Joint survival models: a Bayesian investigation of longitudinal volatility

Dirk W. Bester, Wolfson College

Department of Statistics, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy in Statistics

---

Michaelmas Term 2014

Abstract

In this thesis, we investigate joint models of longitudinal and time-to-event data. We extend the current literature by developing a model that assigns subject-specific variance to the longitudinal process and links this variance to the survival outcome. During development we provide the theoretical definition of the model and its properties, and explore the practical implications for estimating the parameters. We use Markov Chain Monte Carlo (MCMC) methods, and compare the different samplers used in similar models in the literature with our custom MCMC algorithm, written in C++.

We use the Deviance Information Criterion to perform model comparisons, and we formalise suggestions from the literature to use posterior predictive model checking to construct a goodness-of-fit test for our model. We use the model on two real-world datasets to investigate claims relating to the importance of blood pressure volatility on stroke risk, and examine the consequences of ignoring measurement error.

We amend our model to accommodate competing risk, time-dependent baseline hazard rates, and bivariate longitudinal processes — at which point we update our MCMC samplers and identify the issues. Finally, we use our code in a separate, but related, collaboration with other researchers to analyse repeated counts data.

Acknowledgements

This thesis would not have been possible without the guidance of my supervisor, Dr. David Steinsaltz. I am also grateful that the aforementioned guidance extended far beyond matters of pure academic nature. Furthermore I would like to thank the Rhodes Trust, for providing funding to undertake my studies, and to Mary Eaton and the Warden at Rhodes House for their help in particular. Special thanks to Susan R. Hutchinson and Ruth Ripley, for the technical support without which I would not have been able to write the necessary code, and Joshua M. Curk for editing my thesis.

Lots of gratitude to my parents, Dr. Frederick Bester and Jeannette Bester, who provided an enriching environment for all my studies preceding this degree. And thank you to Juanita Bester, my loving wife, for her patience, love, and support through this journey.

Contents

Introduction

Longitudinal measurements recorded alongside time-to-event data occur in various settings, including HIV clinical trials (Pawitan and Self, 1993; Tsiatis et al., 1995; Wulfsohn and Tsiatis, 1997) or the Mexican fruit fly data described in Carey et al. (1998). In the past, the longitudinal and survival outcomes were analysed separately, but with recent advances in estimation techniques and computing resources, methods to jointly model them have become popular in statistical literature. Modelling these outcomes jointly can give new insights to the factors underlying mortality (Henderson et al., 2002) as well as the dynamics of longitudinal processes in the wake of informative censoring.

Initially, joint models were developed to model longitudinal processes affected by informative right censoring (Schluchter, 1992). It soon became clear, however, that a useful consequence of these models was the ability to measure how a longitudinal marker affected survival — especially in clinical research. For instance, it can contribute to a better understanding of medical conditions and it can also identify which markers are related to survival or disease progression, which in turn can aid physicians with treatment planning. It can also evaluate current treatment guidelines to determine whether they are the best reflection of our current knowledge of a condition.

Consider the 'usual blood pressure hypothesis', as defined by Rothwell (2010):

> The theoretical true underlying level of blood pressure, which cannot be measured with total precision, but which is widely considered to be the most important component of blood pressure, determining its adverse

> effects and accounting for the benefits of antihypertensive drugs. Risk relations between measurements of blood pressure and risk of vascular events can be corrected for inaccuracy in estimation of usual blood pressure by adjustment for regression-dilution bias.

Rothwell (2010) is suggesting, here, that blood pressure-related risk is usually determined using mean blood pressure, corrected for measurement error. The author argued that the blood pressure-related risk of vascular events should not be measured purely based on the underlying level of blood pressure, but to take the volatility of blood pressure into account as well. He supported this statement by showing evidence that emerged from studies regarding blood pressure risk, but with different objectives in mind than establishing the importance of blood pressure volatility.

Consequently, it is insufficient to fully prove the argument, but it is enough to warrant further investigation into the matter. Since having a stroke is the outcome most strongly related to hypertension, it is an appropriate subject for scrutiny when examining blood pressure movements as an indicator of vascular risk.

In a different paper, Rothwell et al. (2010) also investigated the importance of blood pressure volatility on stroke risk. We specifically draw attention to table 1, showing results from a Cox model. Other than the caption reproduced below, the exact model specification is not given. The table presents results from multiple Cox models, using individual mean and volatility of Systolic Blood Pressure (SBP) as covariates. Volatility is calculated as Standard Deviation (SD) or Coefficient of Variation (CV). Since individuals in the study provided different numbers of blood pressure measurements, the tests were repeated using 2, 4, 6, 8, and 10 measurements. Each time the models include individuals with at least n longitudinal observations, so the model using $n = 2$, for instance, also includes the individuals with 4 or more observations. We notice that the confidence intervals for the Hazard Ratio (HR) calculated using $n = 2$ and $n = 10$ measurements do not overlap. We expect that these estimates are tainted by ignoring measurement error, which we will consider in

more detail below.

Rothwell's claim about the importance of the blood pressure volatility was the main driving force behind this thesis, and we investigate it using joint survival models. We spend chapter 1 performing an exploratory analysis of the two joint survival datasets that feature prominently in this thesis, focussing on the assumptions we will use later. In chapter 2 we review the survival modelling literature, in order to understand how the likelihood is constructed in survival models, and we look at the Cox proportional hazards model. We then review relatively new facets of survival analysis, namely joint models of longitudinal and time-to-event data, spending time to consider the model fitting and estimation techniques. Continuing with the literature, we use chapter 3 to explain the principles of Bayesian statistics, the basics of MCMC methods, and how the two collaborate to allow sampling-based inference. We also review some of the standard software packages available for doing Bayesian analysis with MCMC.

Using SD SBP

	Mean SBP			SD SBP		
n	HR	(95% CI)	p-val	HR	(95% CI)	p-val
2	2.44	(1.53 ; 3.89)	<0.0001	1.15	(0.73 ; 1.81)	0.55
4	2.44	(1.39 ; 4.29)	0.002	1.51	(0.86 ; 2.66)	0.16
6	2.49	(1.24 ; 4.97)	0.01	2.02	(0.97 ; 4.22)	0.061
8	1.85	(0.84 ; 4.10)	0.13	6.01	(1.72 ; 20.96)	0.005
10	1.44	(0.58 ; 3.57)	0.43	13.04	(1.66 ; 102.60)	0.015

Using CV SBP

	Mean SBP			CV SBP		
n	HR	(95% CI)	p-val	HR	(95% CI)	p-val
2	2.67	(1.74 ; 4.11)	<0.0001	1.09	(0.73 ; 1.62)	0.67
4	2.82	(1.67 ; 4.76)	<0.0001	1.50	(0.90 ; 2.48)	0.12
6	3.07	(1.62 ; 5.83)	0.001	1.98	(1.05 ; 3.77)	0.036
8	2.68	(1.29 ; 5.56)	0.008	5.00	(1.75 ; 14.30)	0.003
10	2.26	(0.98 ; 5.17)	0.055	13.05	(1.74 ; 97.66)	0.012

Table 1. The table given in Rothwell et al. (2010, p. 897): 'Table 1: Hazard ratios (top vs bottom quintile) for risk of subsequent stroke (*i.e.*, after the measurement period) in the UK-TIA trial from a model combining mean SBP and visit-to-visit variability in SBP (SD or CV or VIM), repeated with increasingly precise estimates of both variables.'

Chapter 4 explains the model we use to investigate Rothwell's claim: a joint survival model that links the longitudinal volatility with the survival outcome. We define the model, and offer a parameter estimation scheme using standard software, as well as a custom MCMC algorithm. We use the model to analyse Rothwell's dataset, and present our findings, along with techniques for model comparison. We explain the design of the custom MCMC algorithm in chapter 5, and we use it to investigate the use of Metropolis-Hastings updates in parameter estimation of joint survival models.

In chapter 6, we extend the work of other authors on joint survival models, formalising their goodness of fit methods for use with our own joint model. In the same chapter, we see how ignoring covariate measurement error can bias results, and we examine the meaning of 'regression dilution', a term that often occurs in blood pressure studies. Chapter 7 continues our contribution to the literature, by showing extensions of the model from chapter 4. Specifically, we introduce competing risks, time-dependent baseline hazard rates, and a bivariate longitudinal process.

Chapter 8 discusses a problem that is not directly related to our joint survival models, but which can be investigated using the same code and techniques. Here, we direct our attention to a study with repeated count measures, where multiple events per subject were possible. This data originated from a study on the behaviour of schoolchildren, and whether it could be improved using vitamin supplements.

Chapter 1

Exploratory data analysis

1.1 UK-TIA

The United Kingdom Transient Ischaemic Attack (UK-TIA) data, as discussed in Farrell et al. (1991), is one of the studies mentioned by Rothwell (2010) in support of his argument for the importance of blood pressure volatility. The dataset consists of 2,435 patients who had suffered a transient ischaemic attack or minor ischaemic stroke. They had been considered at risk of having another vascular event and were monitored from 1979 to 1985. The study recorded blood pressure at irregular intervals, with a maximum of 23 readings — figure 1.1 shows a histogram of n_i, the number of measurements provided by patient i. The time of any serious vascular-event or death was recorded and no patients were lost to follow up.

Blood pressure was measured as SBP over Diastolic Blood Pressure (DBP). Figure 1.2 shows data for the first 20 subjects in the UK-TIA dataset. Plotting the survival data in this manner is not very informative, so we also provide a plot of the Kaplan–Meier estimate of the survival function in figure 1.3. While developing our model, we initially assumed that the SBP and DBP were independent longitudinal processes of the form:

$$Y_i(t) = \mu_i + \sigma_i \varepsilon(t) \tag{1.1}$$

for the ith subject, where $\varepsilon(t)$ is a random variable, uncorrelated with time, with mean 0, and constant variance. That is, we assumed observations were generated by a subject-specific mean and variance, both of which remain constant throughout a subject's life.

We tested the assumption of heteroskedasticity using the Bartlett Test, with H_0: all samples have the same variance. For the SBP, the test returned a p-value smaller than 0.0001, validating our assumption of unequal variances. This was also the case for DBP. After establishing heteroskedasticity, we used the Kruskal-Wallis Rank Sum Test with H_0: the location parameters of the distribution of $Y_i(t)$ are the same in each group, to test our assumption of different μ_i parameters for individuals. This assumption was also validated with p-values smaller than 0.0001 for both SBP and DBP. We did not have enough observations per subject to consider the mean and

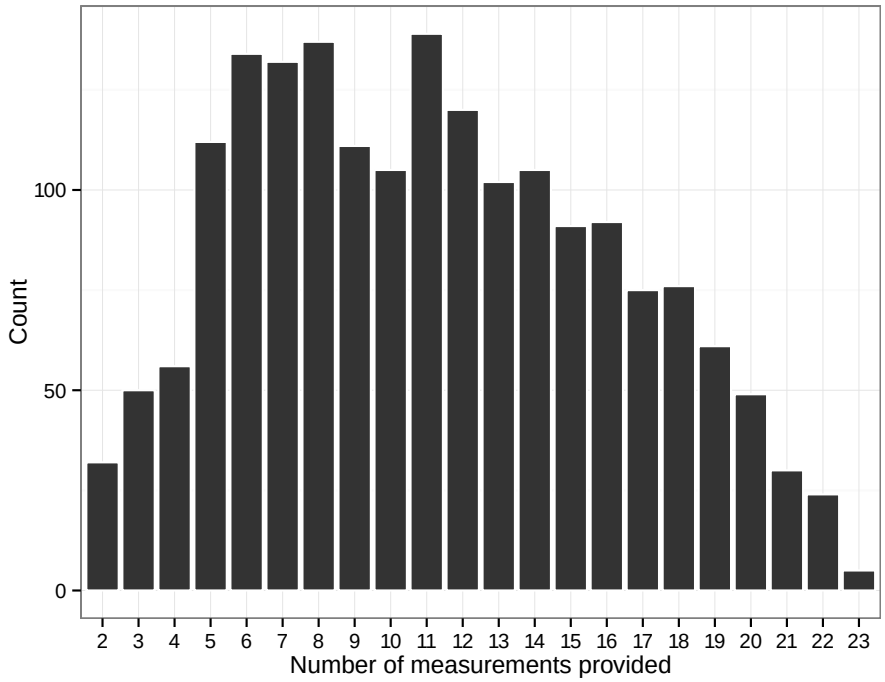


Figure 1.1. Histogram of the number of observations provided by individuals in the UK-TIA dataset.

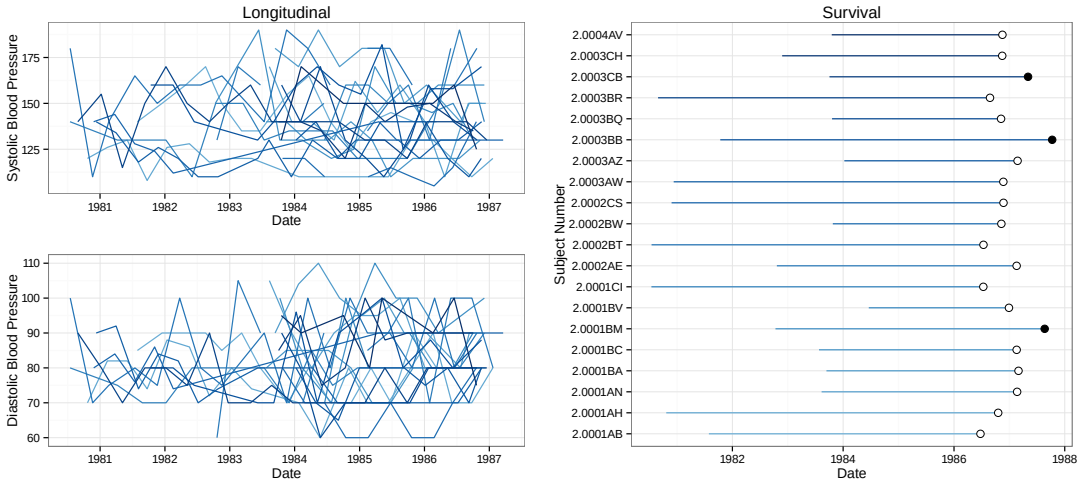


Figure 1.2. The observations for the first 20 subjects in the UK-TIA dataset.

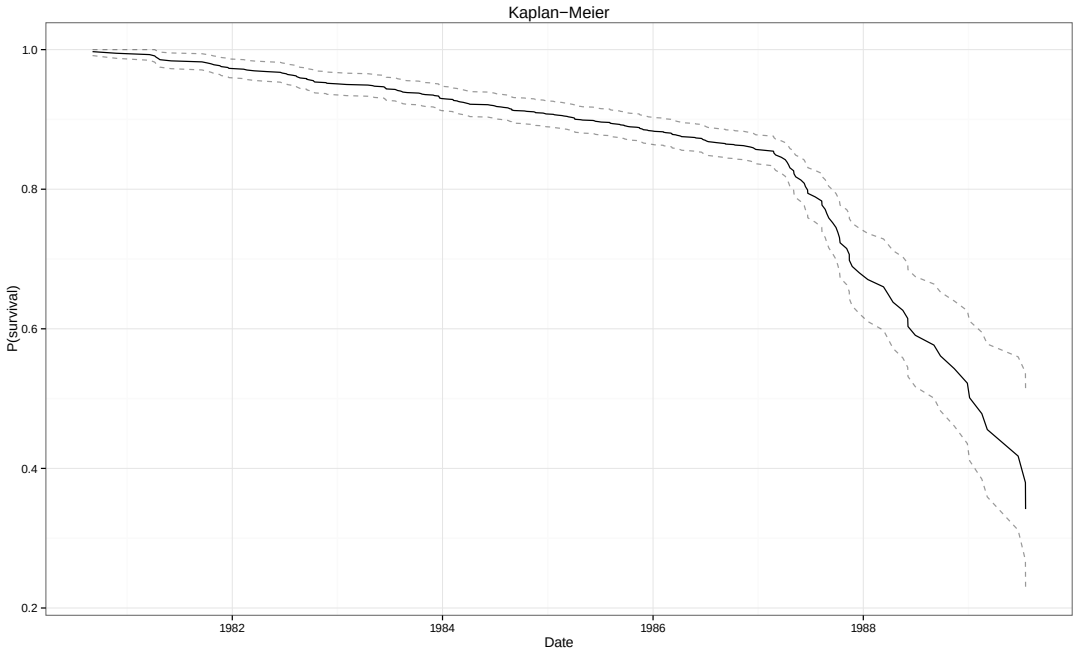


Figure 1.3. Kaplan-Meier estimate of the survival function for the UKTIA data, 1838 individuals with 229 stroke-related events.

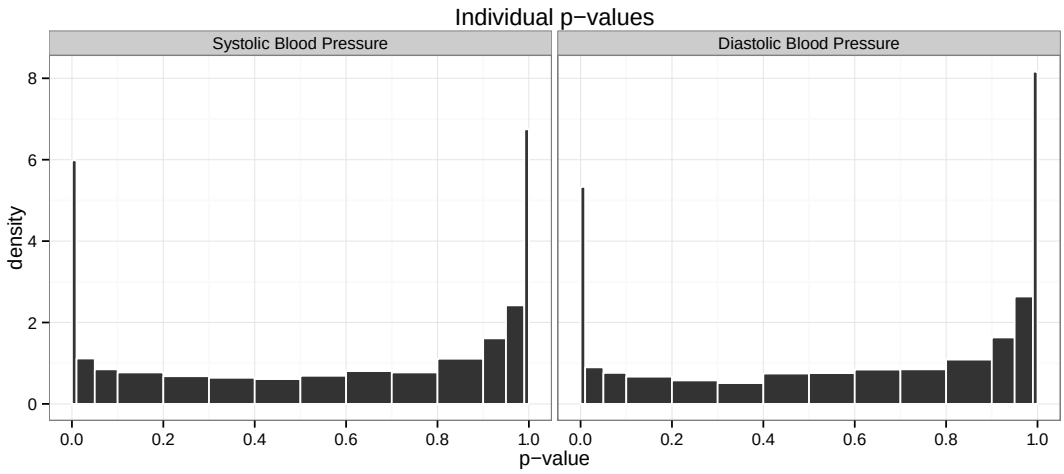


Figure 1.4. Histograms of the longitudinal process slopes, calculated for each individual over time.

variance to be time-dependent processes $\mu_i(t)$ and $\sigma_i(t)$, respectively.[1]

An implicit assumption of (1.1) is the lack of an individual slope in the data, which we investigated using a permutation test. We re-ordered the blood pressure measurements 1000 times for each individual, and calculated the slope of each random permutation. The amount of times the real slope is smaller than that of the permuted measurements can be treated as a p-value. If there truly is no individual slope, the p-values across all individuals will have a uniform distribution between zero and one. Figure 1.4 shows histograms, for SBP and DBP, of the individual p-values from the permutation test. The distribution is not uniform, so there is some evidence of individual slopes being present, but we made a simplifying assumption that there were none.

McCullagh (2013) presented an argument for temporal realignment of longitudinal process. This involves modelling the longitudinal observations not as they arrive over time, but rather as time-until-event. The process is called a revival process, and it is defined as

$$Z_i(s) = Y_i(T_i - s)$$

[1]This is a natural extension of the models in this thesis for situations with higher frequency longitudinal data.

where T_i is the event time for individual i for time $s < T_i$, so $Z_i(s)$ is the state of the process of patient i at time s prior to failure. We have that $Z_i(T_i) = Y_i(0)$ and, although $Z_i(s)$ occurs in real time, it is not observable until the subject has an event.

Motivation for the reversal is to realign the process for effective signal extractions. The reasoning is that if some feature of the longitudinal process — such as the slope — is indeed driving events, we should see some pattern emerge if we look at the revival process. In this scenario the longitudinal processes of different subjects may not have a clear pattern visible over age, so we stand to benefit from the realignment. Following the example of McCullagh (2013), we aligned the blood pressures from UK-TIA according to days-in-study, as well as days-until-event. Figure 1.5 gives the realigned times for the 229 individuals in the UK-TIA who had a stroke. They can be compared to figure 2 from McCullagh (2013), but this type of plot is difficult to interpret. The point of the realignment is to see whether there is a clear pattern before the event that could be used to predict it. There are no obvious patterns in figure 1.5, and to ascertain we created line plots with subsets of the 229 individuals. Figure 1.6 shows the SBP for 10 individuals with and without realignment. We do not see evidence of a signal becoming apparent when we realign the longitudinal processes of the SBP, and the same held true for the DBP.

For our purposes, we will assume the blood pressure processes each follow a normal distribution, meaning $\varepsilon(t) \sim N(0, 1)$ in (1.1) or equivalently $Y(t) \sim N(\mu_i, \sigma_i^2)$. Crucially, we assume that each subject generates blood pressure measurements according to an individual normal distribution — where everyone has their own mean and variance for the two processes. To test our assumption, we perform the Shapiro-Wilk test of normality for the 1,691 subjects with 5 or more non-identical measurements. We give a histogram of Shapiro-Wilk's p-values in figure 1.7, and we did not reject normality for 70% of the SBP processes and 51% of the DBP processes.

Finally, we investigated the assumption of the two longitudinal processes being

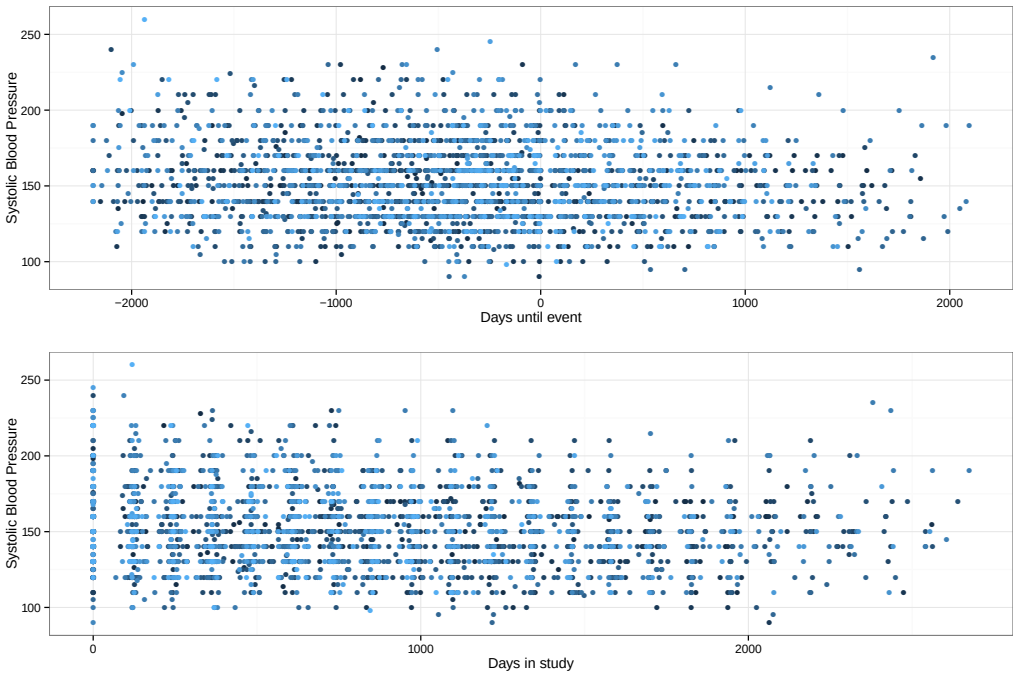


Figure 1.5. SBP, original process vs temporal realignment for the 229 individuals who had a stroke.

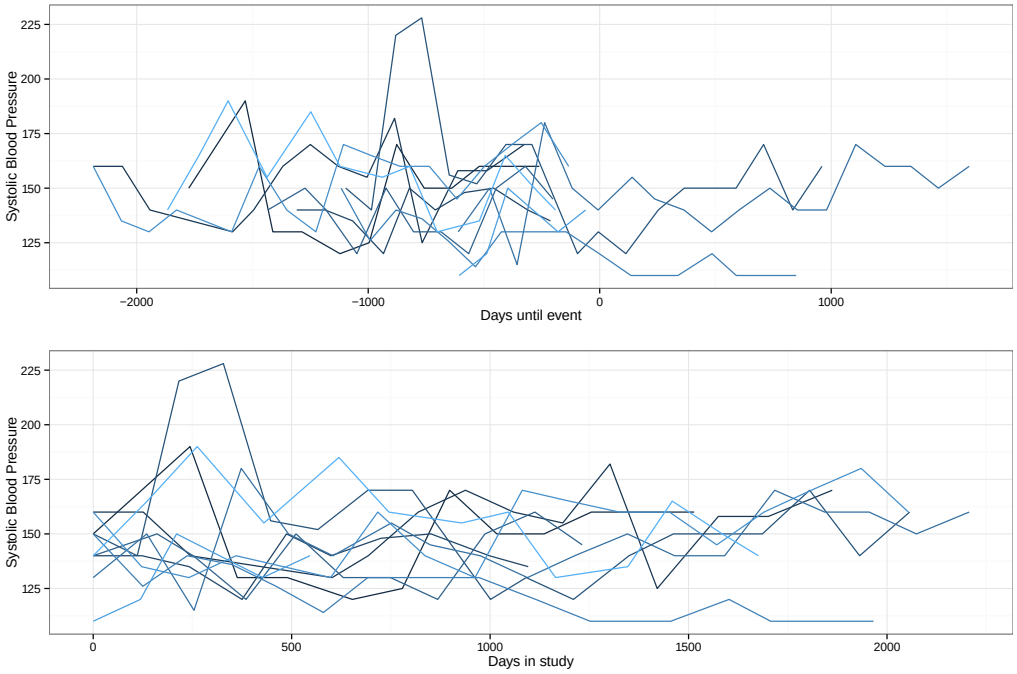


Figure 1.6. SBP, original process vs temporal realignment for 10 of the 229 individuals who had a stroke.

independent. We calculated the within-individual correlation for the 1,809 individuals with non-zero variances, given in figure 1.8. The average correlation is 0.53, and 92% of the individuals have a correlation larger than zero. With this in mind, the assumption that the SBP and DBP processes are independent does not appear to hold. We will initially analyse the SBP and DBP in separate models, with the aim of joining them in a multivariate model later in the thesis. Until then we will regard the independence assumption (or that there is only one process to include) as a simplifying assumption that will allow us to develop a simple model. Our goal is to develop a simple base model as a foundation for further model development.

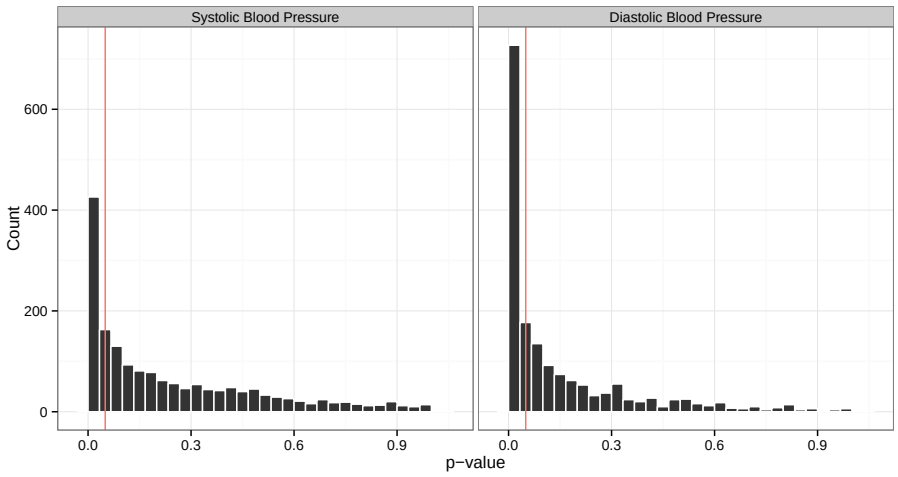


Figure 1.7. Individual p-values for the Shapiro-Wilk test. The 0.05 level is marked by the red line.

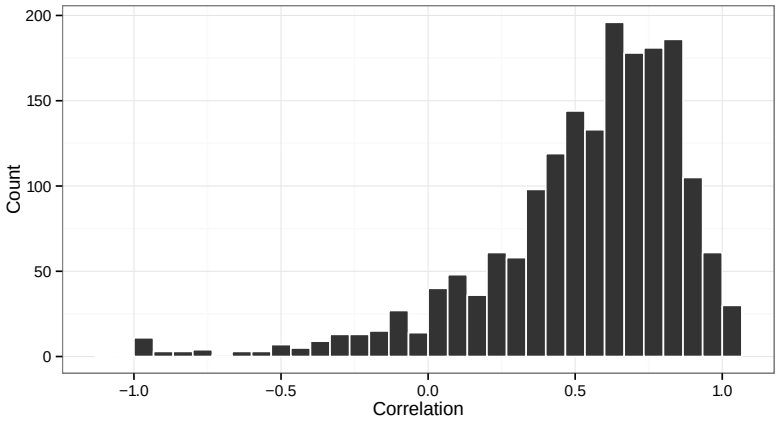


Figure 1.8. Histogram of within-individual correlations.

1.2 NHANES

The second dataset we will use in this thesis is the National Health and Nutrition Examination Survey (NHANES) from the National Center for Health Statistics (NCHS) (2013). This dataset contained about 18,000 observations, ten times that of the UK-TIA dataset. There were around 1000 heart related events, as well as deaths due to injury and cancer. Unlike the earlier study, however, subjects did not provide blood pressure readings over time; three measurements were taken at the start of the study along with various other markers. Further data were also available about the race, sex, and age at entry of study participants. Figure 1.9 shows the longitudinal data for the first 20 individuals in the NHANES dataset, alongside a plot of the Kaplan-Meier estimate of the survival function for all the individuals in the dataset.

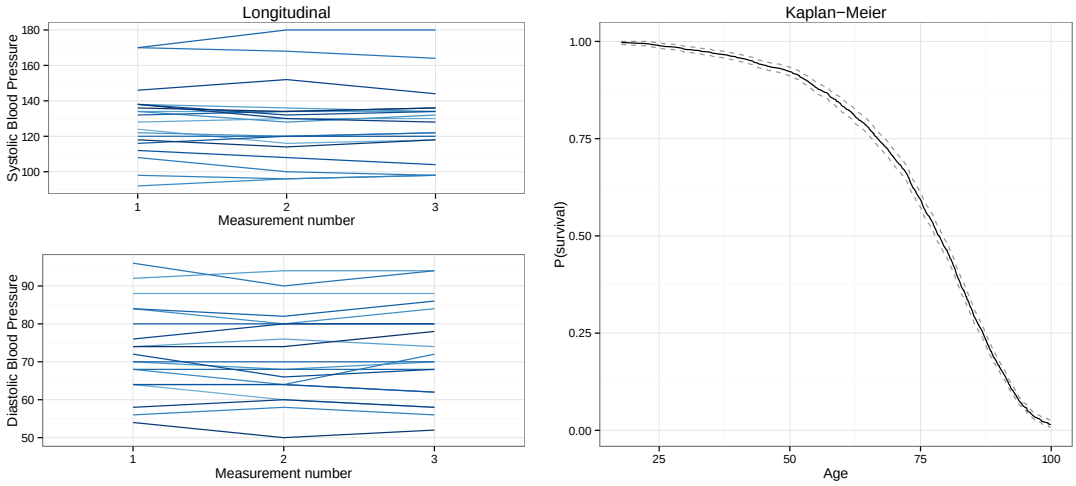


Figure 1.9. The blood pressure observations for the first 20 subjects in the NHANES dataset.The Kaplan-Meier estimate of the survival function is for 16995 individuals with 2588 deaths (all causes of mortality).

We will again use (1.1) for the longitudinal process,[2] and therefore we once again used the Bartlett Test to establish whether heteroskedasticity was present in the individual variances, followed by the Kruskal-Wallis Test to check the assumption of

[2]Since we only have 3 measurements at the start of the study, the blood pressure process in the NHANES dataset is technically not a longitudinal process. We use the term, however, to relate it to the other joint models in this thesis.

differing individual means. All assumptions were validated for both SBP and DBP, with all tests returning p-values smaller than 0.0001.

To test whether there was indeed a lack of an individual slope present in the longitudinal data, we again performed a permutation test. Each patient only provided three measurements, however, so we could not perform the exact same test as in section 1.1. Thus, we grouped subjects according to their number of unique measurements. Then we calculated all possible slopes for permutations of each subject's measurements. For instance, a person with three unique measurements will have six possible slopes. A person with only one unique measurement will only have one possible slope, so we exclude those individuals from this test. We numbered the possible slopes from small to large, and wrote down the number of each individual's real slope. Under the null hypothesis of no individual slope being present we expect the slope numbers to have a uniform distribution. Figure 1.10 shows the histograms — for SBP and DBP — of the slope numbers for each individual. The plots do not appear to exhibit uniform distributions, but we will not check it formally. We again made the simplifying

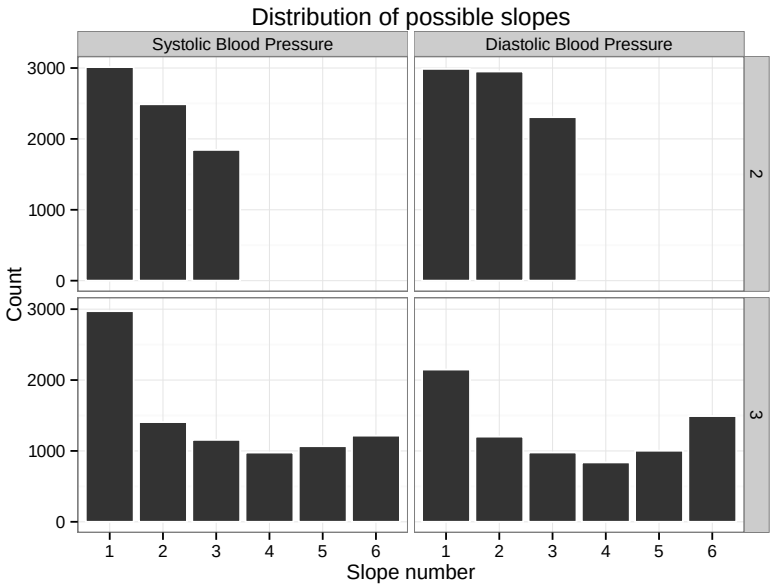


Figure 1.10. Histograms of the slopes, calculated for each individual by fitting a line through the three blood pressure measurements.

assumption that there was no individual slope present. This allowed us to focus on the estimation algorithm for a simple model, where assumptions could be relaxed after successful implementation.

As with the UK-TIA data, we assumed the SBP and DBP each have a normal distribution and we therefore performed a Shapiro-Wilk normality test for each individual. We had only 3 observations of SBP and DBP for each individual, so although we could still perform the test, its power was very low. For the SBP, we did not reject normality for 50% of the individuals, and the corresponding figure for the DBP was 48%. We checked the independence of the SBP and DBP processes by calculating the correlation between each individual's SBP and DBP measurements. Figure 1.11 shows a histogram of the within-individual correlations, and we have no reason to suspect a correlation between three SBP and DBP measurements taken a few minutes apart on the same day. We will revisit this idea in chapter 7.

Mild tendency for observers to prefer certain last digits in reporting blood pressure measurements has been reported in blood pressure studies (Ostchega et al., 2003). The last-digit preference in our NHANES data is substantial, with about 26% of

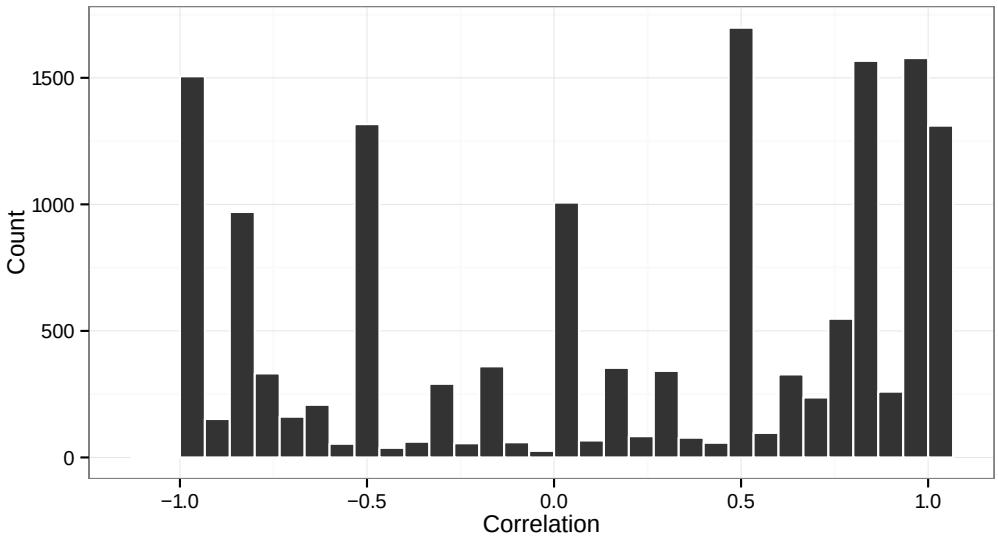


Figure 1.11. Histogram of within-individual correlations.

	0	2	4	6	8
SBP	0.240	0.198	0.158	0.169	0.235
DBP	0.267	0.187	0.160	0.159	0.227

Table 1.1. Fraction of measurements with each final digit.

all the DBP measurements ending in 0, but only about 30% ending in 4 or 6. One expects approximately 20% of the reported observations to end in each of the possible digits 0,2,4,6, and 8. These will be slightly modified by the overall distribution of blood pressure measurements. Since the true blood pressure values are spread over multiple decades — the range from first to ninth decile of DBP measurements is (60,90), and for SBP is (104,156) — this effect will be very small. Instead of an equal distribution we see the proportions indicated in table 1.1. These distortions agree broadly with those previously reported for a later wave of NHANES blood pressure measurements, but they are substantially more extreme (Ostchega et al., 2003). By increasing the clustering of the observations, this preference for reporting last digits 0 and 8 will somewhat reduce the variance of observations. Since the effect will be small, we ignored it in our main analysis, but it will be relevant in our analysis below of pseudo-replication.

Pseudo-replication

After performing initial analyses on the NHANES data, we received a further part of the dataset which included measurements taken at subjects' homes. This meant we had two sets of three measurements to analyse for each patient. We discovered a difference between the measurements taken at home, and the measurements taken at the clinic. Multiple lines of evidence independently show that some of the examiners — possibly most of them — either intentionally or inadvertently duplicated measurements, rather than recording three independent measurements. Our first attempt was to estimate how many individuals in the sample would be expected to have three identical

	SBP		DBP	
	Mean	SD	Mean	SD
m	128	0.167	74	0.1
τ	0.0023	0.000026	0.0073	0.00009
r	6.072	0.45	2.82	0.078
λ	120	10.28	55	2

Table 1.2

.

or two identical measurements, if they had three independent measurements. The complication here is that three measurements for the same person are more likely to match than three random blood pressure measurements from different people. We thus used a simulation study, with a parametric distribution on the blood pressure measurements. We fitted a model to individual means and variances for SBP and DBP, using the same assumptions we will use later in this thesis:

$$\begin{aligned} Y_i(t) &\sim N(\mu_i, \tau_i^{-1}) \\ \mu_i &\sim N(m, \tau^{-1}) \\ \tau_i &\sim \Gamma(r, \lambda), \end{aligned} \tag{1.2}$$

where r and λ are the shape and rate of the Gamma distribution, respectively. Parameter estimates are given in table 1.2, we will discuss this model and its parameter estimation in chapter 4. We used these assumptions to execute the following algorithm:

1. Sample n independent SBP means and variances and n independent DBP means and variances.

2. For each mean and variance sample three independent blood pressure values normally distributed with that mean and variance.

3. Round the blood pressure values to even integers. To account for the last-digit preferences, we assign the digits 0,2,4,6,8 to values ending in the ranges (8.8,1.5), (1.5,3.2), (3.2,4.9), (4.9,6.6), (6.6,8.8) respectively for DBP, and ranges (9.0,1.4),

(1.4,3.4), (3.4,5.0), (5.0,6.7), (6.7,9.0) for SBP.

4. Count the number of triples that are either all identical, have two matches, or have three distinct values.

5. Repeat 10 times, and average.

For the home data, 5.0% of the subjects had three identical measurements reported for SBP, and 6.4% for DBP, only 2.2% and 3.0% respectively should have been expected to have three identical measurements according to our model, implying that more than half of the subjects with reported zero variance are spurious. The clinic data agreed with our predictions, having 2.5% and 2.8% triple matches, respectively. There is also more direct evidence that duplications are partly due to examiner error.

If each individual had three independent measurements, we would expect that when there are two identical measurements reported, the odd one is equally likely to be any one of the three. If there was a trend in the measurements — for example, a tendency of patients to relax and lower their blood pressure, or the reverse — we would expect the second measurement to be the least likely to be distinct. Instead, we find the counts and proportions given in table 1.3. The disparity is huge, and it is not what we would expect from independent measurements. It is, however, very much

Which one distinct	1	2	3
SBP	41.9%	33.9%	24.2%
DBP	40.0%	35.8%	24.2%

Table 1.3. Fraction that the first, second, or third measurement is the distinct one of the three, for the home data.

Which one distinct	1	2	3
SBP	36.8%	32.6%	30.5%
DBP	35.7%	33.0%	31.3%

Table 1.4. Fraction that the first, second, or third measurement is the distinct one of the three, for the clinic data.

consistent with what we would expect to see if some examiners at some times took two independent measurements, and then copied one of the first two measurements for the third. If we think of the 24.2% in category 3 (measurements 1 and 2 identical) as representing the true rate of matching, then we would conclude that about 27% of the apparent matches are spurious, which agrees with the inference we drew from the simulations. Table 1.4 shows the proportions for the clinic data, and they are more in line with our expectations.

The last piece of evidence emerges from the results recorded by different examiners. There are 12 different examiners coded in the data, as well as various versions of missing data. If the triple matches were an accurate reflection of the similarity of the subjects' blood pressure measurements, we would expect that all examiners would have approximately equal numbers of such results. Instead, we find the results in figure 1.12. Three examiners have much higher fractions of identical blood pressure measures than all ten others, and it is the same three for both SBP and DBP measures. If we exclude those three, the results for the other ten examiners are consistent (according to a chi-squared test) with the hypothesis that they all have the identical frequency of three identical measurements, 3.9% for SBP and 5.4% for DBP. The former matches closely the proportions that we expect from our model, but the latter is still substantially higher than what we expect.

There is very little variation among the examiners in the proportion of subjects reported with two matching measurements. If we look at the position of the non-matching measurement, however, there is considerable variation, as we see in the ternary plot figure 1.13. Every single examiner had fewer matches between the first and second measurements than between first and third or between second and third. The three examiners who stood out for their proportion of triple matches (marked green in figure 1.13) also have the highest proportion of matches between the second and third measurements. Only one of the examiners has proportions that do not fail a

chi-squared test at the 0.05 level for equality of the proportions of the three different possible matches.

In principle, there might be an argument for removing all the individuals with zero SD in either SBP or DBP from the study population, to remove the contamination. We have carried this out during our analysis of the NHANES data, and found results that are similar to those we report in this thesis. Since we found the effect to be small, we have chosen to include as many of the subjects as possible in our final analysis.

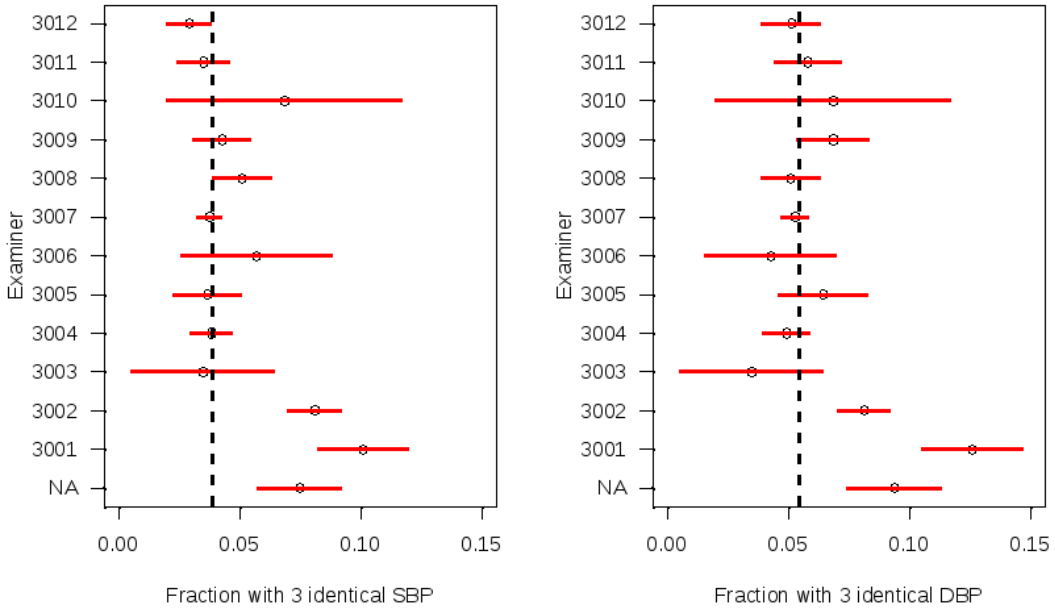


Figure 1.12. Fraction of subjects with three identical blood pressure measurements reported by each examiner, with 95% confidence intervals. Black dashed line is the average of all examiners other than 3001, 3002, and NA (not given, 0 or 88888).

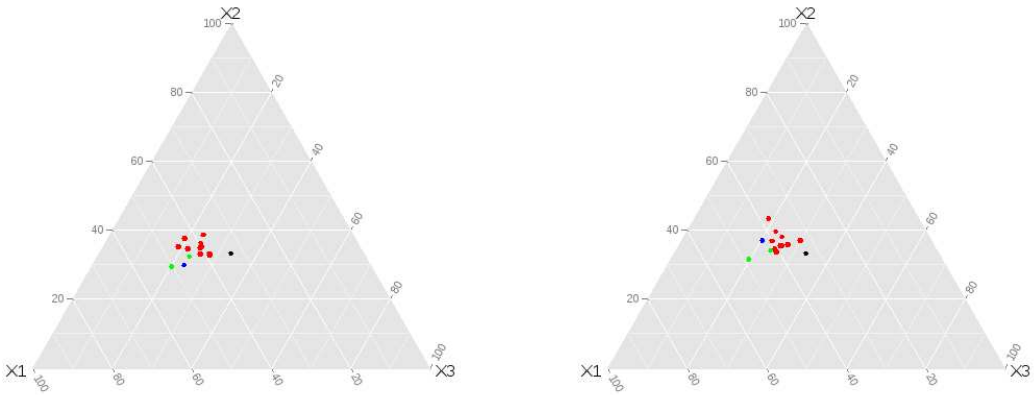


Figure 1.13. Proportion of subjects who had matching results for the first two measurements (X3), last two (X1) or the first and third (X2). Left is SBP, right DBP. The black dot is the point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The green dots correspond to the two examiners who had exceptionally high proportions of triple matches. The red dots correspond to the other ten examiners. The blue dot includes all results where the examiner was missing.

Chapter 2

Information from mortality

Survival analysis, or the broader definition of time-to-event analysis, occupies itself with the study of waiting times. These can be waiting times to a single event, such as the time to death, or time spent in a particular state, such as the progression of a disease. They can also be the inter-arrival times of applicable events, such as modelling the number of offences committed by prison inmates. The study of survival and event arrival is not limited to humans, as it can be used in any setting involving an unknown waiting time before an event of interest. Thus, applicable candidates for which the dynamics can be modelled include automotive parts, light bulbs, mortgage payments, or anything that changes or develops and is monitored over time.

This chapter introduces survival analysis and the difficulties with modelling time-to-event data. We shall see that, although we can make use of likelihood methods, there are special qualities about survival data that need to be taken into consideration. Most notable is the occurrence of censoring, which often takes place in studies with a limited monitoring period. Furthermore, although each subject in the population in question experiences the risk of the event happening, this risk cannot be measured in any single individual.

2.1 Classical survival analysis

2.1.1 The survival function

Time-to-event data is analysed through the survival function, that is, the probability that an event occurs after time t (the entity survives beyond time t), defined as

$$S(t) = P(T > t).$$

For a continuous random variable T, the survival function is a strictly decreasing function. Furthermore, $S(t)$ is the complement of the cumulative distribution function $F(t) = P(T \leq t)$, or in other words $S(t) = 1 - F(t)$. The survival function can also be expressed as the integral of the probability distribution function,

$$S(t) = P(T > t) = \int_t^\infty f(x)dx,$$

thus

$$f(t) = -\frac{dS(t)}{dt}.$$

In accordance with the features of a probability distribution function, $f(t)$ is a non-negative function with the area under $f(t)$ being equal to one. However, when working with time-to-event data, we can regard $f(t)$ as the 'approximate' probability that an event will occur at time t.

2.1.2 The hazard rate

In addition to the survival function, the hazard rate is another fundamental concept in survival analysis. It is defined by means of a conditional probability. Statisticians call it the hazard rate, but since it has been the subject of analysis in many fields it is known by other terms as well (Klein and Moeschberger, 2003; Steinsaltz et al.,

2012), such as the force for mortality to demographers, the mortality rate to biologists, the conditional failure rate in reliability studies, the intensity function in stochastic processes, the age specific failure rate in epidemiology, or the inverse of Mill's ratio to economists.

The hazard rate is the probability of an event occurring within a small time frame from t to $t + dt$, given that it has not occurred at time t, or

$$h(t) = \lim_{dt \to o} \frac{P\left(t \leq T < t + dt | T \geq t\right)}{dt}. \tag{2.1}$$

This can be written as

$$h(t) = -\frac{d}{dt} \ln\left[S(t)\right].$$

Then, through integration and using $S(0) = 1$, we can solve for the survival function

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t h(x)dx\right) \\ &= \exp\left(-H(t)\right), \end{aligned}$$

where $H(t) = \int_0^t h(x)dx$ is known as the integrated hazard rate. The key characteristic of the hazard rate is that it uniquely defines the survival function. We can use it to completely specify a survival model.

A common choice used in queuing problems, as well as reliability theory, is the unit hazard rate, given as $h(t) = 1$. This amounts to exponential waiting times and the approach is popular due to the memoryless property of the exponential distribution. While this property is useful in queuing problems, it is sometimes inadequate for analysing survival in biological populations. Thus, models with more flexible hazard rates have been developed, such as hazard rates with polynomial or exponential growth. A specification that demonstrates the latter is the popular Gompertz-Makeham law,

which has an age-independent component (Makeham, 1860) and an age-dependent component (Gompertz, 1825), giving the hazard rate

$$h(t) = \alpha e^{\beta t} + \lambda. \tag{2.2}$$

The popularity of this hazard function is due to the seemingly exponential increase in human mortality, especially after the age of 30. Some authors have reported that the exponential hazard is inadequate for humans of advanced age, due to late-life mortality deceleration (Greenwood and Irwin, 1939). However, Gavrilov and Gavrilova (2011) investigated the phenomenon, giving a historical overview and emphasising the challenges of estimating hazard rates at extremely old ages. They found that the Gompertz law adequately describes mortality up to the ages of 102-105 and suggested that the earlier findings of late-life mortality deceleration appeared to be artefacts of mixing together different cohorts and using cross-sectional — rather than longitudinal — data.

Before moving on to likelihood construction, it is worth noting the Nelson-Aalen estimator, used to estimate the integrated hazard rate non-parametrically. It is defined as

$$\tilde{H}_{NA}(t) = \sum_{t_i \leq t} \frac{\delta_i}{n_{t_i}}$$

for observed times $t_1 \dots t_N$ with δ_i being the number of events at t_i and n_{t_i} being the total individuals at risk at t_i. This method is used to gain insight to the hazard rate without making parametric assumptions. It is a well known quantity in survival analyses.

2.1.3 Likelihood

Before we attempt to construct a likelihood for modelling purposes, we need to look at the factors that make survival data special. These inhibit us from using the well-

developed regression methods available for continuous and discrete data. We might consider waiting times to be no more than data points that can be studied through simple linear regression or generalised linear models. This is not possible with survival data, since it is almost sure to contain censoring. The reason being that, since we are studying the time-to-event, we will have to wait for the event to occur during the data collection period. However, studies have limited resources, so the event of interest might not occur during the available period.

For example, during a period of observing divorces, we find that some couples divorce, while others do not and thus experience censoring at the end of the study. While the latter is a positive matter for the couples involved, it has consequences for the statistician. We cannot simply ignore non-events, since the fact that it did not occur within a specific period still provides information about the waiting time. This is the most common type of censoring, known as right censoring. It can also occur during the study if an individual leaves due to an event unrelated to the study. Continuing the previous example, the latter could occur when a marriage ends due to reasons other than divorce, or if a couple migrates and as a result gets lost to follow up. Figure 2.1 shows the usual pattern observed in time-to-event studies for calendar time, as well as time spent in study by each subject. The only thing we learn from a right-censored lifetime, is that the event time is larger than the censoring time. Other types of censoring include interval censoring and left censoring. We do not consider them here, but an explanation of likelihood construction in settings with these types of censoring, as well as truncation, is given in Klein and Moeschberger (2003). They also outline a simple method for likelihood construction when dealing with right censoring as described here — type I censoring, as they refer to it.

For a lifetime T experiencing the hazard rate $h(t)$, let δ be the indicator taking 1 if the lifetime is censored, and 0 otherwise. Let C be the censoring time and V the time we observe. If the event is observed, V will take the value of T, and C if it is

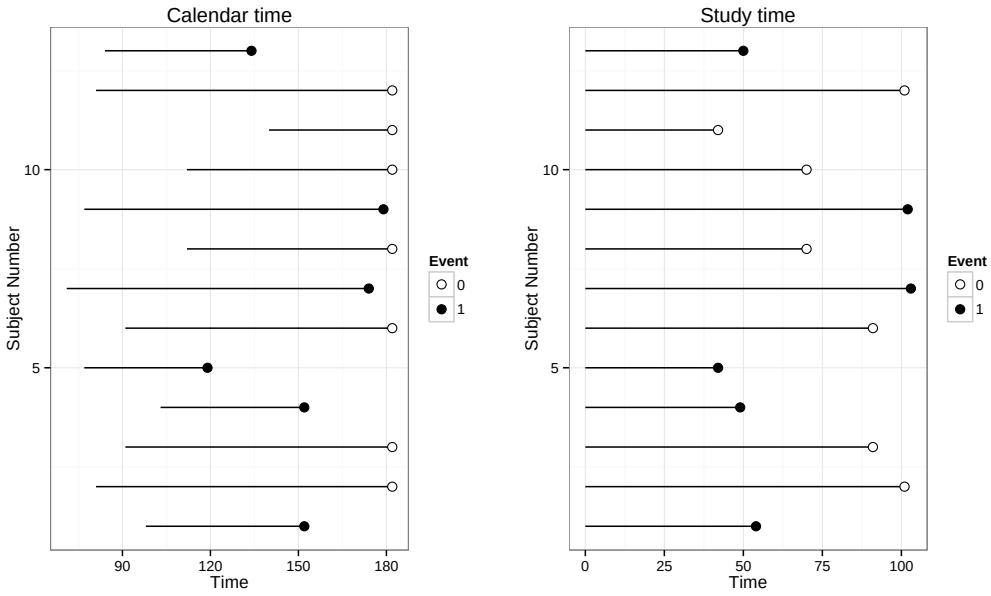


Figure 2.1. Showing right censoring for subjects with different stariting times, and the same subjects with starting times backed up to 0.

censored, that is $V = \min(C, T)$. Thus, for each individual in the study, we observe the pair of random variables $\{V, \delta\}$, with a likelihood contribution of

$$\left[h(v_i)^{\delta_i}\right] \exp\left(-H(v_i)\right).$$

Then for a group of N subjects with observation pairs $\{v_i, \delta_i\}$, the complete likelihood can be written as

$$P(\boldsymbol{v}, \boldsymbol{\delta}) = \left[\prod_{i=1}^{N} h(v_i)^{\delta_i}\right] \exp\left(-\sum_{i=1}^{N} H(v_i)\right), \tag{2.3}$$

giving the log likelihood

$$\log(L) = \sum_{i=1}^{N} \delta_i \log h(v_i) - \sum_{i=1}^{N} \{H(v_i)\}. \tag{2.4}$$

The likelihood in (2.4) resembles that of a Poisson process. It is not surprising, since time-to-event data can be successfully treated using counting processes, and thorough

explanations are given in Klein and Moeschberger (2003, c. 3.6) and Aalen et al. (2008, c. 5). It is important to keep in mind that the likelihood in (2.3) is constructed with the assumption that censoring is non-informative, or random. In other words, the censoring time is independent of the survival time. This assumption is often appropriate, and in other cases it is regarded as a necessary simplifying assumption, though care should be taken as it does not always hold. Lagakos (1979) gives three examples in clinical trials where this assumption is questionable, mostly relating to cases where censoring may have taken place due to ill health and thus censored individuals may be expected to have shorter life expectancies.

2.1.4 Cox proportional hazards model

A notable development in survival modelling came after the seminal paper by Cox (1972), introducing the proportional hazards model — also known as the Cox model. It revolutionised the literature on survival modelling (Sinha et al., 2003) and to date it has been cited in almost 35,000 papers. This model makes use of the assumption that the hazard rate at time t is given by

$$h(t) = h_0(t) \exp(\boldsymbol{X}\boldsymbol{\beta}), \tag{2.5}$$

where $h_0(t)$ is a baseline hazard rate at time t, experienced by all individuals, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters. The $n \times p$ matrix $\boldsymbol{X}$ contains the covariate vectors $x'_{(1)}, x'_{(2)}, \ldots x'_{(n)}$ for the n individuals in the model. The semi-parametric form of (2.5) makes it difficult to use ordinary likelihood methods to estimate the parameters (Aalen et al., 2008, p. 134). Rather, these are estimated using the partial likelihood derived by Cox (1975)

$$\prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{x}'_{(i)}\boldsymbol{\beta})}{\sum_{j \in R(v_i)} \exp(\boldsymbol{x}'_{(j)}\boldsymbol{\beta})} \right)^{\delta_i} \tag{2.6}$$

for the observed data set

$$\{(v_i, \delta_i, \boldsymbol{x}'_{(i)}) : i = 1, 2, \ldots, n\},$$

where v_i is the observed time, δ_i is the event indicator and $R(t) = \{i : v_i \geq t\}$ is the set of individuals at risk at time t.

Using the partial likelihood eliminates the need to estimate $h_0(t)$, simplifying the problem. This is especially alluring in studies interested only in the influence of a certain covariate on event risk, where no knowledge or inference is required about the baseline hazard rate $h_0(t)$. Should we need it, however, Breslow (1972) showed that the cumulative baseline hazard rate can also be estimated after fitting the Cox model, by using

$$\begin{aligned}\hat{H}_0(t) &= \sum_{v_i \leq t} \hat{h}_0(v_i) \\ &= \sum_{v_i \leq t} \frac{1}{\sum_{j \in R(v_i)} \exp(\boldsymbol{x}'_{(j)}\boldsymbol{\beta})},\end{aligned}$$

known as the Breslow estimator of the cumulative baseline hazard rate.

We can use (2.3) to write the full likelihood that would apply to the Cox model using the hazard rate in (2.5), as

$$L = \prod_{i=1}^{N} \left[\left(h_0(v_i) e^{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}} \right)^{\delta_i} \exp\left(- \int_0^{v_i} h_0(u) e^{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}} du \right) \right]. \tag{2.7}$$

This likelihood is more complicated than (2.6) and parameter estimation is not trivial. However, Guo and Carlin (2004) showed that parameters in (2.7) can be successfully estimated in a Bayesian model using MCMC, and they also showed that this can be undertaken using readily available software packages.

A benefit of using the full likelihood, as opposed to the partial likelihood in (2.6),

is the ability to estimate $h_0(\cdot)$ and $\boldsymbol{\beta}$ simultaneously. We can assign a parametric form to $h_0(t)$, such as (2.2), or we can model it non-parametrically by assuming that it takes on a point mass of $h_0(t_1), h_0(t_2), \ldots, h_0(t_m)$ at the m distinct failure times $t_1, t_2, \ldots, t_m$. This leads to a semi-parametric model, since the hazard rate consists of the non-parametric baseline hazard and the parametric assumption of $e^{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}}$. In this case, the likelihood can be written as

$$\begin{aligned} L &= \prod_{i=1}^{N} \left[\left(h_0(v_i) e^{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}} \right)^{\delta_i} \exp\left(- \int_0^{v_i} h_0(u) e^{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}} du \right) \right] \\ &= \prod_{i=1}^{N} \left[\left(\prod_{j=1}^{m} h_0(t_j)^{I(v_i=t_j)} e^{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}\delta_i} \right) \exp\left(-e^{\boldsymbol{x}'_{(j)}\boldsymbol{\beta}} \sum_{j=1}^{m} h_0(t_j) I(t_i \leq v_i) \right) \right], \end{aligned}$$

where $I(\cdot)$ is the indicator function. When there is no covariate measurement error, this approach of maximising the full likelihood is known to lead to the same estimates for β and $h_0(\cdot)$ as the maximum partial likelihood and the Breslow estimators, respectively (Hu et al., 1998). However, covariate measurement error complicates matters and leads to estimation bias. We will investigate the extent of estimation bias in the case of joint longitudinal and survival models in chapter 6.

2.2 Multiple events per subject

In the preceding sections we dealt with situations where having an event terminated a subject's participation in the study. Some situations, however, might allow for multiple events per subject requiring the analysis of multiple event times. We can use similar methods as described before, assuming events arrive according to a non-homogeneous Poisson Process. That is, the number of events at time t, $N(t) : t \geq 0$ is a non-homogeneous Poisson Process with rate $h(t)$. If the rate is constant, that is $h(t) = \lambda$, then we have a homogeneous Poisson Process.

The consequence is that $N(t)$, the number of events arriving by time t, has a

Poisson distribution with rate $H(t) = \int_0^t h(x)dx$. The rate used here is analogous to the hazard rate discussed in section 2.1.2, and it signifies the probability of an event occurring in the infinitesimal period dt between times t and dt.

An individual with hazard rate $h(t)$ under observation between times v and w, who had n events at times t_j, $j = 1 \dots n$ where $v \leq t_1 < t_2 < \dots t_n \leq w$ contributes an amount to the likelihood equal to

$$\begin{aligned}
&h(t_1)e^{-\int_v^{t_1} h(x)dx} h(t_2)e^{-\int_{t_1}^{t_2} h(x)dx} h(t_3)e^{-\int_{t_2}^{t_3} h(x)dx} \dots\ h(t_n)e^{-\int_{t_{n-1}}^{t_n} h(x)dx} e^{-\int_{t_n}^{w} h(x)dx} \\
&= \left[\prod_{j=1}^{n} h(t_j)\right] e^{-\left(\int_v^{t_1} h(x)dx + \int_{t_1}^{t_2} h(x)dx + \int_{t_2}^{t_3} h(x)dx + \ \dots\ + \int_{t_{n-1}}^{t_n} h(x)dx \int_{t_n}^{w} h(x)dx\right)} \\
&= \left[\prod_{j=1}^{n} h(t_j)\right] e^{-\int_v^w h(x)dx}
\end{aligned}$$

or a log likelihood

$$\sum_{j=1}^{n} \log h(t_j) - \int_v^w h(x)dx.$$

This can easily be extended to a case with multiple individuals each having a hazard rate $h_i(t)$ and n_i event times $v_i \leq t_{i1} < t_{i2} < \dots t_{in_i} \leq w_i$ with v_i and w_i being study entry and exit times, as before. Then the total likelihood becomes

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{N} \left(\left[\prod_{j=1}^{n_i} h_i(t_{ij})\right] e^{-\int_{v_i}^{w_i} h_i(x)dx} \right) \\
&= \left[\prod_{i=1}^{N}\prod_{j=1}^{n_i} h_i(t_{ij})\right] \exp\left(-\sum_{i=1}^{N}\int_{v_i}^{w_i} h_i(x)dx\right) \qquad (2.8)
\end{aligned}$$

and hence the log likelihood

$$\log \mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log h(t_{ij}) - \sum_{i=1}^{N}\int_{v_i}^{w_i} h_i(x)dx, \qquad (2.9)$$

which is similar to (2.4), meaning we can easily modify code designed for single-event

studies to accommodate multiple events per subject.

2.3 Joint survival models

Recent years have seen papers with informative reviews on the subject of joint models of survival and longitudinal data, such as Tsiatis and Davidian (2004) or Neuhaus et al. (2009). A detailed discussion of joint models from a Bayesian perspective can be found in Ibrahim et al. (2005, c. 7).

The first paper that used random effects to join the longitudinal response with informative drop-out, is Wu and Carroll (1988). They combined a linear mixed effects model with a general distribution $M(t)$ conditional on the random effects for each subject to take censoring into account. Next, they used numerical integration to estimate parameters by maximising a pseudo-likelihood. Since then, this type of modelling has received attention from various sources, leading to diverse model specifications and estimation procedures. We will elaborate on a few of the influential examples encountered in the literature, starting with earlier developments.

Similar to Wu and Carroll (1988), Schluchter (1992) was interested in modelling the dynamics of a longitudinal process in the presence of informative right censoring. This was done using a simple transformation model in which time to drop-out was jointly modelled with subject random intercept and slope, as a multivariate normal distribution. Using a joint normal model avoided the integration over random effects and its accompanying difficulties. The paper mentioned that the survival time appeared to affect the slope of an individual's longitudinal development. A referee then pointed out that it is more natural to argue in the other direction — the slope influences the survival time. This is one of the first instances in statistical literature suggesting that longitudinal markers may predict survival. Another paper that used the joint normal modelling approach was De Gruttola and Tu (1994). Although this

approach simplifies implementation through avoiding integration, it places strong distributional assumptions on the survival outcome, which may not always hold in practice. Pawitan and Self (1993) used a fully parametric model, with a Weibull regression for the survival times and a generalised linear model for the longitudinal data. They estimated parameters using numerical integration, maximising the likelihood through an optimisation algorithm.

The first review of the joint modelling framework was Little (1995), who introduced two general classifications of the models used in the preceding literature. To illustrate this, let $\boldsymbol{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{in_i}\}$ be the vector of longitudinal observations and v_i the event time of the survival outcome for the ith subject. The classification then follows from the factorisation. The first is called a pattern-mixture model and conditions on the survival time:

$$f(\boldsymbol{y}, \boldsymbol{v}) = f(\boldsymbol{y}|\boldsymbol{v})f(\boldsymbol{v}). \tag{2.10}$$

The second is referred to as a selection model, and it conditions on the longitudinal process:

$$f(\boldsymbol{y}, \boldsymbol{v}) = f(\boldsymbol{v}|\boldsymbol{y})f(\boldsymbol{y}). \tag{2.11}$$

To this end, the type of data and objectives of a study may influence whether a modeller uses pattern-mixture or selection models. Little (1995) explained these classifications in a setting where characteristics of the longitudinal process were usually the aim of the study, rather than the time to event — denoted by R_i, the interval number in which a subject dropped out. When the longitudinal process is of interest, pattern-mixture models are frequently used since exact specification of the drop-out mechanism is not required (Guo et al., 2004).

The later review on joint models by Neuhaus et al. (2009) extended the classifications of (2.10) and (2.11). They stated that, in addition to mixture and selection models, there is also the shared parameter approach, attained through assuming

conditional independence of $\boldsymbol{y}$ and v_i:

$$f(\boldsymbol{y}_i, v_i|\boldsymbol{\theta}_i) = f(\boldsymbol{y}_i|\boldsymbol{\theta}_i)f(v_i|\boldsymbol{\theta}_i).$$

Assuming a distribution for the parameters $\boldsymbol{\theta}_i$ we end up with the following joint model for N independent subjects:

$$\begin{aligned} f(\boldsymbol{y}, v) &= \prod_{i=1}^{N} \int f(\boldsymbol{y}_i, v_i|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &= \prod_{i=1}^{N} \int f(\boldsymbol{y}_i|\boldsymbol{\theta}_i) f(v_i|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \end{aligned} \tag{2.12}$$

The assumption that $\boldsymbol{\theta}_i \sim N(0, \Sigma_\theta)$ is often used. Building on (2.12) we need to further denote c_i as the possible censoring time for subject i, with $v_i^* = \min(c_i, v_i)$ and event indicator $\delta_i = I(v_i \leq c_i)$ to allow us to write the contribution to the likelihood for each subject. For subject i, we will observe $(\boldsymbol{y}_i, v_i^*, \delta_i)$ as well as a possible vector of observation times $\boldsymbol{t_i} = \{t_{i1}, t_{i2}, \ldots, t_{in_i}\}$. In the shared random effects scenario, the survival part of the distribution is often specified through the hazard rate:

$$h(t|\boldsymbol{\theta}_i) = h_0(t) \exp\left[g(\boldsymbol{X}_i(t), \boldsymbol{\beta}, \boldsymbol{Z}_i(t), \boldsymbol{\theta}_i)\right],$$

where $\boldsymbol{X}_i(t)$ is a matrix of time varying covariates for the vector $\boldsymbol{\beta}$ of corresponding fixed effects. In the same way, $\boldsymbol{Z}_i(t)$ is a matrix corresponding to random effects $\boldsymbol{\theta}_i$ and $h_0(t)$ denotes the baseline hazard. The exact specification of g is not important at this stage, but rather to note that we end up with a likelihood of the form:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int \left\{ \prod_{j=0}^{n_i} f(y_{ij}|\boldsymbol{\theta}_i) \right\} h(v_i^*|\boldsymbol{\theta}_i)^{\delta_i} \exp\left[- \int_0^{v_i^*} h(t|\boldsymbol{\theta}_i) dt \right] f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \tag{2.13}$$

that can be used to estimate the parameters. The integrals in (2.13) are often intractable, requiring the use of numerical techniques for parameter estimation.

Expectation-maximization (EM) algorithms were most prevalent amongst early authors, but recently Bayesian frameworks using MCMC have become popular (Ibrahim et al., 2005).

Apart from the classification as either a shared parameter or mixture and selection model, Diggle et al. (2008) added three further characteristics. First, a model can incorporate random effects, as in (2.13). Second, models can be semi-parametric (Song et al., 2002). Third, as mentioned earlier, integration can be simplified through the use of a transformation model, such as Schluchter (1992) or De Gruttola and Tu (1994).

Now that we are familiar with the classifications, we can proceed to look at more examples of joint models in the literature, paying attention to the specification of the longitudinal process, the assumptions underlying time to event or subject drop-out, and the technique used to estimate the parameters. The observed values $\boldsymbol{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{in_i}\}$ at times $\boldsymbol{t_i} = \{t_{i1}, t_{i2}, \ldots, t_{in_i}\}$ may be subject to measurement error. Thus, it is usual to give a specification of the true underlying longitudinal process as another process plus error. If we denote the observed process as $Y_i(t_{ij})$ being y_{ij} observed at time t_{ij}, we can write $W_i(t_{ij})$ as the true process:

$$Y_i(t_{ij}) = W_i(t_{ij}) + e_i(t_{ij}), \tag{2.14}$$

where $e_i(t) \sim N(0, \sigma_e^2)$ is the measurement error, assumed to be independent of the other random effects in the model. This signifies a general measurement error for all observations across all individuals. Sufficient care should be taken to allow both for further within-subject and between-subject variance. If this is not done, we may find significant autocorrelations present in the values of $e_i(t)$, especially when the time between observations is small. When observations are taken far apart, it may be safe to assume that they are independently distributed.

Tsiatis and Davidian (2004) gave an excellent summary of different longitudinal

specifications encountered in various papers. We have slightly deviated from their notation, with the aim of expanding their list of examples. A standard approach taken when using subject-specific random effects is to characterise the true longitudinal process as a simple linear model:

$$W_i(t) = \alpha_{0i} + \alpha_{1i}t \tag{2.15}$$

This is the approach used by Pawitan and Self (1993), De Gruttola and Tu (1994), Tsiatis et al. (1995), Faucett and Thomas (1996) and Wulfsohn and Tsiatis (1997). Here, α_{0i} and α_{1i} refers to subject-specific intercept and slope. Brown and Ibrahim (2003) use a quadratic form:

$$W_i(t) = \alpha_{0i} + \alpha_{1i}t + \alpha_{2i}t^2 \tag{2.16}$$

and they use a semi parametric approach in order to relax the distributional assumptions on $\boldsymbol{\alpha}_i$. These authors specified (2.15) and (2.16) as polynomial functions of t but more flexible specifications can also be constructed:

$$W_i(t) = f(t)^T\alpha_i\,, \tag{2.17}$$

where $\alpha_i = (\alpha_{0i}, \alpha_{1i}, \alpha_{2i}, \ldots, \alpha_{pi})^T$ is a vector of random effects and $f(t)^T$ is a vector of functions of time t, which may be non-linear. Ding and Wang (2008) used a single multiplicative random effect and regarded $f(t)$ as the unknown mean function, which they modelled and estimated non-parametrically. Tseng et al. (2005) used an accelerated failure time model for the survival process together with a general polynomial form as (2.17), also used by Hsieh et al. (2006) in conjunction with a time-dependent Cox proportional hazards model. Brown et al. (2005) applied the polynomial form to a case where every subject i had multiple biological markers

recorded at each time point, thereby defining a multivariate framework.

The approach of (2.17) can also include fixed effects. This is denoted by a fixed effects (possibly time varying) covariate vector $x_i(t)$, corresponding to a vector β of fixed effects, both of length q, to produce:

$$W_i(t) = f(t)^T \alpha_i + x_i(t)^T \beta, \tag{2.18}$$

similar to the form used by Chi and Ibrahim (2006) and Rizopoulos et al. (2009). Models with the linear mixed effects structure of (2.18) can be fitted using the R (R Core Team, 2012) joint model package, `JM`, provided by Rizopoulos (2010). Furthermore, Guo and Carlin (2004) showed the differences between separate and joint modelling using standard computer packages. They used WinBUGS (Spiegelhalter et al., 2004) to fit models of the form (2.18).

Further extensions by Henderson et al. (2000), Wang and Taylor (2001) and Xu and Zeger (2001) consider a longitudinal model of the form:

$$W_i(t) = f(t)^T \alpha_i + U_i(t), \tag{2.19}$$

where $U_i(t)$ is stochastic process. Wang and Taylor (2001) take $U_i(t)$ to be an integrated Ornstein-Uhlenbeck (IOU) process. Roberts and Sangalli (2010) deviated from the forms using $f(t)^T$ and instead opted to use a latent diffusion model, based on the diffusion process satisfying the stochastic differential equation (SDE):

$$\begin{aligned} dW_i(t) &= \beta(W_i(t), \theta) + \sigma dB_t \qquad t \geq 0 \\ W_i(0) &= 0, \end{aligned} \tag{2.20}$$

where Θ (given as θ) is a random variable with values in $\mathbb{R}^d$ and σ is assumed constant and known. Specifying models using forms such as (2.19) and (2.20) allow for the

trend to vary with time and introduces within-subject autocorrelations.

Neuhaus et al. (2009) explained the three different focuses that result in the application of joint models. Firstly, there are joint models focused on serial trends of the longitudinal process. Secondly, the focus can be on the event time, with the aim of better understanding how the longitudinal process affects survival. Thirdly, models can have equal focus on both outcomes. The focus of the model will, to a large extent, affect how the survival and longitudinal parts are specified. Similarly, Tsiatis and Davidian (2004) discussed philosophical considerations regarding the structure of the longitudinal process and its dependence on the focus of the study. Although complicated specifications may mimic the true processes more accurately, they tend to have greater implementation difficulties. With this in mind, we will now proceed to look at the available approaches that have been used to estimate the parameters.

2.3.1 Model fitting and estimation techniques

Due to the complex form of the likelihood in (2.13), usually involving integrals with no closed form solution, parameter estimation requires numerical techniques. In fact, even settings such as the simple transformation model discussed in section 2.3 — where the aim is to avoid complex integrals — would often still require numerical optimisation (see De Gruttola and Tu, 1994, for example). Numerical techniques in the literature can be classified into two groups: the frequentist and Bayesian approaches, employing the EM algorithm, and MCMC methods.

Most prevalent is the frequentist setting, using the EM algorithm (Dempster et al., 1977). A long list of authors used the EM algorithm, including De Gruttola and Tu (1994), Wulfsohn and Tsiatis (1997), Hogan and Laird (1997, 1998), Song et al. (2002), Henderson et al. (2000), Guo et al. (2004), Diggle et al. (2008), and Rizopoulos et al. (2009). A Monte Carlo EM algorithm (Wei and Tanner, 1990) was used by Tseng et al. (2005) and Ding and Wang (2008). To estimate the standard error of the parameters,

Tseng et al. (2005) and Hsieh et al. (2006) used bootstrap methods in succession to applying the EM algorithm.

To our knowledge, the first authors to use a Bayesian framework for parameter estimation in a joint survival model were Faucett and Thomas (1996). This method involves taking priors on the parameters and using the resulting posterior distributions for inference. For a detailed discussion about joint survival models from a Bayesian perspective, refer to Ibrahim et al. (2005). Other authors who used MCMC Gibbs sampling include Wang and Taylor (2001), Brown and Ibrahim (2003), Brown et al. (2005), Chi and Ibrahim (2006), Xu and Zeger (2001), and Roberts and Sangalli (2010). In the next chapter we will discuss Bayesian methods and how they work together with MCMC techniques.

Chapter 3

Bayesian analysis and Markov Chain Monte Carlo

3.1 The Bayesian choice

The term 'Bayesian statistics' is used to describe the process of solving statistical problems — such as prediction, inference, model choice, signal processing, or monitoring — using Bayes' theorem. In its simplest form, this is a manipulation of conditional probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.1}$$

This formula is not exclusive to Bayesians, as frequentists also use conditional probabilities. It is, however, central to all Bayesian inference.

3.1.1 The posterior distribution

Constructing a model using the Bayesian framework involves expressing the uncertainty in model quantities — such as parameters — probabilistically, by assigning a distribution to unknown values. It provides a method to update subjective beliefs with the arrival of new information. For a set of observations Y, from a distribution

with parameters θ, inference requires setting up a joint distribution of $p(Y, \theta)$ which will consist of a prior distribution $p(\theta)$, and a likelihood $p(Y|\theta)$, such that

$$p(Y, \theta) = p(Y|\theta)p(\theta),$$

known as the full probability model. However, Bayesian inference uses the posterior distribution

$$\begin{aligned} p(\theta|Y) &= \frac{p(Y|\theta)p(\theta)}{p(Y)} \\ &= \frac{p(Y|\theta)p(\theta)}{\int p(Y|\theta)p(\theta)d\theta} \end{aligned} \tag{3.2}$$

which follows form applying Bayes' theorem in (3.1). To a Bayesian, there is no fundamental difference between the observations and the parameters in a model, as both are treated as random quantities.[1] Inference is based on the conditional probability of the quantity of interest, given the observed data.

The likelihood $p(Y|\theta)$, or $L(\theta|Y)$ by another notation, is a quantity of interest that is similar for both Bayesians and frequentists. Up to the process of parameter estimation the model assumptions that lead to the likelihood function are the same for both. This also holds true for models that appear intuitively Bayesian, such as hierarchical models. The analysis only becomes Bayesian when we assume a prior distribution on the unknown parameters and use the posterior for inference.

3.1.2 Model comparison and hypothesis testing

We can also use (3.1) to express uncertainty in the model itself by assigning a prior distribution on the range of possible models. Let M denote the model, and let $p(M)$ be

[1]Here lies the fundamental disagreement between Bayesians and frequentists. Frequentists believe that a parameter, for example the mean height of someone in a population, cannot be considered random since there has to be a true, albeit unknown, value.

the prior probability that a model is correct. We then arrive at a posterior probability for a model

$$p(M|Y) = \frac{p(Y|M)p(M)}{p(Y)}. \tag{3.3}$$

In the case where the researcher has to choose from a closed set of N models $\mathcal{M} = \{M_i, M_2, ...M_N\}$, the above expression (3.3) becomes

$$\begin{aligned} p(M_i|Y) &= \frac{p(Y|M_i)p(M_i)}{p(Y)} \\ &= \frac{p(Y|M_i)p(M_i)}{\sum_{j=1}^{N} p(Y|M_j)p(M_j)}. \end{aligned}$$

Here, $p(Y|M_i)$ is the marginal likelihood of the data given the model, given by

$$p(Y|M_i) = \int p(Y|\theta_i, M_i)p(\theta_i)d\theta_i$$

where θ_i is the parameters of model M_i. Notice that this is an expected value on θ_i:

$$p(Y|M_i) = E_\theta\Big[p(Y|\theta_i, M_i)\Big] \tag{3.4}$$

which opens up the possibility of using Monte Carlo integration — a useful property, as we shall see in the next section.

Comparisons are done using Bayes factors. Given two models M_0 and M_1, the Bayes factor is calculated as (Jackman, 2009):

$$B_{10} = \frac{p(Y|M_1)}{p(Y|M_0)} \tag{3.5}$$

$$= \frac{p(M_1|Y)}{p(M_0|Y)} \Big/ \frac{p(M_1)}{p(M_0)} \tag{3.6}$$

and it is a measure of the evidence for M_1 against M_0 given the data. A larger Bayes

factor represents more evidence for model M_1. In MCMC settings we can calculate the quantities in (3.5), using (3.4). Jeffreys (1961) provided the scale given in table 3.1 for interpreting Bayes factors.

In the same way, Bayes factors can also be used for hypothesis testing. To test the hypothesis of H_1 vs H_0, we calculate the Bayes factor

$$B_{10} = \frac{p(Y|H_1)}{p(Y|H_0)} \tag{3.7}$$

$$= \frac{p(H_1|Y)}{p(H_0|Y)} \Big/ \frac{p(H_1)}{p(H_0)} \tag{3.8}$$

and again use table 3.1 for interpretation. Wakefield (2009) explained: 'The Bayes factor is a summary measure that provides an alternative to the p-value for the ranking of associations, or the flagging of associations as "significant".' Thus, Bayesians use Bayes factors for hypothesis testing, as opposed to the p-values used in frequentist analysis.

During hypothesis testing — as opposed to model comparison — it is usually easier to calculate the quantities in (3.8), than (3.7). For instance, to test the hypothesis $H_1 : \theta > 0$ vs $H_0 : \theta < 0$, we see that $p(H_1) = p(\theta > 0)$ is simply the prior probability, while $p(H_1|Y) = p(\theta > 0|Y)$ is the posterior probability. Both of these should be available in some form after fitting a Bayesian model. Jackman (2009, c. 1.8) contrasted Bayesian and frequentist hypothesis testing by means of an example; the author analysed the same data using both frequentist and Bayesian methods, and

Bayes Factor	evidence for M_1
< 1	negative (supports M_0)
$1 < 3$	barely worth mentioning
$3 < 12$	positive
$12 < 150$	significant
> 150	highly significant

Table 3.1. Interpretation of Bayes factors

showed the subtle differences between the interpretations thereof.

Deviance Information Criterion

Spiegelhalter et al. (2002) introduced a complexity measure p_D for the effective number of parameters in a model. They combined this with poster mean deviance to form the Deviance Information Criterion (DIC). They define the deviance function as

$$D(\theta) = -2\log\{p(y|\theta)\} - 2\log\{g(y)\}$$

where $g(y)$ is a function of the data alone, such that $2\log\{f(y)\}$ becomes a constant that is the same for all models to be compared. If $\boldsymbol{\theta}$ and $\boldsymbol{y}$ are the entire collections of parameters and data, respectively, then p_D is defined as:

$$p_D = E_{\boldsymbol{\theta}|\boldsymbol{y}}\big[D(\boldsymbol{\theta})\big] - D\left(E_{\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{\theta}]\right) = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}). \tag{3.9}$$

Guo and Carlin (2004) used the 'Bayesian Central Limit Theorem' to explain the why (3.9) approximates the true number of parameters in the model.

The DIC is obtained by adding the posterior mean deviance to (3.9)

$$DIC = \overline{D(\boldsymbol{\theta})} + p_D$$

or equivalently

$$DIC = D(\bar{\boldsymbol{\theta}}) + 2p_D,$$

highlighting the connection to the Akaike Information Criterion (AIC). As with the AIC, models with a lower DIC should be favoured to models with a higher DIC. It is a comparative measure that does not make sense on its own. Both $\overline{D(\boldsymbol{\theta})}$ and $D(\bar{\boldsymbol{\theta}})$ can be computed with ease in models where posterior distributions of parameters are obtained using MCMC simulation — making the DIC a useful tool in such cases.

3.1.3 Prediction

Bayesians use the posterior predictive distribution for prediction purposes. It is defined as the distribution of a new data point, given the observed data:

$$\begin{aligned} p(y^{new}|\boldsymbol{y}) &= \int p(y^{new}, \boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\ &= \int p(y^{new}|\boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\ &= \int p(y^{new}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}. \end{aligned} \tag{3.10}$$

We obtain the last line since y^{new} is independent of $\boldsymbol{y}$ when conditioned on $\boldsymbol{\theta}$. The posterior likelihood of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{y})$, appears in this formula. We can sample from the posterior predictive distribution through composition: sample a value $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta}|\boldsymbol{y})$, then sample y^{new*} from $p(y^{new}|\boldsymbol{\theta}^*)$. This is easy in MCMC settings, since samples of $\boldsymbol{\theta}^*$ are already available at each iteration.

The posterior predictive distribution is not only used for prediction, but also for model checking — as we will see in section 6.1 — and for data augmentation in missing data problems, as discussed by Jackman (2009, c. 5.2). In chapter 4, we will propose a joint survival model, and use the techniques presented in this chapter to estimate its parameters.

3.2 Markov Chain Monte Carlo

MCMC gives us a practical approach to fit complicated theoretical models. Formerly, in order to make a complicated model useful, it either had to be mathematically tractable or a piece of software had to be specifically designed for a chosen problem. The arrival of MCMC made it possible to use a set of methods on a wide array of problems, thereby greatly extending the possible range of models that could realistically be applied. In the rest of this chapter, we will explore the characteristics that make

MCMC so versatile. Gilks et al. (1996b) stated that 'MCMC is essentially Monte Carlo integration using Markov chains', which is an appropriate starting point for this section, where we will show how problems with complex integrals can be investigated using MCMC methods.

Throughout this chapter, we use X as a random variable of interest, with a density of $\pi(x)$, which might have a complex form, or only be known in part, for example, up to a normalising constant. The density can also be multidimensional, such that $X \in \mathbb{R}^d$. The goal is to conduct inference using $\pi(x)$.

3.2.1 Calculating expectations with Monte Carlo

Bayesian inference uses the posterior distribution (3.2), and all of the quantities of interest during inference can be expressed in terms of posterior expectations of functions of θ, written as $g(\theta)$. The posterior expectation of $f(\theta)$, given a set of observed data Y, is

$$E[g(\theta)|Y] = \int g(\theta)p(\theta|Y)d\theta. \tag{3.11}$$

The evaluation of the integral in this equation is usually a difficult matter, especially in high dimensional problems. This problem is not unique to Bayesian analysis; frequentist inference often requires evaluation of complex integrals as well. Methods that focus on complicated integrals include numerical integration (although it can be costly in high dimensional problems), Laplace approximations and Monte Carlo integrals.

Monte Carlo integrals are the building blocks of Markov Chain Monte Carlo. They provide a method to evaluate an expectation $E[g(X)]$ by drawing samples X_i, for

$i = 1 \dots G$, from its distribution $\pi(\cdot)$, and then using the law of large numbers

$$E[g(X)] \approx \frac{1}{G} \sum_{i=1}^{G} g(X_i),$$

which approximates the integral

$$E[g(X)] = \int g(x)\pi(x)dx$$

by simulating from $\pi(x)$. Here, the modeller chooses G, it has no relation to the number of observations in the model, often denoted by N. Complex integrals can be numerically evaluated if we are able to simulate from the appropriate distribution, $\pi(x)$.

The simplest case occurs when we sample from a distribution function with an invertible cumulative distribution function $\mathcal{F}(x) = \int_0^x \pi(w)dw$. Simulating from $\pi(\cdot)$ can then be done by drawing $U \sim \mathcal{U}[0, 1]$ and taking $\mathcal{F}^{-1}(U)$. This is known as the inverse transform method. More specific methods for simulating from chosen standard distributions are well developed (Ripley, 2009), and they are useful in the cases where conjugacy can be achieved. However, a wide array of problems require us to simulate from non-standard, or unfamiliar, densities.

3.2.2 Practical sampling with Markov Chains

Developing algorithms to sample directly from non-standard distributions can be challenging, even in one-dimensional cases. Furthermore, creative researchers are providing statisticians with an ever growing and diverse set of models, more often than not extending the inference to multiple dimensions. Thus, even though we have innovative methods to approximate high dimensional expectations through sampling, it would be a Herculean task to come up with a direct sampling method for every

new model. Fortunately, the task is also unnecessary, given methods that rely on Markov Chains. It involves running a Markov Chain that has $\pi(\cdot)$ as a stationary distribution, as a way to obtain an approximate sample from $\pi(\cdot)$. These are the methods commonly known as MCMC.

A sequence of random variables $\{X^{(1)}, X^{(2)}, X^{(3)}, \ldots\}$ is said to possess the Markov property if, given the current value or present state, the next state does not depend on the rest of the history. The sequence is then known as a Markov Chain. Formally, for $t \geq 0$, $X^{(t)}$ is a Markov Chain if

$$\begin{aligned} P(X^{(t+1)} = x | X^{(1)} = x^{(1)}, X^{(2)} = x^{(2)}, X^{(3)} = x^{(3)}, \ldots, X^{(t)} = x^{(t)}) \\ = P(X^{(t+1)} = x | X^{(t)} = x^{(t)}). \end{aligned}$$

A thorough treatment of Markov Chains and important properties for MCMC applications such as irreducibility, recurrence, ergodicity, and convergence can be found in Robert and Casella (2004, c. 6).

Arguably, the most famous MCMC algorithm is the Metropolis-Hastings algorithm. Proposed by Metropolis et al. (1953) and later extended by Hastings (1970), it provides a straight-forward method to move from one value in a chain to the next. Sampling from a density $\pi(\cdot)$ is done by sampling from a proposal density, and then accepting or rejecting the proposal. It is similar to rejection sampling, but it provides an added benefit of setting up 'local' proposals, that is, proposals of each new value can depend on the last accepted one. In line with earlier notation, let X be the random value we want to sample, with a distribution $\pi(X)$. Let $q(x|x')$ be a proposal distribution on the same space as X, that is, for any $x' \in \mathbb{R}^d$ we have $q(x|x') > 0$ and $\int_{\mathbb{R}^d} q(x|x')dx = 1$. Then the Metropolis-Hastings algorithm involves choosing an initial starting point $X^{(0)}$ and iterating for $t = 1, 2, 3, \ldots$ as in the following steps.

1. Sample $X \sim q(X|X^{(t-1)})$.

2. Compute the acceptance ratio

$$\alpha\Big(X|X^{(t-1)}\Big) = \min\left(1, \frac{\pi(X)q(X^{(t-1)}|X)}{\pi(X^{(t-1)})q(X|X^{(t-1)})}\right).$$

3. Set $X^{(t)} = X$ with probability $\alpha\Big(X|X^{(t-1)}\Big)$, otherwise keep the previously sampled value, that is, set $X^{(t)} = X^{(t-1)}$.

Convergence properties of the Metropolis-Hastings algorithm have been extensively studied (Robert and Casella, 2004, c. 7, for instance), concluding that the resulting Markov Chain $\{X^{(t)}, t = 1, 2, 3, \ldots\}$ can be regarded as a sample from $\pi(\cdot)$ and used to study its properties. Although the algorithm allows for freedom in the choice of the conditional proposal $q(\cdot|\cdot)$, care should be taken to prevent the algorithm getting stuck at certain values of X, causing what is known as *slow mixing*. A notable advantage of the algorithm is the fact that we do not need to know the exact form of $\pi(x)$, we just need to know it up to a normalising constant $\tilde{\pi}(x) \propto \pi(x)$, since

$$\frac{\pi(x)q(x^{(t-1)}|x)}{\pi(x^{(t-1)})q(x|x^{(t-1)})} = \frac{\tilde{\pi}(x)q(x^{(t-1)}|x)}{\tilde{\pi}(x^{(t-1)})q(x|x^{(t-1)})}$$

which is all we need to generate new samples.[2]

The Metropolis-Hastings algorithm serves as a theoretical cornerstone of MCMC methods,but it has some practical limitations. For instance, even though it can handle high dimensional densities

$$\pi(\boldsymbol{x}) = \pi\left(\{x_1, x_2, x_3, \ldots, x_d\}\right),$$

it is usually slow to mix, since the proposed value $X \in \mathbb{R}^d$ will have a low acceptance probability if each of the individual dimensions do not update favourably. The

[2]This proportionality occurs regularly in practice, especially in Bayesian statistics where inference is carried out using the posterior distribution proportional to the prior multiplied by the likelihood, or $p(\theta|x) \propto f(x|\theta)p(\theta)$.

algorithm favours new values if the ratio $\pi(X)/q(X|X^{(t-1)})$ is increased. In scenarios with multidimensionality, it can take many proposals before this condition is met, and thus the chain mixes slowly. Furthermore, it might be hard to select an adequate proposal distribution that updates all the variables in a multidimensional model simultaneously.

Metropolis et al. (1953) originally proposed a framework that — rather than updating a multivariate X simultaneously at each step — divided X into simpler components $\{X_{.1}, X_{.2}, X_{.3}, \ldots, X_{.h}\}$, where the components had differing dimensions, and then updated them one by one. This venture is not only simpler to implement mathematically, but also more efficient computationally, and it is referred to as *Single-component Metropolis-Hastings* (Gilks et al., 1996b).

3.2.3 Gibbs sampling

A special case of the Single-component Metropolis-Hastings is the Gibbs sampler. A well known technique, it is sometimes referred to as the 'workhorse' of the MCMC world (Robert and Casella, 2004). The paper responsible for the name of Gibbs sampling was based on earlier techniques in statistical physics by Geman and Geman (1984). The article by Gelfand and Smith (1990) introduced Gibbs sampling into mainstream statistics, and it is usually accredited with responsibility for the rise in its implementation. Early papers such as Ripley (1979), however, proposed similar solutions that did not garner the same response from the statistical community, as stated by Robert and Casella (2004, c. 10), who give further examples of other papers that have gone largely unnoticed.

Applying Gibbs sampling to update $X \in \mathbb{R}^d$ involves updating one element at a time, while keeping the others fixed at their current values. We choose a starting position $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, \ldots, x_d^{(0)})$, then update for a given $\boldsymbol{x}^{(t)} =$

$(x_1^{(t)}, x_2^{(t)}, x_3^{(t)}, \ldots, x_d^{(t)})$ at each iteration

$$
\begin{aligned}
1) \quad & X_1^{(t+1)} \sim \pi_1(x_1^{(t)} | x_2^{(t)}, x_3^{(t)}, \ldots, x_d^{(t)}) \\
2) \quad & X_2^{(t+1)} \sim \pi_2(x_2^{(t)} | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_d^{(t)}) \\
3) \quad & X_3^{(t+1)} \sim \pi_3(x_3^{(t)} | x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_d^{(t)}) \\
& \vdots \\
d-1) \quad & X_{d-1}^{(t+1)} \sim \pi_{d-1}(x_{d-1}^{(t)} | x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)} \ldots, x_d^{(t)}) \\
d) \quad & X_d^{(t+1)} \sim \pi_d(x_{d-1}^{(t)} | x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)} \ldots, x_{d-1}^{(t+1)}).
\end{aligned}
$$

The densities $\pi_1, \pi_2, \pi_3, \ldots, \pi_d$ are known as *full conditional densities*. These are the only densities we need to simulate from the joint density $\pi(\boldsymbol{x})$, when using the Gibbs sampler. Some or all of these densities may be univariate, providing an effective way to greatly simplify the problem of simulating from a high dimensional density. The Gibbs sampler is in fact a special case of the Metropolis-Hastings algorithm where the proposal density is defined as

$$q(x_j | \boldsymbol{x}') = \pi_{X_j | X_{-j}}(x_j | \boldsymbol{x}'_{-j}),$$

and the acceptance ratio is always one.

In the Gibbs sampler, the conditional distributions contain enough information to successfully sample from the full joint posterior distribution. This allows inference on complicated likelihoods. In frequentist statistics, attempting to maximise a likelihood over many dimensions using optimization algorithms causes problems. There is no guarantee that the algorithm will converge to the global maximum, and may get stuck in a saddlepoint.[3]

[3]The Bayesian framework aided by MCMC methods does not completely remove this problem, since there is also uncertainty about convergence. This is known as pseudo-convergence (Geyer, 2011).

The extraordinary property of being able to perfectly recover a joint distribution from its conditionals is known as the Hammersley-Clifford Theorem.

Theorem 3.1 (Hammersley-Clifford)**.** *Under the positivity condition, the joint distribution π satisfies*

$$\pi(x_1, x_2, \ldots, x_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X_{-j}}(x_j|x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_d)}{\pi_{X_j|X_{-j}}(z_j|x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_d)}$$

for any $\boldsymbol{z} = \{z_1, \ldots, z_d\}$ that satisfies $\pi(\boldsymbol{z}) > 0$.

The proof is given in Robert and Casella (2004, p. 377). As we will see later, this property joins conveniently with Bayesian hierarchical modelling to allow a complete methodology from model design, model fitting, model diagnostics, prediction, and hypothesis testing for a variety of models.

The Gibbs sampler reveals its strength when it is easy to manipulate the conditional distributions relevant to the model, as is the case with hierarchical models. Despite this, we should proceed with caution since the Gibbs sampler is not without drawbacks. Geyer (2011) mentions that the sampler is sometimes used because it requires no further choices to be made and that it is 'automatic' in this sense. As a counterpoint, in the case of Metropolis-Hastings, for example, the modeller would have to carefully consider which proposal distribution to use. As a closing remark Geyer states that Gibbs updates should only be used if the resulting samplers work well, otherwise it will be better to use a different technique.[4]

[4] A good treatment of Gibbs sampling in practice can be found in Gilks et al. (1996a), while the theoretical implications of Gibbs sampling and MCMC methods in general appear in Robert and Casella (2004) and Brooks et al. (2011).

3.3 Bayes and MCMC

Bayesian statistics and the application of Bayesian ideas have seen tremendous growth over the past twenty years. This can be attributed to the widespread adaptation of simulation-based inference. Prior to that, a statistical problem could rarely be investigated without finding a solution in closed form. Solutions to more complicated statistical problems remained possible only through ingenious numerical integration methods. Likewise, in the Bayesian universe, a model rarely had much practical value without the use of conjugate priors. This again limited us to cases that had a mathematically tractable solution.

The arrival of simulation-based inference brought with it the concept that we can learn about the properties of a posterior distribution by simulating from it, even though we might not be able to derive it analytically. We might also be unable to maximise a likelihood to estimate a parameter of interest, but we can assume a prior on it and simulate from the resulting posterior. Research in MCMC methods provides tools to simulate from very complicated posterior distributions, opening up a broad new spectrum of possibilities for statistical models.

The practical benefit of using MCMC in a Bayesian framework is presented well by Spiegelhalter (1998), highlighting how graphical models and MCMC provide a direct path between model specification and estimation. The framework furthermore includes ways to handle missing data, prediction, and model diagnostics. 'Graphical models' describes the visual representation of the conditional dependence in a statistical model. This is usually done using a Directed Acyclic Graph (DAG) that describes the structural dependence through a set of connected nodes and arrows, such as figure 3.1 reproduced from Spiegelhalter (1998).

A graphical model is no more than a set of structural assumptions presented using a DAG. Furthermore, a graphical model makes assertions on dependence structure, it

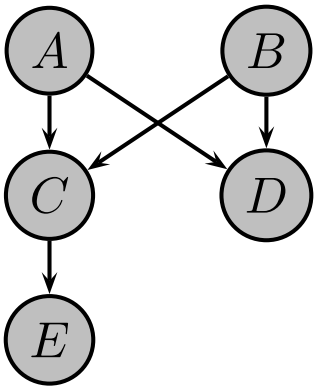


Figure 3.1. A DAG representing conditional independence of five random quantities.

does not necessarily have a probabilistic interpretation. We can go one step further in pursuit of a full probability model, by assuming the joint distribution of all the random variables is fully specified by the conditional distributions of all individual variables, given their parents

$$P(\boldsymbol{x}) = \prod_{x_i \in \boldsymbol{x}} P(x_i | parents[x_i]), \tag{3.12}$$

where $P(\cdot)$ is a probability distribution. Thus, the full probability model is defined through the conditional distributions. Continuing with the example from Spiegelhalter (1998), the full probability model resulting from the dependencies shown in figure 3.1 would be

$$P(A, B, C, D, E) = P(C|A, B)P(D|A, B)P(E|C)P(A)P(B),$$

highlighting that the joint distribution factorises conveniently into terms representing the 'local' dependence structure. This works naturally with Gibbs sampling, since we would only need to sample from all the full conditional distributions, and they are

readily available for each node

$$\begin{aligned} P(x_i|\boldsymbol{x}_{-i}) &= \frac{P(x_i, \boldsymbol{x}_{-i})}{P(\boldsymbol{x}_{-i})} \\ &\propto P(x_i, \boldsymbol{x}_{-i}) \\ &\propto \text{terms in } P(\boldsymbol{x}) \text{ containing } x_i \\ &= P(x_i|parents[x_i]) \prod_{x_j \in child[x_i]} P(x_j|parents[x_j]). \end{aligned} \tag{3.13}$$

The proportionality is a strong feature, since it will allow us to move through the steps of the Gibbs sampler while doing only 'local' computations around the current nodes. The simplification that thus arises from the proportionality considerably reduces the computational burden of sampling from the full conditional distributions. In (3.13), it is usual for $P(x_i|parents[x_i])$ to be the prior distribution of the component x_i, while $\prod P(x_j|parents[x_j])$ will arise as the contribution to the likelihood of each child of x_i. It is important to note that we do not sample from the distributions assumed on the nodes in the model. Samples are drawn from the full conditionals, which result from the full posterior distribution.

The techniques discussed in this chapter have brought us to the point where we can outline a general methodology for modelling, that ranges from defining a model to setting up an MCMC sampler, and thus estimating the parameters of interest. It consists of the following steps.

Methodology 3.1 (Developing an MCMC sampler)**.**

1. *Define the model structure and assumptions. This includes defining the conditional dependence, and calculating the likelihood contribution of observations.*

2. *Select appropriate prior distributions.*

3. *Determine the full probability model, as well as the full conditional distributions.*

4. *Devise methods to sample from the full conditional distributions, and write the code for the algorithm.*

5. *Choose adequate starting positions for the chains.*

6. *Run the sampler for a set amount of iterations, do convergence diagnostics, and rerun as required.*

7. *Use the resulting samples for inference — they can be regarded as samples from the full posterior distribution.*

Step 1 does not require us to be Bayesian, since a particular model structure and the underlying probabilistic assumptions of observed quantities are the same whether we are using a Bayesian or frequentist approach. A frequentist model can thus be altered to be fitted in the MCMC framework by regarding the parameters as random quantities, and assuming priors on them in step 2. Step 3 entails writing the model out analytically, to determine the factorisations, and hence arrive at the full conditional distributions. Step 4 requires the most consideration. This is usually where a modeller aims to attain conjugacy. If we select appropriate conjugate prior distributions in each of the arising equations of the form (3.13), we will end up with full conditional distributions of a familiar form, allowing us to draw samples using well developed methods, as discussed in Ripley (2009).

Step 5 need not be arduous, since it is possible to choose a random starting position and throw away the first part of the MCMC chain — known as burn-in. As Geyer (2011) stated, however, 'Burn-in is only one method, and not a particularly good method, of finding a good starting point.' A chain started near the centre of its equilibrium distribution will not require any burn-in. This is not always easy, since the equilibrium distribution is the unknown feature that we would like to study, and it can grow complicated as the number of dimensions increases. When the sampler runs quickly, burn-in will not be costly, but a slow sampler with highly correlated

values might require a long burn-in period, making it a costly endeavour. We can choose any value that we might expect to see as a sample, or at least choose points that exist in the distribution being studied.[5]

Steps 6 and 7 are the final part of the process, where the resulting samples can be used to determine whether convergence can be assumed, and thus be useful to perform inference. Once we have completed the process for a particular model and have determined that the code works, we can adapt the code for more complicated models. This methodology has proven useful since, as shown by Spiegelhalter (1998), a real example that is only moderately challenging can result in full probability models that are extremely difficult to treat analytically. This results in the only feasible way to carry out complex integration being through simulation-based methods.

3.4 Sampling from unfamiliar densities

MCMC estimation often involves sampling from unfamiliar densities.[6] Through the structure of the Gibbs sampler, a multidimensional posterior distribution can be broken into simpler parts, allowing us to sample from multiple densities of a lesser dimension.

This does not completely solve the initial problem of sampling from a complex full posterior density, since many of the arising full conditional distributions might also be unfamiliar, and difficult to sample from. The result is having multiple low-dimensional — albeit unfamiliar — distributions, as opposed to one unfamiliar but tediously high-dimensional distribution. Where the latter may appear near impossible to sample from, the former is usually much easier to deal with computationally.

In many cases it will be possible to choose conjugate prior distributions that result in the same familiar posterior distribution thereby making sampling easier. However,

[5]Poorly chosen starting positions lead to problems that resemble coding errors.

[6]We use the term 'unfamiliar' to refer to any density that does not have a well known sampling procedure. These procedures are discussed in Ripley (2009).

an adequate conjugate prior distribution might not exist. If it exists, it might not be a realistic assumption — the only justification for adopting it might be that it provides a convenient full condition distribution. In both cases, we stand to benefit from a flexible technique or method to sample from unfamiliar densities. Since unfamiliar densities are often encountered in MCMC applications (Gilks, 1996), techniques to sample from them have been developed and studied, and we present some of the popular ones in the next sections.

3.4.1 Adaptive rejection sampling

The Adaptive Rejection Sampler (ARS) as originally described by Gilks and Wild (1992) and later updated in Gilks (1992), is the first method developed to sample from unfamiliar densities in a Gibbs sampler. The former paper featured the introduction of the ARS method, while the latter extended the technique to 'derivative-free' ARS. Both of these papers describe how to use the ARS method to sample from unknown distributions with log-concave densities. It enables us to sample from an unfamiliar univariate density $\pi(x_i|\boldsymbol{x}_{-i})$ if we can evaluate some function $f_i(x_i)$ that is proportional to it. It starts by finding points to the left and right of the mode of the conditional distribution, so it might involve some initial search — controlled by setting a scale parameter. Consequently, the technique requires some calibration in order to avoid large computation times, but fortunately the scale parameter can be 'retrospectively' adjusted without influencing the equilibrium distribution.

ARS also makes use of an 'accept-reject' framework and each reject allows the model to update itself and thus draw future samples more quickly. Unfortunately this feature is not very useful in the Gibbs sampling framework. We only need one sample from the conditional distribution sample at each Gibbs iteration. The next time we need to sample from this distribution, it will have changed because the other parameters in the model were updated. Gilks et al. (1995) further extended

the method to Adaptive Rejection Metropolis Sampling (ARMS), which can manage densities without log-concavity.

3.4.2 Slice sampling

The Slice sampler (Neal, 2003) is another widely used technique which places very few requirements on the densities to be sampled. As with the ARMS sampler, to sample from $\pi(x_i|\boldsymbol{x}_{-i})$ we only need a function $f_i(x_i)$, where $f_i(x_i) \propto \pi(x_i|\boldsymbol{x}_{-i})$. To draw a sample from $f_i(x_i)$, we introduce an auxiliary variable y, and then define a joint density that is uniform over the region $U = \{(x, y) : 0 < y < f_i(x)\}$. Formally, the joint density for (x, y) is

$$p(x, y) = \begin{cases} 1/Z & \text{if } 0 < y < f_i(x) \\ 0 & \text{otherwise} \end{cases}$$

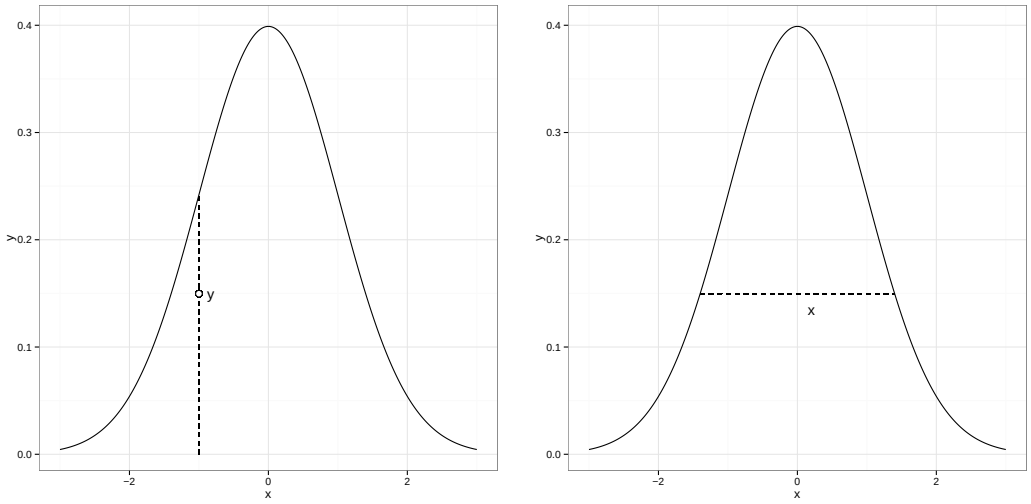
where $Z = \int f_i(x)dx$. This way, we achieve the desired result for the marginal density of x

$$p(x) = \int_0^{f_i(x)} (1/Z)dy = f_i(x)/Z.$$

To sample from x, we sample from the joint density and ignore y. The joint density can be sampled using a Gibbs sampler; the distribution of y given the current x is uniform over $(0, f_i(x))$, while the distribution of x given the current y is uniform over the region $S = \{x : y < f(x)\}$, called the 'slice'. See figure 3.2 for an illustration.

By introducing the auxiliary variable, we have simplified the problem of sampling from an unfamiliar distribution $\pi(x_i|\boldsymbol{x}_{-i})$ to sampling from two uniform distributions. The most difficult part of the Slice sampling scheme is to find the region $S = \{x : y < f(x)\}$, since it is not well defined. An implemented Slice sampler usually involves estimating this region through the following steps:

1. Draw a value y uniformly from $(0, f_i(x_0))$. Define the horizontal 'slice' as



(a) Drawing y on the uniform interval $[0, f_i(x_0)]$.

(b) Drawing x_1 uniform on the region $S = \{x : y < f(x)\}$.

Figure 3.2. The two main steps of Slice sampling.

$S = \{x : y < f(x)\}$.

2. Find an interval $I = (L, R)$ that contains all, or much, of the slice. Usually the only feasible way to achieve this is through exploration of the area around x_0.

3. Sample x_1, the new point, uniform on the interval $I = (L, R)$.

Neal (2003) further explained that it is usually safer to work with $g_i(x) = \log(f_i(x))$ than $f_i(x)$ to avoid the possibility of numerical underflow. Furthermore, the author provided alternative suggestions to perform the exploration to find the interval $I = (L, R)$, together with illustrations. The paper also showed how Slice sampling can be extended to the multivariate case.

3.4.3 Metropolis-Hastings

Another course of action is to make use of the Metropolis-Hastings algorithm to sample from an unfamiliar full conditional distribution (Brooks, 1998). This is often referred to as *Metropolis-within-Gibbs*, but the term causes confusion since the Gibbs sampler is already a special case of the Metropolis-Hastings algorithm (Geyer, 2011).

Using a Metropolis-Hastings update requires choosing an adequate proposal distribution. This method is therefore not ideal for use in standard Gibbs sampling packages such as JAGS or BUGS, since it requires a fair amount of thought and input from the user. The first two methods introduced can be regarded as 'black box' sampling methods. They only require evaluation of a function that is proportional to the required distribution, rather than knowledge about the shape or characteristics of the distribution — which are some of the factors that go into constructing a successful Metropolis-Hastings algorithm. There exists a trade-off — the 'black box' samplers are easier to implement, but their sampling efficiency might not compare to a well designed Metropolis-Hastings algorithm. Conversely, Metropolis-Hastings is more difficult to implement and thus not the preferred method if implementation speed is important.

3.5 BUGS and JAGS

Bayesian inference Using Gibbs Samling (BUGS), by Lunn et al. (2000), is a software package that implement Gibbs sampling. Since the project's commencement in 1989 at the Medical Research Council Biostatistics Unit in Cambridge, UK, it has been tremendously successful and has been linked to the growth of the application of Bayesian techniques (Lunn et al., 2009). The software has its own language, designed to mimic the statistical specification of the dependencies in the model. The BUGS language is a declarative language. The order of the statements does not matter, as the system parses all the code before the model is constructed internally. This allows the syntax to be intuitive, but it has certain drawbacks as well, for instance, `if-then-else` statements are not possible.

BUGS takes care of the bulk of the work described in methodology 3.1. Through the language, it provides an easy way to describe the model structure and prior

distributions, simplifying steps 1 and 2. Next, it takes care of steps 3 and 4 by identifying the full conditional distributions and making use of pre-programmed methods to sample from them. The modeller does not need to be concerned with the choice of sampling algorithms, which is often the most difficult part of constructing an MCMC sampler. BUGS has a well developed arsenal of algorithms to tackle sampling from full conditional distributions. This is the software's chief strength, since steps 3 and 4 are unlikely to be directly related to the problems a modeller is trying to solve. Scollnik (2000) gave a comprehensive introduction to BUGS through an example. The author showed how to construct an MCMC algorithm by writing out the full probability model and deriving the full conditional distributions, and covering in detail how the software simplifies the venture by performing these tasks automatically.

By reducing the time and expertise required to develop a working Gibbs sampler, BUGS has made the technique available to a large audience. There is a caveat, however: it is easier for inexperienced users to make mistakes that go unnoticed by the software, to the extent that it still produces seemingly sound results.[7] Some of these mistakes may result when re-using the code of experienced users without understanding the ramifications, and Lunn et al. (2009) highlighted some well-known pitfalls.

Just Another Gibbs Sampler (JAGS), by Plummer (2003), is a software package featuring the same model description language as BUGS, but written in a different programming language. Its chief aim is to be a cross-platform engine for the BUGS language.[8] It was our preferred package since it is easy to implement across GNU/Linux, Windows, and Mac. BUGS and JAGS have interfaces with R, through the packages `BRugs` (Thomas et al., 2006) and `rjags` (Plummer, 2012), respectively. These allow for easy integration with R.

Both the ARS and Slice sampler are used in the standard Gibbs sampling packages.

[7]This is the main reason for the BUGS manual (Spiegelhalter et al., 2004) starting out with the warning 'Beware: MCMC sampling can be dangerous!'

[8]See the website for a short description of JAGS, along with the other two aims of the software `http://mcmc-jags.sourceforge.net/`

BUGS checks for log-concavity and applies ARS if it is present, or otherwise defaults to Slice sampling (Congdon, 2003, c. 2). JAGS applies the Slice sampler in all situations involving unfamiliar densities. For a detailed comparison of the different sampling algorithms and their characteristics, see Neal (2003).

3.6 Stan

Like JAGS and BUGS, Stan (Stan Development Team, 2014b) is a programming language implementing full Bayesian statistical inference.[9] Development started in 2010 with the aim of producing an MCMC sampling package that would address the known issues encountered when using Gibbs sampling in settings with highly correlated posteriors (Stan Development Team, 2014c). This required a more efficient sampler, rather than more efficient implementation and led to adopting *Hamiltonian Monte Carlo* and developing the No U-Turn Sampler (Hoffman and Gelman, 2014).

Stan also features its own modelling language, with some similarities to those of JAGS and BUGS — it is also declarative. Where these languages are interpreted, however, Stan is compiled and then run. Despite the more efficient samplers, the Stan Development Team mentions that the very first model ever run performed several orders of magnitude slower than BUGS. Stan has since evolved considerably, but it is still a work in progress. The Stan website lists a wide array of models and examples that have been successfully implemented.[10] Like JAGS and BUGS, Stan also has an R interface, `rstan` (Stan Development Team, 2014a), allowing it to be invoked directly from R.

In this thesis we make use of both JAGS and Stan. Due to the non-standard nature of our models, however, we were unsure about the ability of the standard software packages to successfully implement them. This, together with the fact that

[9] Stan takes it name from Stanislaw Ulam, co-inventor of Monte Carlo Methods (Metropolis and Ulam, 1949).

[10] `http://mc-stan.org`

the standard packages were slow to converge, motivated us to write our own sampler in C++ using methodology 3.1. Furthermore, the standard packages were unable to deal with time-dependent hazard rates and JAGS could not handle multivariate longitudinal processes. Writing our own sampler gave us more room for modifying our models in which we could fully control the sampling algorithms, and the reduced runtime meant quicker implementation and testing. Finally, having the same output from independent MCMC algorithms validated our results, as well as the ability of the standard sofware-packages to handle non-standard models.

Chapter 4

A joint model for longitudinal volatility

This chapter defines a model to study the effect of longitudinal volatility on event risk. This chapter gives the model assumptions, and uses them to construct the likelihood, then explains how the model can be treated in both frequentist and Bayesian frameworks. This chapter also shows how the same model can be presented as a Bayesian hierarchical model, which makes parameter estimation possible using the techniques from the previous chapter. We then fit the model and present the results.

4.1 Model assumptions and structure

We define our model by demonstrating its longitudinal and survival parts. We then assume priors to allow a Bayesian framework. The model does not have to be applied in a Bayesian setting, since the assumptions regarding conditional dependence can be applied by Bayesians and frequentists alike.

4.1.1 Longitudinal

Assume longitudinal process is independent of the survival outcome, given the parameters. The process that generates the longitudinal measurements can be defined as follows: let $Y_i(t_j)$ be the observed blood pressure for patient i at time t_j, where $i = 1 \dots N$ and $j = 1 \dots n_i$, that is, N patients each with their own number of measurements, n_i. We assumed a Gaussian model for the longitudinal response Y_i at time t:

$$\begin{aligned} Y_i(t) &= \mu_i + \sigma_i \varepsilon_t \\ &= \mu_i + \tau_i^{-\frac{1}{2}} \varepsilon_t, \end{aligned} \tag{4.1}$$

where μ_i and τ_i represent the longitudinal mean and precision respectively, for the ith individual and $\varepsilon_t \sim N(0, 1)$. We use the precision $\tau_i = {}^1/_{\sigma_i^2}$ to simplify the notation and programming. The precision — as opposed to the variance — is also used by the Gibbs sampling software packages, JAGS and BUGS, when specifying distributions.

The volatility parameter τ_i includes subject volatility as well as measurement error, and it will be different for each subject. Lange et al. (1992) used a similar specification for their longitudinal process, allowing for heterogeneity with respect to subject variance. Theirs is the only example in the literature that, similar to our model, assigns each subject an individual variance parameter. We added another level of hierarchy to our model, however, and also allowed subject variance to influence survival.

Our assumption in (4.1) is equivalent to saying $Y_i(t) \sim N(\mu_i, \tau_i^{-1})$. We furthermore assume distributions on the mean and precision of individual blood pressures:

$$\begin{aligned} \mu_i &\sim N(m, \tau^{-1}) \\ \tau_i &\sim \Gamma(r, \lambda), \end{aligned} \tag{4.2}$$

where r and λ are the shape and rate of the Gamma distribution, respectively. This treats the individual mean and precision as random effects. Using the Gamma distribution departs from usual random effects models; standard joint survival models usually assume that random effects follow a Normal distribution. During data collection, observations were rounded to the nearest multiple of five. We accounted for the rounding simply by adding random 'jitter' to the data, but concluded it did not affect the results.

4.1.2 Survival

We specified a hazard rate, in similar fashion to the Cox proportional hazards model (Cox and Oakes, 1984):

$$h(t|\mu_i, \tau_i) = h(t|\boldsymbol{\theta}_i) = h_0(t) \exp\{\beta_0 + \beta_1 \mu_i + \beta_2 \tau_i\}, \tag{4.3}$$

where $\boldsymbol{\theta}_i$ is the set of parameters for the ith individual. The hazard rate can never contain random quantities — it will always be calculated assuming μ_i and τ_i are given. This specification allowed us to measure the significance of the mean and precision (and therefore the variance) on the risk of having a stroke.

Initially, we assumed a constant hazard rate and let $h_0(t) = 1$, but this can be extended to allow for more complex models. If we let $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2\}$, we can write each patient's survival contribution to the likelihood as

$$f(v_i, \delta_i|\mu_i, \tau_i, \boldsymbol{\beta}) = [h(v_i|\boldsymbol{\theta}_i)]^{\delta_i} \exp\left(-\int_0^{v_i} h(t|\boldsymbol{\theta}_i)dt\right), \tag{4.4}$$

where δ_i is the event indicator and v_i is the observed event time for the ith patient. Note that (4.4) assumes random, uninformative censoring. If censoring during the study was related to any of the covariates in the model, this assumption would not hold

and we would have to extend the model to accommodate competing risks.[1] Finally, taking $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\tau}_i, m, \tau, r, \lambda\}$ to be the set of all the parameters, gives the following joint likelihood:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{\delta}) = \prod_{i=1}^{N} \left(\prod_{j=1}^{n_i} f(y_{ij}|\mu_i, \tau_i) \right) f(v_i, \delta_i|\mu_i, \tau_i, \boldsymbol{\beta}) f(\mu_i|m, \tau) f(\tau_i|r, \lambda), \qquad (4.5)$$

from which we can draw inference. Particular interest lies with the significance of β_1 and β_2, which measure the effect of the mean and precision on the survival outcome. We can illustrate the parameter dependencies using a Directed Acyclic Graph (DAG), as shown in figure 4.1.

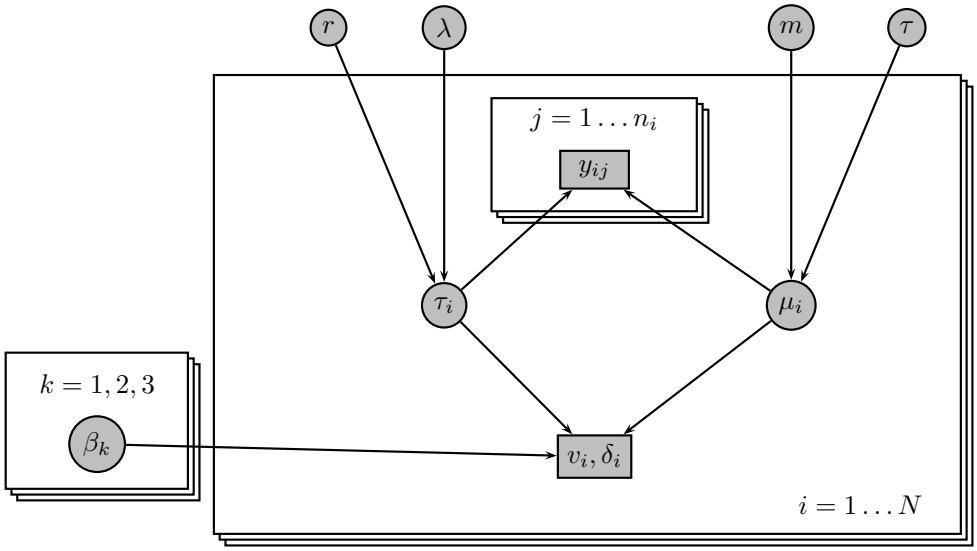


Figure 4.1. Directed acyclic graph

[1]For an in-depth discussion of this assumption and its properties, refer to Tsiatis and Davidian (2004).

4.1.3 Bayesian interpretation

The parameters we need to estimate are r, λ, τ, m, and $\boldsymbol{\beta}$. In the Bayesian setting, we assume priors on them that give the posterior distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{v},\boldsymbol{\delta}) \propto \prod_{i=1}^{N}\left(\prod_{j=1}^{n_i} f(y_{ij}|\mu_i,\tau_i)\right) f(v_i,\delta_i|\mu_i,\tau_i,\boldsymbol{\beta})\times \\ f(\mu_i|m,\tau)f(\tau_i|r,\lambda)f(\boldsymbol{\beta})f(m)f(\tau)f(r)f(\lambda)\end{aligned} \tag{4.6}$$

from which we can generate samples of $\boldsymbol{\theta}$ using MCMC. Initially, we assume vague, but proper, priors on all of the parameters:

$$\begin{aligned}\beta_i &\sim N(\mu_\beta, \tau_\beta^{-1}) \\ m &\sim N(\mu_m, \tau_m^{-1}) \\ \tau &\sim \Gamma(r_\tau, \lambda_\tau) \\ r &\sim \Gamma(\alpha_r, \beta_r) \\ \lambda &\sim \Gamma(\alpha_\lambda, \beta_\lambda).\end{aligned} \tag{4.7}$$

We can also use prior knowledge from other studies to borrow information and strengthen our inference. The choice of these priors results in the full conditionals of the MCMC sampler being conjugate, and this was the main reason for adopting them.

We regard the two sets of parameters $\{\mu_1 \dots \mu_N\}$ and $\{\tau_1 \dots \tau_N\}$ — the mean and precision parameters of each individual's blood pressure, respectively — as nuisance parameters, which we will also need to estimate even though we are not directly interested in them. Thus, the full set of parameters to be sampled in the MCMC algorithm is $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \mu_1 \dots \mu_N\ , \tau_1 \dots \tau_N\ , m, \tau, r, \lambda\}$ and we are mainly interested in $\boldsymbol{\beta}$.

Another way to look at the model is to consider it as being Bayesian hierarchical.

To do this, we express the assumptions as a hierarchy of prior distributions. Let

$$Y_i(t) \sim N(\mu_i, \tau_i^{-1}),$$

with prior distributions

$$\mu_i \sim N(m, \tau^{-1})$$

$$\tau_i \sim \Gamma(r, \lambda).$$

The prior parameters m, τ, r, and λ are known as *hyperparameters*. To complete our hierarchical model, we assume a final layer of priors, given in (4.7), on these hyperparameters as well. When we assume a distribution on a hyperparameter, rather than specifying it directly, it is known as a *hyperprior*. The resulting model is exactly the same as that defined by the structural dependencies in figure 4.1, and defining the model in terms of the hierarchical structure provides us with exactly the declarative statements needed to program the model in JAGS or BUGS.

We used a Bayesian framework in this setting for three reasons. First, it uses the same assumptions and leads to the same full posterior distribution, but the interpretation is more intuitive. Second, the Bayesian model naturally leads to a course for parameter estimation through MCMC Gibbs sampling. We can easily extend the framework to hierarchical Bayes — allowing us to relax the prior assumptions made on μ_i and τ_i by estimating the prior parameters from the data using non-informative hyperpriors.

Third, longitudinal studies, such as the blood pressure study, often suffer from data scarcity. There are few observations available per individual to use for estimating the parameters of the longitudinal distribution. In well studied phenomena like blood pressure, however, we usually have information about the distribution from other sources. Adopting a Bayesian approach allows us, in theory, to use this information by including it using a prior distribution — in our case through the hyperpriors. If we

had knowledge about the specific distribution of individual mean and variance blood pressure, we would abandon our vague hyperpriors for more informative hyperpriors.

4.1.4 Frequentist interpretation

For the sake of completeness we also explain this model from a frequentist point of view. To do this, we need to express the longitudinal model as

$$\begin{aligned} y_{ij} &= \mu_i + \sigma_i \varepsilon_{ij} \\ &= \mu_i + \tau_i^{-\frac{1}{2}} \varepsilon_{ij} \end{aligned} \tag{4.8}$$

where $\varepsilon_{ij} \sim N(0,1)$. Then we regard μ_i and τ_i as random effects, having distributions as given in (4.2). The survival model has the same format as (4.3). To construct the likelihood, we use the shared parameter approach and assume conditional independence of the longitudinal and survival process:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{\delta}) &= \prod_{i=1}^{N} f(\boldsymbol{y}_i, v_i, \delta_i) \\ &= \prod_{i=1}^{N} \iint f(\boldsymbol{y}_i, v_i, \delta_i | \mu_i, \tau_i) p(\mu_i) p(\tau_i) d\mu_i d\tau_i \\ &= \prod_{i=1}^{N} \iint f(\boldsymbol{y}_i | \mu_i, \tau_i) f(v_i, \delta_i | \mu_i, \tau_i) p(\mu_i) p(\tau_i) d\mu_i d\tau_i \\ &= \prod_{i=1}^{N} \iint \left(\prod_{j=1}^{n_i} f(y_{ij} | \mu_i, \tau_i) \right) f(v_i, \delta_i | \mu_i, \tau_i) f(\mu_i) f(\tau_i) d\mu_i d\tau_i. \end{aligned} \tag{4.9}$$

A frequentist needs to maximise (4.9) to estimate parameters. We will not consider this setting further, but point out that the structural dependence shown in figure 4.1 is not unique to the Bayesian model. There is no example in the literature of a model with the aim of measuring the effect of longitudinal variance on the survival, such as the one presented in this thesis.

4.2 Parameter Estimation

In this section we will use our Bayesian model, along with the techniques given in the previous chapter, to set up a parameter estimation scheme. We fitted the model using a Gibbs sampler, implementing it in two ways, first using JAGS and Stan, and then through constructing a custom Gibbs sampler, written in C++ (Stroustrup, 2000).

4.2.1 JAGS

Guo and Carlin (2004) fitted and compared separate and joint models using the BUGS software package. The code used to fit the models in their paper can be found on Brad Carlin's software page.[2] The page also contains alternative model code, which was submitted after the publication of their paper.

We will be using a modification of code suggested by Michael Sweeting, featuring an application of the WinBUGS 'ones trick' (Spiegelhalter et al., 2004). Our preferred package was JAGS and the code used to fit our model is given in appendix B.1. The goal of these tricks is to increment the log likelihood within the sampler each iteration.

Although these tricks allowed us to fit the model, it is a workaround to implement a model in JAGS for which it was not originally designed. The nature of the tricks might also cause concerns about numerical accuracy, since they involve using logarithms of declared functions, rather than declaring the logarithms of the functions directly in the code. Another property of the Gibbs sampler that might be affected by the workaround is the speed of the algorithm — we are not declaring the new distributions and samplers directly, so we cannot ensure that the most effective samplers are being used. Fitting this model in JAGS was possible due to the simple form of the hazard rate (4.3), and more complex hazard rates would require numerical integration that cannot be managed by JAGS.

[2] `http://www.biostat.umn.edu/~brad/software.html`

While there are obvious benefits that accompany the use of existing software which has been tested, reviewed and modified by the wide scientific community, we believed that it was worthwhile to construct a purpose-built Gibbs sampler. It allowed for more direct control over the implementation of the model, along with exact knowledge of how estimation was performed. The simple model fitted in JAGS can furthermore be used as a stepping stone, corroborating the output through comparison with JAGS before aspiring to fit more complex survival models with the custom sampler. This way, we can build upon the simple model's structure with the confidence that the estimation techniques are correct. The simple model also serves the purpose of assessing the ability of JAGS or BUGS to handle joint survival models.

4.2.2 Stan

The Stan language is similar to that of BUGS and JAGS, but it requires more explicit declarations of the parameters and data in the model. We modified the JAGS code to fit the model in Stan, and we did not have to use any tricks. Stan has an `increment_log_likelihood()` method which has the same internal result as the tricks, but with simpler implementation on the user's part. Our Stan code is given in appendix B.2, presenting Stan's ability to fit joint survival models.

4.2.3 Custom sampler

Our full probability model has the posterior distribution given in (4.6), and we used this to set up full conditional distributions to sample from, as explained in 3.13. Writing $\boldsymbol{\theta}_{/\mu_i}$ as all the elements in $\boldsymbol{\theta}$ excluding μ_i and using (3.13) to get the full

conditional distribution of the parameter μ_i, we have

$$\begin{aligned} f(\mu_i|\boldsymbol{\theta}_{/\mu_i}, \boldsymbol{y}_i, v_i, \delta_i) &= \frac{f(\mu_i, \boldsymbol{\theta}_{/\mu_i}|\boldsymbol{y}_i, v_i, \delta_i)}{f(\boldsymbol{\theta}_{/\mu_i}|\boldsymbol{y}_i, v_i, \delta_i)} \\ &\propto f(\mu_i, \boldsymbol{\theta}_{/\mu_i}|\boldsymbol{y}_i, v_i, \delta_i) \\ &= f(\boldsymbol{\theta}|\boldsymbol{y}_i, v_i, \delta_i) \\ &= \text{terms in } f(\boldsymbol{\theta}|\boldsymbol{y}_i, v_i, \delta_i) \text{ containing } \mu_i \\ &= \prod_{j=1}^{n_i} \Big(f(y_{ij}|\mu_i, \tau_i) \Big) f(\mu_i|\tau_i, m, \tau) f(v_i, \delta_i|\mu_i, \tau_i, \boldsymbol{\beta}) \end{aligned}$$

$$= \exp\left[-\frac{\tau_i n_i + \tau}{2} \left(\mu_i - \frac{\tau_i \sum_{j=i}^{n_i} y_{ij} + \tau m}{\tau_i n_i + \tau} \right)^2 \right] [h(v_i)]^{\delta_i} \exp\left[-H(v_i)\right].$$

Following the same steps, we can derive all of the conditional distributions.

4.2.4 Full conditional distributions

μ_i:

$$\begin{aligned} f(\mu_i|\text{All others}) &\propto \underbrace{\prod_{j=i}^{n_i} \Big(f(y_{ij}|\mu_i, \tau_i) \Big) f(\mu_i|\tau_i, m, \tau)}_{\text{similar to conjugate normal posterior}} \underbrace{f(v_i, \delta_i|\mu_i, \tau_i, \boldsymbol{\beta})}_{\text{survival likelihood}} \\ &= \exp\left[-\frac{\tau_i n_i + \tau}{2} \left(\mu_i - \frac{\tau_i \sum_{j=i}^{n_i} y_{ij} + \tau m}{\tau_i n_i + \tau} \right)^2 \right] [h(v_i)]^{\delta_i} \exp\left[-H(v_i)\right] \end{aligned} \tag{4.10}$$

τ_i:

$$\begin{aligned} f(\tau_i|\text{All others}) &\propto \underbrace{\prod_{j=1}^{n_i} \Big(f(y_{ij}|\mu_i, \tau_i) \Big) f(\tau_i|\mu_i, r, \lambda)}_{\text{similar to conjugate gamma posterior}} \underbrace{f(v_i, \delta_i|\mu_i, \tau_i, \boldsymbol{\beta})}_{\text{survival likelihood}} \\ &\propto \tau_i^{(r+\frac{n_i}{2})-1} \exp\left[-\tau_i \left(\lambda + \frac{\sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2}{2} \right) \right] [h(v_i)]^{\delta_i} \exp[-H(v_i)] \end{aligned} \tag{4.11}$$

β_j:

$$\begin{aligned} f(\beta_j|\text{All others}) &\propto \prod_{i=1}^{N} f(v_i, \delta_i|\mu_i, \tau_i, \boldsymbol{\beta}) f(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{N} \left[h(v_i)\right]^{\delta_i} \exp[-H(v_i)]\, f(\boldsymbol{\beta}) \end{aligned} \tag{4.12}$$

Where

$$\begin{aligned} h(t) &= \exp\{\beta_0 + \beta_1\mu_i + \beta_2\tau_i\} \\ H(t) &= \int_0^t h(x)dx \\ &= t \times \exp\{\beta_0 + \beta_1\mu_i + \beta_2\tau_i\} \end{aligned}$$

For numerical reasons, we actually use $\frac{\mu_i}{100}$ in the model, meaning that we reparametrise $\beta_1 = \frac{\beta_1^*}{100}$.

m:

$$\begin{aligned} \mu_i|\tau_i &\sim N(m, \tau) \\ m &\sim N(\mu_m, \tau_m^{-1}) \\ \tau &\text{ known} \\ f(m|\text{All others}) &\propto \prod_{i=1}^{N} f(\mu_i|\tau_i, m, \tau) f(m) \\ \text{which becomes } N &\left(\frac{\tau_m\mu_m + \tau \sum_{i=1}^{N} \mu_i}{\tau_m + \tau N}, (\tau_m + \tau N)^{-1} \right) \end{aligned} \tag{4.13}$$

τ:

$$\begin{aligned}
\mu_i|\tau_i &\sim N(m,\tau) \\
\tau &\sim \Gamma(r_\tau,\lambda_\tau) \\
f(\tau|\text{All others}) &\propto \prod_{i=1}^{N} f(\mu_i|\tau_i,m,\tau)f(\tau) \\
\text{which becomes } \Gamma &\left(r_\tau + \frac{N}{2}, \lambda_\tau + \frac{\sum_{i=1}^{N}(\mu_i - m)^2}{2} \right)
\end{aligned} \tag{4.14}$$

r:

$$\begin{aligned}
\tau_i &\sim \Gamma(r,\lambda) \\
r &\sim \Gamma(\alpha,\beta) \\
f(r|\text{All others}) &\propto \left(\prod_{i=1}^{N} \frac{\lambda^r \tau_i^{r-1}}{\Gamma(r)} \right) \frac{\beta^\alpha r^{\alpha-1} e^{-r\beta}}{\Gamma(\alpha)} \\
&\propto \frac{\lambda^{Nr} \prod_{i=1}^{N} \tau_i^{r-1}}{\Gamma(r)^N} \frac{r^{\alpha-1}e^{-r\beta}}{1} \\
&= \frac{\lambda^{Nr} \left(\prod_{i=1}^{N} \tau_i \right)^{r-1} r^{\alpha-1}e^{-r\beta}}{\Gamma(r)^N}
\end{aligned} \tag{4.15}$$

λ:

$$\begin{aligned}
\tau_i &\sim \Gamma(r,\lambda) \\
\lambda &\sim \Gamma(\alpha_\lambda,\beta_\lambda) \\
f(\lambda|\text{All others}) &\propto \lambda^{Nr+\alpha_\lambda-1} e^{\sum \tau_i + \beta_\lambda} \\
\text{which becomes } \Gamma &\left(Nr + \alpha_\lambda, \sum_{i=1}^{N} \tau_i + \beta_\lambda \right)
\end{aligned} \tag{4.16}$$

The resulting distributions in (4.10) to (4.16) are the full conditionals that we need to sample from at each iteration of the MCMC algorithm. The parameters m, τ, and λ have conjugate distributions that are Normal (4.13), Gamma (4.14), and Gamma (4.16), respectively. We can sample from them with the techniques described in Ripley (2009). The rest of the distributions do not adhere to standard forms — they are what

we labelled as unfamiliar distributions in section 3.4.

The ARS sampler is the most prevalent in the joint modelling literature (Brown and Ibrahim, 2003; Brown et al., 2005; Chi and Ibrahim, 2006). Chi and Ibrahim (2006) also encountered densities that were not log-concave and thus employed ARMS. In their handbook, Ibrahim et al. (2005) stated that ARS can be used in the joint modelling setting, but do not mention the prerequisite of log-concave densities.[3] As all of our unfamiliar densities were log-concave we employed the ARS sampler, as available from Gilks' website,[4] which featured a tested and working version of the algorithm. In doing so, we avoided coding errors, which in turn allowed us to focus fully on the structure of Gibbs sampler. Later, we replaced ARS with the Slice sampler, in order to compare speed and validate output. The Slice sampler also allowed for densities that were not log-concave, but we did not extend the model to use this.

Despite the popularity of the ARS sampler in the literature, we found no reason to favour it over the Slice sampler. In our study the Slice sampler improved the speed of the program — as given in appendix C. As we saw in section 3.4.2 the dynamics behind the Slice sampler are easy to understand and implement, while placing few restrictions on the density to be sampled.

4.3 Results

We ran the model on the dataset described in section 1.1, and used the R package `coda` (Plummer et al., 2006) to analyse the output from our C++ program. The

[3] It is worthwhile to briefly elaborate on log-concavity of densities and its appearance in joint survival models of the form (4.6), and we include some comments in appendix A.1.

[4] `http://www1.maths.leeds.ac.uk/~wally.gilks/`

unknown parameters in our model are given in (4.7), and we used the priors

$$\begin{aligned} \beta_i &\sim N(0, 0.0001^{-1}) \\ m &\sim N(20, 0.0001^{-1}) \\ \tau &\sim \Gamma(0.0001, 0.0001) \\ r &\sim \Gamma(0.0001, 0.0001) \\ \lambda &\sim \Gamma(0.0001, 0.0001). \end{aligned} \tag{4.17}$$

We chose to assume vague priors on the parameters m, τ, r, and λ, rather than specifying them directly. This was to prevent our beliefs about the distributions of μ_i and τ_i from affecting the results. To ascertain that our priors were not influencing our results, we also tested prior sensitivity by experimenting with different values for (4.17). All of our runs produced the same results. These priors, however, aren't realistic. We chose them with intention to be vague, taking guidance from the BUGS language examples. Although they did not influence our analysis negatively, we should have used more realistic priors, which reflect our knowledge, or expert knowledge, of the subject under scrutiny. The $\Gamma(0.0001, 0.0001)$ distribution favours small numbers of r and λ, whereas a better vague prior would rather decrease slowly on the positive real line, such as $\Gamma(1, 0.0001)$. There is no reason to use vague priors, however, since blood pressure is a well studied phenomenon and we should have used priors for m, τ, r, and λ that reflected this. The prior for m, for instance, has 0.5 probability of being less than zero, which is impossible.

Lunn et al. (2009) warn to use caution when setting priors for the variance components of unobserved parameters. The prior of $\sigma^{-2} = \tau \sim \Gamma(\epsilon, \epsilon)$ becomes biased away from $\sigma = 0$ as ϵ tends away from 0 (Gelman, 2006). This is a problem if we want to allow for the possibility of all individuals having the same μ_i or τ_i parameters. However, our model specifically assumes the individuals have different means and

variances for their blood pressure processes, and therefore our use of the $\Gamma(\epsilon, \epsilon)$ prior is justified.

4.3.1 Systolic blood pressure

For numerical reasons, we fitted the model using the hazard rate:

$$h(t|\mu_i, \tau_i) = \exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 \tau_i\right\}. \tag{4.18}$$

The SBP was mentioned by Rothwell et al. (2010) and Rothwell (2010) as a strong predictor of subsequent stroke, so we considered it first. Results for $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2\}$ are given in table 4.1 and figure 4.2, as produced by `coda`. The effective sample size is the proportion of independent MCMC samples in the chain, after correcting for autocorrelation. We notice from table 4.1 and figure 4.2 that the MCMC chains for $\boldsymbol{\beta}$ have large autocorrelations, leading to small effective sample sizes. This was the case for all variants of the MCMC algorithms we used, except for Stan (table C.4). The high autocorrelations were indicative of either slow mixing, or a lack of MCMC convergence. We arrived at the same results with multiple MCMC implementations, however, and also performed convergence tests which did not reject convergence. These are given in appendix C. Furthermore, of our multiple implementations Stan did not experience the slow mixing, adding further confidence in our results.[5]

In the tables, the 95% highest posterior density (HPD) interval is the shortest interval in the parameter space that contains 95% of the posterior probability. We used Bayes factors to test the significance of the parameters, with results in table 4.3. Table 4.2 gives the MCMC results for m, τ, r, and λ. The results for our constant hazard rate model support the claims made by Rothwell (2010) that it is not only a

[5]Stan was slower to run in real time than the other samplers, but it was computationally more effective, achieving more effective samples with shorter MCMC runs. See table 5.2 for a comparison of average seconds per sample.

	Mean	SD	HPD interval (95%)	Effective sample size
β_0	-4.7	0.85	(-6.4 ; -3.1)	0.0034
β_1	1.3	0.48	(0.38 ; 2.24)	0.0037
β_2	-147	46	(-239 ; -59)	0.0116

Table 4.1. Posterior distribution MCMC results for the model prior parameters, from a chain of 200,000 iterations, SBP.

	Mean	SD	HPD interval (95%)	Effective sample size
m	145.7	0.4130	(144.9 ; 146.5)	0.798
τ	0.0035	0.0001	(0.0033 ; 0.0038)	0.639
r	3.28	0.17	(2.95 ; 3.63)	0.022
λ	589	37	(518 ; 663)	0.021

Table 4.2. Posterior distribution MCMC results for the model prior parameters, from a chain of 200,000 iterations, SBP.

H_1	$p(H_1\|\boldsymbol{Y})$	Bayes Factor
$\beta_0 < 0$	> 0.9999	> 10000
$\beta_1 > 0$	0.9959	244
$\beta_2 < 0$	0.9999	8332

Table 4.3. Hypothesis testing for the parameters of interest, SBP. Tests assumes prior probability of H_1 and H_0 are equal.

Model	$h(t)$	p_D	DIC
Precision	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 \tau_i\right\}$	3571	172907
CV	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 \frac{\sigma_i}{\mu_i}\right\}$	3564	172913
SD	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 \sigma_i\right\}$	3566	172924
Mean only	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100}\right\}$	3569	172937

Table 4.4. DIC and p_D values for models with varying hazard rates, SBP. The p_D can be interpreted as the effective number of parameters in the model.

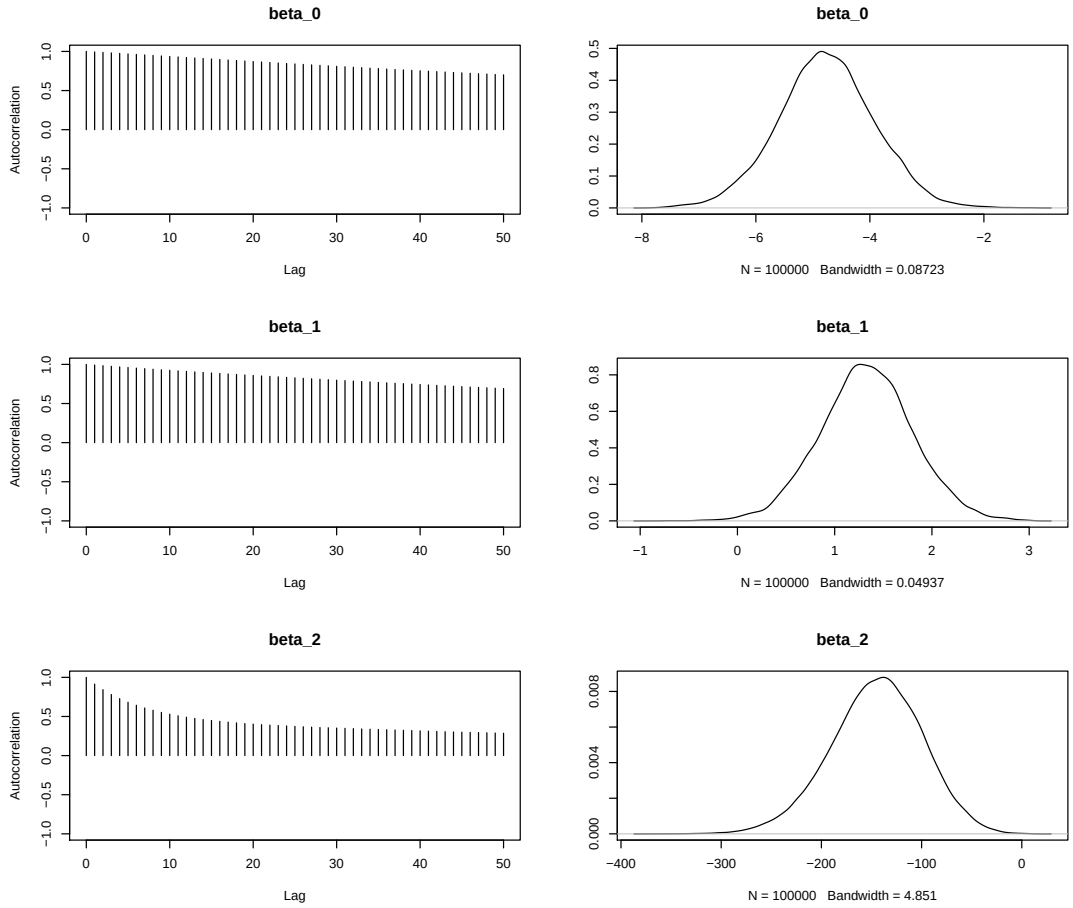


Figure 4.2. MCMC chain autocorrelation and kernel density plot of the β parameters, SBP

patient's mean blood pressure that affects stroke risk, but that the volatility should be taken into account as well. All of the $\boldsymbol{\beta}$ parameters are significant. A higher mean leads to a higher hazard rate, while a smaller variance (larger precision) translates to a lower hazard rate. We can demonstrate this in an intuitive manner, by calculating

the expected time to stroke:

$$
\begin{aligned}
E(T_i|\mu_i,\tau_i) &= \int_0^\infty S(t|\mu_i,\tau_i)dt \qquad \text{where } S(t) \text{ is the survival function} \\
&= \int_0^\infty e^{-\int_0^t h(x|\mu_i,\tau_i)dx}dt \\
&= \int_0^\infty e^{-\int_0^t \exp\left(\beta_0+\beta_1\frac{\mu_i}{100}+\beta_2\tau_i\right)dx}dt \\
&= \frac{1}{e^{\left(\beta_0+\beta_1\frac{\mu_i}{100}+\beta_2\tau_i\right)}}\left[-e^{-t\exp\left(\beta_0+\beta_1\frac{\mu_i}{100}+\beta_2\tau_i\right)}\right]_0^\infty \\
&= e^{-\left(\beta_0+\beta_1\frac{\mu_i}{100}+\beta_2\tau_i\right)} \qquad (4.19)
\end{aligned}
$$

The constant hazard rate assumed in this model is responsible for the simple form of (4.19). Using this equation, we can see that the expected time to stroke for a person with $\mu_i = 145.7$ (at the median of the μ_i distribution) and $\tau_i = 0.0036$ ($\sigma_i = 16.6$, 30th percentile of the τ_i distribution) is about 29 years. A person having the same μ_i but with $\tau_i = 0.0067$ ($\sigma_i = 12.2$, 70th percentile of the τ_i distribution) has an expected time to stroke of 45 years. A more direct comparison can be drawn using the hazard rates for different values of μ_i and τ_i, but looking at the expected time to event gives an intuitive overview. The two expected times calculated above may seem high, especially considering that in order to qualify for the study a patient had to have suffered a mild stroke and would therefore most likely be of advanced age. These times, however, are expectations of time to stroke and should not be regarded as expected future lifetime in general. A person with a high expected time to stroke will probably succumb to a different form of mortality, rather than having a stroke. They do illustrate the point that a higher blood pressure variance increases stroke risk.

Our findings are thus in accordance with those of Rothwell (2010), as well as those of Rothwell et al. (2010). Their paper, however, did not account for measurement error in the longitudinal observations. Simply using the raw measurements can lead

to biased estimates of model parameters, as discussed by Prentice (1982), Hughes (1993), Raboud et al. (1993), and Hu et al. (1998). Furthermore our concern is not measurement error in the observations. Our concern lies with error in estimating the mean and variance using a small number of observations. That is, whereas in the standard setting the random error is a disturbance, in this setting the random error is a real phenomenon, part of what we are actually trying to measure: the process variance. Our model, however, allows for uncertainty in the subject mean and variance.

Rothwell et al. (2010) used three measures to calculate the blood pressure variability of subjects. These measures were the SD, the Coefficient of Variation (CV), and the Variation Independent of Mean (VIM). The VIM was calculated as being proportional to $^{SD}/_{\bar{x}}$, with x being a blood pressure measurement derived from curve fitting. They grouped the individuals into quintiles according to their estimated mean and variability, and calculated hazard ratios of top vs bottom quintile for risk of subsequent stroke. Since the subjects had different numbers of blood pressure readings, the analysis was done separately for subjects with two, four, six, eight, and ten readings. This was done for all three measures of variability. The subjects with the most readings were expected to provide the most accurate estimates for variability, but they made up only a small portion of the data, thus not providing many events to analyse. The results for their SD and CV model are reproduced in table 1. The estimated hazard ratios for the models using numbers of measurements were different. Moreover, the 95% confidence intervals of the hazard ratios calculated for the subjects with two, and ten measurements did not overlap. Using the CV and VIM made the situation worse. We can expect such a bias when measurement error in the covariates is not properly accounted for. In the next chapter, we attempt to calculate the exact extent of this bias.

Like the study by Rothwell et al. (2010), our model can also use the longitudinal

SD or CV in the hazard rate, or it can be extended to their VIM measure. It provides a way to model the longitudinal and survival processes jointly, and we avoid the practice of using the output of one model as the input for another. The Bayesian model also allows us to account for longitudinal measurement error. We have a choice of using either SD, CV, or the precision, and we will use the model-comparison methods discussed in section 3.1.2 to determine which hazard rate provides the best fit. Table 4.4 contains the values of p_D and DIC calculated using MCMC output of 30,000 iterations for models with hazard rates as indicated. We did not use VIM, since they did not specify how they calculated it.

Since p_D should estimate the effective number of parameters in the model, we expect it to be near 3,600, and table 4.4 shows this to be true.[6] According to the DIC, the model that achieves the best fit is the one that uses the precision in the hazard rate, as given in (4.18). We also see that all three models that include a variance effect are superior to the model that only includes a mean effect.

In order to draw a comparison to our results, we fitted a Cox proportional hazards model as used by Rothwell et al. (2010), but with the observed continuous covariates rather than the categorical variables. Moreover, we used the precision in the hazard rate, for comparability:

$$h(t) = h_0(t) \exp\left(\beta_1 \frac{MU_i}{100} + \beta_2 PREC_i\right) \tag{4.20}$$

where MU_i and $PREC_i$ are the observed mean and precision ($1/sD_i^2$) for the ith individual, using SBP. Results are given in table 4.5 and the precision parameter shows a similar pattern to that of table 1 — confidence intervals in the models that used low number of blood pressure measurements do not overlap with those of models that used a high number of measurements.

[6]Our model includes 1,838 μ_is, 1,838 τ_is, as well as 7 other parameters, so we expect p_D to be around 3,683.

	Mean SBP			Precision SBP		
n	β_1	(95% CI)	p-val	β_2	(95% CI)	p-val
2	1.66	(1.07 ; 2.24)	<0.0001	0.57	(-0.52 ; 1.66)	0.3070
4	1.49	(0.81 ; 2.16)	<0.0001	-2.68	(-10.04 ; 4.69)	0.4759
6	1.59	(0.75 ; 2.43)	0.0002	-13.07	(-30.17 ; 4.03)	0.1340
8	1.48	(0.42 ; 2.53)	0.0061	-62.26	(-107.88 ; -16.64)	0.0075
10	1.11	(-0.20 ; 2.43)	0.0976	-137.67	(-221.64 ; -53.69)	0.0013

Table 4.5. Results for the Cox model using the continuous covariates.

4.3.2 Diastolic blood pressure

Rothwell et al. (2010) found DBP to be a weak predictor of subsequent stroke, whereas Rothwell (2010) made no mention of it other than stating that more research is needed 'for which troughs in DBP could be key.' We investigated the effect of the DBP on survival using the same model as in the previous section with the hazard rate (4.18), but this time using DBP in the longitudinal process. MCMC results for the parameters of interest are given in table 4.6, and for the prior parameters in table 4.7. Conversely to the results of Rothwell et al. (2010), we found that the DBP also had a significant effect on the survival outcome, as seen in the hypothesis test of table 4.8, testing $H_1 : \beta_2 < 0$. Although the Bayes factor of 59 for the precision parameter is not as high as the 2,316 calculated for the SBP, it still falls well within the region of significance evidence for H_1 according to table 3.1.

We again used (4.19) to interpret the results by calculating the expected future lifetime for pairs of μ_i and τ_i. This also allowed us to compare the results for the DBP with those of the SBP for consistency. Using the same percentiles as for the SBP future lifetime calculations, we calculated that a person with $\mu_i = 85.3$ (the median of the μ_i distribution) and $\tau_i = 0.012$ (or $\sigma_i = 9.3$, the 30th percentile of the τ_i distribution) has an expected future lifetime of 31 years. A person with the same μ_i but with $\tau_i = 0.019$ (or $\sigma_i = 7.2$, the 60th percentile of the τ_i distribution) has an expected future lifetime of 39 years. Both of these expected future lifetimes are within 15% of those calculated for the SBP. Similar to the previous section, the model with

the precision in the hazard rate was the best model according to the DIC, as shown in table 4.9.

4.4 Independent corroboration

All results shown in section 4.3 are from the custom sampler. These results were corroborated using JAGS and Stan. Output from both these programs are given in appendix C.3 for the SBP model using the hazard rate (4.18).

The independent corroboration served a dual purpose. Firstly, it showed that this type of model, despite its complexity and the necessary use of 'tricks', can be successfully undertaken using standard Gibbs sampling software such as JAGS and BUGS. Secondly, and more importantly, it reinforced confidence in our results: we arrived at the same answer using two independent parameter estimation techniques that, though similar in framework, are not identical in implementation.[7] The custom sampler serves as a stepping stone. We know that the code is correct up to this point, therefore we can extend it to accommodate more complicated hazard rates that go beyond the capabilities of JAGS and BUGS.

Although the idea of using blood pressure volatility to measure stroke risk has previously been discussed in medical literature, our use of the longitudinal volatility to assess risk in a joint survival model is new. In this chapter we used it to substantiate the claims of Rothwell (2010), presented in our introduction. In the next chapter we will show methods to assess goodness of fit for our models, as well as looking at how our model can be extended.

[7] We also corroborated these results using a third estimation scheme, as we will see in chapter 7.

	Mean	SD	HPD interval (95%)	Effective sample size
β_0	-6.28	1	(-8.23 ; -4.25)	0.0024
β_1	3.77	1.03	(1.65 ; 5.78)	0.0026
β_2	-31.5	15.6	(-62.2 ; -1.4)	0.0211

Table 4.6. Posterior distribution MCMC results for the model prior parameters, from a chain of 200,000 iterations, DBP.

	Mean	SD	HPD interval (95%)	Effective sample size
m	85.3	0.18	(85 ; 85.7)	0.711
τ	0.02	0.0008	(0.018 ; 0.021)	0.520
r	4.6	0.28	(4.03 ; 5.13)	0.012
λ	284	20	(246 ; 324)	0.012

Table 4.7. Posterior distribution MCMC results for the model prior parameters, from a chain of 200,000 iterations, DBP.

H_1	$p(H_1\|\boldsymbol{Y})$	Bayes Factor
$\beta_0 < 0$	> 0.9999	> 10000
$\beta_1 > 0$	0.9998	5404
$\beta_2 < 0$	0.9834	59

Table 4.8. Hypothesis testing for the parameters of interest, DBP, assuming prior probability of H_1 and H_0 are equal.

Model	$h(t)$	p_D	DIC
Precision	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 \tau_i\right\}$	3579	150304
SD	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 \sigma_i\right\}$	3576	150308
CV	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 \frac{\sigma_i}{\mu_i}\right\}$	3582	150317
Mean only	$\exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100}\right\}$	3584	150340

Table 4.9. DIC and p_D values for models with varying hazard rates, SBP. The p_D can be interpreted as the effective number of parameters in the model.

4.5 Alternative approach

In this chapter we defined a model that had a similar structure to the Cox proportional hazards model. The logistic regression is another approach that can be employed to analyse data of this form. Peduzzi et al. (1987) compared the Cox proportional hazards model with the logistic regression, in studies where patients are observed for a fixed period of time. Using the logistic regression, events are considered to come from a Bernoulli process:

$$\delta_i \sim \text{Bern}(p_i)$$

and a link function is used to relate the linear predictor of the covariates to the parameter:

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \beta_1 \mu_i + \beta_2 \tau_i \\ p_i &= \text{logit}^{-1}(\beta_0 + \beta_1 \mu_i + \beta_2 \tau_i). \end{aligned}$$

We are not limited to the logit link function. Since we expect p_i to be small, and μ_i and τ_i to be positive, the $\boldsymbol{\beta}$ parameters will be estimated to compensate for the shape of the curve in addition to containing information about the effects of the covariates, as shown in figure 4.3. This is not necessarily wrong, but it will be in our interest to choose a link function that reflects our knowledge of positive covariates and small event probabilities. We are using MCMC methods so we are not dependent on link functions that provide mathematical tractability. One option for a link function is the inverse of the lognormal cumulative distribution function, which will lead to the covariates being mapped from $[0, \infty)$ to $[0, 1)$. This function can be implemented in the standard MCMC software. We do not consider logistic regressions in this thesis, but it is an alternative to the models we investigate.

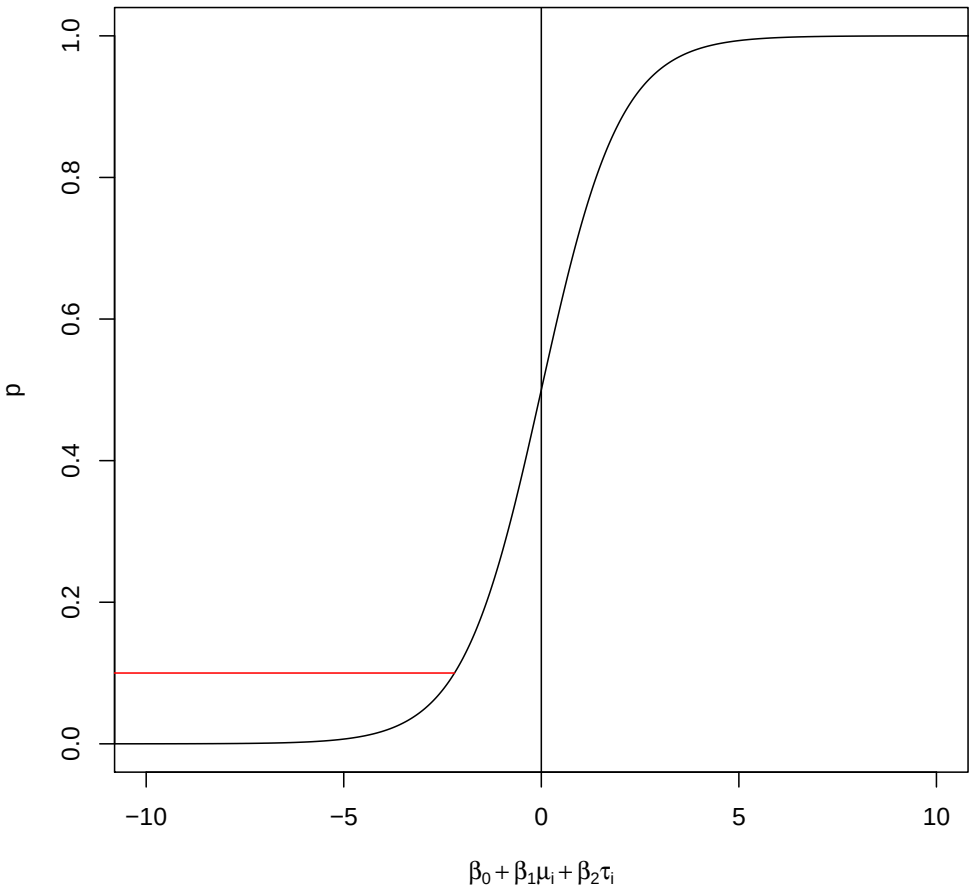


Figure 4.3. The inverse of the logit function, showing how the covariates will be mapped to small values of p. Since μ_i and τ_i will be positive, the β_i parameters will be estimated to adhere to the curve.

Chapter 5

Metropolis-Hastings for joint models

In this chapter we use the software described in chapter 4 to explore sampling from the unfamiliar densities arising in joint survival models. We specifically examine the use of Metropolis-Hastings, and explain why we opted to write our own sampler.

5.1 Smart sampling with a custom sampler

The sampling procedure we describe in this section will hold for all joint survival models specified using the shared parameter approach, as explained in section 2.3. A joint model defined using this approach will have the following posterior:

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{v}) \propto \prod_{i=1}^{N} \Big[f(\boldsymbol{y}_i|\boldsymbol{\theta}) f(v_i|\boldsymbol{\theta}) \Big] p(\boldsymbol{\theta})$$

for a Bayesian model with N subjects. Now consider a subject-specific parameter $\mu_i \in \boldsymbol{\theta}$ that appears in both the survival and longitudinal parts of the model. Let $\boldsymbol{\theta}_{/\mu_i}$ be all the parameters in $\boldsymbol{\theta}$ excluding μ_i. In the Gibbs sampling framework, we need

to sample from the full conditional distribution:

$$\begin{aligned} p(\mu_i|\boldsymbol{\theta}_{/\mu_i}, \boldsymbol{y}_i, v_i) &= \frac{p(\mu_i, \boldsymbol{\theta}_{/\mu_i}|\boldsymbol{y}_i, v_i)}{p(\boldsymbol{\theta}_{/\mu_i}|\boldsymbol{y}_i, v_i)} \\ &\propto p(\mu_i, \boldsymbol{\theta}_{/\mu_i}|\boldsymbol{y}_i, v_i) \\ &= p(\boldsymbol{\theta}|\boldsymbol{y}_i, v_i) \\ &= \text{terms in } p(\boldsymbol{\theta}|\boldsymbol{y}_i, v_i) \text{ containing } \mu_i \\ &= \prod_{j=i}^{n_i} \Big[f(y_{ij}|\mu_i, \boldsymbol{\theta}_{/\mu_i}) \Big] f(v_i|\mu_i, \boldsymbol{\theta}_{/\mu_i}) f(\mu_i|\boldsymbol{\theta}_{/\mu_i}). \end{aligned} \tag{5.1}$$

This is the form we see in (4.10) and (4.11). Both of these consist of three parts — the longitudinal, the survival, and the prior:[1]

$$\underbrace{\prod_{j=i}^{n_i} \Big[f(y_{ij}|\mu_i, \boldsymbol{\theta}_{/\mu_i}) \Big]}_{\text{longitudinal}} \underbrace{f(v_i|\mu_i, \boldsymbol{\theta}_{/\mu_i})}_{\text{survival}} \underbrace{f(\mu_i|\boldsymbol{\theta}_{/\mu_i})}_{\text{prior}}.$$

If we specify the prior distribution as the conjugate prior of the longitudinal part of (5.1), we obtain a familiar distribution multiplied by a survival part:

$$\overbrace{\underbrace{\prod_{j=i}^{n_i} \Big[f(y_{ij}|\mu_i, \boldsymbol{\theta}_{/\mu_i}) \Big]}_{\text{longitudinal}} \underbrace{f(\mu_i|\boldsymbol{\theta}_{/\mu_i})}_{\text{prior}}}^{\text{familiar distribution}} \underbrace{f(v_i|\mu_i, \boldsymbol{\theta}_{/\mu_i})}_{\text{survival}} \tag{5.2}$$

$$= \underbrace{p(\mu_i|\boldsymbol{y}_i, \boldsymbol{\theta})}_{\substack{\text{familiar conjugate} \\ \text{distribution}}} \underbrace{f(v_i|\mu_i, \boldsymbol{\theta}_{/\mu_i})}_{\text{survival}}. \tag{5.3}$$

The familiar distribution takes the form of a conjugate posterior distribution of μ_i.

Wang and Taylor (2001) presented a useful method for sampling from these distributions, using the Metropolis-Hastings update: 'If a conjugate prior is used, we

[1]The prior in this equation is not the actual prior on μ_i, since there may be higher levels in the hierarchy. The prior is actually the distributional assumptions on μ_i in this level of the hierarchy, but we use the terminology since it helps to illustrate our point of conjugacy.

sample a new value directly from the conditional distribution. If a non-conjugate prior is used, we use the standard distribution from the likelihood as the proposal density in the Metropolis-Hastings step.' Their 'standard distribution' is our familiar conjugate distribution in (5.3). This method involves choosing the familiar distribution in (5.3) as the proposal density for the Metropolis-Hastings update. If we want to sample from a density $\pi(X)$, we need to propose a new value X from a distribution $q(X|X^{(t-1)})$ and then calculate the acceptance probability as

$$\alpha\Big(X|X^{(t-1)}\Big) = \min\left(1, \frac{\pi(X)q(X^{(t-1)}|X)}{\pi(X^{(t-1)})q(X|X^{(t-1)})}\right).$$

We will use the familiar distribution as our independent proposal distribution:[2]

$$\begin{aligned} q(\mu_i) &= p(\mu_i|\boldsymbol{y}_i, \boldsymbol{\theta}) \\ \pi(\mu_i) &= p(\mu_i|\boldsymbol{y}_i, \boldsymbol{\theta}) f(v_i|\mu_i, \tau_i, \boldsymbol{\beta}) \\ &= q(\mu_i) \underbrace{f(v_i|\mu_i, \tau_i, \boldsymbol{\beta})}_{\text{survival}} \end{aligned}$$

and the acceptance probability becomes

$$\alpha\Big(\mu_i|\mu_i^{(t-1)}\Big) = \min\left(1, \frac{\pi(\mu_i)q(\mu_i^{(t-1)})}{\pi(\mu_i^{(t-1)})q(\mu_i)}\right) \tag{5.4}$$

$$= \min\left(1, \frac{f(v_i|\mu_i, \tau_i, \boldsymbol{\beta})}{f(v_i|\mu_i^{(t-1)}, \tau_i, \boldsymbol{\beta})}\right). \tag{5.5}$$

The crucial part of their method lies in (5.5). Compare it to (5.2): the function that would have to be evaluated by generic sampling algorithms such as ARMS or the Slice sampler. Both are functions of μ_i, but (5.5) has no product term and thus requires fewer function evaluations than (5.2).

A univariate model may not warrant implementation, but when the longitudinal

[2] An independent proposal distribution occurs when new values are proposed independently from previous values, or $q(X|X^{(t-1)}) = q(X)$

model is multivariate — such as the model in chapter 7 — this drastically reduces the complexity of samplers and the computations required. In a multivariate model $\pi(\mu_i)$, will be a matrix distribution. But if we use the right choice of prior on μ_i, the proposal will be a familiar matrix distribution, and we can use standard methods to sample from it. Then, we only have to evaluate

$$\frac{f(v_i|\mu_i, \tau_i, \boldsymbol{\beta})}{f(v_i|\mu_i^{(t-1)}, \tau_i, \boldsymbol{\beta})}$$

at each iteration of the Metropolis-Hastings sampler. This technique proposes and updates an entire matrix at a time, thus not requiring us to split matrices and sample them with a further Gibbs sampler. We still have the advantage of proportionality in the Metropolis-Hastings acceptance. If $\tilde{\pi}(x) \propto \pi(x)$, then

$$\frac{\pi(x)q(x^{(t-1)}|x)}{\pi(x^{(t-1)})q(x|x^{(t-1)})} = \frac{\tilde{\pi}(x)q(x^{(t-1)}|x)}{\tilde{\pi}(x^{(t-1)})q(x|x^{(t-1)})}$$

and thus we only need (5.3) to a normalising constant.

This novel sampling method was suggested by Wang and Taylor (2001). Despite their paper, the ARS sampler has remained the most prevalent way to sample from unfamiliar densities in the joint survival literature. Other papers that included a multivariate longitudinal process were Brown et al. (2005) and Chi and Ibrahim (2006). Both of these papers used Gibbs sampling, and both of them used ARS (or ARMS) to sample from their unfamiliar distributions, even though both of them cite Wang and Taylor (2001). We reran our model from chapter 4 using this smart Metropolis-Hastings update, with the results given in table 5.1. Like JAGS and Stan, it also corroborated our earlier results. The smart Metropolis-Hastings update ran faster than the model with ARS, but slower than the model with Slice sampling. Table 5.2 gives a comparison of the time taken by each of the programs to draw one effective sample, for the β_1 and β_2 parameters. Exact run times are given in appendix C.2.

	Mean	SD	HPD interval (95%)	Effective sample size	$P(\theta > 0)$
β_0	-4.6	0.86	(-6.27 ; -2.87)	0.003	< 0.0001
β_1	1.24	0.49	(0.25 ; 2.17)	0.004	0.9937
β_2	-150	46	(-239 ; -60)	0.013	0.0001
m	146	0.4137	(144.9 ; 146.5)	0.824	
τ	0.0035	0.0001	(0.0033 ; 0.0038)	0.647	
r	3.27	0.17	(2.93 ; 3.62)	0.021	
λ	588	37	(517 ; 663)	0.021	

Table 5.1. MH Smart sampling: 50,000 iterations (after burn-in).

	β_1	β_2
JAGS	33.2	8.43
Custom — Slice sampling	6.2	1.84
Custom — ARS	14.45	4.29
Stan (excluding warmup)	1.16	1.33
Stan (including warmup)	3.48	3.98
Metropolis-Hastings	10.16	2.99

Table 5.2. The average seconds per effective sample of the various programs.

.

5.2 A bespoke MCMC algorithm

For the purposes of this thesis, we spent three months developing, debugging, and testing the C++ code for our custom sampler, until we had produced output similar to JAGS. Independent corroboration of our results was an important factor in the decision to write our own MCMC algorithm, but using JAGS and Stan would have been sufficient for that. We needed more than independent corroboration. We wanted to investigate MCMC for joint survival models in detail, and compare the different samplers: ARS, Slice, and the Metropolis-Hastings update described in section 5.1. We ported the code provided by the authors of ARS and Slice sampler to run in C++, with function signatures similar to the following:

```
double ars( double x_prev,
            double (* f_logdensity)(double x, void * data),
            void * data ,
            ... // other configuration
          );
```

```
double slicesample( double x_prev,
                    double (* f_logdensity)(double x, void * data),
                    void * data ,
                    ... // other configuration
                   );
```

which made them interchangeable in our code. The 'other configuration' included upper and lower bounds of the density to be sampled from, as well as other calibration parameters needed by the algorithms. Since the Metropolis-Hastings sampler required specific assumptions on the prior distributions, it required replacement of the above functions.

We furthermore required the ability to specify complicated hazard rates, beyond the abilities of the standard MCMC packages. Alongside this, we needed greater execution speed than JAGS and Stan could provide, to experiment with a variety of models. Writing our own sampler also enabled us to sample survival times using the following quality. If $H(t) = \int_0^t h(x)dx$ is a cumulative hazard rate and X is a random variable from an exponential distribution with parameter 1, then $H^{-1}(X)$ is a random variable with hazard rate $h(t)$. This is an effective way of sampling survival times, and it could only be exploited using a custom sampler. It was useful for posterior predictive sampling, and allowed us to experiment with posterior predictive model checks.

We made extensive use of the standalone Rmath library (R Core Team, 2012) and its random number generators. The sampler was set up to run for 1000 iterations, and recorded all the parameter values and the random seeds of the last iteration, which were used as the start state of the next chain. This meant we could run arbitrarily long chains, without reserving computational resources specifically for this task, since we could stop and start the chain as necessary. We also ran chains in parallel by setting different starting values and random seeds, and starting the program as a new thread. We wrote our own functions for reading and writing the data to file. The first

version of our sampler used standard C++ arrays, but we later rewrote the program using Standard Template Library (STL) vectors. The former was 20% faster, but the latter provided code that was easier to comprehend and extend. The result of our efforts and all the iterations of our code is an MCMC Gibbs sampler that fit joint survival models with a variety of hazard rates. The code in its current form requires a how-to manual explaining the set up and format of the data, and the structure of the program. Fitting a new model requires the C++ source code to be changed, and the program to be compiled and run. Alterations to the hazard rate are easy to implement as they do not change the distributional assumptions of the model, but changes to the longitudinal process require extensive changes in the code. The code we wrote for this thesis is too problem specific to be worked into a standalone modelling package like JAGS. It could be written into a library of functions, leaving the hazard rate specification to the end user, but the longitudinal specification will remain limited to what is currently coded into the structure of the program. If our longitudinal model becomes the norm for longitudinal survival models, writing the MCMC algorithm into a package will be a worthwhile venture. Another possibility is to extend the functionality of JAGS, to handle joint models without the use of tricks. Wabersich and Vandekerckhove (2014) provided a useful tutorial on extending the capabilities of JAGS to new distributions.

Chapter 6

Diagnostics for joint models and measurement error

Gauging model fit can be difficult when using MCMC methods to fit complex models. Since the models are not in standard form, we cannot use standard methods. Literature on joint survival modelling that suggest methods for model comparison and model fit includes Brown et al. (2005) and Guo and Carlin (2004), who used the deviance information criterion (DIC) to compare models. Diggle et al. (2008) mentioned graphical model checks using Kaplan-Meier curves, while Henderson et al. (2000, 2002) extended this to posterior predictive model checks. We opted to use posterior predictive model checking to check the longitudinal and survival parts of the model.

Stern and Sinharay (2005) gave a comprehensive overview of popular methods used to assess goodness of fit in Bayesian models. Most of their paper was dedicated to posterior predictive sampling, an idea introduced and discussed by Gelman et al. (1996). In recent years it has been further developed in papers such as Gelman and Meng (1996), Gelman et al. (2000), and Gelman (2003).

6.1 Posterior predictive model checking

The technique involves sampling values from the posterior predictive distribution (3.10), and seeing how well generated data correspond to the original data. In essence, this involves using our fitted model to predict new values and then checking how well they resemble the original observations. We adopt the same notation used by Stern and Sinharay (2005), and Gelman et al. (1996).

Let $\boldsymbol{y}$ be the observed data and $\boldsymbol{\theta}$ the set of parameters in the model, and let $\boldsymbol{y}^{rep}$ denote replicated or predicted data that would be observed if the process that generated $\boldsymbol{y}$ is replicated with the same value of $\boldsymbol{\theta}$ that generated the observed data (Stern and Sinharay, 2005). As noted in Gelman (2003), 'The basic idea is to expand from

$$p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \text{ to } p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})$$

where $\boldsymbol{y}^{rep}$ is a replicated data set of the same size and shape as the observed data $\boldsymbol{y}$. All model checking (both "exploratory" and "confirmatory") can then be interpreted as comparisons between $\boldsymbol{y}$ and $\boldsymbol{y}^{rep}$.' The full Bayesian model thus becomes $p(\boldsymbol{y}, \boldsymbol{y}^{rep}, \boldsymbol{\theta})$ and all posterior calculations, including model checks, are done using $p(\boldsymbol{y}^{rep}, \boldsymbol{\theta}|\boldsymbol{y})$.

To test model fit, we define *discrepancy measures* or *test statistics* $D(\boldsymbol{y}, \boldsymbol{\theta})$, as in Gelman et al. (1996). This is a test statistic like in the classical sense, but it depends on the parameters as well. Thus, $D(\boldsymbol{y}, \boldsymbol{\theta})$ is a function that can be calculated for values of $\boldsymbol{y}$ and $\boldsymbol{\theta}$, which we will use for comparison. We compare $D(\boldsymbol{y}, \boldsymbol{\theta})$ with $D(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$, that is, the discrepancy of the observed data with that of the simulated data. This is done using the reference distribution of $D(\boldsymbol{y}, \boldsymbol{\theta})$, which is derived from the joint posterior distribution of $\boldsymbol{y}^{rep}$ and $\boldsymbol{\theta}$ as

$$p(\boldsymbol{y}^{rep}, \boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}). \tag{6.1}$$

A significant difference between the two discrepancies would indicate lack of model fit. Analytical comparison of these distributions will only be possible for a few simple cases, so for complex models we make use of Monte Carlo simulations. Gelman et al. (1996) mentioned that this method does not lead to an increased computational burden, since we are already drawing samples from the posterior distribution of $\boldsymbol{\theta}$ during model estimation.

We can estimate a posterior predictive p-value by drawing N different datasets $\boldsymbol{y}_i^{rep}$ and $\boldsymbol{\theta}_i$, and calculating the portion of cases in which the simulated discrepancy variable exceeds the realised value. The posterior predictive p-value is defined as

$$p_b = P\big(D(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq D(\boldsymbol{y}, \boldsymbol{\theta}) \big| \boldsymbol{y}\big)$$

and we can estimate it using draws from $p(\boldsymbol{y}^{rep}, \boldsymbol{\theta}|\boldsymbol{y})$ as

$$\hat{p}_b = \frac{1}{N} \sum_{j=1}^{N} I\left[D(\boldsymbol{y}_j^{rep}, \boldsymbol{\theta}_j) \geq D(\boldsymbol{y}, \boldsymbol{\theta}_j)\right], \tag{6.2}$$

where $I[\cdot]$ is the indicator function (Stern and Sinharay, 2005). Although the realised discrepancy is not observable, p_b is well defined and computable. In some cases the discrepancy measure will depend only on the data and not the parameters as well, such that $D(\boldsymbol{y}, \boldsymbol{\theta}) = D(\boldsymbol{y})$, and it is then called a test statistic as in the classical sense.

Considering (6.1), the full set of observations that corresponded to our joint model was $\boldsymbol{y} = \{\boldsymbol{w}_1, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_N\}$, where $\boldsymbol{w}_i$ is the set of observations for each person. This set contained the survival time v_i, the censor indicator δ_i, and the longitudinal observations x_{ij} for $j = 1 \ldots n_j$. Together, they constituted the $\boldsymbol{w}_i$ for the ith individual,

$$\boldsymbol{w}_i = \{v_i, \delta_i,\ x_{i1}, x_{i2} \ldots, x_{in_i}\}.$$

Furthermore, the full set of parameters used in our model was

$$\boldsymbol{\theta} = \{\beta_1, \beta_2, \beta_3, m, k, r, \tau,\ \tau_1, \ldots, \tau_N,\ \mu_1, \ldots, \mu_N\}.$$

The presence of both longitudinal and survival observations for each individual presented some difficulties for posterior predictive model checking, as it complicated the construction of a discrepancy measure. The difference in nature of longitudinal and survival data makes it difficult to construct a $D(\boldsymbol{y}, \boldsymbol{\theta}_i)$ or even $D(\boldsymbol{y})$ that takes both into account at the same time. In addition to this, the survival observations were subject to censoring, further complicating matters. A logical first step was to compare the observed longitudinal data to the generated longitudinal data, and to do the same for the survival data.

6.1.1 Longitudinal

For the longitudinal data, we used the mean, standard deviation, and maximum as discrepancy measures (or rather, test statistics, since they did not make use of the parameters). We used $D(\boldsymbol{y}, \theta) = D(\boldsymbol{y}) = \text{mean}(\boldsymbol{y})$ and then repeated the test using $\text{SD}(\boldsymbol{y})$ and $\max(\boldsymbol{y})$. We simulated 100 sets of $\boldsymbol{y}_i^{rep}$, and compared $D(\boldsymbol{y}_i^{rep})$ with $D(\boldsymbol{y})$ as in (6.2). For each subject, we required a set of longitudinal observations $\boldsymbol{x}_i = \{x_{i1}, x_{i2} \ldots, x_{in_i}\}$. Although n_i, the number of longitudinal observations observed for subject i can be regarded as random, we did not sample it. Instead, we sampled the same number of longitudinal observations for each patient as observed. We needed to sample n_i values from from

$$f(x_{ij}^{rep}|\boldsymbol{y}) = \int_\theta f(x_{ij}^{rep}|\mu_i, \tau_i) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta},$$

where $\boldsymbol{y}$ is all the observed data and $\boldsymbol{\theta}$ is all the parameters. At each iteration we had posterior samples μ_i^* and τ_i^*, so we sampled from the normal distribution $f(x_{ij}^{rep}|\mu_i^*, \tau_i^*)$. The results are given in figures 6.1 to 6.3. Looking at the histograms and p-values, it appears that simulated data replicated the features of the actual observations well.

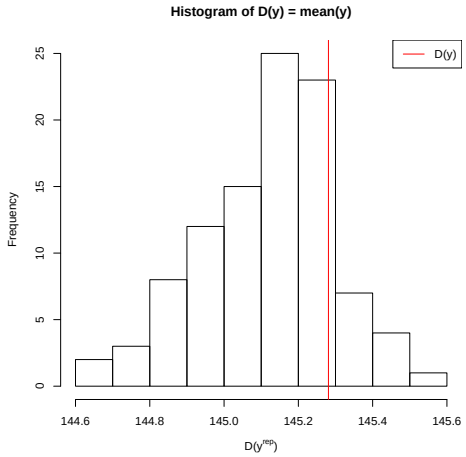


Figure 6.1. Using the mean as discrepancy measure gives a p-value of 0.85.

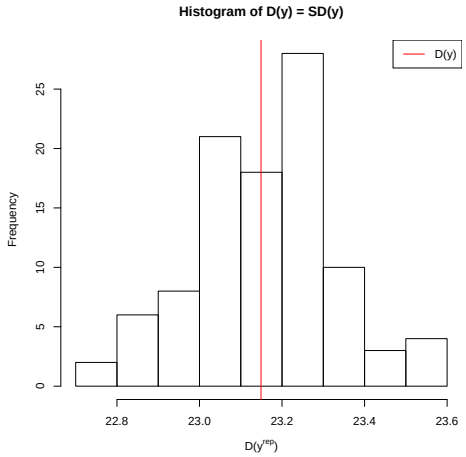


Figure 6.2. Using the standard deviation as discrepancy measure gives a p-value of 0.44.

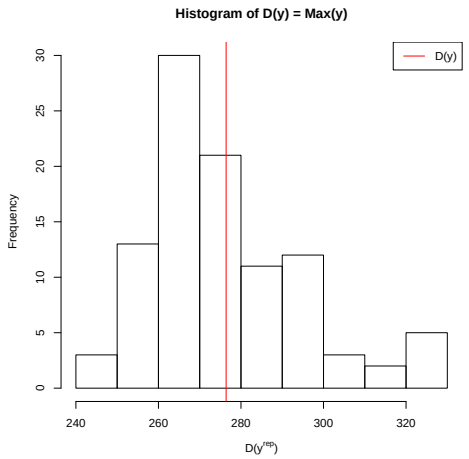


Figure 6.3. Using the maximum as discrepancy measure gives a p-value of 0.60.

6.1.2 Survival

The survival data are more complicated than the longitudinal data. For each subject, we had a pair of observations (v_i, δ_i). We needed to sample from

$$f({v_i}^{rep}|\boldsymbol{y}) = \int_\theta f({v_i}^{rep}|\boldsymbol{\beta}, \mu_i, \tau_i) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta},$$

where

$$f(v_i^{rep}|\boldsymbol{\beta}, \mu_i, \tau_i) = h(t|\mu_i, \tau_i) e^{-H(t|\mu_i, \tau_i)}. \tag{6.3}$$

Again, $\boldsymbol{y}$ is the set of all observations and we had samples of $\boldsymbol{\beta}^*, \mu_i^*$ and τ_i^* from the posterior. We could simply sample v^{rep*} from the exponential distribution with the appropriate hazard rate. Though, since the sampled values of v_i^{rep} have not been subject to censoring they cannot be directly compared to the observed values of v_i. We needed to choose a discrepancy measure that would enable us to effectively compare the replicated and observed data; when dealing with censored data, simple statistics such as $D(\boldsymbol{y}, \boldsymbol{\theta}) = \text{mean}(\boldsymbol{y})$ do not provide meaningful comparisons.

A method used in classical regression diagnostics to assess the fit of a Cox model on survival data is Cox-Snell residuals (Klein and Moeschberger, 2003). Introduced by Cox and Snell (1968), it is based on the fact that for a random event time T with survival function $S(t)$, the variable $Y = -\log(S(T))$ has an exponential distribution with hazard rate $h(t) = 1$ (Sun, 2006). Since the Cox-Snell residuals are a function of the observations as well as the parameters, they seemed a fitting subject for a discrepancy measure. Thus, if we use the hazard rate of our fitted model and calculate the Cox-Snell residuals $i = 1 \ldots N$ as

$$\begin{aligned} r_i = H(v_i|\mu_i, \tau_i) &= \int_0^{v_i} h(t|\mu_i, \tau_i) dt \\ &= v_i e^{\beta_0 + \beta_1 \mu_i + \beta_2 \tau_i} \end{aligned} \tag{6.4}$$

they should resemble a censored sample from the unit exponential distribution. Furthermore, to test how well the r_is corresponded to the unit exponential distribution, we could calculate the Nelson-Aalen estimate of their cumulative hazard function

$$\tilde{H}_{NA}(t) = \sum_{t_i \leq t} \frac{\delta_i}{n_{t_i}}$$

for observed times $t_1 \dots t_N$ with δ_i the number of events at t_i and n_{t_i} the total individuals at risk at t_i. This method is among those suggested by Aalen et al. (2008, p. 169). If the model fits the data well, a plot of this estimated hazard rate would reveal a slope that is equal or near one. A drawback of this method is that the Cox-Snell residuals can only measure the overall goodness of fit, they cannot reveal how the model differs from reality. If the slope of the estimated hazard rate is not equal or near one, we cannot use the Cox-Snell residuals to gain insights about how to change the model.

To this end, we let our discrepancy measure be the slope of the Nelson-Aalen estimated cumulative hazard rate. We calculated the Cox-Snell residuals using (6.4), then we calculated and plotted the Nelson-Aalen estimate of the cumulative hazard rate and, finally, we estimated the slope of this integrated hazard rate. Our discrepancy measure was the least squares estimator of the slope of $\tilde{H}_{NA}(t)$:

$$D(\boldsymbol{y}, \boldsymbol{\theta}) = \frac{\sum_{i=1}^{N} \tilde{H}_{NA}(r_i) r_i - \frac{1}{N} \sum_{i=1}^{N} r_i \sum_{i=1}^{N} \tilde{H}_{NA}(r_i)}{\sum_{i=1}^{N} r_i^2 - \frac{1}{N} \left(\sum_{i=1}^{N} r_i \right)^2}, \tag{6.5}$$

where r_i is the Cox-Snell residuals as in (6.4).

We simulated $k = 1 \dots 1000$ sets of $\boldsymbol{y}_k^{rep}$ from the posterior predictive distribution, and $\boldsymbol{\theta}_k$ from the posterior distribution of the parameters. This allowed us to calculate a thousand sets of Cox-Snell residuals, using the observed $\boldsymbol{y}$ and $\boldsymbol{\theta}_k$ to calculate each r_i in the kth set, and using the simulated $\boldsymbol{y}_k$ and $\boldsymbol{\theta}_k$ to calculate each r_i^{rep} in the kth set.

Each time, we used the Nelson-Aalen estimator to estimate the cumulative hazard rate function. The slopes of the Nelson-Aalen cumulative hazard rate calculated for the sets of r_is will be near one if the model fits well. Unfortunately, if the slope deviates from one, it does not reveal how the model is inadequate. We estimated the slopes for these cumulative hazard rates using (6.5), and compared the discrepancies using the scatterplot in figure 6.4, as it was done in Gelman et al. (1996) and Gelman (2003). The estimated p-value was 0.087, so we did not have reason to believe the generated values were significantly different from the observed values. Considering that this was a model with simple assumptions on both the longitudinal process and the hazard rate, it is fortunate that it is already adequate according to the posterior predictive measures.

There is room for improvement, however, since we ideally want the scatterplot in figure 6.4 to exhibit a similar amount of points above and below the 45° line. The

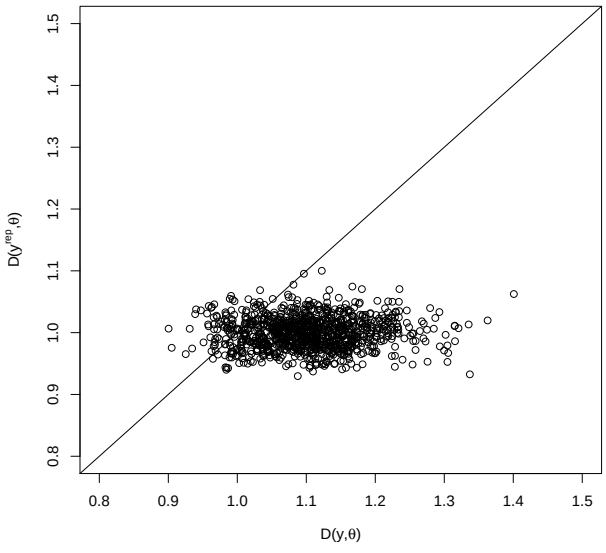


Figure 6.4. Posterior predictive check for the discrepancy variable $D(\boldsymbol{y}, \boldsymbol{\theta})$ as the slope of the estimated hazard rate, defined in equation (6.5). The scatterplot displays 1,000 simulations of $\{D(\boldsymbol{y}^{rep}, \boldsymbol{\theta}), D(\boldsymbol{y}, \boldsymbol{\theta})\}$, where $\boldsymbol{y}$ is the observed data and $\boldsymbol{y}^{rep}$, $\boldsymbol{\theta}$ are the simulated values. The p-value is estimated as the proportion of points above the 45° line.

values of $D(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ will always be around one, since we calculated them using the set of r_i^{rep}s, sampled from the exponential distribution with hazard rate $h(t|\mu_i, \tau_i)$. If the model fits well, the values of $D(\boldsymbol{y}, \boldsymbol{\theta})$ will also be spread around one. To illustrate this, we simulated survival data according to a joint survival model, and used the true model to estimate the parameters with our MCMC algorithm. We then used our posterior predictive model check to gauge the fit for the survival. Figures 6.5 to 6.7 are the scatterplots of $D(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ and $D(\boldsymbol{y}, \boldsymbol{\theta})$ for the simulated data, subjected to different amounts of censoring. When censoring is high (figure 6.5), $D(\boldsymbol{y}, \boldsymbol{\theta})$ has greater variance than when it is low (figure 6.7), but all of the values are still centred around one. Thus, when the model used for estimation is true, it is successfully identified as adequate by our chosen discrepancy measure. Furthermore, censoring increases the variance of our discrepancy measure, but it does not change the conclusion. We tested other discrepancy measures, but they were too sensitive to censoring.

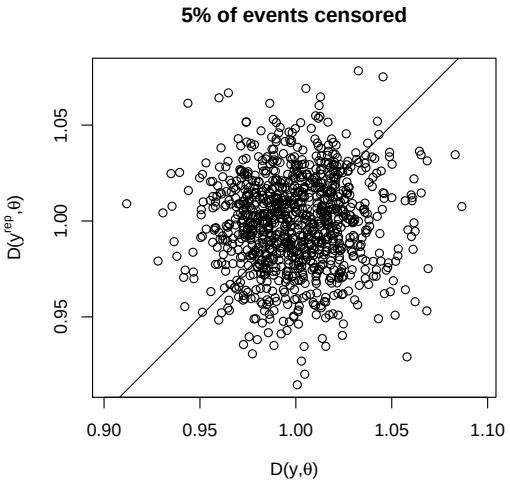


Figure 6.5. Posterior predictive check for the discrepancy in (6.5). True model used for estimation. The p-value was 0.5.

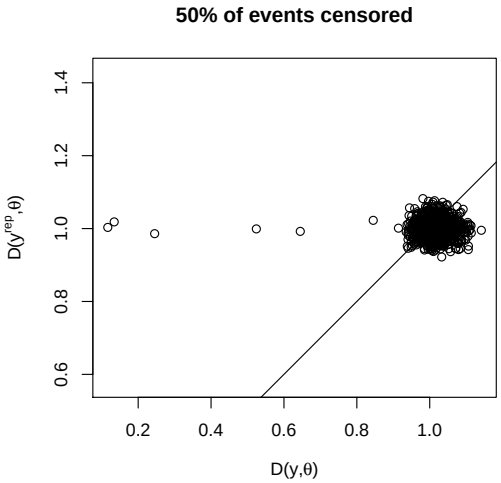


Figure 6.6. Posterior predictive check for the discrepancy in (6.5). True model used for estimation. The p-value was 0.35.

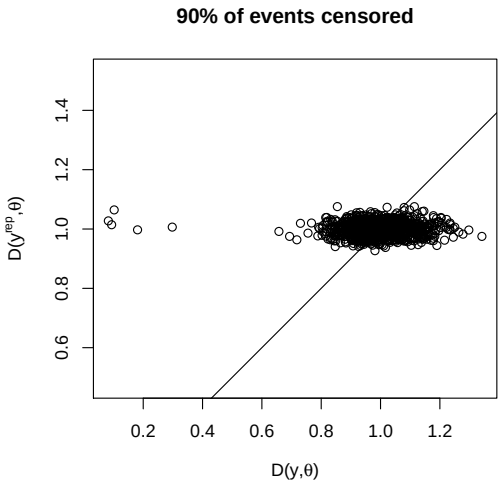


Figure 6.7. Posterior predictive check for the discrepancy in (6.5). True model used for estimation. The p-value was 0.49.

6.2 Regression dilution

Rothwell (2010), on the limitations of the usual blood pressure hypothesis, stated that adjustment for regression dilution can be used to correct risk relations between measurements of blood pressure and risk of vascular events for inaccuracy in estimation of usual blood pressure. The author gave MacMahon et al. (1990) as a reference for regression dilution, but this article only mentioned that regression dilution can be done either parametrically or non-parametrically, without giving exact methodology. In a different article also on the topic of blood pressure volatility, Howard and Rothwell (2003) also investigated regression dilution and gave Clarke et al. (1999) as a reference for calculating the Regression Dilution Ratio (RDR). The results table on p. 4 of Rothwell (2010) mentioned that they used the RDR as calculated 'from the baseline measurement and the visit 7 (2-year) measurement.' It corresponds with the method provided in Clarke et al. (1999, p. 2) for calculating the RDR non-parametrically, so we assumed it was the method used to calculate regression dilution in Rothwell (2010):

> With the non-parametric method of estimating and correcting for the regression dilution associated with a particular exposure period, pairs of measurements of the relevant risk factor that are separated by an appropriate interval ... are subdivided into a few groups according to the value of just the first measurement in each pair. The range of the initial values (r_i) is defined as the difference between the means of these first measurements in the groups with the lowest and highest values. Even though the value of the second measurement in each pair did not determine in which group that pair belonged, the means of the second measurements provide unbiased estimates of the 'usual' levels of the risk factor in each group during the particular exposure period, against which the disease rates can be compared. Generally, the range of these mean usual values (r_u) will be substantially narrower than the range of the initial values (r_i), so the ratio of these two ranges ($R = r_u/r_i$) will be substantially less than 1.

This is a crude method of correcting for the effects of covariate measurement error. Additionally, it only provided a method for correcting error associated with the 'usual

blood pressure level' — understood as the mean blood pressure. Since the variation is harder to measure, especially with only a few measurements, we will have to employ other methods. Rothwell (2010) realised this: 'Measurements of variability in blood pressure are generally less precise than are estimates of usual blood pressure, and risk relations could in theory be adjusted for error in estimation of usual variability.' No suggestion was given, however, of how to make this adjustment.

This absence of a canonical method to correct for the measurement error associated with the variance is possibly the reason why Rothwell et al. (2010) did not attempt any form of correction. Their paper applied a Cox proportional hazards model using mean and SD of SBP as covariates, attempting to assess the influence of blood pressure variation on cardio-vascular risk. In Rothwell (2010) this influence was not measured using covariates, as the author only looked at how higher mean blood pressure affected stroke risk for two groups of patients, namely low visit-to-visit variability and high visit-to-visit variability. Blood pressure measurements were adjusted using the RDR, and parameter estimates were estimated for both adjusted and unadjusted blood pressure measurements.

6.2.1 Covariate measurement error

With this in mind, we attempted to calculate a rough estimate of the bias experienced in Rothwell et al. (2010). Prentice (1982) showed that the naïve approach of using the observed values as the true covariates does not lead to convergence of parameter estimates to the true values, even when the covariate estimates are themselves unbiased. Covariate measurement error in survival is was further discussed by Raboud et al. (1993), Hughes (1993), and Hu et al. (1998), with an overview given by Steinsaltz et al. (2012).

Hughes (1993) gave a method for correction, which is appropriate in the proportional hazards model when the event rate is low. Assuming that we want to use a

covariate W in the proportional hazards model, but that we can only observe the covariate through a marker X prone to measurement error ε, as

$$X = W + \varepsilon,$$

let β^* be the coefficient associated with using the true covariate in the model, and let $\hat{\beta}$ be the naïve estimate of the coefficient estimated when using the values of X in the model. The bias resulting from measurement error can be calculated as

$$\frac{\beta^*}{\hat{\beta}} = (1 + \lambda), \tag{6.6}$$

where $\lambda = {}^{\mathrm{Var}(\varepsilon)}/_{\mathrm{Var}(W)}$ and $\mathrm{Var}(\varepsilon)$ is the error variance and $\mathrm{Var}(W)$ is the underlying variance of W in the sample.

The most informative description of the statistical methodology used in Rothwell et al. (2010), is the caption from table, reproduced in table 1: 'Hazard ratios (top vs bottom quintile) for risk of subsequent stroke (ie, after the measurement period) in the UK-TIA trial from a model combining mean SBP and visit-to-visit variability in SBP (SD or CV or VIM), repeated with increasingly precise estimates of both variables.' They do not give the exact specification for the hazard rate in their Cox model, but we assumed that they used:

$$h(t) = h_0(t) \exp\left(\beta_{mean} MUgroup_i + \beta_{SD} SDgroup_i\right), \tag{6.7}$$

where $MUgroup_i$ and $SDgroup_i$ are the quintile numbers of the ith patient's SBP mean and standard deviation respectively, and $h_0(t)$ is the baseline hazard rate. Then, they compared the hazard ratios of the top and bottom quintiles. This model is the most likely candidate, given the format of results in their paper.

In pursuit of mathematical simplicity we will look at the estimation bias that

would occur in a Cox model that uses one of the following hazard rates, similar to that of Rothwell et al. (2010):

$$h(t) = h_0(t) \exp(\beta_{mean} X_{i1}) \tag{6.8}$$

or

$$h(t) = h_0(t) \exp(\beta_{SD} X_{i2}), \tag{6.9}$$

where X_{i1} and X_{i2} are the estimates of the sample mean and standard deviation of the ith individual's SBP respectively, for $i = 1 \dots N$, for a dataset with N individuals. To use (6.6), we have to consider models containing only one covariate, since Hu et al. (1998) and Hughes (1993) did not investigate models with multiple, possibly dependent, covariates measured with error. The covariates X_{i1} and X_{i2} are calculated as

$$X_{i1} = \bar{Y_i} = \frac{1}{n_i} \sum_{i=1}^{n_i} Y_{ij} \tag{6.10}$$

and

$$X_{i2} = s_i = \sqrt{\frac{1}{n_i - 1} \sum_{i=1}^{n_i} (Y_{ij} - \bar{Y_i})^2} \tag{6.11}$$

where Y_{ij} for $j = 1 \dots n_i$ are the observed SBP values of the ith individual. As before, assume $Y_{ij} \sim N(\mu_i, \sigma_i^2)$.

We further assume that X_{i1} and X_{i2} are i.i.d. random variables and that they give insight to the true mean and variance of the underlying process respectively, but are measured with error

$$\begin{aligned} X_{i1} &= \mu_i + \varepsilon_{i1} \\ X_{i2} &= \sigma_i + \varepsilon_{i2}. \end{aligned} \tag{6.12}$$

This is similar to Hu et al. (1998) who assumed that the conditional density $f_{X|W}(X|W)$

is known, for example, with an additive error model $X_i = W_i + \varepsilon_i$ and ε_i normal with mean zero and known variance $\text{Var}(\varepsilon_i)$. The assumption states that the observed covariate has a distribution around the true underlying value, and it allows us to make use of the result in Hughes (1993) to estimate the arising bias when the covariate measurement error is ignored.

We can use (6.10) and (6.11) to calculate the overall variance of X_{i1} and X_{i2}, assuming that μ_i and σ_i each follow some distribution with existing first and second moments. We have

$$X_{i1} = \bar{Y}_i \sim N(\mu_i, \frac{\sigma_i^2}{n_i}) \tag{6.13}$$

$$X_{i2} = s_i \sim \frac{\sigma}{\sqrt{n-1}} \cdot \chi_{n-1} \tag{6.14}$$

and then we know that

$$s_i^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}, \tag{6.15}$$

where μ_i and σ_i are also random variables in our model. Using the law of total variance, we can calculate

$$\begin{aligned}
\text{Var}(X_{i1}) &= \text{E}_{\sigma_i^2}\Big[\text{Var}_{X_{i1}|\sigma_i^2,\mu_i}(X_{i1}|\sigma_i^2,\mu_i)\Big] \\
&\quad + \text{E}_{\mu_i}\Big(\text{Var}_{\sigma_i^2|\mu_i}\Big[\text{E}_{X_{i1}|\sigma_i^2,\mu_i}(X_{i1}|\sigma_i^2,\mu_i)\Big]\Big) \\
&\quad + \text{Var}_{\mu_i}\big(\text{E}_{X_{i1}|\mu_i}(X_{i1}|\mu_i)\big) \\
&= \text{E}_{\sigma_i^2}\left[\frac{\sigma_i^2}{n_i}\right] + \text{E}_{\mu_i}\Big[\text{Var}_{\sigma_i^2|\mu_i}(\,\mu_i\,|\mu_i)\Big] + \text{Var}_{\mu_i}(\mu_i) \\
&= \text{E}_{\sigma_i^2}\left[\frac{\sigma_i^2}{n_i}\right] + \text{Var}_{\mu_i}(\mu_i)
\end{aligned} \tag{6.16}$$

and since we have from (6.12) that

$$\text{Var}(X_{i1}) = \text{Var}(\mu_i) + \text{Var}(\varepsilon_{i1}),$$

we conclude $\mathrm{Var}(\varepsilon_{i1}) = \frac{1}{n_i}\,\mathrm{E}\,(\sigma_i^2)$. See appendix D for details on the variance decomposition, as well as a step-by-step explanation of (6.16). Similarly, using he law of total expectation, for X_{i2} we have

$$\begin{aligned}
\mathrm{Var}(X_{i2}) &= \mathrm{E}[\mathrm{Var}(X_{i2}|\sigma_i)] + \mathrm{Var}[\mathrm{E}(X_{i2}|\sigma_i)] \\
&= \mathrm{E}\left[\frac{\sigma_i^2}{n_i-1}\left((n_i-1) - 2\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)^2\right)\right] \\
&\quad + \mathrm{Var}\left[\frac{\sigma_i}{\sqrt{n_i-1}}\sqrt{2}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)\right] \\
&= \mathrm{E}\big[\sigma_i^2\big]\left(1 - \frac{2}{n_i-1}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)^2\right) \\
&\quad + \mathrm{Var}[\sigma_i]\left(\frac{2}{n_i-1}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)^2\right).
\end{aligned} \tag{6.17}$$

Combining this with

$$\mathrm{Var}(X_{i2}) = \mathrm{Var}(\sigma_i) + \mathrm{Var}(\varepsilon_{i2}),$$

which results from (6.12), we conclude

$$\mathrm{Var}(\varepsilon_{i2}) = \left(\mathrm{E}\big[\sigma_i^2\big] - \mathrm{Var}[\sigma_i]\right)\left(1 - \frac{2}{n_i-1}\left(\frac{\Gamma(n_i/2)}{\Gamma(n_i-1/2)}\right)^2\right)$$

where Γ is the Gamma function. See appendix D for a step-by-step explanation of (6.17).

With this in mind, we can quantify the bias for a model using the mean blood pressure as a covariate

$$\begin{aligned}
\frac{\beta^*_{mean}}{\hat{\beta}_{mean}} &= 1 + \frac{\mathrm{Var}(\varepsilon_{i1})}{\mathrm{Var}(\mu_i)} \\
&= 1 + \frac{1}{n_i}\frac{\mathrm{E}\,(\sigma_i^2)}{\mathrm{Var}(\mu_i)}.
\end{aligned} \tag{6.18}$$

Similarly, a model using the measured variance of the blood pressure as the covariate

will have a bias of

$$\begin{aligned}\frac{\beta^*_{var}}{\hat{\beta}_{var}} &= 1 + \frac{\mathrm{Var}(\varepsilon_{i2})}{\mathrm{Var}(\sigma_i^2)} \\ &= 1 + \left(1 - \frac{2}{n_i - 1}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)^2\right)\frac{\mathrm{E}[\sigma_i^2] - \mathrm{Var}[\sigma_i]}{\mathrm{Var}(\sigma_i)}. \qquad (6.19)\end{aligned}$$

Rothwell et al. (2010) made use of multiple models, each selecting individuals based on the number of longitudinal observations they provided. It was clear that selecting individuals with more longitudinal observations provided more accuracy within-subject, but it also meant having less data in the model. There was a trade-off between including more data in the model and including more longitudinal observations per subject. In the above expressions, n_i will be this chosen number of observations in the applicable model.

The exact bias is difficult to calculate, since μ_i and σ_i are the quantities that are not directly observed. We can assume that the parameters follow the posterior distributions from our Bayesian model to gain some insight into the bias. Using samples from the posterior predictive distributions, we get the following:

$$\begin{aligned}\mathrm{E}(\sigma_i^2|\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{\delta}) &= 238.315 \\ \mathrm{Var}(\sigma_i|\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{\delta}) &= 22.26212 \\ \mathrm{Var}(\sigma_i^2|\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{\delta}) &= 31881.05 \\ \mathrm{Var}(\mu_i|\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{\delta}) &= 289.0917.\end{aligned}$$

This allowed us to calculate bias in the proportional hazards models of Rothwell et al. (2010), given in table 6.1. In studies where repeated measurements are available for the covariates, it might be possible to calculate estimates for $\mathrm{Var}(X)$ and $\mathrm{Var}(\varepsilon)$, in turn providing insights to the moments of W — the hidden, true covariate — and

allowing us to calculate the bias using (6.18) and (6.19) exactly.[1]

Rothwell et al. (2010) used categorical covariates, as in (6.7), rather than the observed continuous covariates. Norris et al. (2008) used a similar method, wherein they divided responses into quartiles and made use of the arising categorical variables, rather than using the actual values. We tried to replicate the analyses given in Rothwell et al. (2010), using a Cox proportional hazards model with the hazard rate as specified in (6.7) on a similar set of the UK-TIA data, with the results given in table 6.2. We also adapted our joint Baysian model to use the categorical covariates, rather than the continuous values of the mean and the standard deviation. This allowed us to fit our Bayesian model with a hazard rate similar to (6.7) for comparison. Results are given in table 6.3.

We compared the results from the classical and Bayesian models in figure 6.8. We do not know what the true values should be, but we see from the confidence intervals that the Bayesian model's results are more stable, and are nearly similar for 2,4,6,8, and 10 observations. Although our confidence intervals for the hazard ratios associated with the volatility measures are not as wide as those shown in Rothwell et al. (2010), they do illustrate the same pattern. The approach of dividing covariates into categories has the drawback of discarding a lot of information. It also makes

[1]If we were worried about the strict assumptions of our Bayesian model, we could instead use bootstrap methods to calculate the mean and the variance of μ_i or σ_i.

	Bias when using:	
n_i	Mean	Var
2	1.41	4.89
4	1.21	2.62
6	1.14	2.01
8	1.10	1.74
10	1.08	1.58

Table 6.1. The bias experienced during parameter estimation, $^{\beta^*}/_{\hat{\beta}}$ where β^* is the parameter estimated using the true covariates and $\hat{\beta}$ is the estimate obtained using the covariates observed with error. The number of observations provided per individual is n_i.

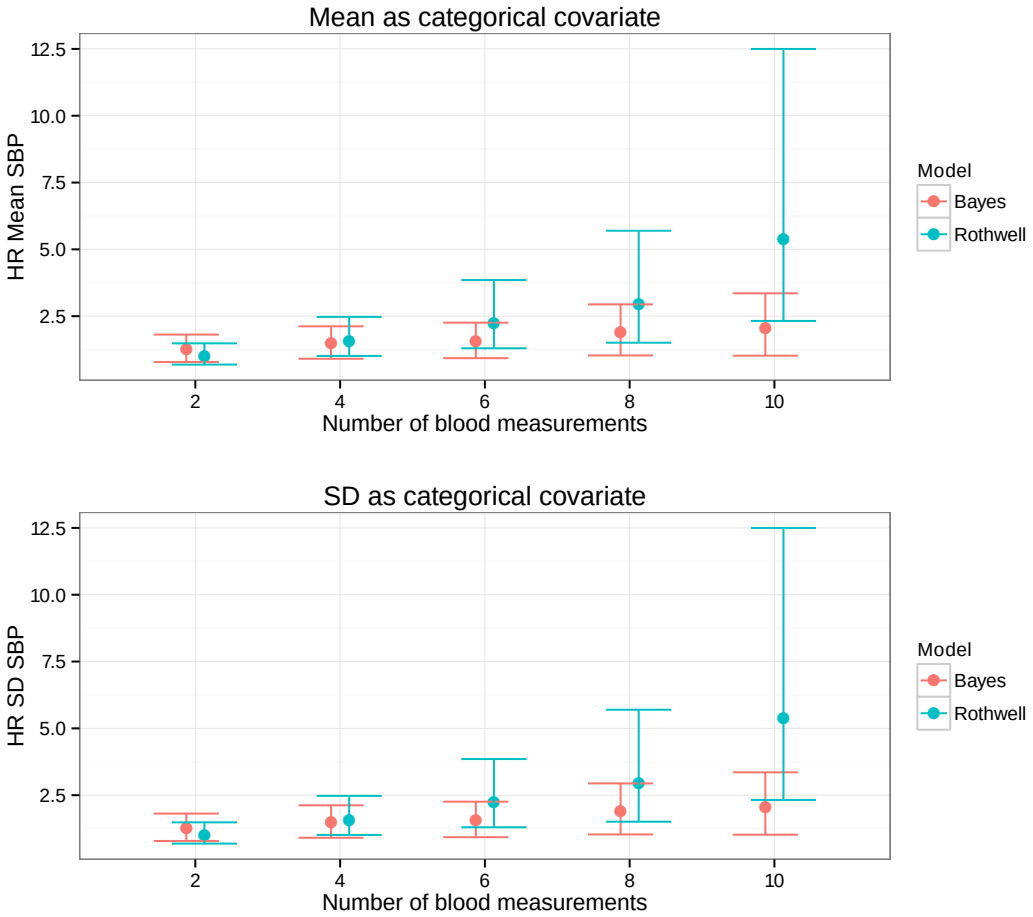


Figure 6.8. Comparison of joint Bayesian and classical model for the UK-TIA data. We used our interpretation of Rothwell et al. (2010) for the classical model. The error bars represent 95% confidence intervals for the classical model, and 95% HPD intervals for the Bayesian model.

Using SD SBP

	Mean SBP			SD SBP		
n	HR	(95% CI)	p-val	HR	(95% CI)	p-val
2	2.86	(1.89 ; 4.31)	<0.0000	1.05	(0.72 ; 1.54)	0.8048
4	2.13	(1.38 ; 3.28)	0.0006	1.61	(1.04 ; 2.49)	0.0325
6	1.74	(1.02 ; 2.96)	0.0419	2.68	(1.56 ; 4.61)	0.0004
8	1.60	(0.83 ; 3.06)	0.1582	2.96	(1.52 ; 5.77)	0.0015
10	1.68	(0.75 ; 3.73)	0.2065	5.38	(2.28 ; 12.66)	0.0001

Using CV SBP

	Mean SBP			CV SBP		
n	HR	(95% CI)	p-val	HR	(95% CI)	p-val
2	2.88	(1.92 ; 4.31)	<0.0001	1.04	(0.72 ; 1.51)	0.8337
4	2.33	(1.55 ; 3.52)	0.0001	1.51	(1.00 ; 2.27)	0.0492
6	2.24	(1.37 ; 3.67)	0.0014	2.25	(1.37 ; 3.69)	0.0014
8	2.13	(1.19 ; 3.84)	0.0116	2.57	(1.41 ; 4.67)	0.0021
10	2.81	(1.37 ; 5.76)	0.0048	3.94	(1.87 ; 8.28)	0.0003

Table 6.2. Our attempt at replicating Rothwell's table — given on p. 3 of this thesis. Hazard ratios (top vs bottom quintile) for risk of subsequent stroke in the UK-TIA trial from a model combining mean SBP and visit-to-visit variability in SBP, using (6.7). Results given for SD and CV.

it hard to quantify the resulting bias. The Bayesian model we developed can use continuous values, thus providing a way to use all the available information, while at the same time taking the measurement error into account.

The idea of unbiasedness can be potentially misleading in a Bayesian framework, especially in hierarchical models. As Gelman et al. (2003, c. 4) explain, it is 'often not possible to estimate several parameters at once in an even approximately unbiased manner.' Therefore, we should not regard our Bayesian model as a way of correcting

	Mean SBP		SD SBP	
n	HR	(95% HPD)	HR	(95% HPD)
2	2.30	(1.46 ; 3.21)	1.27	(0.79 ; 1.81)
4	2.16	(1.34 ; 3.08)	1.48	(0.91 ; 2.12)
6	2.23	(1.27 ; 3.18)	1.55	(0.93 ; 2.26)
8	2.45	(1.30 ; 3.72)	1.89	(1.03 ; 2.94)
10	1.93	(0.90 ; 3.04)	2.07	(1.02 ; 3.36)

Table 6.3. Hazard ratios (top vs bottom quintile) for risk of subsequent stroke in the simulated dataset, using the joint Bayesian model and categorical covariates.

the bias, but rather as a different strategy towards investigating the problem.

6.2.2 Simulation study

To further investigate the effects of ignoring measurement error by using empirical estimates of the mean and standard deviation, we performed a simulation study. We simulated 10,000 datasets using the assumptions in (4.1) and (4.2), with parameter values:

$$\begin{aligned} \mu_i &\sim N(m = 145.3, \tau^{-1} = 0.0035^{-1}) \\ \tau_i &\sim \Gamma(r = 3.259, k = 547.7) \\ \beta_0 &= -4.85 \\ \beta_1 &= 1.302 \\ \beta_2 &= 0.085. \end{aligned} \tag{6.20}$$

These values were inspired by those estimated in the SBP model in chapter 4, since we attempted to simulate datasets similar to UK-TIA. We used the hazard rate

$$h(t|\mu_i, \tau_i) = \exp\left\{\beta_0 + \beta_1 \frac{\mu_i}{100} + \beta_2 (\tau_i)^{-\frac{1}{2}}\right\}.$$

which uses the standard deviation rather than the precision, and therefore we used a β_2 parameter to reflect a similar relationship between risk and variability as the model estimated in chapter 4. Then, we estimated the parameters using a Cox model with the hazard rate

$$h(t) = h_0(t) \exp\left\{\beta_1 \frac{MEAN_i}{100} + \beta_2 SD_i\right\}, \tag{6.21}$$

where $MEAN_i$ and SD_i were empirical estimates of the mean and SD of the generated $Y_i(t)$ values. Despite using a model that included both the mean and the variance — as opposed to the simple hazard rates in (6.8) and (6.9) — the bias we gave in table

6.1 still applied.

Figure 6.9 shows how the parameter estimates differ from the actual values used to produce the simulations, both before and after correcting for the bias, using the values in table 6.1. We can successfully correct the bias using these values. The β_2 parameter associated with the standard deviation of the SBP — our measure for volatility — is more adversely affected by the estimation bias than the β_1 associated with the mean SBP.[2]

Bringing our simulation study closer to the model that appeared in Rothwell et al. (2010), we used the categorical covariates in the Cox proportional hazards model, rather than the continuous measurements. Again, we simulated 10,000 datasets using the parameters in (6.20) and used the hazard rate given in (6.7) to fit the model used in Rothwell et al. (2010) on each simulated dataset. As before, we estimated the hazard ratio between the top and bottom quintile for each covariate. The distributions of the estimated hazard ratios obtained are given in figure 6.10. To get a sense of what to expect, we can compare the risk of a subject in the top quintile (90th, percentile) with a subject in the bottom quintile (10th percentile). The person in the top quintile has a hazard ratio of 1.75 and 2.41 for the mean and standard deviation respectively, when compared to the person in the bottom quintile. We also corrected the HR estimates for each simulation for bias using the values in table 6.1; these are given in figure 6.10. The bias correction overcompensates for the underestimation of the hazard ratio, and we do not observe a bias as large as table 6.1 would have us believe. The true correction factors for models using categorical covariates will be smaller than the values in table 6.1, but they are harder to calculate.

Using categorical covariates seemed to slightly alleviate bias — it was not as big as the expected bias for the continuous covariates, but this approach discards information. Moreover, grouping the individuals according to their number of measurements further

[2]The intercept, β_0, is absorbed into the baseline hazard rate of the Cox model, which is why it does not form part of our output.

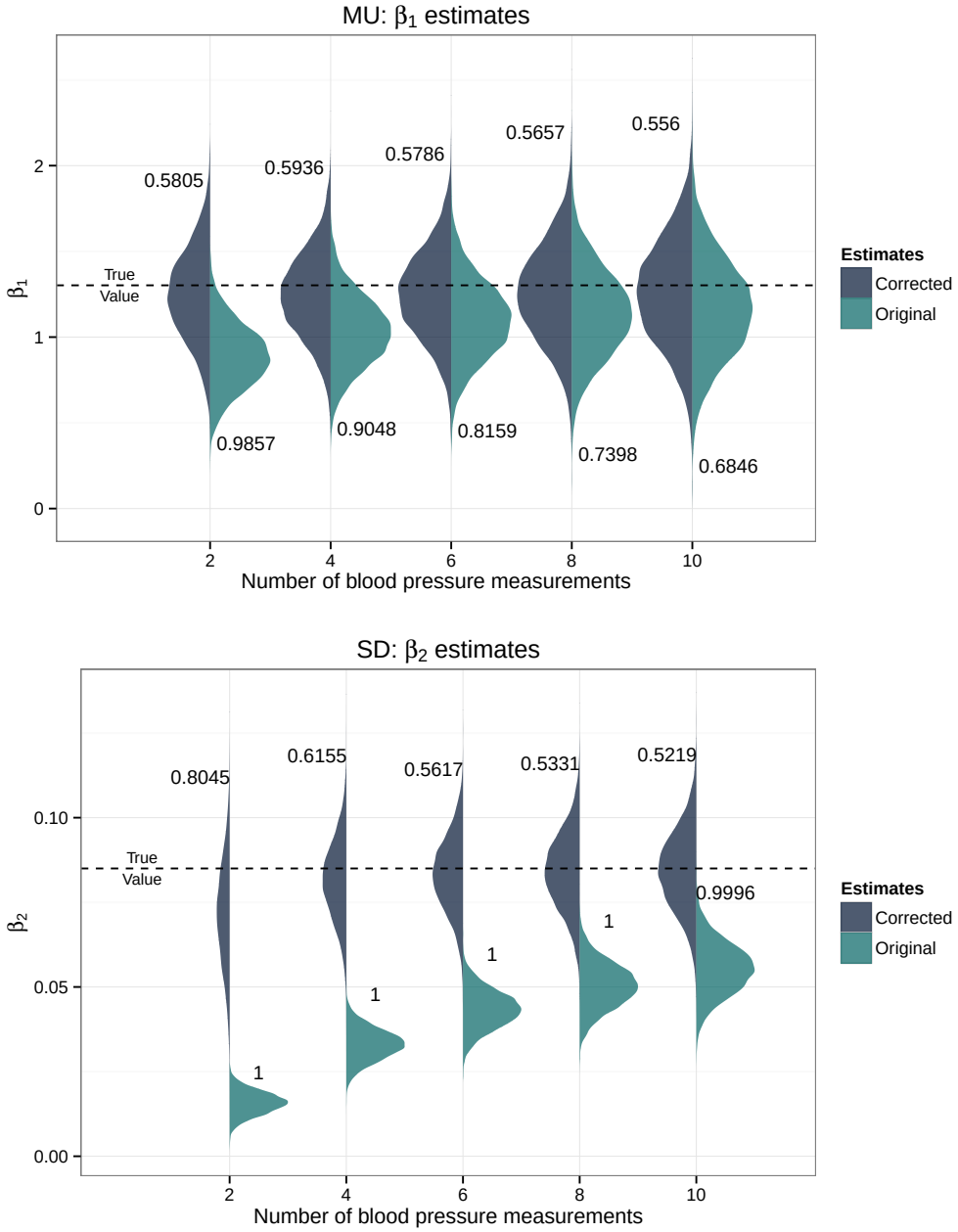


Figure 6.9. Fitting a classical model similar to Rothwell et al. (2010) on the simulated dataset, but using continuous covariates . The bias corrections were done using the values in table 6.1. The numbers show the proportion of estimates below the true value, for 10,000 simulation-estimation routines.

erodes the information available for each estimate. Using categorical variables gives a small probability of making a big mistake. Using continuous variables gives a larger probability of making a small mistake. Our joint Bayesian model provides a way to use all the information at once, while taking the measurement error into account, and

it allows us to include our assumptions about the underlying dynamics more explicitly.

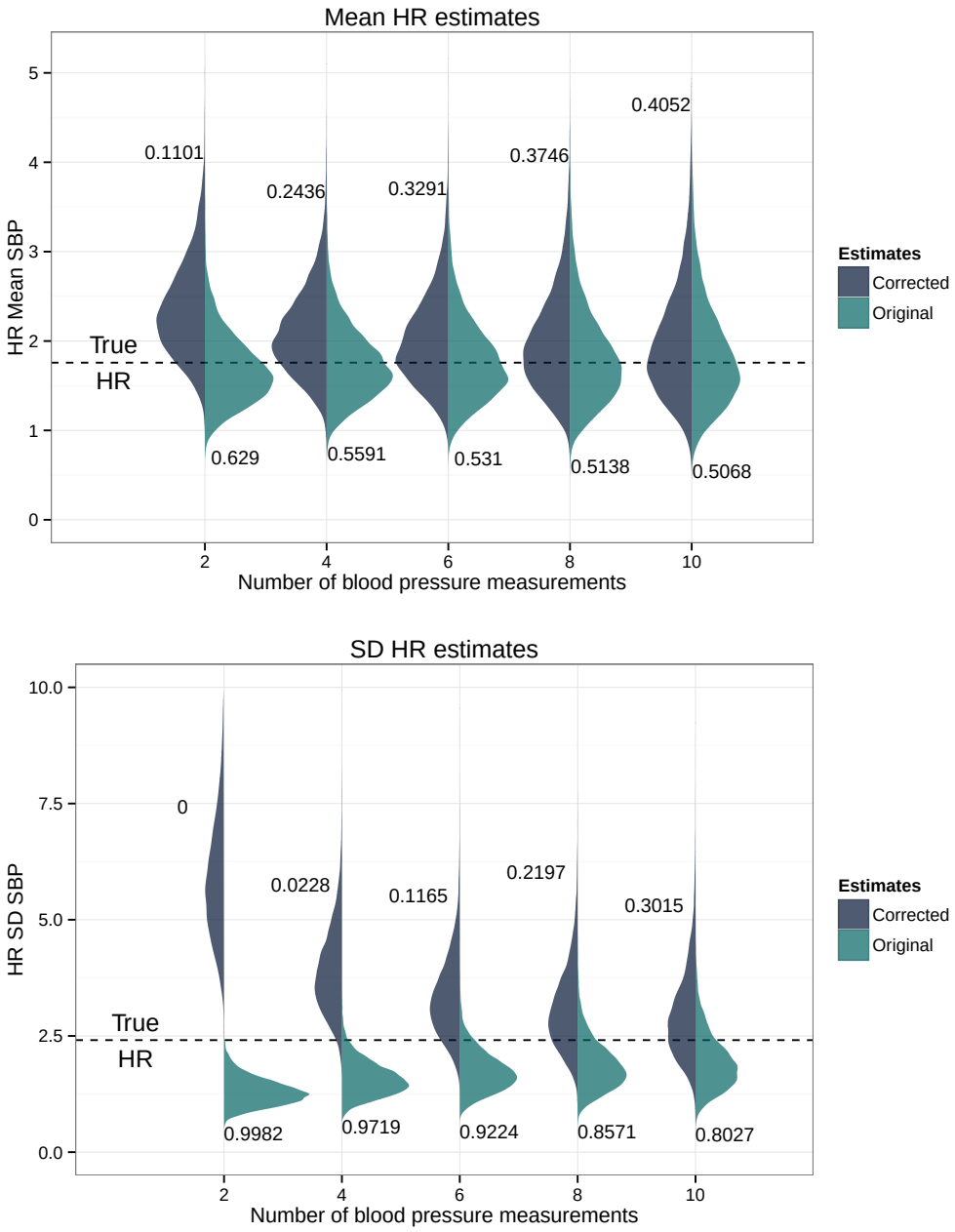


Figure 6.10. Hazard rate estimates for 10,000 simulations, using the Rothwell et al. (2010) model that compares the top and bottom quintiles. Dotted lines are the true hazard ratios for someone in the middle of the top quintile vs someone in the middle of the bottom quintile. The numbers on the plot show the proportion of the estimates that fell below the dotted line.

Chapter 7

Extending the model

Following the successful implementation of our model, we reached out to other researchers, hoping to find similar data and verify our findings. This led to collaboration with a group of researchers from Stanford university, working on analysis of the NHANES dataset.

Despite there only being three measurements for each individual taken at the outset of the study, we could still use this to measure the effect of blood pressure mean and volatility and relate it to event risk.[1] Due to the information available about the race, sex, and age at entry of study participants, we opted to extend our model to allow for different baseline hazard rates. We describe this extension in this chapter, as well as considering methods to allow for multivariate longitudinal processes, in order to measure the effect of SBP and DBP together.

Other papers that have featured multivariate longitudinal processes include Brown et al. (2005), who used a cubic B-spline model for the longitudinal process, along with proportional hazards for the survival, and Chi and Ibrahim (2006), who specified both the longitudinal and survival parts of their model as multivariate. Both of these papers used MCMC techniques to estimate their parameters.

[1] We later also gained access to 3 additional measurements for each subject, taken at the subject's home.

7.1 Competing risks

The first addition to our model from chapter 4 is the ability to handle multiple events. This is related to the topic of censoring. For example, if we have information about other events — different ways of exiting the study — and we believe them to be related to the covariates we are studying, we have a violation of the random-censoring assumption. In other words, if we are only interested in heart events, we could regard events due to cancer or accidents as a form of censoring. We need to be able to include these events in our model, at the very least in order to ascertain whether they are in fact independent of the covariates we wish to link to cardio-vascular mortality.

In the case with competing risks, the arising likelihood is an extension of (2.3). Consider a scenario with p different event types, or ways of exiting the study. The observed data for subject i will be $\{v_i, \delta_{i1}, \delta_{i2}, \ldots, \delta_{ip}\}$, where v_i is the observed event time, and δ_{ij} is the indicator variable for the jth event type, each taking thec value 0 or 1. Assuming each subject can only exit the study once, all but one of the δ_{ij}s will be zero. Finally, assume each type of event has its own hazard rate $h_j(t)$. Within this scenario, the ith individual contributes a quantity of

$$\begin{aligned} &h_1(v_i)^{\delta_{i1}} e^{-H_1(v_i)} h_2(v_i)^{\delta_{i2}} e^{-H_2(v_i)} h_3(v_i)^{\delta_{i3}} e^{-H_3(v_i)} \ldots h_p(v_i)^{\delta_{ip}} e^{-H_p(v_i)} \\ &= \prod_{j=1}^{p} \left[h_j(v_i)^{\delta_{ij}} e^{-H_j(v_i)} \right] \end{aligned} \tag{7.1}$$

to the likelihood. The total likelihood becomes

$$P(\boldsymbol{v}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_p) = \left[\prod_{i=1}^{N} \prod_{j=1}^{p} h_j(v_i)^{\delta_{ij}} \right] \exp\left(- \sum_{i=1}^{N} \sum_{j=1}^{p} H(v_i) \right), \tag{7.2}$$

which is a generalisation of (2.3). Since only the survival part of the likelihood is updated, competing risks can easily be added to a working MCMC algorithm.

7.2 Time-dependent baseline hazard rate

The UK-TIA data did not contain individual covariates, such as race or gender, nor did it contain information about the subjects' age at outset of the study. Although we could argue that the selection procedures for those data ensured that individuals in the study were fairly homogeneous, the same cannot be said for the NHANES dataset. Due to the heterogeneous nature of the data, specifically the different ages at outset of the study, we needed to discuss how baseline hazard rate can be included in our earlier model.

We mentioned in section 2.1.4 that the Cox model uses the partial likelihood, which eliminates the need to estimate the baseline hazard rate. This was useful for our model, since we were not directly interested in the shape of the baseline hazard rate. The partial likelihood method essentially compares all individuals at risk at each event time. In an MCMC setting, however, this does not work. Comparing all the individuals at risk is an excellent method that allows us to ignore the shape of the baseline hazard rate, but it comes at a computational cost. If we were to use the partial likelihood in our joint model, the survival part of the likelihood would become

$$\prod_{i=1}^{N} \left(\frac{\exp(\boldsymbol{x}'_{(i)}\boldsymbol{\beta})}{\sum_{j \in R(v_i)} \exp(\boldsymbol{x}'_{(j)}\boldsymbol{\beta})} \right)^{\delta_i} \tag{7.3}$$

as opposed to

$$\prod_{i=1}^{N} [h(v_i)]^{\delta_i} \exp\left(- \int_0^{v_i} h(t)dt \right), \tag{7.4}$$

where $h(t)$ is the entire hazard rate, including the baseline hazard. The hazard rate in (7.3) is only comprised of the individual level covariates, which will include the longitudinal parameters in the case of joint survival models. The problem lies in the denominator of (7.3). The parameters in the hazard rate that link the longitudinal and survival processes will usually be sampled using techniques that explore the density,

such as ARMS or the Slice sampler. Thus, at each iteration, we have to evaluate either (7.3) or (7.4) multiple times. The sum in the denominator of (7.3) presents a large computational burden. It scales as N, the number of individuals in the model, gets larger, since $R(t)$, the number of individuals at risk at time t, will be large.

The computational burden is a problem, and it is exacerbated in models with individual level parameters appearing in the hazard rate, as the μ_i and τ_i parameters in our model from chapter 4. At each iteration of the Gibbs sampler, we will have to sample new updates for μ_i and τ_i, for $i \in 1 \dots N$, and each of those will require multiple evaluations of (7.3). We will furthermore need thousands of MCMC iterations, in addition to burn-in. Using the partial likelihood proved to be impractical for our model and we made the process of parameter estimation more manageable by assuming a parametric form for the baseline hazard rate.[2]

One approach to model the baseline hazard rate is to assume a non-parametric model, resulting in a semi-parametric survival model. The name divulges that the model has a parametric part, usually the part with the proportional hazard parameters in the Cox model, and a non-parametric part, which models the baseline hazard rate. Recall the Cox model hazard rate (2.5):

$$h(t) = h_0(t) \exp(\boldsymbol{X}\boldsymbol{\beta}).$$

The semi-parametric model involves discretising $h_0(t)$, with intervals at all the event times. Thus, $h_0(t)$ becomes a non-negative polygonal function with vertices at the event times $t_1 < t_2 < \ldots < t_D < t_{D+1}$ where D is the number of events. The function takes the values $h_0 = 0 < h_1 < h_2 < \ldots < h_D < h_{D+1}$, which need to be determined. In a Bayesian model, we can estimate them along with the other parameters by assuming a prior structure, as shown by Mostafa and Ghorbal (2011). Ibrahim et al. (2005, chap. 3.1) mention that this semi-parametric model is sometimes referred to as

[2]Some single iterations of the MCMC algorithm took more than 30 minutes.

a *piecewise exponential model*, and they show the likelihood construction and options for the prior structure.

Our chosen approach was a baseline hazard rate according to the Gompertz law, with the form similar to (2.2), specified as

$$h_0(t) = Be^{\theta t}$$

where t is the subject's age, or

$$h_0(t) = Be^{\theta(x+t)}$$

where t is the time in the study and x is the age at outset. Our specification furthermore allowed for different values of B and θ among gender and different race-groups.[3] These parameters were estimated along with the rest of the parameters in the model, using our MCMC algorithm. We used the Gompertz law for the baseline hazard rate to model the NHANES dataset, in a model that also accounted for the location the measurements were taken. For each patient, we had three measurements taken at home, and three taken at the clinic. Let Y_{Hij} be the jth blood pressure measurement taken at home for patient i, and Y_{Cij} be the jth blood pressure measurement taken at the clinic. Then we assumed

$$Y_{Cij} \sim N(M_i - \Delta_i, \tau_{Ci}^{-1})$$
$$Y_{Hij} \sim N(M_i + \Delta_i, \tau_{Hi}^{-1})$$

[3]For instance, if we have three race-groups and two genders, we will have $2 \times 3 = 6$ groups leading to 12 more parameters for our model.

We further assumed distributions on these, treating them as random effects:

$$M_i \sim N(m_1, \tau_1)$$
$$\Delta_i \sim N(m_2, \tau_2)$$
$$\tau_{Ci} \sim \Gamma(r_1, \lambda_1)$$
$$\tau_{Hi} \sim \Gamma(r_2, \lambda_2)$$

This led to the following priors that were easily added to our MCMC algorithm:

$$M_i - \Delta_i \sim N\left(m_1 - m_2, \frac{1}{\tau_1} + \frac{1}{\tau_2}\right)$$
$$M_i + \Delta_i \sim N\left(m_1 + m_2, \frac{1}{\tau_1} + \frac{1}{\tau_2}\right).$$

We used the hazard rate:

$$h(t|\boldsymbol{\theta}_i) = B_k e^{\theta_k(x+t)} \exp\left[\beta_0|\Delta_i| + \beta_1 M_i + \beta_2(\sigma_{Ci} - c) + \beta_3(\sigma_{Hi} - c)\right], \tag{7.5}$$

where $\boldsymbol{\theta}_i$ is the set of all parameters for the ith individual. We used different values of B_k and θ_k for gender (male/female) and race (white/black/other), $c = 1/\sqrt{0.05}$, $\sigma_{Ci} = \tau_{Ci}^{-1/2}$, and $\sigma_{Hi} = \tau_{Hi}^{-1/2}$. The age at the start of the study is x, and t is the time in study before an event. In this model, Δ_i acts as a measure of volatility over time: it represents the mean difference between the home and clinic measurements. The parameters σ_{Hi} and σ_{Ci} are the variance of the home and clinic measurements, respectively, and they will include measurement error. A DAG of this model is given in figure 7.1.

The absolute value in the hazard rate introduced a discontinuity that hindered Stan's ability to draw samples effectively. The Stan manual (Stan Development Team, 2014c, sec. 31.6) warns against the use of step-like functions applied to parameters.

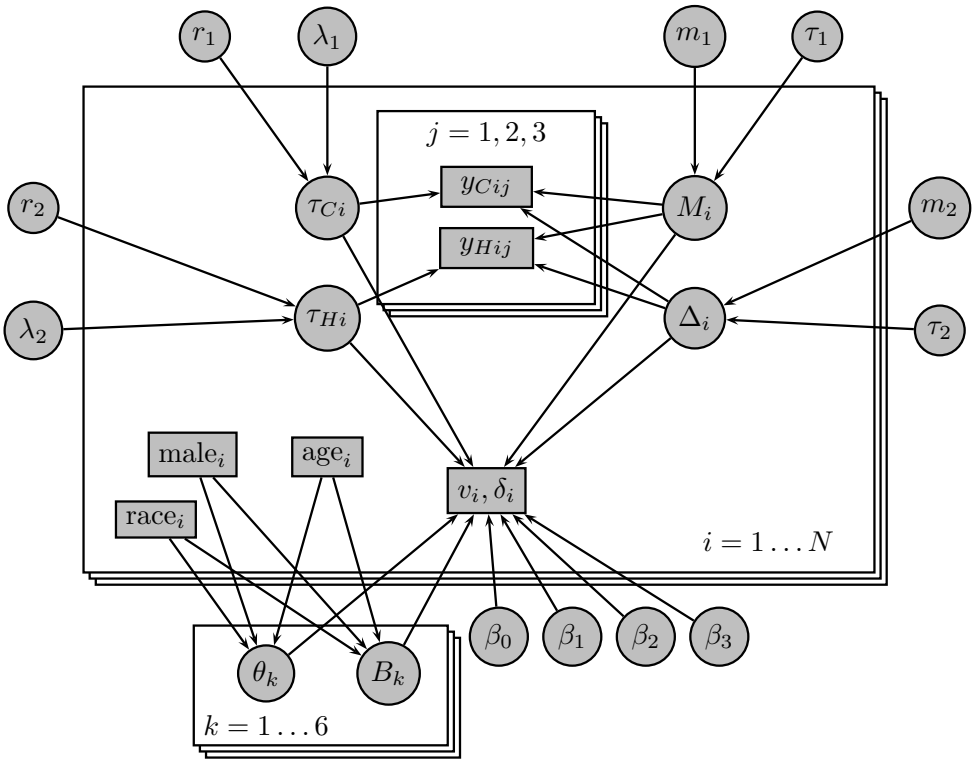


Figure 7.1. Directed acyclic graph of the NHANES data with the measurements taken at home, and the Gompertz baseline hazard rate. The hyperpriors on the top-level parameters are not shown.

The complicated hazard rate was difficult to implement in JAGS, so we used our custom sampler to estimate the parameters Results are given in table 7.1 for the SBP with heart-related mortality, and the Gompertz parameters are in table 7.2. In accordance with the conclusion from chapter 4, we see that the volatility of blood pressure is an important factor for risk of heart-related events. Here, the volatility of both the within-visit measurements taken at the clinic, and the volatility of between-visit measurements are important for measuring event risk. We obtained similar results for the SBP, as well as other measures of mortality, such as all cardiovascular events, all-cause mortality, and cardiovascular events excluding heart attacks.

This model was more complex than the UK-TIA model we used in chapter 4, since we had more information about subjects in the NHANES dataset. We ran the

	Mean	SD	HPD interval (95%)	Effective sample size	BF $H_1 : \theta > 0$	BF $H_1 : \theta < 0$
$\lvert\Delta_i\rvert$	0.024	0.006	(0.0126 ; 0.0359)	0.15	>9999	< 0.0001
M_i	0.004	0.002	(0 ; 0.009)	0.008	34.9389	0.0286
σ_{Ci}	-0.39	0.145	(-0.66 ; -0.13)	0.01	0.0004	2856.1429
σ_{Hi}	-0.06	0.084	(-0.23 ; 0.09)	0.072	0.2733	3.6593
m_1	125	0.2	(124.6 ; 125.2)	0.92		
τ_1	0.0028	0.00001	(0.0027 ; 0.0028)	0.84		
r_1	5.8	1.7	(4.44 ; 10.55)	0.0006		
λ_1	109	38	(79.59 ; 217.44)	0.0006		
m_2	1.34	0.06	(1.2188 ; 1.4552)	0.93		
τ_2	0.02	0.0003	(0.0191 ; 0.02)	0.37		
r_2	10.2	1.3	(7.5 ; 12.9)	0.0002		
λ_2	382	55	(272 ; 491)	0.0002		

Table 7.1. Posterior distribution MCMC results for the parameters of interest, from 4 chains of 10,000 iterations, SBP and heart-related mortality.

	Mean	SD	HPD interval (95%)	Effective sample size
$B_{\text{white-male}}$	0.000008	0.000004	(0.000002 ; 0.000017)	0.007
$B_{\text{white-female}}$	0.000002	0.000002	(0.0000002 ; 0.000006)	0.006
$\theta_{\text{white-male}}$	0.104	0.006	(0.092 ; 0.116)	0.007
$\theta_{\text{white-female}}$	0.113	0.008	(0.097 ; 0.129)	0.006
$B_{\text{black-male}}$	0.00009	0.00005	(0.00002 ; 0.0002)	0.014
$B_{\text{black-female}}$	0.00002	0.00002	(0.000002 ; 0.00005)	0.015
$\theta_{\text{black-male}}$	0.073	0.007	(0.059 ; 0.085)	0.025
$\theta_{\text{black-female}}$	0.086	0.008	(0.071 ; 0.102)	0.02
$B_{\text{other-male}}$	0.00005	0.00003	(0.000007 ; 0.00012)	0.024
$B_{\text{other-female}}$	0.000007	0.000006	(0.0000004 ; 0.000019)	0.018
$\theta_{\text{other-male}}$	0.074	0.008	(0.058 ; 0.089)	0.025
$\theta_{\text{other-female}}$	0.1	0.01	(0.08 ; 0.12)	0.016

Table 7.2. Posterior distribution MCMC results for the Gompertz parameters, from 4 chains of 10,000 iterations, SBP and heart-related mortality.

diagnostics, which we developed in chapter 6, for the survival part of the model, with the results plotted in figure 7.2. The posterior predictive p-value was 0.017, so our model failed to adequately explain the survival dynamics. Unfortunately, the Cox-Snell residuals cannot reveal how our model is inadequate. We decided to investigate another extension to our model.

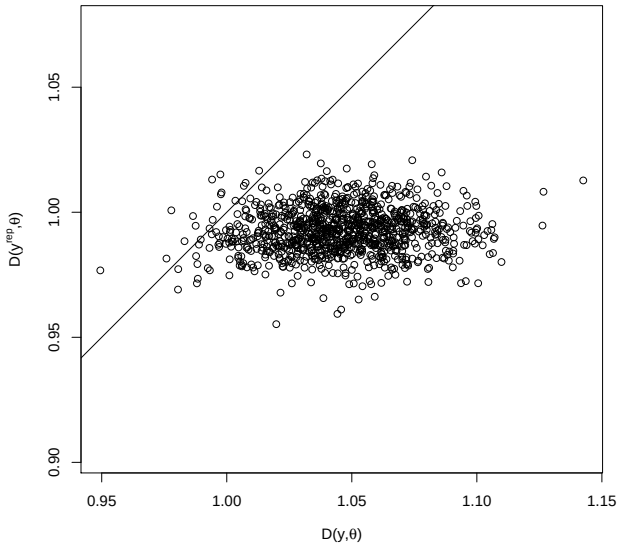


Figure 7.2. Posterior predictive check for the discrepancy variable $D(\boldsymbol{y}, \boldsymbol{\theta})$ as the slope of the estimated hazard rate, defined in equation (6.5). The scatterplot displays 1,000 simulations of $\{D(\boldsymbol{y}^{rep}, \boldsymbol{\theta}), D(\boldsymbol{y}, \boldsymbol{\theta})\}$, where $\boldsymbol{y}$ is the observed data and $\boldsymbol{y}^{rep}$, $\boldsymbol{\theta}$ are the simulated values. The p-value is estimated as the proportion of points above the 45° line.

7.3 Bivariate longitudinal process

To this point, we have only considered models with a univariate longitudinal process. This mimicked Rothwell (2010), who tested SBP and DBP in separate models. If we want to use the SBP as well as the DBP in the same model, we might assume that they are independent:

$$Y_{i1}(t) \sim N(\mu_{i1}, \tau_{i1}^{-1}) \quad , \quad Y_{i2}(t) \sim N(\mu_{i2}, \tau_{i2}^{-1}), \tag{7.6}$$

where $Y_{i1}(t)$ and $Y_{i2}(t)$ are the SBP and DBP, respectively. This will make it easy to add to our existing MCMC algorithm.

If we believe, however, that the two blood pressure processes are correlated, we need a model that can handle multiple processes at once. There is some substance to this belief. The correlation coefficient for all of the available SBP-DBP pairs in the

NHANES dataset is 0.52, see figure 7.3. The second half of this chapter deals with a model that has a bivariate longitudinal process. Consider the model for the blood pressure. Instead of a single process $Y_i(t)$, we have a bivariate process

$$\boldsymbol{Y}_i(t) = \begin{bmatrix} Y_{i1}(t) \\ Y_{i2}(t) \end{bmatrix}.$$

In line with our assumptions from chapter 4, we assume this follows a bivariate normal distribution:

$$\begin{bmatrix} Y_{i1}(t) \\ Y_{i2}(t) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix}, \begin{bmatrix} \tau_{i11} & \tau_{i12} \\ \tau_{i21} & \tau_{i22} \end{bmatrix}^{-1} \right) \quad \text{or} \tag{7.7}$$

$$\boldsymbol{Y}_i(t) \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i^{-1}\right)$$

where $\boldsymbol{\Lambda}_i$ is the precision matrix. This process then generates each individual's blood pressure measurements. Further adhering to the model structure from chapter 4, we

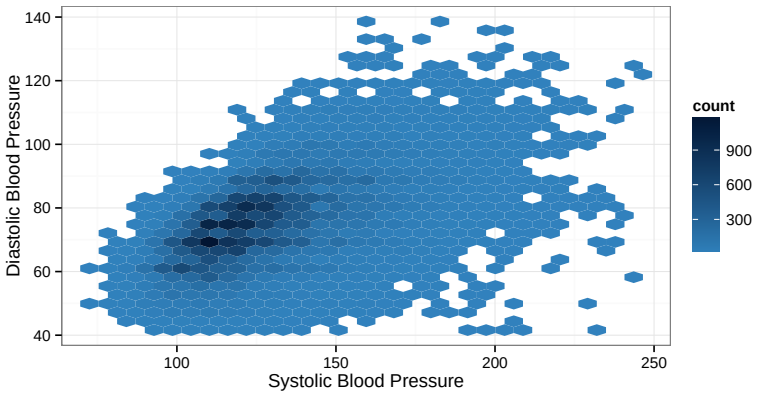


Figure 7.3. Heat-map scatterplot of all of the available SBP and DBP pairs.

assume distributions on $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$ that result in conjugacy:

$$\begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}^{-1} \right) \text{ or alternatively} \tag{7.8}$$

$$\boldsymbol{\mu_i} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}\right)$$

and

$$\begin{bmatrix} \tau_{i11} & \tau_{i12} \\ \tau_{i21} & \tau_{i22} \end{bmatrix} \sim Wishart\left(\begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}, \nu \right) \text{ or} \tag{7.9}$$

$$\Lambda_i \sim Wishart\left(\boldsymbol{V}, \nu\right).$$

Writing the jth observation of individual i as $Y_{i1}(t_{ij}) = y_{i1j}$ and similarly $Y_{i2}(t_{ij}) = y_{i2j}$, where t_{ij} is the time of this observation, we can examine the process that generates

the data according to this model:

$$
\begin{array}{lcl}
 & Wishart\left(\boldsymbol{V},\nu\right) & \\
\mathcal{N}\left(\boldsymbol{\mu}_0,\boldsymbol{\Lambda_0}\right) & & \\
\downarrow \qquad \downarrow & & \\
i=1:\mathcal{N}\left(\begin{bmatrix}\mu_{11}\\ \mu_{12}\end{bmatrix},\begin{bmatrix}\tau_{111} & \tau_{112}\\ \tau_{121} & \tau_{122}\end{bmatrix}^{-1}\right) & \longrightarrow & \begin{bmatrix}y_{111}\\ y_{121}\end{bmatrix}\begin{bmatrix}y_{112}\\ y_{122}\end{bmatrix}\begin{bmatrix}y_{113}\\ y_{123}\end{bmatrix}\cdots\begin{bmatrix}y_{11n_1}\\ y_{12n_1}\end{bmatrix}\\
i=2:\mathcal{N}\left(\begin{bmatrix}\mu_{21}\\ \mu_{22}\end{bmatrix},\begin{bmatrix}\tau_{211} & \tau_{212}\\ \tau_{221} & \tau_{222}\end{bmatrix}^{-1}\right) & \longrightarrow & \begin{bmatrix}y_{211}\\ y_{221}\end{bmatrix}\begin{bmatrix}y_{212}\\ y_{222}\end{bmatrix}\begin{bmatrix}y_{213}\\ y_{223}\end{bmatrix}\cdots\begin{bmatrix}y_{21n_2}\\ y_{22n_2}\end{bmatrix}\\
\vdots & \vdots & \vdots\\
\mathcal{N}\left(\begin{bmatrix}\mu_{i1}\\ \mu_{i2}\end{bmatrix},\begin{bmatrix}\tau_{i11} & \tau_{i12}\\ \tau_{i21} & \tau_{i22}\end{bmatrix}^{-1}\right) & \longrightarrow & \begin{bmatrix}y_{i11}\\ y_{i21}\end{bmatrix}\begin{bmatrix}y_{i12}\\ y_{i22}\end{bmatrix}\begin{bmatrix}y_{i13}\\ y_{i23}\end{bmatrix}\cdots\begin{bmatrix}y_{i1n_i}\\ y_{i2n_i}\end{bmatrix}\\
\vdots & \vdots & \vdots\\
i=N:\mathcal{N}\left(\begin{bmatrix}\mu_{N1}\\ \mu_{N2}\end{bmatrix},\begin{bmatrix}\tau_{N11} & \tau_{N12}\\ \tau_{N21} & \tau_{N22}\end{bmatrix}^{-1}\right) & \longrightarrow & \begin{bmatrix}y_{N11}\\ y_{N21}\end{bmatrix}\begin{bmatrix}y_{N12}\\ y_{N22}\end{bmatrix}\begin{bmatrix}y_{N13}\\ y_{N23}\end{bmatrix}\cdots\begin{bmatrix}y_{N1n_N}\\ y_{N2n_N}\end{bmatrix},
\end{array}
$$

where each individual has $j = 1 \ldots n_i$ observations; different numbers of observations between individuals are allowed. Our model assumed each individual has a distinct $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$, drawn from the global distributions $\mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Lambda_0}\right)$ and $Wishart\left(\boldsymbol{V}, \nu\right)$, respectively. The individual's $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$ then generated the blood pressure observations we saw. These were measured with error, but we did not express it explicitly; the precision matrix $\boldsymbol{\Lambda}_i$ contained the subject's underlying variance as well as the measurement error.

Compared to the model of (7.6) that assumes two independent processes, (7.7) has an extra parameter per individual, $\tau_{i21} = \tau_{i21}$, accounting for the individual covariance between the two longitudinal processes. The same is true for the distribution of $\boldsymbol{\mu}_i$,

the individual mean vector: there is an extra covariance parameter $\tau_{12} = \tau_{21}$. This raises a question about where the correlation in figure 7.3 comes from. It could be due to a correlation between μ_{i1} and μ_{i2}, a between-subject correlation of mean SBP and mean DBP. Or a correlation between Y_{i1} and Y_{i2}, a within-subject correlation of the actual measurements. Establishing which of these are driving the correlation could be a topic for future research. A Bayesian equivalent of a multivariate ANOVA is required if we want to compare the within-subject covariance between the two observed measurements with the between-subject covariance between the two hidden parameters μ_{i1} and μ_{i2}. One possible course of action is to use the MCMC samples from our joint model, and check the significance of $\tau_{i21} = \tau_{i21}$ and $\tau_{12} = \tau_{21}$. If we had prior knowledge of the source of variance, however, we could include that in the model and make the inference stronger.

The extension to multivariate distributions was not difficult to implement, since the overall structure of the model remained the same as that in chapter 4, except that the distributions were now multivariate. As with the univariate model, we made use of conjugacy to aid the construction of the MCMC algorithm. We assumed bivariate normal data $\boldsymbol{Y}_i(t) \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i^{-1}\right)$, and then we assumed the priors in (7.8) and (7.9), which led to the posterior distributions

$$\boldsymbol{\mu}_i | \boldsymbol{\Lambda}_i \, , \boldsymbol{Y}_i \sim \mathcal{N}\left[\left(\boldsymbol{\Lambda}_0 + n_i \boldsymbol{\Lambda}_i\right)^{-1} \left(\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + n_i \boldsymbol{\Lambda}_i \bar{\mathbf{x}}_i\right), \left(\boldsymbol{\Lambda}_0 + n_i \boldsymbol{\Lambda}_i\right)^{-1}\right] \tag{7.10}$$

and

$$\boldsymbol{\Lambda}_i | \boldsymbol{\mu}_i \, , \boldsymbol{Y}_i \sim Wishart\left(\left(\mathbf{V}^{-1} + \sum_{j=1}^{n_i} (\mathbf{x_{ij}} - \boldsymbol{\mu}_i)(\mathbf{x_{ij}} - \boldsymbol{\mu}_i)^T\right)^{-1}, \, n_i + \nu\right). \tag{7.11}$$

Similar to Chi and Ibrahim (2006), our choice of priors were motivated by conjugacy considerations; Gibbs sampling is easier when the full conditional distributions are conjugate.

We related the longitudinal process to the survival using the shared parameter approach, specifying the hazard rate:

$$h(t|\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = h_0(t) \exp\left[\beta_0\mu_{i1} + \beta_1\mu_{i2} + \beta_2{\tau_{i11}}^{-\frac{1}{2}} + \beta_3{\tau_{i22}}^{-\frac{1}{2}}\right], \tag{7.12}$$

which was similar to (4.3), but had effects for the mean and variance of both longitudinal processes. Thus far, the model had the same structure as our univariate model from chapter 4. We continued that by adding another hierarchy to the parameters in (7.8):

$$\boldsymbol{\mu}_0 \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1000 & 0\\0 & 1000\end{bmatrix}\right) \tag{7.13}$$

and

$$\boldsymbol{\Lambda}_0 \sim Wishart\left(3, \begin{bmatrix}1000 & 0\\0 & 1000\end{bmatrix}\right), \tag{7.14}$$

representing our vague posterior distributions. Contrary to the model in chapter 4 we did not assume a priors on (7.9). Our reason for disregarding the priors on $\boldsymbol{V}$ and ν was due to the difficulty thereof. No conjugate distributions were available. Furthermore, we had to deal with the other full conditional distributions, which were not only unfamiliar, but also bivariate.

The plan was to initially specify $\boldsymbol{V}$ and ν as known, and to use a simulation study to gauge the performance of our multivariate model and its accompanying estimation algorithm. If the model performed adequately, we could either use empirical Bayes to find estimates for $\boldsymbol{V}$ and ν, or specify prior distributions and estimate them along with the rest of the parameters in the model. Initially ignoring them also meant fewer places the model and code could fail, and less debugging.

7.3.1 Full conditional distributions

The full conditional distributions resulting from our model assumptions are presented below. We demonstrate that, while the structure of the dependencies remain the same as our chapter 4 model, the distributions are multivariate. We cannot use the exact same techniques as before to sample from them.

$\boldsymbol{\mu}_i$:

$$\begin{aligned} f(\boldsymbol{\mu}_i|\text{All others}) &\propto \underbrace{\prod_{j=i}^{n_i}\Big(f(\boldsymbol{y}_{ij}|\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i)\Big) f(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0)}_{\text{similar to conjugate multivariate normal posterior}} \underbrace{f(v_i,\delta_i|\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i,\boldsymbol{\beta})}_{\text{survival likelihood}} \\ &= f_{\mathcal{N}}\left(\boldsymbol{\mu}_i|\left(\boldsymbol{\Lambda}_0+n_i\boldsymbol{\Lambda}_i\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0+n_i\boldsymbol{\Lambda}_i\bar{\mathbf{x}}_i\right),\left(\boldsymbol{\Lambda}_0+n_i\boldsymbol{\Lambda}_i\right)^{-1}\right)\times \\ &\qquad [h(v_i)]^{\delta_i}\exp\left[-H(v_i)\right] \end{aligned} \tag{7.15}$$

where $f_{\mathcal{N}}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the density of $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

$\boldsymbol{\Lambda}_i$:

$$\begin{aligned} f(\boldsymbol{\Lambda}_i|\text{All others}) &\propto \underbrace{\prod_{j=i}^{n_i}\Big(f(\boldsymbol{y}_{ij}|\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i)\Big) f(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0)}_{\text{similar to conjugate Wishart posterior}} \underbrace{f(v_i,\delta_i|\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i,\boldsymbol{\beta})}_{\text{survival likelihood}} \\ &f_{Wis}\left(\boldsymbol{\Lambda}_i\middle|\left(\mathbf{V}^{-1}+\sum_{j=1}^{n_i}(\mathbf{x_{ij}}-\boldsymbol{\mu}_i)(\mathbf{x_{ij}}-\boldsymbol{\mu}_i)^T\right)^{-1},\, n_i+\nu\right)\times \\ &\qquad [h(v_i)]^{\delta_i}\exp[-H(v_i)] \end{aligned} \tag{7.16}$$

where $f_{Wis}(\boldsymbol{X}|\boldsymbol{V},\nu)$ is the density of a $Wishart(\boldsymbol{V},\nu)$ distribution.

β_j:

$$\begin{aligned} f(\beta_j|\text{All others}) &\propto \prod_{i=1}^{N} f(v_i,\delta_i|\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i,\boldsymbol{\beta})f(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{N}\left[h(v_i)\right]^{\delta_i}\exp[-H(v_i)]\, f(\boldsymbol{\beta}) \end{aligned} \tag{7.17}$$

where the hazard rate is given in (7.12).

$\boldsymbol{\mu}_0$:

$$\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

$$\boldsymbol{\mu}_0 \sim \mathcal{N}(\boldsymbol{\mu}_\mu, \boldsymbol{\Lambda}_\mu^{-1})$$

as given in (7.13).

$$f(\boldsymbol{\mu}|\text{All others}) \propto \prod_{i=1}^{N} f(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) f(\boldsymbol{\mu}_0)$$

which becomes (7.18)

$$\mathcal{N}\left[(\boldsymbol{\Lambda}_\mu + N\boldsymbol{\Lambda}_0)^{-1}\left(\boldsymbol{\Lambda}_\mu\boldsymbol{\mu}_\mu + N\boldsymbol{\Lambda}_0\bar{\boldsymbol{\mu}}_i\right), (\boldsymbol{\Lambda}_\mu + N\boldsymbol{\Lambda}_0)^{-1}\right]$$

$\boldsymbol{\Lambda}_0$:

$$\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

$$\boldsymbol{\Lambda}_0 \sim Wishart(\boldsymbol{V}_\Lambda, \nu_\Lambda)$$

as given in (7.14).

$$f(\Lambda_0|\text{All others}) \propto \prod_{i=1}^{N} f(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) f(\boldsymbol{\Lambda}_0)$$

which becomes (7.19)

$$Wishart\left(\left(\mathbf{V}_\Lambda^{-1} + \sum_{i=1}^{N}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)^T\right)^{-1}, N + \nu_\Lambda\right)$$

$\boldsymbol{V}$:

Specified, or calculated using empirical Bayes.

ν:

Also specified or calculated using empirical Bayes.

7.4 Parameter estimation revisited

The extensions of sections 7.1 and 7.2 are easy to add to the MCMC algorithm from chapter 4 — they simply require altering the likelihood contribution of the survival part. The bivariate longitudinal process, on the other hand, changes the distributions of the Gibbs sampler. Prior to attempting a real dataset, we performed a simulation study, which involved simulating data according to the underlying assumptions and checking whether we could recover the correct parameter estimates.

7.4.1 JAGS

The JAGS software package includes multivariate distributions, and we provided the code to fit the model of section 7.3 with the hazard rate in (7.12) in appendix B.3. When we ran the model, however, it failed with the message:

```
Error in node prec[1,1:2,1:2]
Unable to find appropriate sampler
```

This meant the model failed at sampling $\mathbf{\Lambda}_i$ at $i = 1$, which was the first precision matrix it had to sample. JAGS did not have an appropriate sampler to draw from this distribution. This is one of the non-conjugate distributions in our Gibbs sampler, and requires some special insight to sample from.[4]

7.4.2 Stan

Stan was much better suited for the bivariate longitudinal process, as its language required explicit declaration of covariance matrices. It also seemed to converge successfully, but the program's execution time was impractical. Running the program for 10,000 iterations took 246 hours on a computer with 24 cores[5] and 200GB of

[4]We did not test our model in the BUGS program due to BUGS's platform constraints, but it will have similar limitations.

[5]Intel® Xeon® CPU E5-244 @ 2.40GHz

RAM. This was for a dataset with 2,000 individuals, each providing three longitudinal observations. Furthermore, the chains still had high autocorrelations between samples; the effective sample size of the β_js, was 40 on average, from 10,000 iterations.[6] We provide the Stan code for the model in appendix B.4.

Stan was built with the goal of handling models that slows JAGS and BUGS down to the point where they stop working (Stan Development Team, 2014c, p. vi), and in this instance it was successful: we could fit a model that did not work in JAGS. Waiting upwards of 200 hours for each run, however, was not a feasible time frame. The authors of Stan did state that it is a work in progress, and that they were interested to hear about new models that either worked well or produced problems in Stan. This model — as well as the model from chapter 4 — is certainly a candidate for their scrutiny.

7.4.3 Smart sampling with the custom sampler

The parameters causing issues in the multivariate model are $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$, with full conditionals given in (7.15) and (7.16), respectively. They correspond to μ_i and τ_i in our univariate model from chapter 4, but they are now matrices rather than scalar values. This means at each iteration of our MCMC algorithm, we need to sample the vector $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$, for $i = 1 \dots N$ subjects. Previously, when we had to sample from unfamiliar univariate distributions, we used one of the Slice, ARS, or ARMS samplers. We could theoretically also use this in the multivariate case, sampling the matrices using Gibbs sampling by keeping all its values fixed and updating one cell at a time. However, as stated in section 3.2.3, we should not force the use of a Gibbs sampler, but rather look for alternative methods. Since we are sampling matrices, the Metropolis-Hastings that proposes a matrix as a new update is an obvious candidate. We thus used the smart sampling method explained in chapter 5 to sample $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$.

[6] Calculated using the `effectiveSize()` function, from the `coda` library in R.

For $\boldsymbol{\mu}_i$, we used the part of (7.15) that resembled a multivariate normal distribution as proposal distribution for Metropolis-Hastings updates. Likewise, for $\boldsymbol{\Lambda}_i$, we used the part of (7.16) that resembled a Wishart distribution as proposal distribution.

7.5 Simulation study

In this section we tested our estimation scheme, by simulating data and checking whether we could recover the true parameters. We simulated data according to the assumptions in (7.7), (7.8), (7.9), and the hazard rate:

$$h(t|\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = h_0(t) \exp\left[\beta_0(\mu_{i1} - 128) + \beta_1(\mu_{i2} - 76) + \beta_2 {\tau_{i11}}^{-\frac{1}{2}} + \beta_3 {\tau_{i22}}^{-\frac{1}{2}}\right]. \quad (7.20)$$

It is similar to (7.12), but we centred the $\boldsymbol{\mu}_i$ parameters to aid MCMC convergence. We used values inspired by our actual dataset

$$\begin{aligned} \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 128 \\ 76 \end{bmatrix}, \begin{bmatrix} 0.003384 & -0.003896 \\ -0.003896 & 0.012557 \end{bmatrix}\right) \\ \begin{bmatrix} \tau_{i11} & \tau_{i12} \\ \tau_{i21} & \tau_{i22} \end{bmatrix} &\sim Wishart\left(\begin{bmatrix} 0.04139 & 0 \\ 0 & 0.05325 \end{bmatrix}, 3\right) \end{aligned} \quad (7.21)$$

to produce ten longitudinal measurements for $N = 2000$ individuals. We generated events using (7.20) with

$$\begin{aligned} \beta_0 &= 0.1 \\ \beta_1 &= 0.15 \\ \beta_2 &= 0.2 \\ \beta_3 &= 0.25 \end{aligned} \quad (7.22)$$

and applied random censoring such that around 5% of events were censored.

We then estimated the parameters for each of the datasets using our custom

sampler from section 7.4.3. Our interest lay with the β_j parameters in particular, since they measure the effect of the longitudinal process on the survival. We repeated this simulation-estimation process 100 times, using (7.21) and (7.22) to simulate each distinct dateset. Figure 7.4 shows the posterior mean and the 95% HPD intervals of the β_i parameters and their coverage of the true parameter values for 100 simulations. We notice two clear outliers in the graphs for β_2 and β_3, which are the two parameters associated with the blood pressure precisions. Furthermore, these outliers appear correlated: in the two simulations where β_2 was overestimated, β_3 was underestimated. We do not know whether this was due to the MCMC algorithm used, or whether it was a consequence of the assumptions of the model.

To further assess the estimation, we calculated z-values for the β_i parameters, using the posterior samples. If $\boldsymbol{\beta}_i^{(d)}$ is the set of MCMC samples for β_i from the dth estimation run, then the z-value is

$$z_d = \frac{\beta_i - \text{mean}(\boldsymbol{\beta}_i^{(d)})}{\text{SD}(\boldsymbol{\beta}_i^{(d)})}.$$

This is our estimation error and since we expect it to follow a normal distribution we perform a QQ-plot against the quantiles of the normal distribution given in 7.5. From these results we see a clear estimation bias occurring in the β_i parameters. None of the other parameters displayed this bias, and therefore we omitted their estimation results here.

We suspected that the problems arose due to the correlation of μ_{i1} and μ_{i2}. To test this we ran the simulation-estimation routine again, this time using

$$\begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 128 \\ 76 \end{bmatrix}, \begin{bmatrix} 0.003384 & 0 \\ 0 & 0.012557 \end{bmatrix} \right) \tag{7.23}$$

instead of (7.21), to remove the correlation between the individual mean SBP and

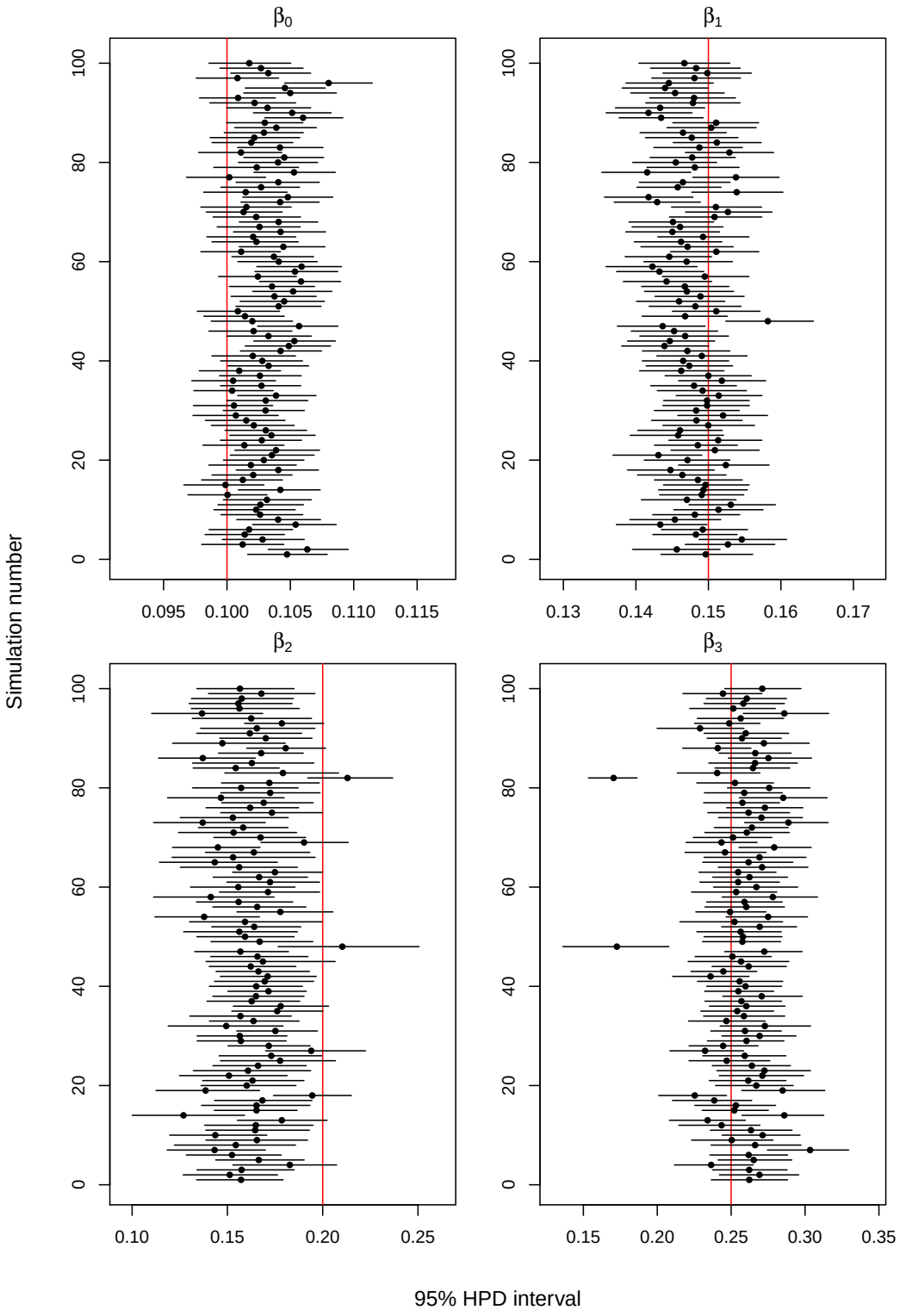


Figure 7.4. Posterior mean (dots) and the 95% HPD intervals (lines) of the β_i parameters for the 100 simulations. The red line shows the true parameter value.

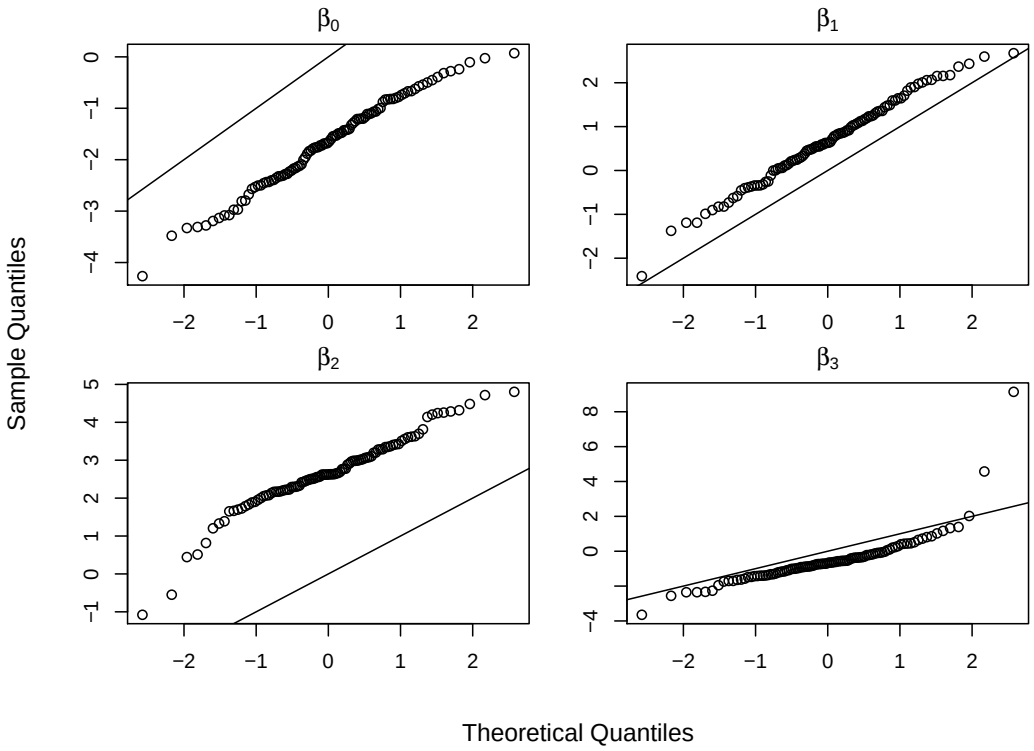


Figure 7.5. Normal QQ-plot of the estimation results, using hazard rate (7.20) and parameters of (7.21) and (7.22).

mean DBP. Other than this, everything was identical to the first simulation-estimation. The QQ-plots of the $z^{(d)}$ are given in figure 7.6. The bias pattern is less apparent, but there is still a small bias visible in the β_2 parameter.

Finally, we repeated another simulation-estimation routine assuming two independent processes:

$$\begin{aligned} \text{Systolic blood pressure :} \qquad & y_{i1j} \sim N(\mu_{i1}, {\tau_{i1}}^{-1}) \\ \text{Diastolic blood pressure :} \qquad & y_{i2j} \sim N(\mu_{i2}, {\tau_{i2}}^{-1}) \end{aligned}$$

where

$$\begin{aligned} \mu_{i1} &\sim N(m_1, \tau_1) & \mu_{i2} &\sim N(m_2, \tau_2) \\ \tau_{i1} &\sim \Gamma(r_1, \lambda_1) & \tau_{i2} &\sim \Gamma(r_2, \lambda_2) \end{aligned}$$

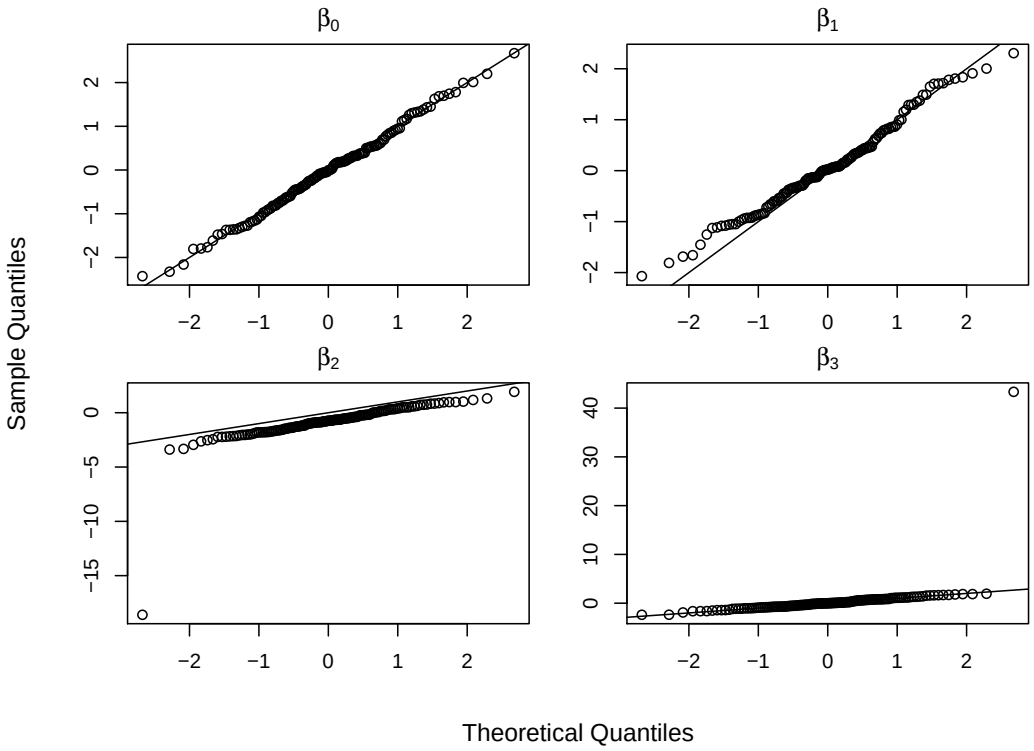


Figure 7.6. Normal QQ-plot of the estimation results, using hazard rate (7.20) and parameters of (7.23) and (7.22).

with

$$\begin{aligned} m_1 &= 128 & m_2 &= 74 \\ \tau_1 &= 0.0023 & \tau_2 &= 0.0073 \\ r_1 &= 6 & r_2 &= 3 \\ \lambda_1 &= 120 & \lambda_2 &= 60 \end{aligned} \tag{7.24}$$

and again using the hazard rate in (7.20). The QQ-plots of the $z^{(d)}$ are given in figure 7.7 and they are exactly what we expected. Thus, the estimation issues witnessed for the multivariate models do not emerge when we assume the longitudinal processes to be independent.

We used this chapter to consider extensions to the model developed in chapter 4, starting with brief discussions on competing risks, and time-dependent baseline hazard rates. Then we extended the model to accommodate a bivariate longitudinal

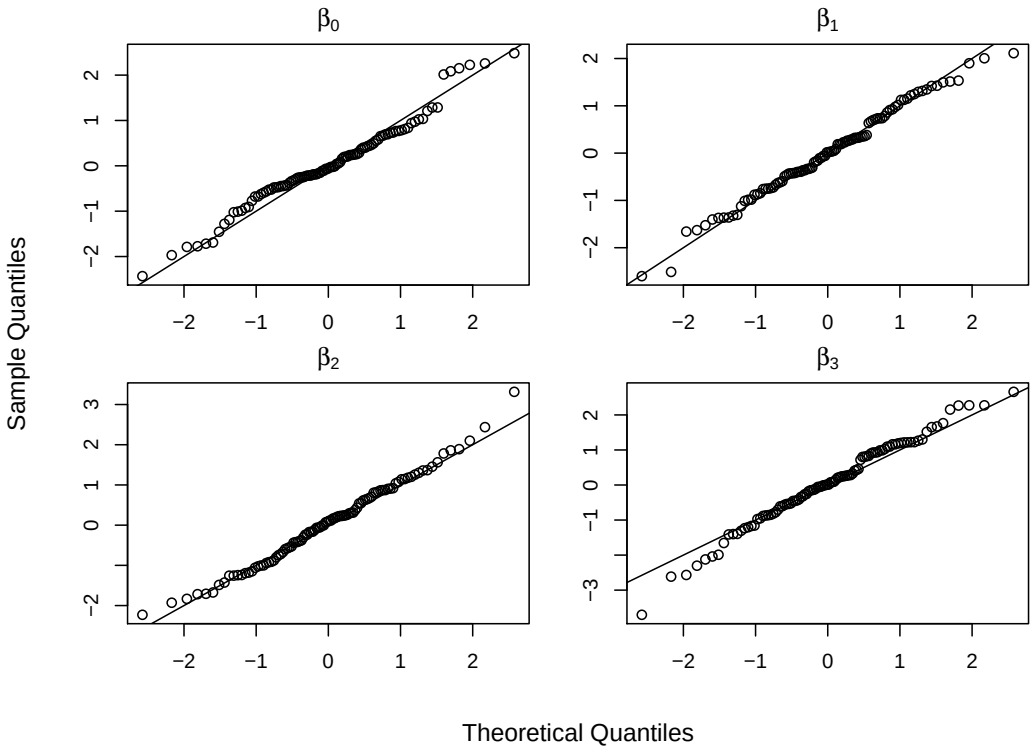


Figure 7.7. Normal QQ-plot of the estimation results, using hazard rate (7.20) and parameters of (7.22) and (7.24).

process. During the simulation study we found a bias in estimation of our bivariate model. We were unable to find an explanation for this. What we did find was that the bias seemed to be exacerbated by a higher correlation between the μ_{i1} and μ_{i2}. When the two longitudinal processes were completely independent, we did not experience the estimation problems of the other bivariate models. Exactly how the independence assumption would affect estimation and inference in a joint model where there is a correlation present is a question for future research.

We also considered a model that used the mean and the difference of the means as covariates, but encountered the same problem. The same held true when we used

$$\begin{bmatrix} u_{i1}(t) \\ u_{i2}(t) \end{bmatrix} = \begin{bmatrix} y_{i1}(t) \\ y_{i1}(t) - y_{i2}(t) \end{bmatrix}$$

to replace $\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix}$ in the model. This makes sense, since if we have

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right)$$

then

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_1 - Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_1 - \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1^2 - \rho\sigma_1\sigma_2 \\ \sigma_1^2 - \rho\sigma_1\sigma_2 & \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2 \end{bmatrix} \right)$$

where the correlation between U_1 and U_2 is still non-zero. If we believe the problem lies with the correlation, a probable solution would be to choose a transformation of $\boldsymbol{y}_i(t)$ that would yield an uncorrelated $\boldsymbol{u}_i(t)$, but then the interpretation of the results becomes difficult. We did not consider such transformations.

Chapter 8

Multiple event arrivals

In this chapter we deviate from investigation on joint models, and consider a study of event arrivals in a heterogeneous population, where an event does not end participation in the study. The software we developed for our Bayesian model, described in chapter 4, proved useful in this study as well after minor adjustments.

8.1 A behavioural study

We examined a statistical analyses of data from the *Oxford Clack Nutrition Study: A randomised control trial investigating the impact of supplementary vitamins, minerals and omega 3 fatty acids on adolescent behaviour and cognition.* The study investigated whether the presence of Omega-3 fatty acids, when introduced to the diet of schoolchildren, had a positive effect on behaviour.

Researchers monitored the offences of 196 children for a baseline period of twelve weeks, after which they were randomly assigned either treatment, in the form of nutritional supplements, or placebo. After a holiday period of four weeks, they again monitored behaviour for twelve weeks in order to establish whether the active treatment had any effect on offending (a representation of the study time-frame is given in figure 8.1). Offences were logged daily in the school's self-designed databases.

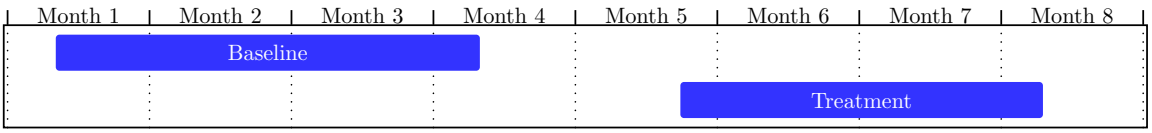


Figure 8.1. The time line of the RCS School Study

Throughout the study, 850 events were recorded. The most reported offence was disruptive behaviour, but aggression, tardiness, and other offences were also noted. There were instances of positive referrals (behaviour that the teacher considered to positive and noteworthy) as well, but only 74 were recorded — too few for meaningful analysis.

8.1.1 Negative binomial distribution

Studies that consider event arrivals over two periods, usually baseline and treatment periods, can be analysed using the assumption that events arrive according to a Poisson distribution. Lawless (1987) explained, however, that count data in studies with heterogeneous subjects often display extra Poisson variation, or overdispersion. Subjects in a study experience a larger variance than can be explained by the Poisson distribution. The author then examined the properties of the Negative Binomial distribution as an alternative to the Poisson distribution when overdispersion is present. This method has since been well developed as a way of dealing with overdispersion in models with count data. Hilbe (2011) showed that it could be derived either as a Poisson-Gamma mixture model, or by considering it as a member of the exponential family of distributions and then modelling it in the generalised linear model framework. The former treats counts as Poisson random variables where the rate parameters have Gamma heterogeneity. The author further provided a detailed overview of the negative binomial regression, its properties, and its various forms.

Gesch et al. (2002) analysed a dataset similar to that of the schoolchildren study, involving prison inmates. As with the children, the inmates received nutritional

supplements after a baseline period where offences were monitored, which were then compared to offences in a treatment period. Due to the presence of overdispersion, the authors used a negative binomial regression — there was a disparity among the observed offence rates of inmates — and provided Lawless (1987) as a reference. The statistical methodology section in Gesch et al. (2002) only contained a short description of their methods, summed up in 289 words, and despite the ease of doing binomial regressions in R we were not able to reproduce their model for use in our study. After extensive communication with the lead author of the paper, however, we were able to obtain the code and further details from the retired designated-statistician who had worked on the study. The model they used was not, in fact, the negative binomial regression as given in Lawless (1987), but rather a model based on a version of the *bivariate negative binomial* distribution.

The first mention of this family of distributions appears in Marshall and Olkin (1985) as a distribution that can be generated from the bivariate Bernoulli distribution. Iwasaki and Tsubaki (2006) provided a detailed derivation of the bivariate negative binomial distribution with the form as it appears in the book by Hilbe (2011), who gave an overview of the model's attributes and commented on its use.[1] He wrote,

> Bivariate probit models are perhaps the most well-known bivariate models. Software such as Stata, R, SAS, LIMDEP, Matlab, and others have bivariate probit as a basic offering. Bivariate Poisson and negative binomial models, however, have not enjoyed such software support. Given their potential value in understanding the relationship of the two related count variables, though, it is somewhat surprising that they have been generally ignored by the research community.

Perhaps equally surprising is the fact that some papers, which indeed made use of this distribution, neglected to mention it. Hilbe (2011) further stated that a number of parametrizations of this distribution has been proposed, however, no closed form

[1] We are referring to the second edition — the bivariate negative binomial distribution does not appear in the first print.

solution exists, and parameter estimation requires numerical methods. Indeed, we found that Gesch et al. (2002) derived their distribution, along with its relevant properties required for inference specifically for their prison study, and wrote code for parameter estimation in Matlab. We were unable to re-implement it.

The bivariate negative binomial model is a useful candidate for comparing the two related count variables in their study. However, the fact that this model has multiple variations, each requiring extensive coding for parameter estimation, makes it difficult to reproduce. This is perhaps one reason for its absenteeism from the literature despite its useful properties. We opted to use a Bayesian hierarchical model to implement the Poisson model with Gamma heterogeneity. Our aim was to do so using available software packages, with reproducibility in mind. We will use the rest of the chapter to present this model, along with the findings from the schoolchildren study.

8.1.2 Overdispersed Poisson regression: a Bayesian hierarchical model

To investigate treatment effect on the offence rate, we assumed that offences arrived according to a Poisson process. Let $N_i(t, s)$ be the number of events observed for pupil i over a period running from time t to time s, where $t < s$. Then,

$$N_i(t, s) \sim Poisson\left(\int_t^s h(x|\lambda_i)dx\right), \tag{8.1}$$

where $h(x|\lambda_i)$ is the hazard rate for pupil i at time x, specified as:

$$h(t|\lambda_i) = \begin{cases} \lambda_i & \text{for } t \text{ in baseline period} \\ \lambda_i e^{\alpha+\alpha_A(A_i)} & \text{for } t \text{ in treatment period,} \end{cases} \tag{8.2}$$

with A_i being an indicator variable taking 1 if subject i is in the active group, and 0 for the placebo group. Furthermore, we assumed the prior

$$\lambda_i \sim \Gamma(r, \theta)$$

with r and θ as the shape and rate of the Gamma distribution, respectively. This results in a likelihood of the form (2.8) and defining the model in terms of its structural dependencies leads to a Bayesian hierarchical model as in section 4.1.3.

Similarly to the model in chapter 4, this model does not have to be Bayesian. We can consider the λ_is as random effects and estimate the parameters using an EM algorithm, for example. The actual model used by Gesch et al. (2002) can be regarded as a frequentist's equivalent of our model. The coding required to estimate the parameters, however, is much simpler in the Bayesian setting, thanks to software packages that implement Gibbs sampling. Following from this is the fact that the Bayesian model can be easily adjusted to add more covariates, without having to rewrite large amounts of code — making the model easy to reproduce and verify.

Our interpretation of the parameters goes as follows: a pupil who had a baseline offence rate of λ_i, demonstrated an offence rate during the treatment period of $\lambda_i e^{\alpha}$ if he was in the placebo group, and $\lambda_i e^{\alpha+\alpha_A}$ if he was in the active group. The individual rate parameter λ_i takes the role of a random effect, and we assume each pupil within the total population has an underlying offence rate, but that we only observed a sample of these. We chose the Gamma distribution due to convenience, but it can be substituted with any appropriate distribution.

The four parameters we needed to estimate were α, α_A, r, and θ and we assumed

vague priors on them, as

$$\begin{aligned}
\alpha &\sim N(0, 0.0001^{-1}) \\
\alpha_A &\sim N(0, 0.0001^{-1}) \\
r &\sim \Gamma(0.0001, 0.0001) \\
\theta &\sim \Gamma(0.0001, 0.0001).
\end{aligned}$$

After running the MCMC sampler, we tested convergence using the Gelman and Rubin (1992) diagnostic. The chains showed convergence for all of our models and all effective sample sizes were larger than 5,000. Diagnostics were done using the `coda` (Plummer et al., 2006) package in R.

The JAGS code that fits the model with the hazard rate in (8.2) is given in appendix B.5, and the results are in table 8.1. To test the significance of the parameters, we

	Mean	SD	HPD interval (95%)	H_0	Bayes Factor
α	0.2837	0.09394	0.101 ; 0.4686	$\alpha < 0$	631.9114
α_A	0.07513	0.1573	-0.233 ; 0.3807	$\alpha_A < 0$	2.164958
r	0.2073	0.028	0.1541 ; 0.2636		
θ	7.123	1.525	4.264 ; 10.14		

Table 8.1. Model results, using data for all of the pupils

made use of Bayes factors according to the guidelines in table 3.1. These results suggest that all the pupils saw a significant increase of $e^{0.28}$ in their underlying offence rates, and that the pupils in the active group saw a further, although insignificant, increase in their offence rates of $e^{0.075}$.

We became aware, however, that the random allocation of treatment followed an unlikely pattern. This is visible in the average baseline offence rates, and we wanted to determine how it affected our results, if at all. Half of the pupils were assigned the active treatment at the outset of the study, while the other half were assigned the

placebo. Since the treatment was not administered during the baseline period, we would expect the two groups to demonstrate similar behavioural patterns. This was not the case. The pupils in the active treatment group had an average offence rate[2] of 0.017, while the pupils in the placebo group had an average offence rate of 0.042. The placebo group exhibited an average offence rate 2.5 times higher that the active group even before the treatment period had started. Using bootstrap samples, we determined the probability of witnessing this allocation of pupils at random as 0.0087, or about one in hundred.

8.2 Blood data

Of the 196 children, 53 also agreed to provide blood samples from which the levels of EPA, DHA, total n3, total n6, n6:n3 ratio, and n3 index were measured. The researchers considered EPA to be of interest, and wanted to use it as a covariate in the model to see whether a change in EPA levels could be linked to behaviour. We also tested DHA and total n3, using the hazard rate

$$h(t|\lambda_i) = \begin{cases} \lambda_i & \text{for } t \text{ in baseline period} \\ \lambda_i e^{\alpha+\beta(\Delta Blood_i)} & \text{for } t \text{ in treatment period.} \end{cases} \tag{8.3}$$

In the model defined above, $\Delta Blood_i$ is the increase of the relevant blood-level measurement, for pupil i, calculated as post measurement minus pre measurement. As before, α is used to measure an overall shift in offence rates from the baseline to treatment period, whereas β would capture a movement caused by a change in blood measurements. An overall shift in the offence rates from one period to the next would be e^{α}, and a movement due to a change in the blood measurement would be equal to $e^{\beta(\Delta Blood)}$, where $\Delta Blood = Bloodpre - Bloodpost$.

[2]Calculated as $^{\text{number of offences}}/_{\text{number of days in period.}}$

The results indicated no significant effect of a change in the underlying blood concentration on the observed behaviour of pupils in the study. This is in accordance with the earlier results, since the pupils in the active group — who received nutritional supplements — saw their blood concentrations significantly change for the better. Being in the active group can thus be used as a proxy variable for having one's blood measurements significantly altered. Pupils in the placebo group did not experience a significant change in their blood concentrations. This held true for DHA, EPA, and total n3. It was also the case when we changed the hazard rate to

$$h(t|\lambda_i) = \begin{cases} \lambda_i e^{\beta(Bloodpre_i)} & \text{for } t \text{ in baseline period} \\ \lambda_i e^{\alpha+\beta(Bloodpost_i)} & \text{for } t \text{ in treatment period,} \end{cases} \tag{8.4}$$

where $Bloodpre_i$ and $Bloodpost_i$ were blood measurements taken at the beginning and end of the treatment period, respectively. The hazard rate in (8.4) links the offence rate during a period to the blood measurement assumed to be in effect over that period.

Up to this point, the conclusion seemed clear: nutritional supplements had no effect on the observed behaviour. We fitted another model, however, with the hazard rate

$$h(t|\lambda_i) = \begin{cases} \lambda_i e^{\beta(Bloodpre_i)} & \text{for } t \text{ in baseline period} \\ \lambda_i e^{\alpha_A(A_i)+\beta(Bloodpost_i)} & \text{for } t \text{ in treatment period,} \end{cases} \tag{8.5}$$

which raised some concerns. This model included a parameter to measure the effect of receiving the treatment, as well as the effect of the blood concentration. The results are given in table 8.2, and have troublesome implications. The positive α_A indicates an increase in offence rate, meaning that those who received the active treatment increased their offence rates during the treatment period. The negative parameter estimate of β indicates that a higher EPA concentration was linked to a lower offence rate. Since only the pupils in the treatment group showed significant changes in

	Mean	SD	HPD interval (95%)	H_1	Bayes Factor
α_A	1.677	0.4743	0.7484 ; 2.6	$\alpha_A > 0$	> 10000
β	-2.35	1.019	-4.326 ; -0.3263	$\beta < 0$	193.5525
r	0.2401	0.05787	0.1346 ; 0.3556		
θ	2.3	1.582	0.2427 ; 5.409		

Table 8.2. Model Results, using data for all of the pupils

their blood concentrations, this result is counter-intuitive. The same results were obtained using the change in blood concentration, $\Delta Blood_i$, instead of the pre and post measurements.

This result suggests that the behaviour of children on the active treatment worsened, but that seeing a large change in blood concentrations had a protective effect. We concluded that the result should be considered with caution. The sample of 53 children who provided blood samples was heavily influenced by the unlucky randomisation, and results were mostly driven by a few outliers. Pupils in the placebo group started out with double the average daily offences of the active group, as shown in table 8.3. Had the active and placebo groups started out with the same baseline offence rates, we would perhaps have been able to draw clearer, more trustworthy conclusions.

	Active	Placebo	Total
Number of pupils	98	98	196
Events	106 - 124	254 - 233	360 - 357
Mean events/day	0.02 - 0.03	0.06 - 0.07	0.04 - 0.05

Table 8.3. Table showing number of pupils in each category, the number of events recorded for each period, and the mean daily offence rates. Numbers given as *Baseline - Treatment*.

8.3 Randomisation and bias

The unfortunate randomisation should not affect the parameter estimation, as the model does not depend on any assumptions regarding effective randomisation. To

demonstrate that the parameter estimates remain unbiased, we simulated 100 datasets using proper allocation and worst case scenario allocation. Then, we used our model to see how well we could estimate the true parameters under both scenarios. 'Proper allocation' assigned the treatments to the children randomly, 'worst case scenario' assigned the 98 highest offending children to the placebo group. The latter usually led to very few events occurring in the treatment group. We see in figures 8.2 and 8.3 that the parameter estimates remained unbiased regardless of how the treatment is allocated. In the worst case scenario simulation, however, we see that the uncertainty around the α_A parameter is larger than proper allocation simulation. This makes sense, since the proper allocation results in more events occurring in the active group, giving us more information with which to estimate the parameters associated with that group.

The implications we faced during this study caused by the unlucky randomisation serve as a reminder that experimental design is crucial. It is crucial that the specific questions to be addressed by the study be identified and then that the sampling or treatment allocation scheme be designed accordingly. Though completely random treatment allocation — as used in this study — might be regarded as the safe option when it comes to explaining to non-technical reviewers, and though the probability of unlucky randomisation is small, when it does happen it can lead to more uncertainty around the parameters. The randomisation could have been done by aligning the pupils according to the number of baseline offences, then grouping them two-by-two and randomly assigning placebo or treatment to each pair of students. This would have kept the integrity of randomisation intact, while ensuring that the baseline and treatment groups started out with similar offence rates. If the model is correct, there is no bias. However, problems with the randomisation cause the results to be more sensitive to model failure.

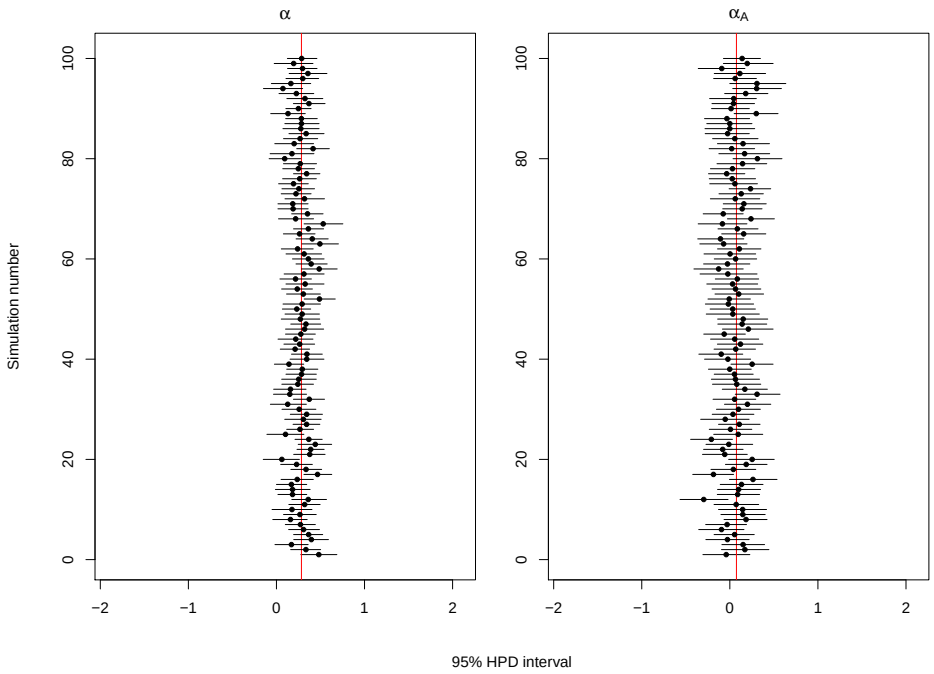


Figure 8.2. Proper allocation. Posterior mean (dots) and the 95% HPD intervals (lines) of the α and α_A parameters, 100 simulations. The red line shows the true parameter value.

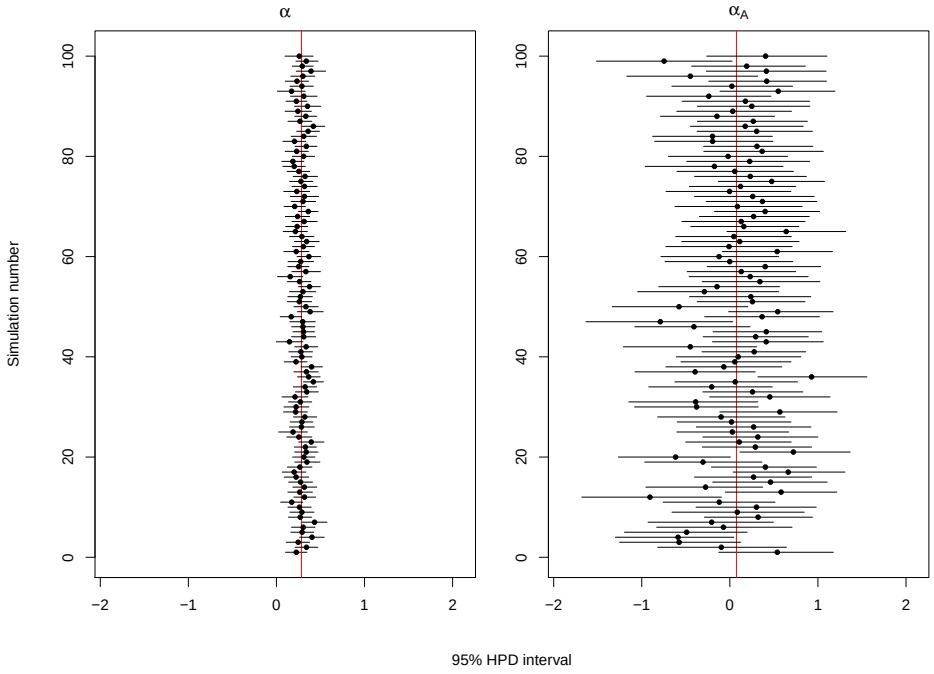


Figure 8.3. Worst case scenario. Posterior mean (dots) and the 95% HPD intervals (lines) of the α and α_A parameters, 100 simulations. The red line shows the true parameter value. This figure has the same scale as figure 8.2.

8.4 Time-dependent hazard rate

During the study, we also fitted a model with time-dependent hazard rates of the form

$$h(t|\lambda_i) = \begin{cases} \lambda_i & \text{for } t \text{ in baseline period} \\ \lambda_i e^{\alpha+\alpha_A(A_i)+t\left(\beta+\beta_A(A_i)\right)} & \text{for } t \text{ in treatment period.} \end{cases} \tag{8.6}$$

This is similar to 8.2, but includes both an intercept (α and α_A) and slope (β and β_A) for the rate movements of the active and placebo group. The rationale behind this was the expectation that the nutritional supplements would have an effect that grows over time as the concentrations of the vitamins in the body steadily rises. This model could not be handled by JAGS, but we were able to modify the code used in chapter 4 to estimate its parameters. Once again, the simpler model provided a stepping stone wherein we could corroborate the output of our custom algorithm with that of JAGS, ensuring bug-free code. The time-dependent hazard rate could be added to our program without affecting the overall structure of the program. Results of this time-dependent hazard rate were nonsensical, however, and suggested overfitting. We did not pursue it further. The exercise did test the flexibility of our custom code, however, proving it useful beyond the original reasons for which it was designed.

8.5 An argument for Bayes

Hilbe (2011) mentioned that, while the univariate negative binomial regression enjoys support in most standard software packages, the same does not hold for the bivariate version. A researcher wishing to analyse data with two repeated counts per subject is burdened with writing custom code for parameter estimation, as we saw in the study by Gesch et al. (2002). In addition to being concise, the JAGS code in appendix B.5 also reveals the exact model structure, thanks to the declarative nature of the

JAGS and BUGS language. Thus, the code can easily be included in an article while providing enough detail to allow understanding and reproduction of the methods used.

The ease of implementation also results in flexibility. Where it is usually a daunting task to extend from the univariate negative binomial model to bivariate case due to the amount of coding and debugging involved, this is not the case with JAGS. In addition, we can even analyse a model with three or more counts,

```
model{
for (i in 1:N) {

  count_in_period_1[i] ~ dpois(rate_in_period_1[i])
  count_in_period_2[i] ~ dpois(rate_in_period_2[i])
  count_in_period_3[i] ~ dpois(rate_in_period_3[i])
  # etc.

  rate_in_period_1[i] <- # link to i's covariates
  rate_in_period_2[i] <- # link to i's covariates
  rate_in_period_3[i] <- # link to i's covariates
  # etc.
  }
# Priors of covariate parameters
}
```

transcending the abilities of the bivariate negative binomial distribution. We can easily add more covariates, such as the blood data.

During the school study we saw some large outliers, suggesting that the variance among individuals may even be larger than the Gamma distribution can explain. We did not consider other distributions to take account of the heterogeneity, but JAGS and BUGS allows this possibility. Our arguments towards adopting a Bayesian approach centres around the ease of parameter estimation, rather than philosophical arguments. Though some authors may find this depressing (Crowder, 2012, c. 3), the openness and ease of use provided by MCMC software certainly has benefits. The JAGS and BUGS developers placed sampling-based inference within reach of all relevant fields. There is no over-emphasising this. The brevity and simplicity of the code in appendix B.5, with which we can fit a model that previously had implementation complexity

as a barrier of entry is a victory for the software. The model is likelihood based and therefore not Bayesian nor frequentist by design, but the readily available estimation procedures of the Bayesian method alongside the ease of reproduction which allows for scrutiny of results should not be overlooked.

In this chapter, we used a parametric model. We mentioned the counter-intuitive results that occurred when using the hazard rate in (8.5), and that unlucky randomisation could have contributed. It is more likely, however, that the result was due to model failure. We made the assumption that events arrived according to a Poisson process, accounting for heterogeneity by allowing rates to follow a Gamma distribution. Despite this, most of the pupils had zero events in both periods — more than could be explained by our parametric model. In retrospect it would have been better to follow an approach that did not make such strong parametric assumptions. For instance, we could compare the distribution of event times in the treatment and placebo groups using a Kolmogorov-Smirnov test. Figure 8.4 shows a QQ-plot of the event times of the placebo and treatment groups. A Kolmogorov-Smirnov test with the H_0 that they are from the same distribution produces a p-value of 0.02. This is of interest in the study; if the treatment is effective we would expect it to take some time for the supplements to work, leading to different distributions of event times for the treatment and placebo groups, which is indeed the case in 8.4. The same tests can be done for the number of offences for each period in each group, without making any specific assumptions on their distributions. It would have been best to start simple and investigate the hypotheses using the least amount of assumptions, and to only add complexity to the model if necessary.

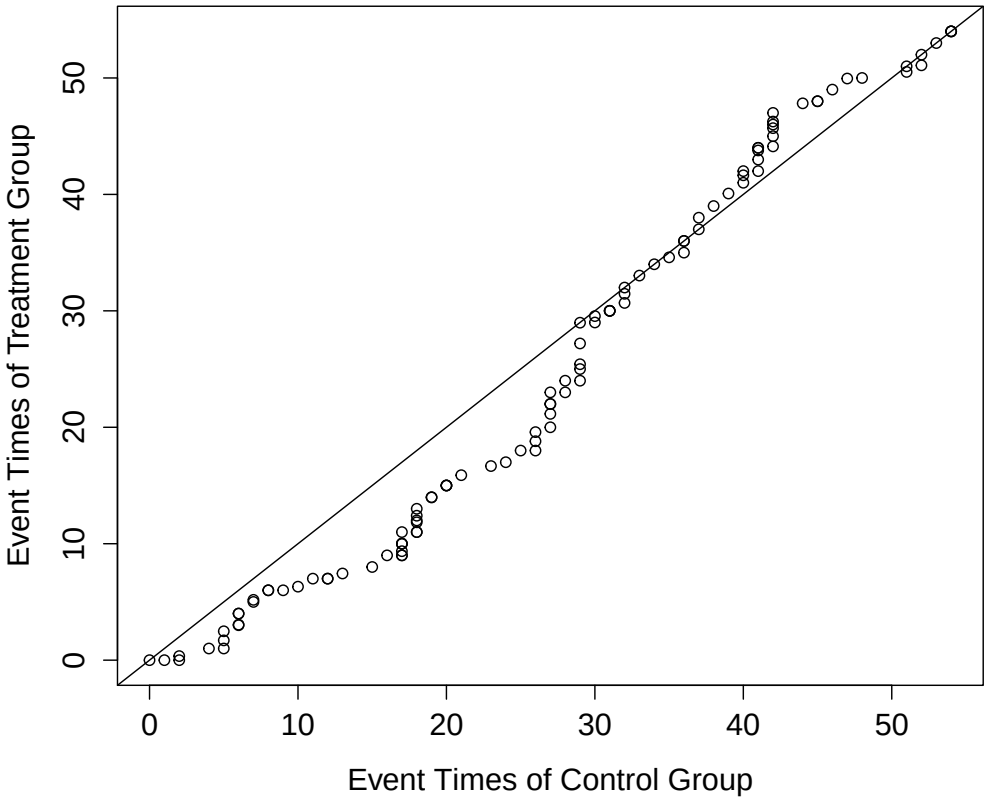


Figure 8.4. Quantile-quantile plot of the even times in the placebo group vs the treatment group.

Conclusion

Rothwell's question about the importance of longitudinal volatility on stroke risk inspired this thesis. In pursuit of an answer, we investigated joint survival models. The literature on these models provided a solid foundation, but none measured the effect of longitudinal volatility on the survival outcome. We therefore developed our own.

Although there were packages available to fit joint models in R, such as `JM` or `joineR`, these were developed for specific cases and we were unable to use them to measure the importance of longitudinal volatility. Guo and Carlin (2004) provided a more flexible option for parameter estimation, using BUGS. We were uncertain, however, about the code they used to estimate parameters. Even after adapting the code provided by Michael Sweeting and obtaining results, we were unsure about the exact workings of the program — particularly the use of BUGS tricks. The only way to control the internal sampling and ensure that the code was doing what we needed, was to write our own MCMC algorithm. Therefore, we scrutinised the working of BUGS and JAGS and wrote our own sampler in C++, chosen for its execution speed. This also allowed us to test different samplers. Although we were required to consider and investigate elements of the C++ language in detail, we do not discuss it. The choice of programming language and its exact implementations are independent of the model and its assumptions.

Once we had an algorithm that produced results, we needed a method to test

goodness of fit. The literature mentioned graphical checks and posterior predictive model checking, but it was necessary for us to lay out a detailed strategy. We combined the suggestions of Henderson et al. (2000, 2002) to use posterior predictive model checking and Diggle et al. (2008) to use graphical model checks, to form a complete strategy for goodness-of-fit testing in joint models. Our method used the slope of the Nelson-Aalen estimator, and it can be extended to more complex joint survival models.

The double-blind placebo-controlled study was unrelated to our original question and we originally intended to mention it only briefly in the thesis. However, we spent considerable time investigating the model and therefore we included the details in chapter 8, keeping it separate from the work on joint survival models. It was a splendid test for the flexibility of our code.

Like the UK-TIA dataset, the NHANES data had blood pressure and survival component, but the blood pressure observations of the latter did not technically constitute a longitudinal process, since they were all collected at the start of the study. Despite this, we could still use joint models of longitudinal and time-to-event data to analyse it. A distribution can be fitted to blood pressure observations whether they arrive over time throughout life, or whether they are collected at the start of the study. We therefore continued using the term 'longitudinal' in chapter 7, to display the parallels with our model from chapter 4, and to avoid confusion. We experimented with multiple extensions to our model, but not all of them had a place in this thesis. One example was our attempt to allow for longitudinal effects on survival to decay over time. Another was assigning subjects with an individual variance as well as a global measurement error, making the total longitudinal variance $\sigma_i^2 + \sigma_\varepsilon^2$ for a subject. These extensions did not yield useful insights. Other extensions include competing risks and the baseline hazard rate, which could be added to our model with minimal effort since they only required changing the hazard rate in the algorithm.

They represent the relaxation of crucial assumptions, which may not hold in practice, therefore we included them in chapter 7. Our final extension allowed a bivariate longitudinal process and it changed the distributional assumptions, requiring us to modify the MCMC sampler. This was an important addition, as it would have allowed us to determine which of the two blood pressure processes were most important for measuring event risk. Using conjugate distributions provided convenient proposal distributions allowing us to sample with the Metropolis-Hastings technique from Wang and Taylor (2001).

A simulation study revealed, however, that the algorithm was producing biased parameter estimates. We were unable to find a solution to this problem within the time frame of this thesis, but provide the details of our model to highlight this caveat of which future researchers should be aware. This bias only occurred when we simulated datasets with a correlation between μ_{i1} and μ_{i2}. Thus, we can successfully apply our model to analyse multiple independent longitudinal processes linked to a survival outcome.

We identify several ventures for future research that presented itself during this thesis.

1. In chapter 8 we examined the difference between the bivariate negative binomial regression and our Poisson model with Gamma heterogeneity. Future research can look at the detailed differences between our model and the different variants of the bivariate negative binomial model, and determine when one should be chosen in favour of another.

2. The specification of our longitudinal process in chapter 4 was simple, and it helped us establish the adequacy of our estimation algorithms. Future researchers can extend the specification of the longitudinal process. An example in cases where high-frequency data are available, is a model with stochastic volatility, with the stochastic volatility linked to the survival.

3. We identified the inability of our algorithm from chapter 7 to recover the correct parameters in our simulation study. Future research should determine whether this is a problem with the MCMC algorithm, or an identifiability problem with the model itself, and whether it can be amended.

4. In this thesis we used general sampling methods, such as ARMS and the Slice sampler, and we also investigated the Metropolis-Hastings sampler for joint survival models, as presented in Wang and Taylor (2001). The current joint survival literature usually defaults to using ARS or ARMS. The circumstances under which one sampling method should be preferred in favour of another should be investigated. In chapter 4, the Slice sampler performed best, but in the Metropolis-Hastings sampler was the only feasible method in chapter 7. We believe that, in a model with enough data to fit stochastic volatility to the longitudinal process, the Metropolis-Hastings sampler will perform better, but this should be formally investigated.

Bibliography

Aalen, O., O. Borgan, and H. Gjessing (2008). *Survival and Event History Analysis: A Process Point of View*. Statistics for Biology and Health. Springer.

Bergstrom, T. and M. Bagnoli (2005). Log-concave probability and its applications. *Economic theory 26*(2), 445–469.

Breslow, N. E. (1972). Contribution to the discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (The Statistician) 47*(1), 69–100.

Brooks, S., A. Gelman, G. L. Jones, and X. Meng (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC.

Brown, E., J. Ibrahim, and V. DeGruttola (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics 61*(1), 64–73.

Brown, R. E. and G. J. Ibrahim (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics 59*(2), 221–228.

Carey, J. R., P. Liedo, H. G. Müller, J. L. Wang, and J. M. Chiou (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of mediterranean fruit fly females. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 53*(4), B245–B251.

Chi, Y. and J. Ibrahim (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics 62*(2), 432–445.

Clarke, R., M. Shipley, S. Lewington, L. Youngman, R. Collins, M. Marmot, and R. Peto (1999). Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *American Journal of Epidemiology 150*(4), 341–353.

Congdon, P. (2003). *Applied Bayesian Modelling*. John Wiley & Sons Inc.

Cox, D. and D. Oakes (1984). *Analysis of Survival Data*. Chapman and Hall.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological) 30*(2), 248–275.

Crowder, M. (2012). *Multivariate Survival Analysis and Competing Risks*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

De Gruttola, V. and X. M. Tu (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics 50*(4), 1003–1014.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Diggle, P. J., I. Sousa, and A. G. Chetwynd (2008). Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture. *Statistics in Medicine 27*(16), 2981–2998.

Ding, J. and J. Wang (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics 64*(2), 546–556.

Farrell, B., J. Godwin, S. Richards, and C. Warlow (1991). The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: final results. *Journal of Neurology, Neurosurgery & Psychiatry 54*(12), 1044–1054.

Faucett, C. L. and D. C. Thomas (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine 15*(15), 1663–1685.

Gavrilov, L. A. and N. S. Gavrilova (2011). Mortality measurement at advanced ages: a study of the social security administration death master file. *North American actuarial journal 15*(3), 432–447.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review 71*(2), 369–382.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis 1*(3), 515–533.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis, Second Edition*. Taylor & Francis.

Gelman, A., Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 49*(2), 247–268.

Gelman, A. and X.-L. Meng (1996). Model checking and model improvement. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in Practice*, Chapter 11, pp. 189–201. Chapman & Hall/CRC.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica 6*(4), 733–760.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), 721–741.

Gesch, C., S. Hammon, S. Hampson, A. Eves, and M. Crowder (2002). Influence of supplementary vitamins, minerals and essential fatty acids on the antisocial behaviour of young adult prisoners. *British Journal of Psychiatry 181*, 22–28.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, pp. 169–193. University Press.

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones, and X. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, pp. 3–48. Chapman & Hall/CRC.

Gilks, W. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J. Bernardo, j. Berger, A. Dawid, and A. Smith (Eds.), *Baysian Statistics 4*, pp. 641–649. Oxford University Press, Oxford.

Gilks, W. (1996). Full conditional distributions. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, Chapter 5, pp. 75–88. Chapman & Hall/CRC.

Gilks, W., N. Best, and K. Tan (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics 44*(4), 455–472.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996a). *Markov Chain Monte Carlo in Practice*. Taylor & Francis.

Gilks, W., S. Richardson, and D. J. Spiegelhalter (1996b). Introducing Markov chain Monte Carlo. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, Chapter 1, pp. 1–19. Chapman & Hall/CRC.

Gilks, W. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics 41*(2), 337–348.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London 115*, 513–583.

Greenwood, M. and J. O. Irwin (1939). The biostatistics of senility. *Human Biology 11*, 1–23.

Guo, W., S. J. Ratcliffe, and T. T. T. Have (2004). A random pattern-mixture model for longitudinal data with dropouts. *Journal of the American Statistical Association 99*(468), 929–937.

Guo, X. and B. P. Carlin (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician 58*(1), 16–24.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Henderson, R., P. Diggle, and A. Dobson (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics 1*(4), 465–480.

Henderson, R., P. Diggle, and A. Dobson (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics 3*(1), 33–50.

Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press.

Hoffman, M. D. and A. Gelman (2014). The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*, 1593–1623.

Hogan, J. W. and N. M. Laird (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine 16*(3), 239–257.

Hogan, J. W. and N. M. Laird (1998). Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research 7*(1), 28–48.

Howard, S. and P. Rothwell (2003). Regression dilution of systolic and diastolic blood pressure in patients with established cerebrovascular disease. *Journal of Clinical Epidemiology 56*(11), 1084–1091.

Hsieh, F., Y. Tseng, and J. Wang (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics 62*(4), 1037–1043.

Hu, P., A. Tsiatis, and M. Davidian (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics 54*(4), 1407–1419.

Hughes, M. D. (1993). Regression dilution in the proportional hazards model. *Biometrics 49*(4), 1056–1066.

Ibrahim, J., M. Chen, and D. Sinha (2005). *Bayesian Survival Analysis*. Springer.

Iwasaki, M. and H. Tsubaki (2006). Bivariate negative binomial generalized linear models for environmental count data. *Journal of Applied Statistics 33*(9), 909–923.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley.

Jeffreys, H. (1961). *The Theory of Probability* (third ed.). Oxford: Clarendon Press.

Klein, J. and M. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.

Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics 35*(1), 139–156.

Lange, N., B. Carlin, and A. Gelfand (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *Journal of the American Statistical Association 87*(419), 615–626.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics 15*(3), 209–225.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association 90*(431), 1112–1121.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine 28*(25), 3049–3067.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing 10*(4), 325–337.

MacMahon, S., R. Peto, R. Collins, J. Godwin, J. Cutler, P. Sorlie, R. Abbott, J. Neaton, A. Dyer, and J. Stamler (1990). Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *The Lancet 335*(8692), 765–774.

Makeham, W. M. (1860). On the law of mortality and the construction of annuity tables. *The Assurance Magazine, and Journal of the Institute of Actuaries 8*(6), 301–310.

Marshall, A. W. and I. Olkin (1985). A family of bivariate distributions generated by the bivariate Bernoulli distribution. *Journal of the American Statistical Association 80*(390), 332–338.

McCullagh, P. (2013). Survival models and health sequences. *arXiv preprint*.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*, 1087.

Metropolis, N. and S. Ulam (1949). The Monte Carlo method. *Journal of the American Statistical Association 44*(247), 335–341.

Mostafa, A. A. and A. B. Ghorbal (2011). Using WinBUGS to Cox model with changing from the baseline hazard function. *Applied Mathematical Sciences 5*(45), 2217–2240.

National Center for Health Statistics (NCHS) (2013). Centers for disease control and prevention (CDC).

Neal, R. M. (2003). Slice sampling. *Annals of Statistics 31*(3), 705–741.

Neuhaus, A., T. Augustin, C. Heumann, and D. Daumer (2009). A review on joint models in biometrical research. *Journal of Statistical Theory and Practice 3*(4), 855–868.

Norris, P. R., S. M. Anderson, J. M. Jenkins, A. E. Williams, and J. A. Morris Jr (2008). Heart rate multiscale entropy at three hours predicts hospital mortality in 3,154 trauma patients. *Shock 30*(1), 17–22.

Ostchega, Y., R. J. Prineas, R. Paulose-Ram, C. M. Grim, G. Willard, and D. Collins (2003). National health and nutrition examination survey 1999-2000: effect of observer training and protocol standardization on reducing blood pressure measurement error. *Journal of Clinical Epidemiology 56*(8), 768–774.

Pawitan, Y. and S. Self (1993). Modeling disease market processes in AIDS. *Journal of the American Statistical Association 88*(423), 719–726.

Peduzzi, P., T. Holford, K. Detre, and Y.-K. Chan (1987). Comparison of the logistic and Cox regression models when outcome is determined in all patients after a fixed period of time. *Journal of chronic diseases 40*(8), 761–767.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Plummer, M. (2012). *rjags: Bayesian graphical models using MCMC*. R package version 3-7.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika 69*(2), 331–342.

R Core Team (2012). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raboud, J., N. Reid, R. Coates, and V. Farewell (1993). Estimating risks of progressing to AIDS when covariates are measured with error. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 156*(3), 393–406.

Raftery, A. E. and S. M. Lewis (1992). Practical Markov chain Monte Carlo: comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statistical Science 7*(4), 493–497.

Ripley, B. (2009). *Stochastic Simulation*. John Wiley & Sons.

Ripley, B. D. (1979). Algorithm AS 137: simulating spatial patterns: dependent samples from a multivariate density. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*(1), 109–112.

Rizopoulos, D. (2010). JM: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software 35*(9), 1–33.

Rizopoulos, D., G. Verbeke, and E. Lesaffre (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 637–654.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods*. Springer-Verlag New York Inc.

Roberts, G. and L. Sangalli (2010). Latent diffusion models for survival analysis. *Bernoulli 16*(2), 435–458.

Rothwell, P., S. Howard, E. Dolan, E. O'Brien, J. Dobson, B. Dahlöf, P. Sever, and N. Poulter (2010). Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *The Lancet 375*(9718), 895–905.

Rothwell, P. M. (2010). Limitations of the usual blood-pressure hypothesis and importance of variability, instability, and episodic hypertension. *Lancet 375*(9718), 938–948.

Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine 11*(14-15), 1861–1870.

Scollnik, D. P. (2000). Actuarial modeling with MCMC and BUGS: additional worked examples. In *Actuarial Research Clearing House*.

Sinha, D., J. G. Ibrahim, and M. Chen (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika 90*(3), 629–641.

Song, X., M. Davidian, and A. Tsiatis (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics 58*(4), 742–753.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn (2004). *WinBUGS user manual*. MRC Biostatistics Unit, Institute of Public Health.

Spiegelhalter, D. J. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 47*(1), 115–133.

Stan Development Team (2014a). RStan: the R interface to Stan, version 2.3.

Stan Development Team (2014b). Stan: a C++ library for probability and sampling, version 2.3.

Stan Development Team (2014c). *Stan modeling language users guide and reference manual, version 2.3*.

Steinsaltz, D., G. Mohan, and M. Kolb (2012). Markov models of aging: theory and practice. *Experimental Gerontology*.

Stern, H. S. and S. Sinharay (2005). Bayesian model checking and model diagnostics. In D. Dey and C. Rao (Eds.), *Bayesian Thinking: Modeling and Computation*, Volume 25, pp. 171–192. Elsevier.

Stroustrup, B. (2000). *The C++ Programming Language*. Addison-Wesley.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer.

Thomas, A., B. O'Hara, U. Ligges, and S. Sturtz (2006). Making BUGS open. *R News 6*(1), 12–17.

Tseng, Y., F. Hsieh, and J. Wang (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika 92*(3), 587–603.

Tsiatis, A. A. and M. Davidian (2004). Joint modelling of longitudinal and time-to-event data: an overview. *Statistica Sinica 14*(3), 809–834.

Tsiatis, A. A., V. DeGruttola, and M. S. Wulfsohn (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association 90*(429), 27–37.

Wabersich, D. and J. Vandekerckhove (2014). Extending JAGS: a tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods 46*(1), 15–28.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology 33*(1), 79–86.

Wang, Y. and J. Taylor (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association 96*(455), 895–905.

Wei, G. C. G. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*(411), 699–704.

Wu, M. C. and R. J. Carroll (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics 44*(1), 175–188.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics 53*, 330–339.

Xu, J. and S. Zeger (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(3), 375–387.

Appendix A

Log-concavity

A.1 Comments on log-concavity of densities

A density $f(x)$ is said to be log-concave if (Bergstrom and Bagnoli, 2005)

$$\frac{d^2 \log f(x)}{dx^2} \leq 0.$$

When using the ARS algorithm we only need to know the relevant density up to a normalising constant. Our joint posterior likelihood (4.6), broken up into the following parts:

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) \propto \overbrace{\prod_{i=1}^{N}\left(\prod_{j=1}^{n_i} f(y_{ij}|\mu_i,\tau_i)\right)}^{\text{longitudinal}} \overbrace{f(v_i,\delta_i|\mu_i,\tau_i,\boldsymbol{\beta})}^{\text{survival}} \times$$
$$\underbrace{f(\mu_i|m,\tau)f(\tau_i|r,\lambda)f(\boldsymbol{\beta})f(m)f(\tau)f(r)f(\lambda)}_{\text{priors}}$$

When the individual parts — in our case the survival and longitudinal — are log-concave, then the entire posterior will be log-concave. This is because the sum of log-concave functions is also a log-concave function. When we need to sample a specific parameter, the functions not containing it will make up part of the normalising

constant. Thus, if we choose our model in such a way that the different parts have log-concave densities, we will be able to sample using the ARS algorithm. Since it will be easy to specify a log-concave distribution for the longitudinal part, we need to supply a hazard function that would make the survival part log-concave in the parameters. From (4.4):

$$\begin{aligned}\log f(v_i, \delta_i|\mu_i, \tau_i, \boldsymbol{\beta}) &= \log\left(\prod_{i=1}^{N}[h(v_i)]^{\delta_i}\exp\left\{-\int_0^{v_i}h(t)dt\right\}\right)\\ &= \sum_{i=1}^{N}\delta_i\log\left[h(v_i)\right] - \sum_{i=1}^{N}\int_0^{v_i}h(t)dt \qquad \text{(A.1)}\end{aligned}$$

and we need $\frac{d^2\log f(\boldsymbol{v}|\mu_i,\tau_i,\boldsymbol{\beta})}{d\beta_i^2} \leq 0$ in β_i, which is the parameter of interest. Since δ_i will either be 0 or 1, (A.1) will not affect the log-concavity of the posterior as a whole if both of the following conditions hold:

- $h(v_i)$ is log-concave in the parameters
- The integrated hazard $H(v_i) = \int_o^{v_i} h(t)dt$ has a second derivative with respect to the parameters that is either positive or zero.

If these two conditions hold and the relevant prior distribution is log-concave in the parameters, we can use the ARS sampler. Bergstrom and Bagnoli (2005) provide useful tables containing most of the standard distribution functions and their log-concave characteristics with respect to parameters. Since these conditions hold for our hazard rate, it is log-concave. Furthermore, all of our conditional posterior densities (4.10 – 4.16) are either conjugate or log-concave, making sampling possible through either the use of standard sampling methods (Ripley, 2009) or the ARS sampler.

Appendix B

Code

B.1 JAGS code: joint survival model

```
data{
      for (i in 1:N){
            ones[i] <- 1
      }
}

model{

for (i in 1:N) {
    for (j in 1:M[i]) {
       X[i, j] ~ dnorm(mu[i],tau[i])
    }
            tau[i] ~ dgamma(r,lambda)
            mu[i] ~ dnorm(m,theta)
    #-------------------------------------------------------#
    # Code that allows the hazard to be related directed    #
    # to intercept, slope, and underlying longitudinal value  #
    # To specify the likelihood, we use the ones trick      #
    #-------------------------------------------------------#

     q[i]<-L[i]/C
     ones[i]~dbern(q[i])
    # Likelihood for survival data
    # event=1 if event, and 0 if censoring
     L[i]<-pow(h[i],event[i])*S[i]

    # Hazard for individual i at their survival time
     h[i] <- exp(B + (B.mu*mu[i]/100) + B.tau*tau[i])
```

```
    # Cumulative hazard H[t] = int_0^t (h[u] du)
    # V[i] is the survival/event time
     H[i] <- V[i]*h[i]
    # Survival probability for individual i
    # at their survival time
     S[i]<-exp(-H[i])
    # Density function
     f[i]<-h[i]*S[i]

   }

C<- 10000000 # part of trick

#priors
#Parameters of interest
 B~dnorm(0,1.0E-6)
 B.mu~dnorm(0,1.0E-6)
 B.tau~dnorm(0,1.0E-6)

#initial parameters for fitting the model
 r~dgamma(0.4370*0.001,0.001)
 lambda~dgamma(1707*0.00001,0.00001)
 m~dnorm(20.0,0.000001)
 theta~dnorm(1.0E-3,0.0000001)
}
```

B.2 Stan code: joint survival model

```
data{
int<lower=0> N;
int<lower=0> M[N];
real X[N,23];
real<lower=0> V[N];
int<lower=0,upper=1> event[N];
}

parameters{
real<lower=0> tau[N];
real mu[N];
real<lower=0> r;
real<lower=0> lambda;
real m;
real<lower=0> tau_top;
real B ;
real B_mu ;
real B_tau ;
}


model{
for(i in 1:N){
   real logL[N];
   real logh[N];
   real H[N];

    for (j in 1:M[i]){
       X[i, j] ~ normal(mu[i], 1/sqrt(tau[i]) );
    }
    tau[i] ~ gamma(r,lambda);
    mu[i] ~ normal(m, 1/sqrt(tau_top) );



    ## Hazard for individual i at their survival time
    logh[i] <- B + (B_mu*mu[i]/100) + B_tau*tau[i];

    ## Cumulative hazard H[t] = int_0^t (h[u] du)
    ## V[i] is the survival/event time
    H[i] <- V[i]*exp( logh[i] );


    ## Likelihood for survival data - event=1 if event, and 0 if censoring
    logL[i] <- event[i]*logh[i] - H[i];
```

```
   # this replaces the JAGS trick
   increment_log_prob(logL[i]);

   }

   #priors
# #Parameters of interest
 B     ~ normal(0,10000);
 B_mu  ~ normal(0,10000);
 B_tau ~ normal(0,10000);

#initial parameters for fitting the model
        r~gamma(0.0001,0.0001);
        lambda~gamma(0.0001,0.0001);
        m~normal(0,10000);
        tau_top~gamma(0.0001,0.0001);
}
```

B.3 JAGS code: bivariate longitudinal process

```
# declare variables used
# This is necessary for multivariate distributions
var X[N,10,2] , N , M[N], prec[N,2,2], V[N], event[N],
lambda[2,2], m_prior_mean[2], m_prior_prec[2,2], mu[N,2],
prec_top_V0[2,2], m[2] , prec_top[2,2]


data{

        for (i in 1:N){
                ones[i] <- 1
        }

# initialize values
r <- 3
lambda[1,1] <- 0.041389997
lambda[1,2] <- 0
lambda[2,1] <- 0
lambda[2,2] <- 0.053251636

m_prior_mean[1] <- 0
m_prior_mean[2] <- 0

m_prior_prec[1,1] <- 0.00001
m_prior_prec[1,2] <- 0
m_prior_prec[2,1] <- 0
m_prior_prec[2,2] <- 0.00001

prec_top_n0 <- 3
prec_top_V0[1,1] <- 0.0001
prec_top_V0[1,2] <- 0
prec_top_V0[2,1] <- 0
prec_top_V0[2,2] <- 0.0001

}

model{

for (i in 1:N) {
    for (j in 1:M[i]) {
       X[i, j,] ~ dmnorm(mu[i,],prec[i,,])
    }
                prec[i,,] ~ dwish(lambda, r)
                mu[i,] ~ dmnorm( m , prec_top )
```

```
    # Code that allows the hazard to be related directed to
    # intercept, slope, and underlying longitudinal value

    # To specify the likelihood, we use the ones trick

    q[i]<-L[i]/C
    ones[i]~dbern(q[i])
    ## Likelihood for survival data:
    #   event=1 if event,
    #   and 0 if censoring
    L[i]<-pow(h[i],event[i])*S[i]

    ## Hazard for individual i at their survival time
    h[i] <- exp(  B.mu  *mu[i,1] +
                  B.mu2 *mu[i,2] +
                  B.tau *( 1/sqrt(prec[i,1,1]) ) +
                  B.tau2*( 1/sqrt(prec[i,2,2]) ) )

    ## Cumulative hazard H[t] = int_0^t (h[u] du)
    ## V[i] is the survival/event time
    H[i] <- V[i]*h[i]
    ## Survival probability for individual i at
    #  their survival time
    S[i]<-exp(-H[i])
    ## Density function
    f[i]<-h[i]*S[i]

   }

C<- 10000000

# Priors

# Parameters of interest
B.mu~dnorm(0,1.0E-6)
B.tau~dnorm(0,1.0E-6)

B.mu2~dnorm(0,1.0E-6)
B.tau2~dnorm(0,1.0E-6)

# initial parameters for fitting the model
m ~ dmnorm(m_prior_mean , m_prior_prec)
prec_top ~ dwish(  prec_top_V0 , prec_top_n0)


}
```

B.4 Stan code: bivariate longitudinal process

R code to run the model:

```
N <- 2000
# make the dataset into a 2x10 matrix for each indivudual

X <- array( 0 ,  # fill with 0
          dim = c(N,  # 2000 indivs
                      10,  # rows
                       2# cols
                     ) )
for(i in 1:N)
X[i, , ] <- matrix(x[i,] , #this is a 1x20 col vector
                   10,2 , byrow=TRUE)


library(rstan)

dwlist <- list(X=X , N=N ,  M=nj, event=event, V=T)

#This version does not save an R object, saving memory
stan(file = 'mystan.stan',
      data = dwlist, iter = 20000, chains = 1,
      sample_file="sampfolder/samples.R")
```

The contents of the `mystan.stan` file:

```
data{
int<lower=0> N;
int<lower=0> M[N];
real X[N,10,2];
real<lower=0> V[N];
int<lower=0,upper=1> event[N];
}

transformed data{ # see page 248 of manual
real r;
matrix[2,2] lambda;
vector[2] m_prior_mean;
matrix[2,2] m_prior_prec;
real<lower=0> prec_top_n0;
matrix[2,2] prec_top_V0;

vector[2] Y[N, max(M)];
```

```
#assign the prior matrices. See page 246 of manual
r <- 3;
lambda[1,1] <- 0.041389997;
lambda[1,2] <- 0;
lambda[2,1] <- 0;
lambda[2,2] <- 0.053251636;

m_prior_mean[1] <- 128;
m_prior_mean[2] <- 76;

m_prior_prec[1,1] <- 0.0001;
m_prior_prec[1,2] <- 0;
m_prior_prec[2,1] <- 0;
m_prior_prec[2,2] <- 0.0001;

prec_top_n0 <- 3;
prec_top_V0[1,1] <- 0.0001;
prec_top_V0[1,2] <- 0;
prec_top_V0[2,1] <- 0;
prec_top_V0[2,2] <- 0.0001;

    #convert the multidim array to a matrix of vectors
      for(i in 1:N){
          for (j in 1:M[i]){
          Y[i,j][1] <- X[i,j,1];
          Y[i,j][2] <- X[i,j,2];
          }
       }
}

parameters{
cov_matrix[2] prec[N];
vector[2] mu[N];
# real<lower=0> r;
# cov_matrix[2] lambda;
# since this is also pos. def.
vector[2] m;
cov_matrix[2] prec_top;
 real B_mu1 ;
 real B_mu2 ;
 real B_tau1 ;
 real B_tau2 ;
}


model{
      for(i in 1:N){
```

```
          real logL[N];
          real logh[N];
          real H[N];

          for (j in 1:M[i]){
          Y[i,j] ~ multi_normal_prec(mu[i] , prec[i] );
          }
          prec[i] ~ wishart(r,lambda);
          mu[i] ~ multi_normal_prec(m, prec_top );



           ## Hazard for individual i at their survival time
           logh[i] <-  (B_mu1*(mu[i,1] -128)) +
                       (B_mu2*(mu[i,2] -76)) +
                          B_tau1*( 1/sqrt(prec[i,1,1] ) ) +
                          B_tau2*( 1/sqrt(prec[i,2,2] ) ) ;

           ## Cumulative hazard H[t] = int_0^t (h[u] du)
           ## V[i] is the survival/event time
           H[i] <- V[i]*exp( logh[i] );


           ## Likelihood for survival data
           #  event=1 if event, and 0 if censoring
           logL[i] <- event[i]*logh[i] - H[i];


          increment_log_prob(logL[i]);

        }

   #priors
 #Parameters of interest
  B_mu1 ~ normal(0,10000);
  B_mu2 ~ normal(0,10000);
  B_tau1 ~ normal(0,10000);
  B_tau2 ~ normal(0,10000);

#initial parameters for fitting the model
        m~multi_normal_prec( m_prior_mean, m_prior_prec ) ;
        prec_top~wishart( prec_top_n0, prec_top_V0 );
}
```

B.5 JAGS and R code: multiple event arrivals

The following code was used in R, after setting up the data:

```
 library(rjags) # To enable JAGS within R

JAGSlist <- list("N" = dim(tblPupils)[1],
"B_ni" = totalBase,  # vector with total
                     # baseline offences for each pupil
"T_ni" = totalTreat, # vector with total
                     #treatment offences for each pupil
"A" = Active,    #vector: TRUE for active and
                 #        FALSE for placebo
"w1" = w1,       #number of days in baseline period per pupil
"w3"= w3         #number of days in treatment period per pupil
)

#Now for the call to JAGS:
RCS.model <- jags.model("RCSJAGS.r",
                        JAGSlist,n.chains = 1,
                        n.adapt = 100)
RCS.results = coda.samples(dwjags,
                           c("alpha","alphaa","r","theta"),
                           n.iter = 50000)
# we can also use c("alpha","alphaa","r","theta","lambda")
# if we are interested in the lambda_i chain for each pupil


RCS.results <- RCS.results[[1]] #To unlist it
summary(RCS.results)
```

Next, the contents of the `RCSJAGS.r` file:

```
 model{
for (i in 1:N) {

B_ni[i] ~ dpois(w1[i]*lambda[i])
T_ni[i] ~ dpois(w3[i]*lambda[i]*exp(alpha + alphaA*A[i]))

lambda[i] ~ dgamma(r,theta)
}

#priors
alpha~dnorm(0,0.0001)
alphaA~dnorm(0,0.0001)
```

```
r~dgamma(0.0001,0.0001)
theta~dgamma(0.0001,0.0001)
}
```

Appendix C

MCMC convergence and software comparison

This appendix includes details about the MCMC convergence of chains generated by the custom sampler, as well as those produced by JAGS. All of the convergence diagnostics were done in R, using the package `coda` (Plummer et al., 2006) and the functions therein. Information on the convergence tests used in this appendix can be found in the help files of `coda`.

C.1 Convergence diagnostics

The chains to be compared consist of 40,000 iterations each, with starting values chosen after running multiple chains for both samplers, and examining plots of the chains on time.

Geweke's convergence diagnostic

As explained in Geweke (1992), it tests the H_0 of equality of means between the first and last parts of the chain.

- Fraction of chain in first window = 0.1

- Fraction of chain in second window = 0.5

	Custom	JAGS
β_0	0.545	0.020
β_0	0.426	0.961
β_0	0.585	0.999
m	0.505	0.678
τ	0.330	0.677
r	0.679	0.001
λ	0.661	0.000015

Table C.1. Geweke's convergence test, p-values.

Heidelberger and Welch's convergence diagnostic

	Custom sampler			JAGS		
	Stationarity test	Start iteration	p-value	Stationarity test	Start iteration	p-value
β_0	passed	1	0.945	passed	8002	0.3068
β_1	passed	1	0.946	passed	8002	0.3854
β_2	passed	1	0.904	passed	8002	0.1033
m	passed	1	0.669	passed	1	0.0857
τ	passed	1	0.858	passed	1	0.7629
r	passed	1	0.144	failed		0.0315
λ	passed	1	0.130	failed		0.0328
	Halfwidth test	Mean	Halfwidth	Halfwidth test	Mean	Halfwidth
β_0	passed	-4.85e+00	1.43e-01	passed	-4.69e+00	1.78e-01
β_1	passed	1.30e+00	8.17e-02	passed	1.24e+00	9.74e-02
β_2	passed	-1.08e+02	4.29e+00	passed	-1.20e+02	8.36e+00
m	passed	1.45e+02	5.07e-03	passed	1.45e+02	6.25e-03
τ	passed	3.46e-03	1.46e-06	passed	3.47e-03	3.99e-06
r	passed	3.26e+00	1.18e-02			
λ	passed	5.48e+02	2.38e+00			

Table C.2. Heidelberger and Welch's convergence diagnostics for both samplers.

Given the convergence and diagnostic results for the custom sampler we can safely assume that all of our chains were mixing, and that we were drawing samples from the underlying equilibrium distributions. The results show that the custom sampler produced reliable output within 40,000 iterations (excluding burn-in), as it failed none of the convergence tests. The $\boldsymbol{\beta}$ were the slowest to mix, so many more chains would be needed for inference. The run-length diagnostic by Raftery and Lewis (1992) reveals that in order to estimate the 0.025 and 0.975 quantiles, we need to execute between 300,000 and 350,000 iterations to achieve a margin of 0.005 accuracy with 95% probability.

The output from JAGS also shows signs of convergence, though more iterations will be needed since the effective sample sizes are much smaller than those of the custom sampler. Some of the JAGS chains also failed the convergence tests presented

here, and may require many more runs in order to produce trustworthy results. The same conclusion is reached using both of the samplers, but the custom sampler is noticeably faster.

C.2 Computation time

In this section, we look at the performance of the two samplers with regard to computation time. We compared JAGS with the custom sampler for both of the sampling algorithms used — ARS and the Slice sampler. All computation was done using an Intel® Core™2 Duo CPU E8600 3.33GHz, with 2 cores and 4GB of RAM. The operating system used was Linux Fedora 16.

Times given are for 10,000 iterations excluding read-and-write time to file. Ten runs were performed and the average time taken by JAGS was 963 seconds. The average time for the custom sampler using ARS was 499 seconds, whereas the Slice sampler displayed an average time of 214 seconds. Using the smart MH sampler described in section 7.4.3 produced an average time of 376 seconds. In Stan, one run of 10,000 samples took 1 hour and 50 minutes, after a 'warm-up' of 3 hours and 40, giving a total of 5 hours and 30 minutes. Stan took significantly longer to run, but it also produced more effective samples, as shown in table C.4.

C.3 Corroborated results

We present the results as reproduced by JAGS (table C.3) and Stan (table C.4). This is the same model and data that produced tables 4.1 and 4.2 in chapter 4. We only reproduce the results for the SBP, since they should suffice as proof that all samplers corroborated each other.

	Mean	SD	HPD interval (95%)	Effective sample size
β_0	-4.69	0.8264	(-6.32 ; -3.08)	0.0025
β_1	1.29	0.47	(0.35 ; 2.18)	0.0029
β_2	-147	46	(-237 ; -57)	0.011
λ	592	38	(517 ; 667)	0.016
m	145.7	0.41	(144.9 ; 146.5)	0.76
r	3.29	0.18	(2.94 ; 3.64)	0.017
θ	0.0035	0.0001	(0.0033 ; 0.0038)	0.6

Table C.3. JAGS: 50,000 iterations (after burn-in).

	Mean	SD	HPD interval (95%)	Effective sample size
β_0	-4.71	0.85	(-6.37 ; -3.06)	0.5
β_1	1.30	0.48	(0.35 ; 2.21)	0.57
β_2	-148	46	(-238 ; -59)	0.5
r	3.28	0.17	(2.96 ; 3.64)	0.34
λ	589	37	(518 ; 663)	0.33
m	145.7	0.41	(144.9 ; 146.5)	1
τ	0.0035	0.0001	(0.0033 ; 0.0038)	1

Table C.4. Stan: 14,000 iterations (after burn-in).

Appendix D

Law of total variance: decomposition

Here we show the variance decomposition formula with two conditioning random variables, using the law of total variance, as well as the law of total expectation.

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{E}_{X_1}\big(\mathrm{Var}_{Y|X_1}(Y|X_1)\big) + \mathrm{Var}_{X_1}\big(\mathrm{E}_{Y|X_1}(Y|X_1)\big) \\
&= \mathrm{E}_{X_1}\big(\mathrm{E}_{X_2|X_1}\big[\mathrm{Var}_{Y|X_1,X2}(Y|X_1,X_2)|X_1\big]\big) \\
&\quad + \mathrm{E}_{X_1}\big(\mathrm{Var}_{X_2|X_1}\big[\mathrm{E}_{Y|X_1,X2}(Y|X_1,X_2)|X_1\big]\big) \\
&\quad + \mathrm{Var}_{X_1}\big(\mathrm{E}_{Y|X_1}(Y|X_1)\big) \\
&= \mathrm{E}_{X_2}\big[\mathrm{Var}_{Y|X_1,X2}(Y|X_1,X_2)\big] \\
&\quad + \mathrm{E}_{X_1}\big(\mathrm{Var}_{X_2|X_1}\big[\mathrm{E}_{Y|X_1,X2}(Y|X_1,X_2)|X_1\big]\big) \\
&\quad + \mathrm{Var}_{X_1}\big(\mathrm{E}_{Y|X_1}(Y|X_1)\big)
\end{aligned}$$

The above makes use of the expansion

$$\begin{aligned}
\mathrm{Var}_{Y|X_1}(Y|X_1) &= \mathrm{E}_{X_2|X_1}\big[\mathrm{Var}_{Y|X_1,X2}(Y|X_1,X_2)|X_1\big] \\
&\quad + \mathrm{Var}_{X_2|X_1}\big[\mathrm{E}_{Y|X_1,X2}(Y|X_1,X_2)|X_1\big]\,.
\end{aligned}$$

This decomposition is not unique, as it would differ if we chose to condition on X_2 in the first line.

A step-by-step explanation of the workings of (6.16) follows.

$$\begin{aligned}
\mathrm{Var}(X_{i1}) &= \mathrm{E}_{\sigma_i^2}\Big[\mathrm{Var}_{X_{i1}|\sigma_i^2,\mu_i}(X_{i1}|\sigma_i^2,\mu_i)\Big] && \text{(D.1)}\\
&\quad + \mathrm{E}_{\mu_i}\Big(\mathrm{Var}_{\sigma_i^2|\mu_i}\Big[\mathrm{E}_{X_{i1}|\sigma_i^2,\mu_i}(X_{i1}|\sigma_i^2,\mu_i)\Big]\Big) && \text{(D.2)}\\
&\quad + \mathrm{Var}_{\mu_i}\big(\mathrm{E}_{X_{i1}|\mu_i}(X_{i1}|\mu_i)\big) && \text{(D.3)}\\
&= \mathrm{E}_{\sigma_i^2}\left[\frac{\sigma_i^2}{n_i}\right] && \text{(D.4)}\\
&\quad + \mathrm{E}_{\mu_i}\Big[\mathrm{Var}_{\sigma_i^2|\mu_i}(\,\mu_i\,|\mu_i)\Big] && \text{(D.5)}\\
&\quad + \mathrm{Var}_{\mu_i}(\mu_i) && \text{(D.6)}\\
&= \mathrm{E}_{\sigma_i^2}\left[\frac{\sigma_i^2}{n_i}\right] && \text{(D.7)}\\
&\quad + \mathrm{Var}_{\mu_i}(\mu_i) && \text{(D.8)}
\end{aligned}$$

We have (D.4) as a result from (6.13),

$$\mathrm{Var}_{X_{i1}|\sigma_i^2,\mu_i}(X_{i1}|\sigma_i^2,\mu_i) = \frac{\sigma_i^2}{n_i}.$$

Similarly, we have (D.5) also from (6.13), since

$$\mathrm{E}_{X_{i1}|\sigma_i^2,\mu_i}(X_{i1}|\sigma_i^2,\mu_i) = \mu_i$$

and then

$$\mathrm{Var}_{\sigma_i^2|\mu_i}(\,\mu_i\,|\mu_i) = 0$$

since it is the variance of a constant.

The step-by-step explanation of the workings of (6.17) is similar:

$$\begin{aligned}
\mathrm{Var}(X_{i2}) &= \mathrm{E}[\mathrm{Var}(X_{i2}|\sigma_i)] && \text{(D.9)}\\
&\quad + \mathrm{Var}[\mathrm{E}(X_{i2}|\sigma_i)] && \text{(D.10)}\\
&= \mathrm{E}\left[\frac{\sigma_i^2}{n_i-1}\left((n_i-1)-2\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)^2\right)\right] && \text{(D.11)}\\
&\quad + \mathrm{Var}\left[\frac{\sigma_i}{\sqrt{n_i-1}}\sqrt{2}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)\right] && \text{(D.12)}\\
&= \mathrm{E}\big[\sigma_i^2\big]\left(1-\frac{2}{n_i-1}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)^2\right) && \text{(D.13)}\\
&\quad + \mathrm{Var}[\sigma_i]\left(\frac{2}{n_i-1}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right)^2\right). && \text{(D.14)}
\end{aligned}$$

We have (D.11) as a result from (6.14) and the variance of a χ_{n-1} random variable:

$$\left((n-1)-2\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)^2\right).$$

We also have (D.12) from (6.14) and the expected value of a χ_{n-1} random variable:

$$\sqrt{2}\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)$$

giving

$$\mathrm{E}(X_{i2}|\sigma_i) = \frac{\sigma_i}{\sqrt{n_i-1}}\sqrt{2}\left(\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{n_i-1}{2})}\right).$$