho-1)[(R-1)\rho+1]},$$

I is the $R \times R$ identity matrix, J is an $R \times R$ matrix of ones and R is the number of stratum (Riebler et al., 2010b). In this case $R = 2$, i.e. males and females. If the gender specific age effects are assumed to be correlated, then the second order random walk prior distribution of the time effects is then replaced by:

$$f(\tilde{\varphi}|C_\varphi, \tau_\varphi) \propto \left(\frac{1}{\tau_\varphi}\right)^{R(I-2)/2} |C^{-1}| exp\bigg(-\frac{1}{2}\tilde{\varphi}^T [C^{-1} \otimes R^{(2)}]\tilde{\varphi}\bigg), \tag{4.7}$$

where $\tilde{\varphi} = (\varphi_1^T, ..., \varphi_R^T)^T$ is the collection of all the age effects of all the stratum and $R^{(2)}$ is the precision of the second order random walk prior distribution in (3.19) and $C^{-1} \otimes R^{(2)}$ is the Kronecker product of the matrices (Riebler et al., 2010b).

Riebler et al. (2010b) uses the general fishers z-transformation which is a variance stabilizing transformation to reparametrize the correlation parameter ρ in to :

$$\rho = \frac{exp(\rho^*) - 1}{exp(\rho^*) + l} \quad \Leftrightarrow \quad \rho^* = log(\frac{1+\rho}{1-\rho})$$

Then, a normal prior with mean zero and precision 0.2 is used for ρ^*. This leads to an approximately uniform prior between -1 and 1 for the correlation parameter ρ and $\pi(\rho > 0) = 0.5$.

Similar to the AP analysis in section **??**, everyone below 40 years and above 80 years, time before 1981 and the number of total participants in the age and time group below 100 are omitted from this analysis due to the limited number of MI observed. Correlated and scaled smoothing prior distribution such as (4.7) with PC prior for the precision $1/\tau$ in (3.21) and correlation ρ parameters are assigned when correlation is assumed. For the uncorrelated models, scaled second order random walk (3.19) are assigned as prior distribution for the time effects with the PC prior for the precision parameter in (3.21). PC prior distribution is also assigned to the precision parameter of the random effects.
The three stage hierarchical Poisson model in (4.4) is as follows:

$$\begin{aligned} \text{Stage 1:} \quad & y_{ijKg}|\eta_{ijkg} \sim poisson(n_{ijkg}\, exp(\eta_{ijkg})) \\ \text{Stage 2:} \quad & \boldsymbol{x}_g = (\boldsymbol{\eta}_g, \alpha_g, \boldsymbol{\varphi}_g, \boldsymbol{\gamma}_g, \boldsymbol{\psi}_g, \boldsymbol{\kappa}_g) \\ & \boldsymbol{x}_g|\boldsymbol{\theta} \sim N(0, \boldsymbol{Q}(\boldsymbol{\theta})) \\ \text{Stage 3:} \quad & \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}), \end{aligned}$$

where $\boldsymbol{x}_g$ is the gender specific latent field, $\boldsymbol{Q}(\boldsymbol{\theta}))$ is the gender specific joint precision matrix of the latent field $\boldsymbol{x}$ and the hyperparamter $\boldsymbol{\theta}$ include the correlation parameters and $\pi(\boldsymbol{\theta})$ is prior distribution of the hyperparameters.

4.3.1 Results

Various Bayesian multivariate APC model with none, one, two or all three effects assumed common for both sexes are fitted to the data. In addition, multivariate APC models with correlation between the gender specific effects (the effects that were not common for males and females) are fitted to the data. In table 4.1, the values of the deviance information criterion (DIC) for the various models are presented. The aPc model with a common period effect is classified as the best model using DIC as the model choice criteria.
Correlation between the gender specific effects are computed for all the time effects that are allowed to vary across gender. The correlation between the time

	APC	**APc**	**ApC**	**aPC**	**Apc**	**aPc**	**apC**	**apc**
Uncorrelated								
DIC	5802	5768.11	5814.38	5724.38	5750.99	**5722.23**	5756.25	5753.1
Correlated								
DIC	-	5769.93	5800.52	5724.01	5738.27	**5721.64**	5742.06	5738.0

Table 4.1: *The Model choice criteria DIC from INLA for of all the multivariate APC models. The upper-case letters A (age), P (period) and C (cohort) represent the effect that was kept common for males and females, while the lower-case letters represent the gender specific effects*

effects are presenter in table 4.2. The estimated values of the correlation parameters remained similar in the various multivariate APC models. Only the correlation between the gender specific period effects was clearly greater than zero. As a result, including the correlation in the aPc model with joint period effects did not improve the results in a significant manner.

	APc	**ApC**	**aPC**	**Apc**	**aPc**	**apC**
A	-	-	0.7(-0.3-0.9)	-	0.7(-0.3-0.9)	0.7(-0.3-0.9)
P	-	0.93(0.8-0.96)	-	0.9(0.7-0.9)	-	0.9(0.7-0.9)
C	0.3(-0.6-0.9)	-	-	0.29(-0.7- 0.9)	0.3(-0.7-0.9)	-

Table 4.2: *Correlation and the 95% credible interval between gender specific effects in lower case for all the multivariate APC models. The upper-case letters denote the effects that were common for both males and female*

In the figure 4.8, the average log relative risks of MI for males compared to females for the aPc model, with the period effect assumed to be common, for the cohort and age effects are displayed. Both the correlated and uncorrelated age and cohort effects are shown. The shaded region shows the 95% credible interval of the time effects with added correlation and the red line in the middle is the posterior mean of the correlated average log relative risk. While, the dashed lines mark

the 95% credible interval of the uncorrelated model, with the black line denoting the posterior mean of the effects. In the presence of a notable correlation, the coredible interval would have been narrowed. Since the correlation between male and female age effects and between male and female cohort effects are not clearly greater than zero, the effect on the credible interval is not significant.

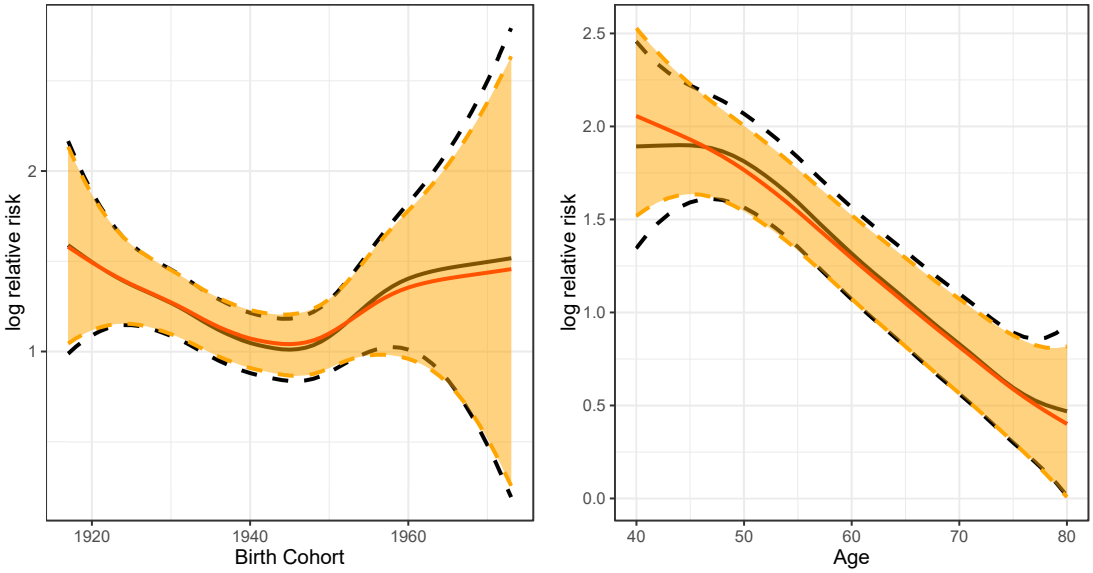


Figure 4.8: *The average log relative risk for males compared to females for the aPc model for the birth cohort (left) and age (right). The shaded region denotes the 95% credible interval for the correlated model and the red line represents the posterior mean of the average log relative risk of the correlated model. While the dashed black lines mark the 95% credible interval for the uncorrelated model and the black line denotes the posterior mean of the average log relative risk the uncorrelated model.*

The average log relative risk for males compared to females with common period effects for the birth cohorts that were before 1925 and after 1955 have very wide credible intervals and cover a large area. While, the risk of MI for males in the birth cohorts in the middle(between 1925 and 1955) was more than twice that of females.

The risk of MI for males compared to females is monotonically decreasing from the age of 40 to 80 (see figure 4.8 (right)). The 40 years old males have more than 3.7 times risk of MI than their female peers in the same age. While this relative risk decreases to below 2.2 having 2.2 risk of MI for 80 years old males compared to their 80 years old female counterparts.
The results are discussed further in chapter 6.

Chapter 5

Survival analysis

In the previous chapter the incidence rate of MI was investigated thoroughly. Next, the focus shifts towards what happens once the individual participants have encountered MI. In this chapter, the focus will be on how the survival time of the Tromsø study participants who had MI evolves with time for the different age groups, sexes and the time of the year the MI.

A brief introduction to the main concepts in survival analysis is provided in section 5.1. Thereafter, the Kaplan-Meier's method and estimation of confidence intervals is presented along with the results of the analysis in section 5.2. In section 5.3, the given data is analysed using one of the most regularly used survival models known as Cox proportional hazards (Cox PH) model. The chapter is then concluded by discussing the results of the analysis in **??**.

5.1 Concepts in survival analysis

Across many fields, the concept of time from entry to an event is often of interest. Time to the onset of a disease in medical sciences, time to divorce in social sciences and time to the failure of an electronic component in engineering are some examples of various fields with such interests. Generally, the time elapsed from a well defined start to the event of interest is referred to as survival time.

Survival time is commonly started at zero and often no event will have occurred at that point (Aalen et al., 2008). In this chapter, the time from the incidence of the MI to death is the variable that will be investigated. In addition, at most only a single event can take place to each participant who has had MI.

5.1.1 Censoring

Survival times can be studied using a wide set of statistical tools. However, the set of tools available can be narrowed if the events are not observed for every individual. In some cases, this can be due to drop out of participants or the time allocated to the data collection is smaller than the largest survival times as shown by the transparent dots in figure 5.1. In others cases, participants can be entered to the study at survival times other than zero commonly known as delayed entry. Such data has therefore both complete and incomplete sets of observations. Figure 5.1 has seven complete and three incomplete observations. Survival analysis models are well suited to dealing with incomplete survival data (Aalen et al., 2008).

Censoring of data is the main reason for incompleteness of survival time data. In figure 5.1, there are three individuals without a registered event due to the end of the study. Those observations are said to be censored. The censoring in figure 5.1 happens towards the right end of the time axis. Hence, it is often called right censoring. Rather than throwing out such observations, they can be part of the participants in the risk set for as long as they are in the study. The total risk set can then be reduced once the censoring takes place (Aalen et al., 2008). According to Liu (2012) there are three types of right censoring; type I, random censoring and type II censoring. Type I censoring happens as a result of the end of the study as in figure 5.1. With random censoring, as the name indicates, the censoring happens at an arbitrary time. In type II censoring, data collection goes on until a fixed number of events are observed.

Survival models are also suitable to analyse truncated survival data. There may be instances where portion of events do not get registered into the dataset. Sup-

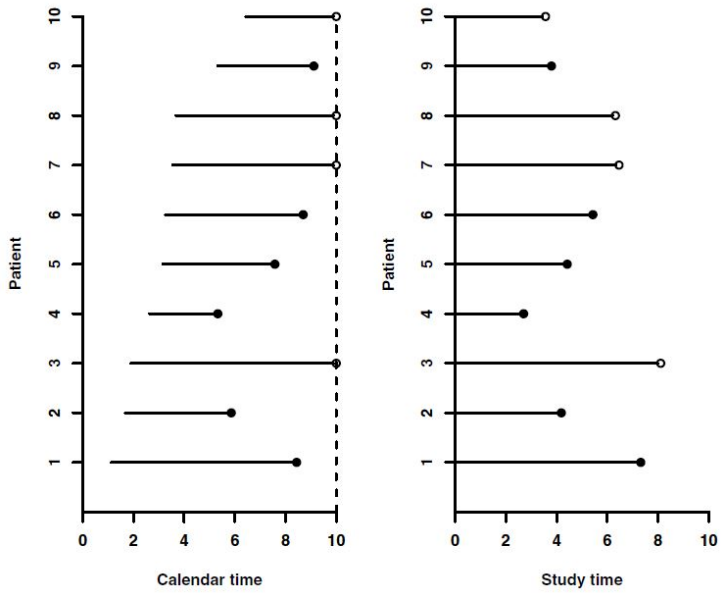


Figure 5.1: *Survival time of 10 patients over 10 time units. The left panel shows the calender time for each patient from enrollment to the event or end of the study. In the right panel, the time axis measures survival time. The opaque dots indecicate time of the event and the transparent dot symbols censoring. Figure from Aalen et al. (2008)*

pose all those who divorced within 6 months of their never never got registered. Thus those events are truncated from the study (Aalen et al., 2008). In this chapter, type I right-censored survival data without truncation will be used to carry out the survival analysis.

5.1.2 Survival and hazard functions

Generally, survival analysis deals with time to event as a continuous random variable. The cumulative distribution of the survival time (T) is the probability of the event taking place prior to time t and is given by :

$$F(t) = \pi(T \leq t) = \int_0^t f(t)dt, \quad t > 0$$

A closely related quantity to the cumulative density function (CDF) of the survival time $F(t)$ is the survival function $S(t)$. It plays a vital role in survival models and is defined as the probability of the event occurring after time t. The survival function is mathematically defined as follows :

$$S(t) = \pi(T > t) = 1 - F(t) = 1 - \int_0^t f(t)dt, \quad t > 0 \tag{5.1}$$
$$S(0) = 1 \text{ and } S(\infty) = 0.$$

Since no event has has taken place at time zero, $S(0)$ is equal to 1. Furthermore, the value decreases to zero with increasing time. This means that no event takes place prior to time zero and given enough time all the subjects will encounter the event. In addition to the survival function, a second function known as the hazard rate $h(t)$ is a crucial part of survival models. The hazard rate is the probability of the event happening in an infinitesimal time interval $[t, t + \Delta t)$ given the event had not occurred prior to time t (Aalen et al., 2008). It is mathematically defined as follows (Liu, 2012):

$$\begin{aligned} h(t) &= \lim_{\Delta t \to \infty} \frac{\pi(t \leq T < t + \Delta t \,|T > t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)} \\ &= -\frac{\partial}{\partial t} log(S(t)). \end{aligned} \tag{5.2}$$

Unlike the survival function, the hazard rate $h(t)$ can have any non-negative value. Based on the hazard rate, the cumulative hazard function $H(t)$ is defined as follows (Liu, 2012):

$$H(t) = \int_0^t h(t)dt = -log\, S(t). \tag{5.3}$$

5.2 The Kaplan-Meier method

5.2.1 Kaplan-Meier survival curves and confidence intervals

Kaplan-Meier's estimator, alternatively known as the product-limit method, is one of the most commonly applied statistical methods in survival analysis. The Kaplan-Meier's method is a non-parametric method from the frequentist framework. Essentially, the Kaplan-Meier method estimates the survival function of complete or right censored data using a product of a sequence of conditional probabilities for mutually exclusive time intervals. The Kaplan-Meier method produces an estimate for the survival function that is constant with time unless an event of interest is encountered. When an event is encountered, the estimated value for the survival function drops abruptly right after the time of an event. As a consequence of that, the Kaplan-Meier method produces the characteristic step function as an estimate for the survival function (Kaplan and Meier, 1958). To compute the Kaplan-Meier estimate, the time scale is sub-divided into smaller time intervals. Then all of the observed complete and incomplete survival times are arranged in order and the survival times are placed in the correct time intervals. Thereafter, interval specific survival rates $\hat{s}(t) = \dfrac{n_j - \delta_j}{n_j}$ are computed, where n_j is the total number of the risk set going into the j^{th} time interval and δ_j is the total number of events in that time interval. Finally, the Kaplan-Meier estimate of the survival function at time t is computed as the product of the interval specific survival rate prior to time t (Kaplan and Meier, 1958). The Kaplan-Meier estimate of the whole survival function $\hat{S}(t)$ is therefore given by

$$\hat{S}(t) = \begin{cases} 1, & \text{if } t < t_1 \\ \prod_{j=1}^{t} \left(\frac{n_j - \delta_j}{n_j} \right), & \text{if } t \geq t_1 \end{cases} \tag{5.4}$$

Here, the censored participants are removed from the risk set for the interval following the censoring time interval, i.e. the censored participants are part of

the total n_j if the censoring took place at the j^{th} time interval and they are removed from the reisk set at the $j + 1^{th}$ time interval (Liu, 2012). In other words, if a patient is censored at day 11, that patient is removed from the risk set starting day 12.

From the Kaplan-Meier estimate, the cumulative hazard function can be derived using (5.3). The estimate of the cumulative hazard function $\hat{H}(t)$ is defined by (5.5) (Liu, 2012).

$$\hat{H}(t) = -log\left[\prod_{j=1}^{t}\left(\frac{n_j - \delta_j}{n_j}\right)\right] \tag{5.5}$$

In general, when the parametric form of the hazard function is not known, the non-parametric Kaplan-Meier estimate of the cumulative hazard function in (5.5) can be useful in testing the assumptions of a parametric form or suggesting a parametric form for the hazard function (Kaplan and Meier, 1958; Liu, 2012).

Once the Kaplan-Meier estimates for the survival functions are computed, then the standard errors and confidence interval of the estimates have to be estimated. However, estimating the standard error and confidence interval associated with the Kaplan-Meier estimate is not straightforward for a number of reasons. One such reason is that the largest survival times can be censored, such that $\hat{S}(t_n) > 0$ and $n_{t_{n+1}} = 0$. Thus, $\hat{S}(t_{n+1})$ is undefined and the standard error estimate in such cases may not be very informative. Nonetheless, for large survival times the aforementioned issue is not of practical importance (Kaplan and Meier, 1958). Another issue related to estimating the confidence interval for the Kaplan-Meier estimates is that survival function is confined between 1 and 0, and the estimated confidence interval should not exceed those limits (Liu, 2012).

Frequently, the variance of the Kaplan-Meier estimate is approximated using the Greenwoods formula which is defined as:

$$\hat{V}[\hat{S}(t)] \approx [\hat{S}(t)]^2 \sum_{j=1}^{t-1} \frac{\delta_j}{(n_j - \delta_j)(n_j)} \tag{5.6}$$

and the standard errors are obtained by taking the squareroot of (5.6) (Kaplan

and Meier, 1958). A more through discussion on the derivation of the estimate is found in Liu (2012); Kalbfleisch and Prentice (2002).

To avoid estimating a confidence interval for the Kaplan-Meier estimate that exceeds the range of the survival function, a log-log transformation of $\hat{S}(t)$ is commonly applied. The confidence interval is then estimated for the log-log survival function, before it is transformed back to the original scale (Kalbfleisch and Prentice, 2002). The confidence interval of the log-log survival function is given by

$$log[-log\hat{S}(t)] \pm z_{1-\alpha/2}\sqrt{\frac{\hat{V}[\hat{S}(t)]}{\hat{S}(t)\, log\hat{S}(t)}},$$

where $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution (Kalbfleisch and Prentice, 2002). The confidence interval of the estimated survival function is given by

$$[\hat{S}(t)]^{1/\theta} \leq [\hat{S}(t)] \leq [\hat{S}(t)]^{\theta}, \tag{5.7}$$

where $\theta = exp\left(z_{1-\alpha/2}\sqrt{\frac{\hat{V}[\hat{S}(t)]}{\hat{S}(t)\, log\hat{S}(t)}}\right)$ (Kalbfleisch and Prentice, 2002). The $1-\alpha$ confidence interval in (5.7) is estimated for each time point separately. Further derivation is needed to estimate a $1-\alpha$ confidence region for the entire Kaplan-Meier estimate of the survival function simultaneously. Such confidence regions are know as simultaneous confidence bands (Liu, 2012). Hall and Wellner (1980) introduced a method for estimating a simultaneous confidence band for random censored data, and survival data of moderate to large size. The confidence band is estimated by first picking $t_L = 0$ and t_U as the largest survival time after the largest complete survival time. Second, a_L and a_u are computed as follows:

$$a_L = \frac{n\hat{V}[\hat{S}(t_L)]}{1 + n\hat{V}[\hat{S}(t_L)]} \quad \text{and} \quad a_U = \frac{n\hat{V}[\hat{S}(t_U)]}{1 + n\hat{V}[\hat{S}(t_U)]}.$$

Then, using a_L and a_U, $\kappa_\alpha(a_L, a_U)$ coefficients for the Hall-Wellner bands can be obtained from tables of confidence coefficients. The resulting Hall-Wellner band gives good results when it is computed with respect to the aforementioned

log-log transformed confidence interval (Liu, 2012). The log-log transformed Hall-Wellner band is then given by:

$$[\hat{S}(t)]^{1/\theta} \leq [\hat{S}(t)] \leq [\hat{S}(t)]^{\theta}, \tag{5.8}$$

where

$$\theta = exp\left(\frac{\kappa_{\alpha}(a_L, a_U)[1 + n\hat{V}(\hat{S})(t)]}{\sqrt{n}\ log[\hat{S}(t)]}\right).$$

The median survival time of the Kaplan-Meier estimates of survival function are often computed as $S(M) = 0.5$ and the confidence interval is given by

$$SE_{median} = SE_{GR}\Big\{(t_{small} - t_{large})/(S(t_large) - S(t_small))\Big\}$$
$$95\%\, CI = (Median - 1.96SE_{median}, Median + 1.96SE_{median}),$$

where SE_{GR} is the standard deviation obtained from Greenwoods variance.

5.2.2 Results

The data available in this chapter is a type I right censored data without truncation. If the Kaplan-Meier method is to produce an adequate estimate for the survival function, the censoring and survival times need to be independent. Here, the type I right censored data depends on the time of the MI, rather than the time at censoring.

In figure 5.2, the Kaplan-Meiers survival curve for the entire Tromsø study population is displayed. The survival time in the x-axis measures the time from MI to death in weeks. Estimation of the survival curve was carried out using the Kaplan-Meier estimator in (5.4) with the pointwise log-log confidence interval in the shaded region estimated by (5.7) and the simultaneous Hall-Wellner confidence band estimated by (5.8).

As expected, the largest drop in the survival function in figure 5.2 takes place in the first week. It then continues to decrease at a slower rate. Since the

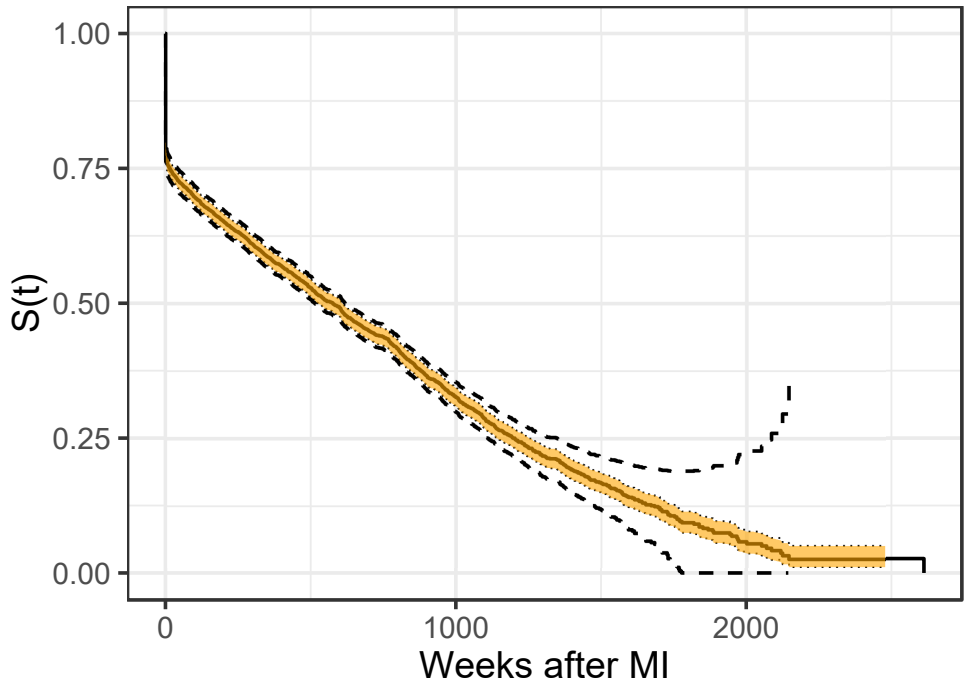


Figure 5.2: *Kaplan-Meier curve of the total population. The shaded region denotes the The shaded region denotes the 95% pointwise confidence interval with the outer dotted line indicating the 95% Hall-Wellner simultaneous confidence band 95% pointwise confidence interval with the outer dotted line indicating the 95% Hall-Wellner simultaneous confidence band*

largest survival time was not censored, the Kaplan-Meiers survival curve goes to zero after 2612 weeks (50 years). The estimated Hall-Wellner simultaneous confidence band follows the pointwise log-log confidence interval in the beginning and gets wider with deceasing size of the risk set and events. In average, the probability of surviving past the first week of MI is 0.79 with (0.77-0.80) 95% log-log confidence interval. The median survival time is 577 weeks (11 years), and the 95% confidence interval is (527,610) weeks.

Separate Kaplan-Meier survival curves for males (blue) and females (red) are displayed in figure 5.3. In general, both survival curves follow the same trend with large decrease in week 1 followed by a slow decrease to zero. However, the

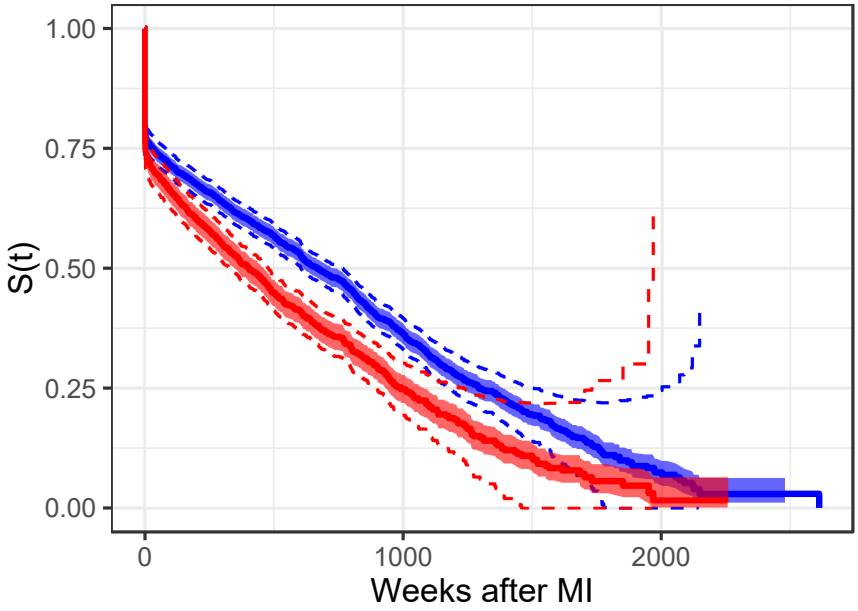


Figure 5.3: *Kaplan-Meier curve of the Male (blue) and female (red) participants separately. The shaded region denotes the log-log 95% pointwise confidence interval with the outer dotted line indicating the 95% Hall-Wellner simultaneous confidence band*

drop in the survival curve in week 1 for females is larger than that of males. The survival function at week 1 for males is $\hat{S}(1) = 0.80$ with a 95% confidence interval that equals (0.78,0.81). In comparison $\hat{S}(1) = 0.76$ for females with a 95% confidence interval that equals (0.74, 0.79). There is also a difference in the median survival time between males and females. While females have median survival time 405 weeks (confidence interval 349,462), the median survival time of their males peers is 669 weeks (confidence interval 616,735). The analysis in section 2.3.1 revealed that there is a significant difference in the average age at the time of MI between males and females. This difference in age can have manifested it self figure 5.3.

In figure 5.4, the Kaplan-Meiers survival curves with the 95% log-log pointwise confidence interval denoted by the shaded regions and the simultaneous Hall-

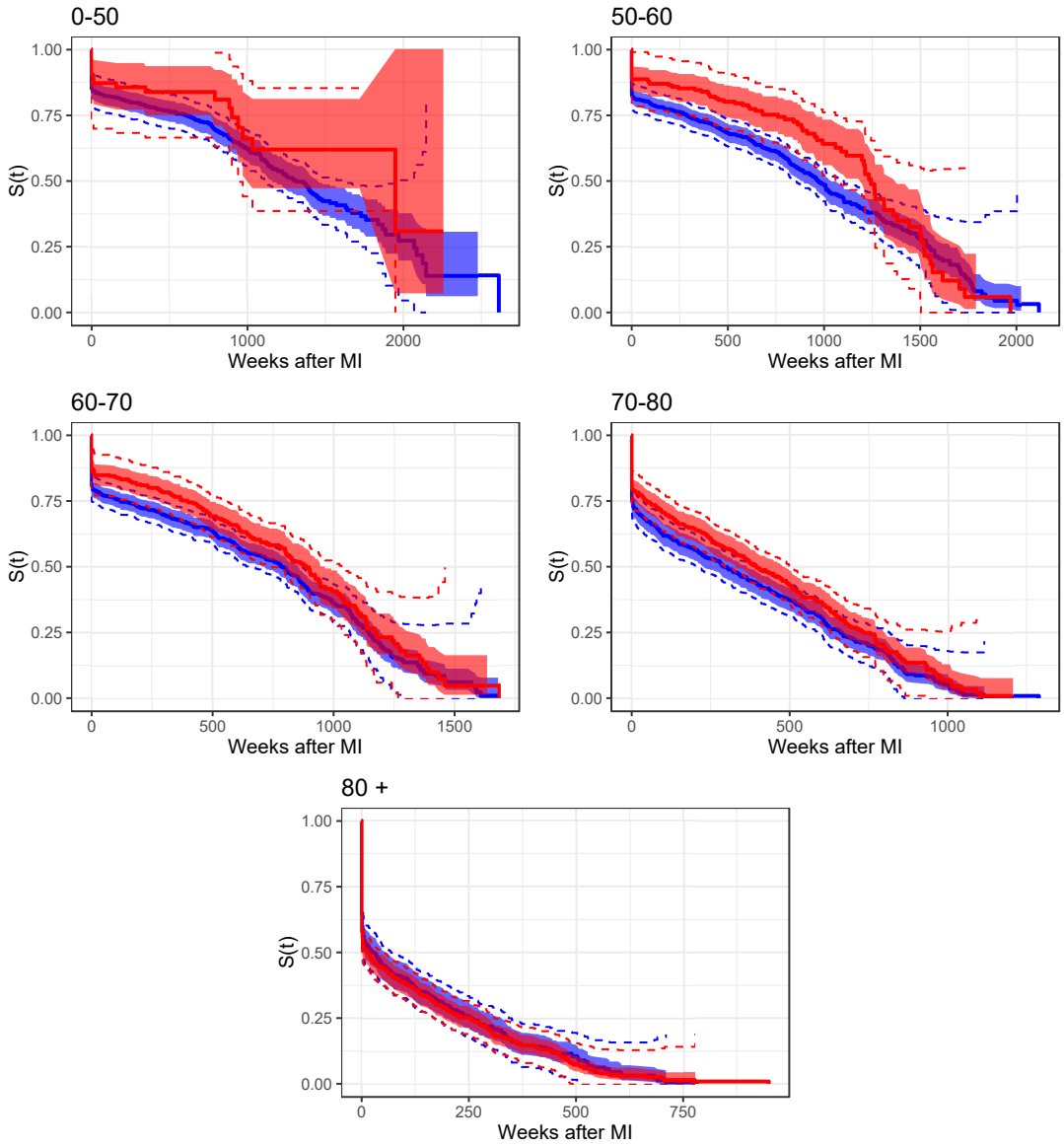


Figure 5.4: *Kaplan-Meiers survival curves of the different age groups, the shaded region denotes the 95% log-log confidence interval(males-blue and females-red) and the dashed lines represent the Hall-Wellner simultaneous confidence bands*

Wellner confidence bands for different age groups and sexes are displayed. Naturally, the probability of survival decreases with increasing age in the age groups. In addition, the proportion of the participants that suffered MI and died within

the first week increases with increasing age.

	Female			Male		
Age gr	$\hat{S}(1)$(95%CI)	Med(95%CI)	Ev.	$\hat{S}(1)$(95%CI)	Med(95%CI)	E
(0 -40]	0.80(0.41-0.95)	793(1--)	4	0.86(0.78-0.92)	1969(1227--)	
(40-50]	0.92(0.81-0.96)	1948(1034--)	17	0.85(0.81-0.88)	1249(1117-1396)	2
(50-60]	0.88(0.83-0.92)	1232(1088-1311)	96	0.83 (0.80-0.85)	961(873-1024)	4
(60-70]	0.87(0.83-0.90)	823(724-932)	178	0.82(0.79-0.85)	781(688-826)	4
(70-80]	0.78(0.73-0.82)	352(274-405)	294	0.75(0.71-0.78)	236(185-286)	4
(80 +]	0.58(0.53-0.63)	13(3-33)	374	0.62(0.56-0.68)	12(2-38)	2
Overall	0.76(0.74-0.79)	405(349-462)	963	0.80(0.78-0.81)	669(616,735)	1

Table 5.1: *The estimated values of the survival function at week 1 and median survival time presented with 95% confidence interval for the various age groups and sexes. Ev. denotes the number of death in the given age group and gender.*

The estimated value of the survival function at week 1 decreased from about 0.86 for the 0-40 male age group to 0.62 for males in the 80 + age group. Similarly, the same estimate decreased from 0.8 for the 0-40 female age group to 0.58 for the 80 + females. These results are summarized in table 5.1 and figure 5.5 (left panel).

The median survival time in weeks is presented in figure 5.5 (right panel) and table 5.1. The longest median survival time is observed in the 40-50 female age group with 1948 weeks after MI and the shortest median survival time was observed in the oldest male age group being 12 weeks .

The differences seen between males and females in figure 5.3 is partly explained by the difference in age among those who have had MI. This can be seen in the the overlap observed in the Kaplan-Meier survival curves of males and females in the different age groups in figure 5.3.

In figure 5.6, Kaplan-Meier survival curves are displayed (top) with the estimated values of the survival function at week 1 (bottom left) and estimated median survival times (bottom right) sub-divided by the season when the MI occurred. In

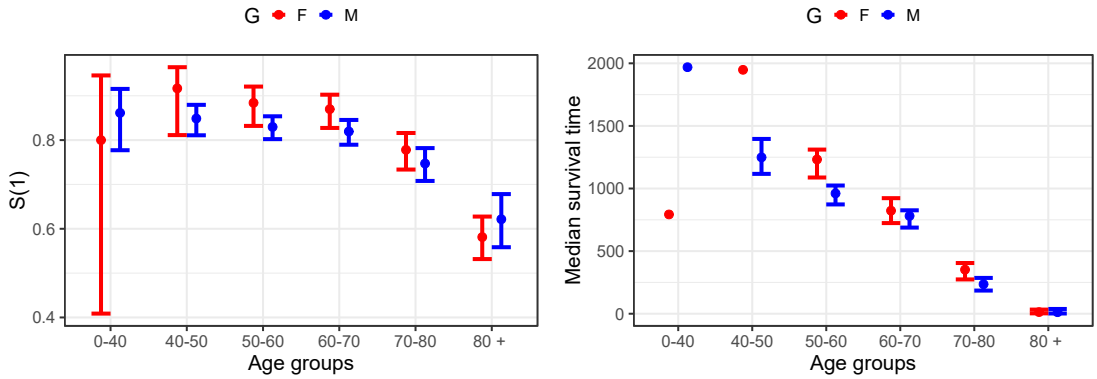


Figure 5.5: *The estimated survival function at week 1 is presented by the dots in the middle, and the vertical lines indicate the extent of the 95% log-log pointwise confidence interval at week 1 (left). Similarly, the median survival time of the various age groups are indicated by the dots and the vertical line indicate the 95% confidence interval (right). Due to the scarce number of events in the first age group for both sexes and the 50-60 age group for females, a reliable upper limit for the confidence interval was not estimated. Males (blue) and female (red)*

section 2.3, some seasonal differences in the MI incidence rate were revealed. However, the Kaplan-Meier survival curves for the seasons in figure 5.6 overlap. Seasonal differences in survival from MI is thus absent. The probability of surviving beyond the first week after MI and median survival times do not depend on the season of occurrence of MI.

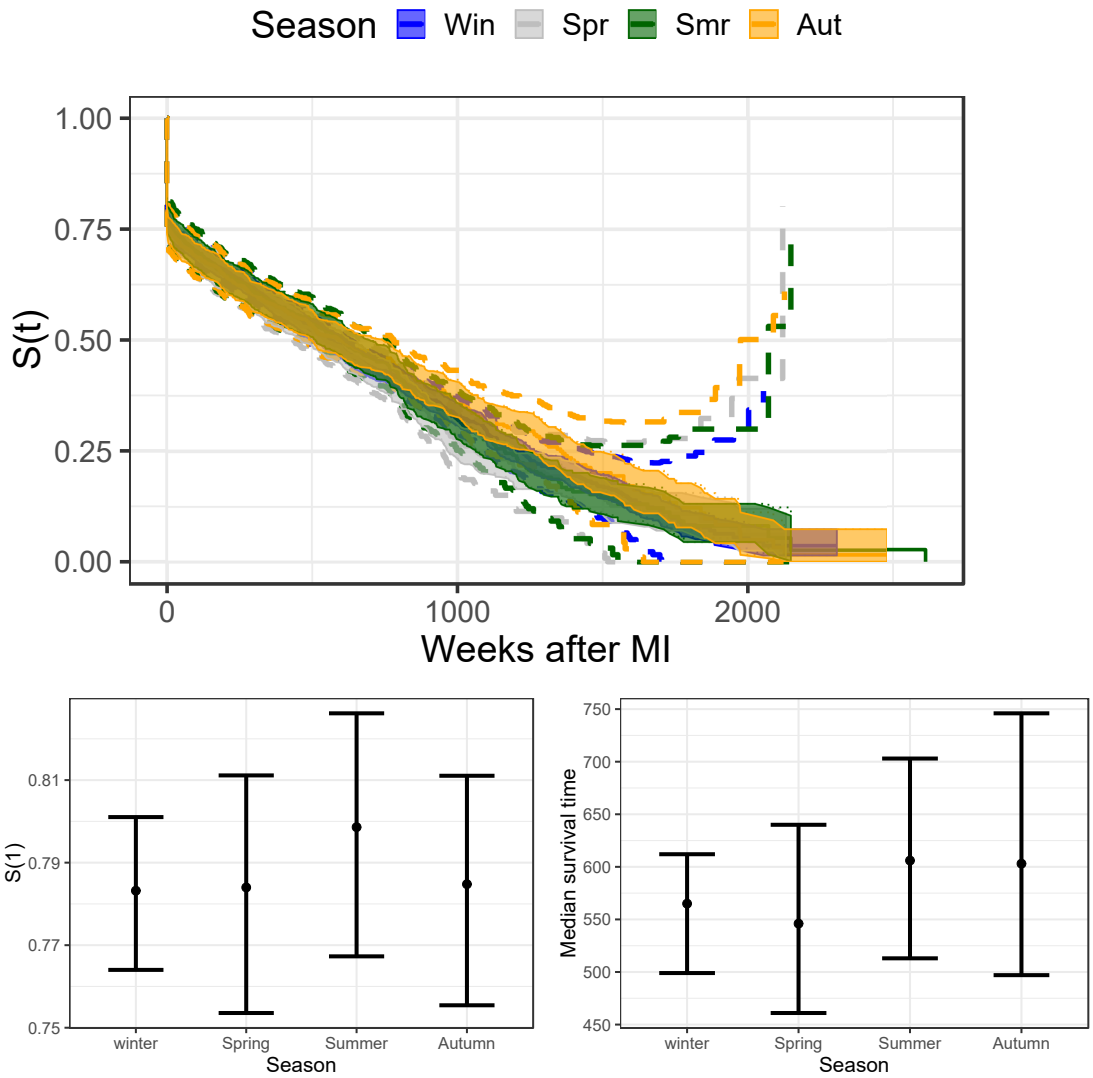


Figure 5.6: *Kaplan-Meier survival curves divided by the season of the incidence of MI (top). The estimated mean values of the survival function at week 1 denoted by the dot in the middle and the log-log pointwise 95% confidence interval marked by the vertical lines (bottom left). Median survival time of participants with 95% confidence interval (bottom right).*

5.3 Proportional hazards models

In the previous section, the descriptive and nonparametric Kaplan-Meier method was introduced, along with the results of the analysis. However, incorporating covariates into the Kaplan-Meier method is not easy. This section presents Proportional hazard rates (PH) models is a commonly applied approach to fit regression models in survival analysis.

PH models assume a common baseline hazard rate of encountering the event of interest at a time t for all the participants and the various individual hazard rates are the given by a product of the baseline hazard rate and a term of the effects of covariates. In general, the PH models are given by

$$h(t) = h_0(t)exp(\sum_{j=1}^{p} \beta_j \boldsymbol{x}_j), \tag{5.9}$$

where $h_0(t)$ is known as baseline hazard rate and $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ are the effects of the multiplicative covariates and $\{\boldsymbol{x}\}$ are the covariates (Liu, 2012).

In a PH modelling framework, the baseline hazards rate $h_0(t)$ depends only on the survival time t, and the term that represents the effects of the covariates is independent of the survival time t. As a direct consequence, the ratio of two individual hazard rates, known as the hazard rate HR is defined as

$$\begin{aligned} HR = \frac{h_i(t)}{h_k(t)} &= \frac{h_0(t)exp(\sum_{j=1}^{J} \beta_j x_{ij})}{h_0(t)exp(\sum_{j=1}^{J} \beta_j x_{kj})} \\ &= exp(\sum_{j=1}^{J} \beta_j (x_{ij} - x_{kj})), \end{aligned}$$

is constant and independent of time. Moreover, two different hazard rates are proportional to each other. Hence, the name proportional hazard rates models. A key assumption in applying PH models to survival data is therefore built on the individual hazard rates being proportional. In addition, the hazard ratio in PH models is regularly used to measure the effects of the covariates on the baseline hazard rate. Hazard ratios that are larger than one imply an increased risk for

individual i compared to k, while the implication of HR less than one is a decrease in risk for the same individual (Liu, 2012; Machin et al., 2006).

In the PH regression models, the parametric assumptions on the hazard rate in (5.9) define the parametric distribution of the survival time t. Based on those assumptions on the hazards rates, there are three types of PH regression models; exponential regression model, Weibull regression model and the semi parametric Cox PH model (Liu, 2012).

Assuming a constant λ as a hazard rate leads to an exponential regression model since

$$h(t) = \lambda \Rightarrow S(t) = e^{-\lambda t}$$
$$F(t) = 1 - e^{-\lambda t}.$$

Implying that the survival time is an exponentially distributed random variable (Liu, 2012). The resulting exponential regression model is then,

$$h(t, x; T \sim Exp) = \lambda exp(\sum_{j}^{p} \beta \boldsymbol{x}_j),$$

where the constant baseline hazard rate λ is also the rate parameter of the exponential distributed survival time (Liu, 2012). If the assumption of the constant hazard rate is accurate, then

$$H(t) = -log[S(t)] = \lambda t \;\; \Rightarrow \;\; log\{-log[S(t)]\} = log(\lambda) + log(t).$$

The right side is in $y = ax + b$ form, where a is represented by $log(\lambda)$, $b = 1$ and x is given by $\log(t)$. Consequently, the log of the cumulative hazard function is linear with $log(\lambda)$ as the intercept, 1 as the slope and $log(t)$ being in the horizontal axis. The Kaplan-Meier estimate of the cumulative hazard function (or the survival function) can be used to verify such assumption (Machin et al., 2006).

In contrast to the constant hazard rate, the Weibull regression model assumes a strictly increasing or decreasing hazard rate. Therefore, this model is well suited

for situation where the hazard rate is either increasing or decreasing. Specifically, the Weibull PH regression model is given by

$$h(t, x; T \sim Weib) = \lambda\kappa(\lambda t)^{\kappa-1} exp(\sum_{j}^{p} \beta \boldsymbol{x}_j),$$

where $\lambda\kappa(\lambda t)^{\kappa-1}$ is the time dependent strictly increasing or decreasing baseline hazard rate $\lambda\kappa(\lambda t)^{\kappa-1}$. The survival time is Weibull distributed with the scale parameter λ and shape parameter κ. Similar to the exponential PH regression model, the log cumulative hazard function can be used to test if the Weibull distribution assumption holds (Machin et al., 2006). The log of the cumulative hazard function is

$$log[H(t)] = (\lambda t)^{\kappa} \;\; \Rightarrow \;\; log[H(t)] = \kappa\, log(\lambda) + \kappa\, log(t).$$

Here, the log of the cumulative hazard function is also linear with $\kappa\, log(\lambda)$ as the intercept, κ as the slope and the x axis in $log(t)$ (Machin et al., 2006).

5.3.1 Cox PH models

Cox PH regression model is a semi-parametric regression model that is often used to analyse survival time. It differs from the Weibull and exponential PH regression models since the model does not place any restriction on the parametric shape of the survival function, apart from the proportionality of the hazard rates. As a result, Cox PH model has wider applicability than the fully parametric regression models (Liu, 2012; Machin et al., 2006).

Estimation of the effects of the covariates in the Cox PH regression models with an arbitrary parametric shape of the baseline hazard rate depends on the notion of partial likelihood. Given all the individuals at risk at time t_i, $\Re(t_i)$, the probability of an individual encountering an event at time t_i is given by

$$P(t_i|\Re(t_i)) = \frac{h_i(t_j)}{\sum_{k\in\Re(t_i)} h_k(t_j)} = \frac{exp(\sum_j^J \beta_j x_{ij})}{\sum_{k\in\Re(t_i)} exp(\sum_j^J \beta_j x_{kj})}.$$

Since the time without events do not provide information about the model parameters, the focus is turned towards the event times (Cox, 1972). the likelihood of the conditional probability above is known as partial likelihood for the $\{\beta\}$ and is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{d} \frac{exp(\sum_{j=1}^{J} \beta_j x_{ij})}{\sum_{k \in \Re(t_i)} exp(\sum_{j=1}^{J} \beta_j x_{kj})}$$

The log likelihood is then

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{d} \Bigg[\sum_{j=1}^{J} (\beta_j x_{ij}) - \sum_{i=1}^{d} log\Bigg(\sum_{k \in \Re(t_i)} exp(\sum_{j=1}^{J} \beta_j x_{kj}) \Bigg) \Bigg],$$

where d is the total number of events encountered. The values of $\{\boldsymbol{\beta}\}$ that maximize the partial log likelihood, give the estimates of parameters in the Cox PH model. Due to the absence of the baseline hazard function in the partial likelihood, making assumption about its parametric form is not necessary (Cox, 1972).

The partial likelihood can be used to compute the effects of the covariates without having to get to the baseline hazard rate. Since the introduction of Cox PH model, there has been several suggestions on how to approximate the baseline hazards rate such as the piecewise log-constant baseline hazard rate introduced by Breslow (1972).

5.3.2 Cox PH in the GLM Framework

In the piecewise log-constant hazards model, the time axis is subdivided into $0 = s_0 < s_1 < \ldots < s_K$, where $s_K > t$, with constant baseline hazard λ_k for $k = 1, \ldots, K$ in each interval. Consequently, the hazard rate in the k^{th} time interval becomes,

$$h_i(t) = \lambda_k \, exp\Big(\sum_{j=1} \beta_j x_{ij} \Big) = exp\Big(\sum_{j=1} \beta_j x_{ij} + b_k \Big) = exp(\eta_{jk}),$$

where b_k is $log\,\lambda_k$ and $\eta_{jk} = \sum_j \beta_j x_{ij} + b_k$ for the i^{th} individual. The log likelihood for a date point in the K^{th} interval is

$$\begin{aligned} l = log(f(t)) = log\Big(h(t)S(t)\Big) &= log(h(t)) - \int_0^t h(u)du \\ &= \delta\eta_K - (t - s_K)\,exp(\eta_K) - \sum_{k=1}^{K-1}(s_{k+1} - s_k)\,exp(\eta_k), \end{aligned}$$

where $\delta = 1$ if the data point represents death and $\delta = 0$ if it is censored. (Breslow, 1972).

The log likelihood of the piecewise log-constant hazards model is the same as the log likelihood of K Poisson distributed points with $K - 1$ of them having rate $\lambda_k = (s_{k-1} - s_k)\,exp(\eta_k)$ and the observation equal to zero and the last observation with rate $\lambda_K = (s_{K-1} - s_K)\,exp(\eta_K)$ and the observation equals to 1 if the survival is not censor and 0 if it is censored (Laird and Oliver, 1981).

The formulation of the Cox PH regression model in the form that has been discussed so far can not be applied using the INLA methodology. However, Martino et al. (2011) introduced a rearranging of the data points into K Poisson distributed data points so that they can be modelled using the LGM framework. There will therefore be K piecewise log-constant baseline hazards $(b_1, \ldots, b_K)$ that are Poisson distributed. Assuming that the difference between any two consecutive baseline hazard rates is iid $N(0, \tau^{-1})$, a scale RW1 smoothing prior in (3.18) assigned as a default to the aforementioned hazard rates (Martino et al., 2011). Using the INLA framework, one can also model non-linear effects of covariates by assigning RW1 and RW2 models. Then, the PC prior in (3.21) is then assigned to the precision parameter τ of the models.

Using INLA the piecewise log-constant Cox PH model is fit to the data in table 2.5. Since some similarity between adjacent age and time period is assumed, the scaled RW2 prior in (3.19) is assigned to the age and time effect. The hyperparameter is then assigned the PC prior in (3.21). The predictor piecewise log-constant cox model used here is

$$h(t) = h_0(t)exp\Big(\alpha + \beta_g + \varphi_{age} + \gamma_{year}\Big), \tag{5.10}$$

where α is the intercept, β_g represents the fixed gender effect, φ_{age} represents the age effect and γ_{year} represents the effect of time period at the incidence of MI . The time axis stretches upto 1808 weeks and it is partitioned into 70 equally space time intervals. Each time interval represents therefore about six months (26 weeks).

5.3.3 Results

Assuming the exponential distribution for the hazard rate implies that the hazard rate does not change with time. This in turn leads to the $log(\hat{H}(t))$ plotted on the $log(t)$ to be linear with a as the slope. While the Weibull distribution allows monotonic changes in the hazard rate with time. This distribution also produces a linear $log(\hat{H}(t))$ curve on a $log(t)$ horizontal axis. In figure 5.7, the $log(\hat{H}(t))$ is plotted against $log(t)$ for different time periods (top left), sexes (top right) and age groups (bottom). None of the curves in the figure are linear. Therefore, the Weibull and exponential distribution can not be used to model the survival time of the participants who suffered MI.

In addition, figure 5.7 displays that the $log(\hat{H}(t))$ curve is parallel for the various age groups (bottom) and sexes (top right). The $log(\hat{H}(t))$ curve for the different period (top left) is less parallel than the to other plots. It must however be noted that age at the time of MI varies substantially in across periods.

The parallel $log(\hat{H}(t))$ curves in figure 5.7 suggests that the hazards in the various groups do not vary with time. This property is essential for modelling the survival time of those who suffer from MI using the proportional hazards model. In addition, figure 5.7 shows that the Weibull and exponential regression models are inappropriate in this case. The semi-parametric Cox PH model with a piecewise log-constant baseline hazards is preferred.

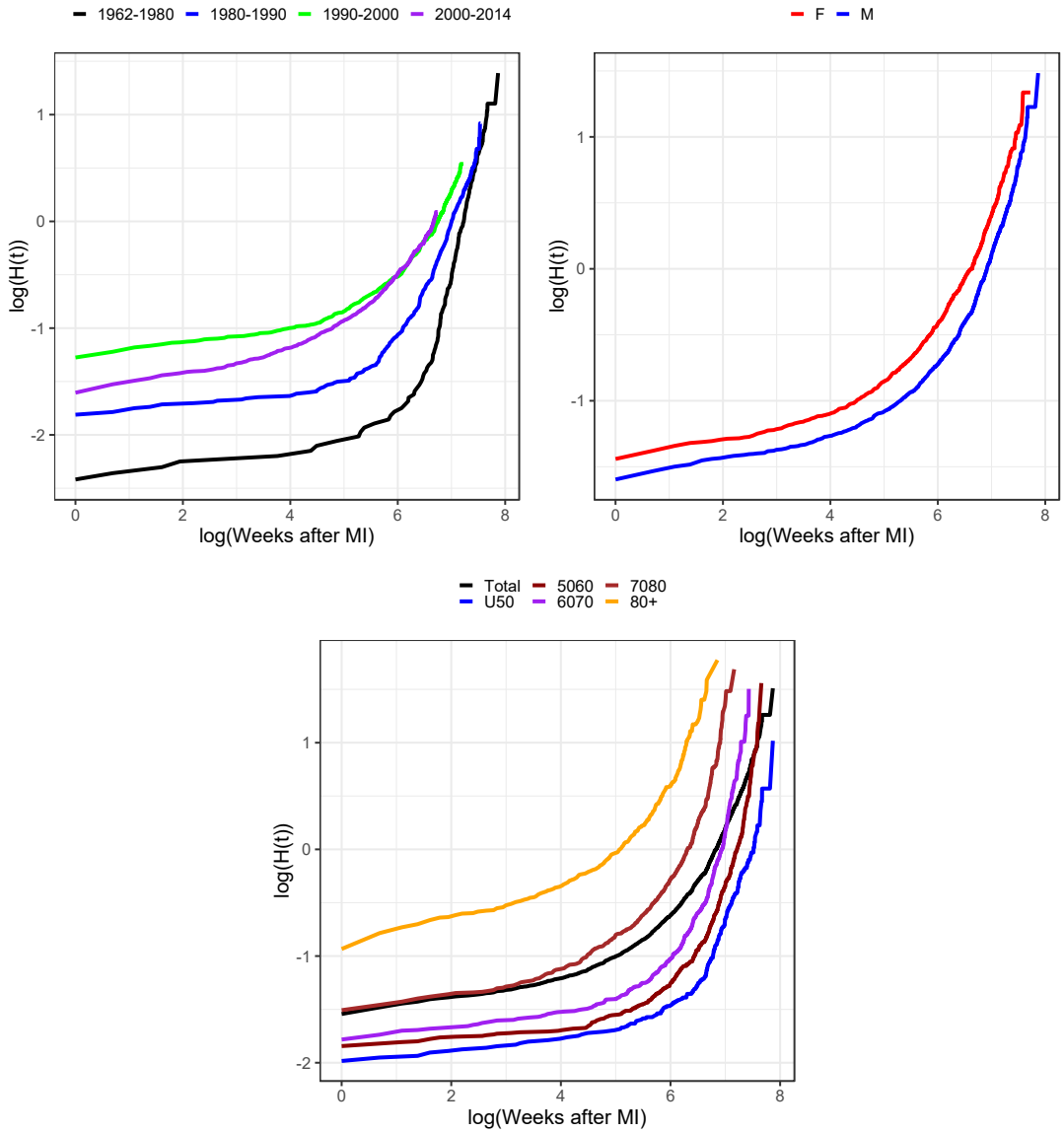


Figure 5.7: *$log(\hat{H}(t))$ curves obtained from the Kaplan-Meier survival curve for the various time periods (top left), sexes (top right) and age groups (bottom)*

In figure 5.8, the marginal posterior mean of the log-baseline hazard (left) and the transformed base line hazard (right) of the of survival data presented in table 2.5 is presented with the solid line. The dashed lines show the 95% credible interval. Each step in figure 5.8 represents 26 weeks, since the time axis was partitioned

into 70 equally space time intervals. The hazard is the highest at the time of MI (first 26 weeks) and drops abruptly in the second 26 weeks. It then stays at the same level for the preceding 1000 weeks. It increase steadily for the next 500 weeks.

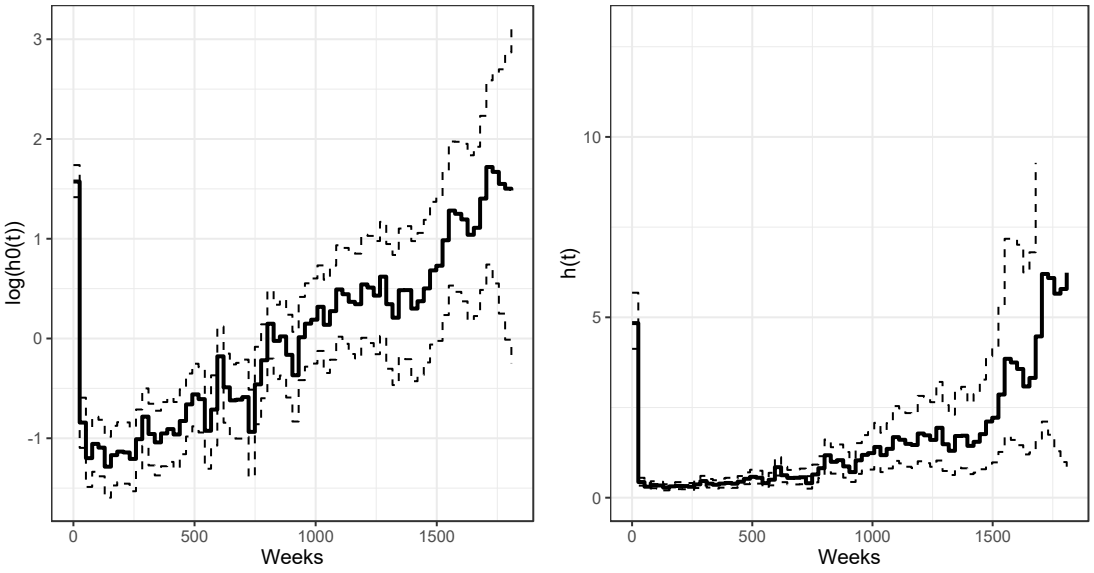


Figure 5.8: *The marginal posterior mean of the log baseline hazard presented by the solid line (left) and the posterior mean of transformed baseline hazard presented by solid line (right). Time axis divided into 70 equally spaced intervals. 95% credible interval marked by the dashed lines*

The marginal posterior mean of the age effect is presented in figure 5.9 with the solid line (left). The shaded region marks the 95% credible intervals of the age effect. An increasing age effect is observed from the age of 50 to 100. To the left in figure 5.9, the marginal posterior mean effect of time of the MI is displayed by the solid line. The 95% credible interval of the effects of the time of MI is marked by the shaded area. A constant trend in the effect of time of MI is observed between 1980 and 1995. It is then followed by a slightly decreasing trend beyond 1995 is observed.

The fixed gender effect β_g is normally distributed with marginal posterior mean

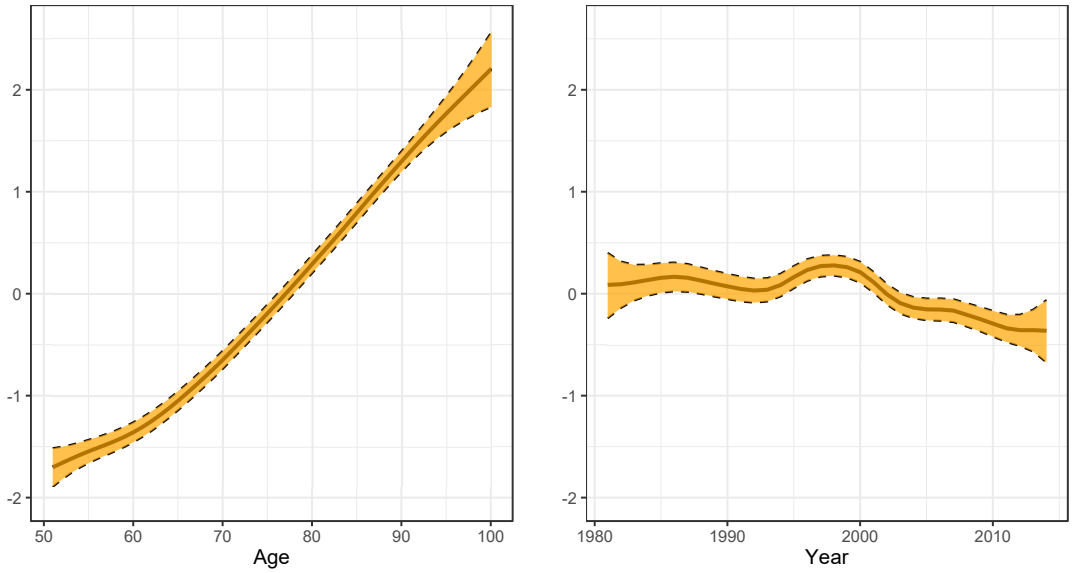


Figure 5.9: *The Posterior mean of the age effect (left) and time of MI effect (right) presented by the solid line and the shaded area display the 95% credible interval intervals.*

equal to 0.23 and the 95% credible interval is (0.15,0.32). Since the hazard ratio (HR) between males and females is

$$\begin{aligned} HR &= \frac{h_0(t)\, exp\,(\beta_m)}{h_0(t)\, exp\,(\beta_f)} = exp\,(\beta_m - \beta_f) \\ &= exp(0.23) = 1.26. \end{aligned}$$

The risk of death after MI is about 25 % higher for the males compared to females The results are discussed further in chapter 6.

Chapter 6

Discussion and concluding remarks

The datasets analysed in this thesis only contained information on gender, age, dates of MI and dates of death. The analysis show a decrease in the incidence rate of MI since about the year 2000. This trend is seen in all the age groups between the ages of 40 and 80 and for both sexes. To further study the causes of this trend, we would need more covariate information, for example tobacco, alcohol consumption, body mass index, other life style diseases and preventative medical drug use . The analysis confirms that there is a higher incidence rate of MI among males than females, and that the mean age at the time of the incidence of MI among males is significantly lower than among females. The multivariate BAPC model showed a high correlation in the period effect of males and females. This indicates that covariate information might influence the incidence rate of MI of both sexes in a similar way. The difference between the sexes in the average risk of MI decrease with increasing age. The analysis did not reveal any birth cohort related effect in the average relative risks.

Covariate information would also be interesting in analysing the survival time after an MI. However, here we can only analyse the hazard rate in terms of gender and the age and year at which the participants had an MI. Using a Cox

PH model within the Bayesian framework, both the age and year effect can be incorporated as non-linear random effects. The results illustrate a close to linear effect for age at the time of MI, while the effect of year at the time of MI is less clear. However, a slight decrease in the hazard rate is observed for those who experienced MI after the year of 2000. As expected, the baseline hazard estimate illustrates a high risk of dying right after the MI and then the hazard rate increases with age. Especially, the oldest age group has a increased risk of dying with the first week of an MI compared to the other age groups. The risk of death after MI is significantly higher for males than females. Note that we do not have information on the actual cause of death.

The statistical methods used in this thesis is versatile and can easily be adapted to study other issues. These could for example include longitudinal data analysis if we had access to time varying covariates. It could also be interesting to study the differences between geographical regions if we had access to data beyond the Tromsø study.

Bibliography

O.O. Aalen, Ø. Borgan, and H.K. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Springer, 2008. ISBN 978-0-387-20287-7.

G Albrektsen, I Hauch, M løchen, DS Thelle, T Wilsgaard, IH Njølstad, and KH Bønna. Gender gap in risk of incident myocardial infarction: the tromsø study. *JAMAInternal Medicine*, 176:1673–11679, 2016.

J.O. Berger, J.M. Bernardo, and D. Sun. The formal definition of reference priors. *The Annals of Statistics*, 37:905–938, 2009.

J.M. Bernardo. The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–121, 1996.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 1994. ISBN 9780470028735.

J. Besag, P. Green, D. Higdon, and K Mengersen. Bayesian computation and stocastic systems. *Statistical Science*, 10:3–41, 1995.

Breslow. Discussion on regression models and life table (by d.r. cox). *Journal of the Royal Statistical Society. Series B*, 34:216–217, 1972.

D Clayton and E Schifflers. Models for temporal variations in cancer rates. ii: Age-period-cohort models. *Statistics in medicine*, 6:469–481, 1987b.

L.K. Cordani and S. Wechsler. Teaching independence and exchangeability. In

A. Rossman and B. Chance, editors, *International Association for Statistics Education, Salvador (Brazil)*, 2006.

M.K. Cowles. *Applied Bayesian Statistics: With R and OpenBUGS Examples*. Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461456964.

D Cox. Regression models and life table. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

L Fahrmeir, T Kneib, S Lang, and B Marx. *Regression: Models, Methods and Applications*. Springer-Verlag Berlin Heidelberg, 2013. ISBN 978-3-642-34332-2.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955.

W.J. Hall and T.A. Wellner. Confidence bands for a survival curve from censored data. *Oxford University Press on behalf of Biometrika Trust*, 67:133–143, 1980.

T Holford. Age-period-cohort analysis. *Encyclopaedia of biostatistics*, pages 105–123, 2005.

L.A. Hopstock, A.S. Fors, K.H. Bønna, J. Mannsverk, I. Njølstad, and T Wilsgaard. The effect of daily weather conditions on myocardial infarction incidence in a subarcticpopulation: the tromsø study 19742004. *Journal of Epidemiology and Community Health*, 66:815–820, 2011.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences.*, 87:453–461, 1945.

J. Jortveit, Digre R., C. Risøe, T. Hole, T. Mannsverk, S.A. Slørdahl, and S. Halvorsen. Hjerteinfarkt i norge i 2013. *Tidsskrift for Den norske legeforening*, 19:1841–1846, 2014.

J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, INC., 2002. ISBN 0-471-36357-X.

E.L Kaplan and P. Meier. Nonparametric estimaion from incompete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

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

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

N. Laird and D. Oliver. Covarianvce analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76:231–240, 1981.

W. Lexis. Einleitung in die theorie der bevölkerungsstatistik. strassburg: Karlj. trbner. 1875.

X. Liu. *Survival Analyisis : Models and Applications*. Higher Education Press, 2012. ISBN 978-0-470-97715-6.

D Machin, Y.B. Cheung, and M.K.B. Parmar. *Survival analysis : a practical approach*. John Wiley and Sons, INC., 2006. ISBN 10 0-470-87040-0.

J Mannsverk, T Wilsgaard, E Mathiesen, M løchen, K Rsmussen, DS Thelle, IH Njølstad, L Hopstock, and KH Bønna. Trends in modifiable risk factors are associated with declining incidence of hospitalised and non-hospitalised acute coronary heart disease in a population. *Circulation*, 133:74–81, 2016.

S. Martino, R. Akerkar, and H Rue. Approximate beysian inference for survival models. *Scandinavian Journal of Statistics*, 38:514–528, 2011.

T.G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with inla: New features. *Computational Statistics and Data Analysis*, 67:68–83, 2013. ISSN 0167-9473.

N.E Miller, D.S. Thelle, O. Førde, and D.O. D Mjos. The tromsø heart study: High-density lipoprotein and coronary heart-disease: A prospective case-control study. *Lancet*, 1:965–968, 1977.

CLJ Murray, D. Phil, and AD Lopez. Global health: Measuring the global burden of disease. *The New England journal of medicine*, 369:448–459, 2010.

A Riebler and L Held. The analysis of heterogeneous time trends in multivariate age-period-cohort models. *Biostatistics*, 11:57–69, 2010a.

A Riebler, L Held, and H Rue. Correlated gmrf priors for multivariate age-period-cohort models. *Proceedings of the 25th International Workshop on Statistical Modelling*, pages 455–460, 2010b.

H Rue and L Held. Gaussian markov random fields: theory and applications, 2005.

H Rue, S Martino, and N Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society*, 71:319–392, 2009.

H Rue, A Riebler, S.H. Sørbye, J.B. Illian, D.P. Simpson, and F.K. Lindgren. Bayesian computing with inla: A review. *The Annual Review of Statistics and Its Applications*, 4:395–421, 2017.

D. Simpson, H Rue, A Riebler, T.G. Martins, and S.H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32:1–28, 2017.

SH Sørbye and H Rue. Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial statistics*, 8:39–51, 2014.

G Sulo, J Igland, and Nygård. Favourable trends in incident of ami in norway during 2001-2009 do not include yoounger adults: a cvdnor project. *European Journal of Preventive Cardiology*, 21:1358–1364, 2014.

G Sulo, J Igland, S.E. Vollset, M. Ebbing, G.M. Egeland, I Ariansen, and G.S. Tell. Trends in incident acute myocardial infarction in norway: An updated analysis to 2014 using national data from the cvdnor project. *European Journal of Preventive Cardiology*, 25:1031–1039, 2018.

D.S. Thelle, E Arnesen, and O. Førde. The tromsø heart study does coffee raise serum cholesterol? *New England Journal of Medicine*, 308:1454–1457, 1983.

JW Tukey. Comparing individual means in the analysis of variance. *International Biometric Society*, 5:99–114, 1949.

R.E. Walpole, R.H. Myers, S.L. Myers, and K.E. Ye. *Probability and Statistics for Engineers and Scientists: Pearson New International Edition.* Pearson Education Limited, 2013. ISBN 9781292037035.