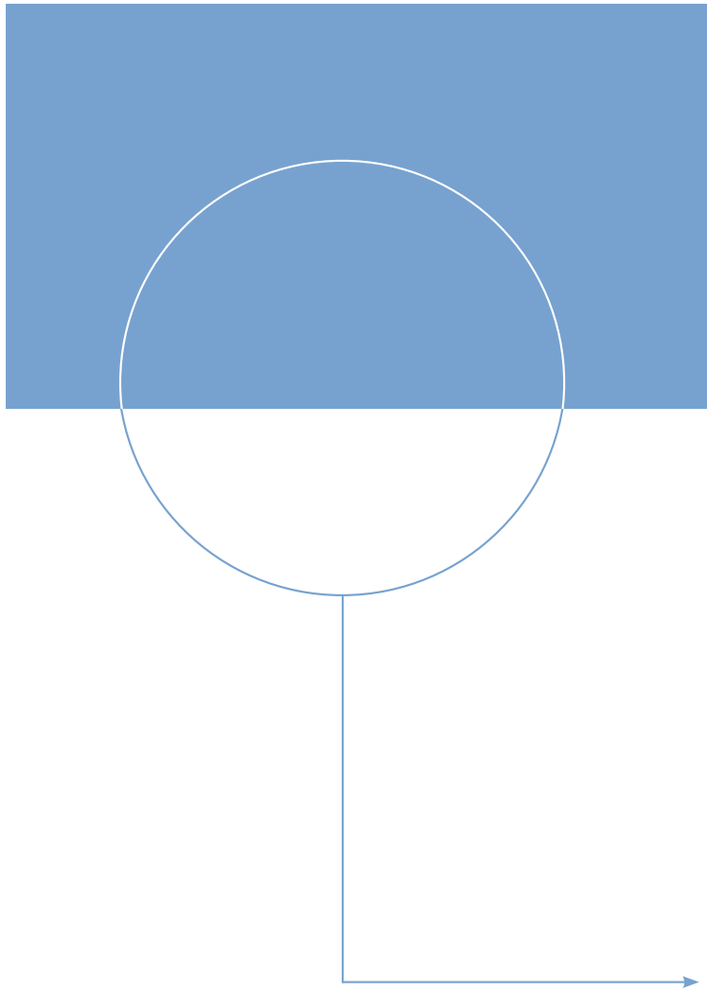

Master's thesis

2020

Lars Fredrik Espeland

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Lars Fredrik Espeland

A shared parameter model accounting for non-ignorable missing data due to dropout

Modelling of blood pressure based on the HUNT Study

June 2020

NTNU
Norwegian University of Science and Technology

NTNU
Norwegian University of Science and Technology

NTNU
Norwegian University of
Science and Technology

A shared parameter model accounting for non-ignorable missing data due to dropout

Modelling of blood pressure based on the HUNT Study

Lars Fredrik Espeland

Applied Physics and Mathematics
Submission date: June 2020
Supervisor: Ingelin Steinsland

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

In this thesis, a shared parameter model is suggested to account for missing data not at random (MNAR) due to dropout in follow-up studies. The proposed model is motivated by and evaluated for a large follow-up study of blood pressure, using data from the Trøndelag Health Study (HUNT). The goal is to draw unbiased inferences about parameters describing the systolic blood pressure in HUNT2 based on data from HUNT1. In order to do so, the fact that some participants drop out before HUNT2 must be taken into account. If the probability of dropout is directly related to the underlying blood pressure in HUNT2, then the data are MNAR and the dropout process needs to be explicitly modelled together with the blood pressure in order to obtain valid inference. The shared parameter model (SPM) proposed is such a joint model. It consists of a linear blood pressure model and a logistic dropout model, connected through a shared individual random effect. Both the blood pressure model and the dropout model are specified with the blood pressure in HUNT1, age and sex as covariates, but age is included through a smooth function in the dropout part. The model is a Bayesian latent Gaussian model, suitable for the integrated nested Laplace approximations (INLA) methodology for approximate Bayesian inference. Inference is obtained using R-INLA.

Parameter estimates obtained from SPM are compared to those obtained from a naive, linear Bayesian blood pressure model with the same covariates as SPM, which ignores the dropout process and assumes that the data are missing at random (MAR) instead of MNAR. Furthermore, two simulation studies are conducted, in which the naive model and SPM are tested on data with known parameters, when missingness is MNAR and MAR, respectively.

Fitting SPM to the HUNT data yields clearly different parameter estimates than the estimates from the model assuming MAR. SPM indicates that participants with a high, underlying blood pressure in HUNT2 have an increased probability of dropout, implying that the data are MNAR. The simulation studies support this. Therefore, a naive model assuming MAR is by all accounts insufficient, and a joint modelling approach is necessary to make unbiased blood pressure inference.

Sammendrag

I denne oppgaven blir en delt-parameter-modell foreslått for å ta hensyn til ikke-tilfeldige manglende verdier (missing not at random, MNAR) i oppfølgingsstudier. Denne modellen er motivert av, og evaluert på, en stor oppfølgingsstudie av blodtrykk, med data fra Helseundersøkelsen i Trøndelag (HUNT). Målet er å forventningsrett estimere parametere som beskriver blodtrykket i HUNT2 basert på data fra HUNT1. For å få til dette må det tas hensyn til at en del deltagere dropper ut av studiet før HUNT2. Hvis sannsynligheten for å ikke møte opp i HUNT2 er direkte relatert til det underliggende blodtrykket man har ved HUNT2, så er dataene MNAR. I så fall må utdroppingsprosessen modelleres sammen med blodtrykket for å kunne få gyldig inferens. Delt-parameter-modellen (shared parameter model, SPM) som blir foreslått er en slik felles modell. Den består av en lineær blodtrykksmodell og en logistisk utdroppingsmodell, sammenkoblet av en delt individuell, tilfeldig effekt. Både blodtrykksmodellen og utdroppingsmodellen er spesifisert med blodtrykket i HUNT1, alder og kjønn som kovariater, men alder blir modellert ikke-lineært i utdroppingsmodellen. Modellen er Bayesiansk, nærmere bestemt en latent Gaussisk modell, noe som gjør den egnet for integrated nested Laplace approximations-metodologien (INLA) for approksimativ Bayesiansk inferens. R-INLA blir brukt til modelltilpasning.

Paremeterestimater fra SPM er sammenlignet med estimater fra en naiv, lineær Bayesiansk blodtrykksmodell med de samme kovariatene som SPM, men som antar at utdropping er betinget tilfeldig (missing at random, MAR) i stedet for MNAR. Videre blir to simuleringsstudier gjennomført, der den naive modellen og SPM blir testet på data med kjente parametere, når manglende verdier er henholdsvis MNAR og MAR.

Tilpasning av SPM til HUNT-dataene gir markant annerledes parameterestimater enn modellen som antar MAR. SPM indikerer at deltagere med et høyt, underliggende blodtrykk ved HUNT2 har større sannsynlighet for å ikke møte opp, noe som gjør at dataene er MNAR. Simuleringsstudiene støtter opp under dette. Derfor er en naiv modell som antar MAR etter alt å dømme utilstrekkelig, og en tilnærming der utdropping modelleres sammen med blodtrykk er nødvendig for å oppnå forventningsrette estimater.

Preface

This thesis concludes my Master of Science (M.Sc.) in Applied Physics and Mathematics, with specialization in Industrial Mathematics, at the Norwegian University of Science and Technology (NTNU). The work was carried out during the spring of 2020 at the Department of Mathematical Sciences, and can be seen as a continuation of the work conducted in my specialization project in the autumn of 2019. Some parts of the theory are therefore inspired by the project.

Working with this thesis has been both exciting and challenging. Most of all, it has allowed me to develop a deeper understanding of many statistical concepts, which I am thankful for having had the opportunity to.

I am very grateful for the support, motivation and helpful guidance provided by my supervisor Ingelin Steinsland. I would also like to thank Emma Ingeström for providing access to and introducing the HUNT Study. Also thanks to Oddgeir Lingaas Holmen and the HUNT Cloud team for their helpfulness. At last, but not least, a big thanks to my family and friends, and to my fellow students for making the years at NTNU great.

Trondheim, June 2020
Lars Fredrik Espeland

Table of contents

1 | Introduction

According to the World Health Organization (2019), more than one billion people worldwide are suffering from hypertension, more commonly known as high blood pressure. Hypertension means that the pressure in the arteries is consistently elevated. It is defined as a systolic blood pressure level above 140 mmHg and/or diastolic blood pressure above 90 mmHg. Having blood pressure above these levels, the benefits of treatment, either through medication or change of lifestyle, clearly outweigh the risks of treatment (Williams et al., 2018). Being hypertensive increases for instance the risks of cardiovascular diseases and chronic kidney disease, and is therefore a prominent cause of premature death (Whelton, 1994; Lewington et al., 2002; Jha et al., 2013). Due to the high prevalence and potentially severe complications, hypertension is regarded as a major health problem worldwide (Kearney et al., 2005). Therefore, from a public health perspective, research on blood pressure related issues is highly important.

To survey the blood pressure of a population is within the main scope of the Trøndelag Health Study, HUNT. As the entire population of former Nord-Trøndelag county has been invited to participate, HUNT is Norway's largest health study. The data are obtained over four decades, from the mid 1980s until today, through four population surveys so far, HUNT1 to HUNT4. The HUNT data consist of questionnaire data and clinical measurements and samples. Originally, the main goals were to address arterial hypertension, diabetes, screening of tuberculosis and quality of life, but the scope has gradually expanded (Krokstad et al., 2013). The four parts of the HUNT study can together be seen as an example of a longitudinal study, in which participants are followed over time.

In this work, the focus is restricted to data from HUNT1 and HUNT2. The data in HUNT1 were collected in 1984-1986, with around 77,000 participants, while the HUNT2 survey was carried out in 1995-1997, with around 65,000 participants. Out of those, 47,000 also participated in HUNT1. See Krokstad et al. (2013) for a full cohort profile.

The data from HUNT1 and HUNT2 are well suited for blood pressure research. The blood pressure, alongside with other information, is provided for a large number of individuals. As the participants of HUNT1 are followed up after around ten years, it is possible to look at the development of blood pressure, and make inference on parameters describing this development. In blood pressure research using data from follow-up studies such as HUNT1 and HUNT2, this is typically of interest (Sparrow et al., 1982; Wilsgaard et al., 2000).

However, not all of those that took part in HUNT1 are also participating in HUNT2. There might be various reasons for such dropout, for example that the person had died, moved out of the area, was too sick to participate or did not have the time to take part. Participants not showing up cause data to be missing. Missing data are common in statistics, and are almost always present in longitudinal studies, and especially frequent in clinical trials and epidemiological studies (Verbeke and Molenberghs, 2000).

Missing data might lead to biased or inefficient inference if the missing values are ignored or inappropriately handled (Little and Rubin, 2002). The most common way to handle missing data is to do a so-called complete-case analysis (Schafer and Graham, 2002), in which only those individuals without any missing values are included in the analysis. Although this is an intuitive and simple procedure, there might be major disadvantages with this method. In particular, parameter estimates obtained from the observed sample might be biased compared to those of the full, underlying population, which in turn might lead to wrong conclusions about the population. Therefore, it is usually vital to consider the missing data.

It is possible to classify the missingness of values depending on how the missing values are related to the data (Little and Rubin, 2002). How one should handle the missing data depends in turn on this missing data mechanism. In the HUNT Study, there might, on the one hand, be no explainable reason why participants of HUNT1 are missing in HUNT2. If the outcome of the study is the blood pressure in HUNT2, then neither observed covariates from HUNT1 nor the underlying blood pressure in HUNT2, which would have been observed if the participant was not missing, are in this case related to the probability of not showing up. The data are then said to be missing completely at random (MCAR). In this case, a complete-case analysis will give unbiased inference. This is however a strong assumption, and is rarely the case in practice (Fitzmaurice, 2008; Wu, 2010).

On the other hand, dropout before HUNT2 might somehow be related to the blood pressure the person has at the time of HUNT2, either directly or indirectly. If the cause of data being missing is unrelated to the missing values themselves, but depends on other observed variables, the data are said to be missing at random (MAR). For example, this is the case if older people tend not to show up in HUNT2. Then the probability of being missing depends on the age, which is observed, and the age might in turn be correlated with the blood pressure. However, conditionally on age, subjects are missing at random. The distribution of the observed sample might be biased compared to that of the full population, but there are many good and efficient methods to handle the missing values when the data are MAR. This includes multiple imputation, thoroughly presented by Van Buuren (2018), and maximum likelihood methods (Little and Rubin, 2002). Together, MCAR and MAR are usually referred to as ignorable missing data mechanisms (Little and Rubin, 2002).

Whether or not participants are missing might also be directly related to the values that would have been observed if the person was not missing. For instance, people with high blood pressure, not exclusively explained by the observed covariates, might be more likely to not show up in HUNT2. If so, the data are missing not at random (MNAR), which is

a non-ignorable missing data mechanism (Little and Rubin, 2002). In longitudinal studies, this class of missing values is also called informative dropout (Diggle and Kenward, 1994). If this is the case, the distribution of the observed sample is different from that of the full population, leading to biased inference. As opposed to data being MCAR or MAR, it is not straightforward to handle data being MNAR. In order to do so, the missing data process needs to be modelled together with the blood pressure measurements.

In this work, the overall goal is to validly model the blood pressure in HUNT2 based on data from HUNT1. A few variables from HUNT1 are therefore selected as covariates. These variables are the systolic blood pressure, age and sex. The systolic blood pressure in HUNT1 is assumed to be highly correlated with the blood pressure in HUNT2 and is therefore a natural choice. Furthermore, age and sex are both well-established influential factors on the blood pressure level (Whelton, 1994). Later on, BMI is also added as a covariate.

However, since missing values caused by dropout are present, it is not straightforward to obtain valid blood pressure inference. As it is not possible to explicitly state the missing data mechanism of the data at hand, the typical approach of formulating a model assuming that the data are MAR and hence ignoring the dropout process is conducted to begin with. If the data actually are MAR, then such a naive approach is sufficient, but it provides biased inference if the data in reality are MNAR. Therefore, a joint model for the blood pressure and the missingness of blood pressure values is proposed. In general, joint modelling of measurements and dropout in longitudinal studies is a common approach (Diggle and Kenward, 1994; Little, 1995; Verbeke and Molenberghs, 2000). In this work, the joint model is a shared parameter model, in which a random effect is shared between the blood pressure model and the dropout model (Wu and Carroll, 1988; Vonesh et al., 2006; Steinsland et al., 2014). Both the shared parameter model and the naive model are Bayesian models, fitted using the integrated nested Laplace approximations (INLA) methodology through R-INLA (Rue et al., 2009). INLA is a fast alternative for approximate Bayesian inference for latent Gaussian models. Latent Gaussian models can contain a large variety of different functions of covariates, including for example the possibility of adding non-linear and random effects, allowing a large degree of flexibility when modelling through INLA. In the shared parameter model, such a non-linear smoothing term is included in the dropout model.

In Chapter 2, background theory behind some of the most important concepts of this work are presented. This includes missing data theory, in which the missing data mechanisms MCAR, MAR and MNAR are further elaborated. In addition, an introduction to Bayesian inference and latent Gaussian models are presented, before the INLA methodology used to perform approximate Bayesian inference for latent Gaussian models are discussed. Then, in Chapter 3, the data used from HUNT1 and HUNT2 are introduced and explored. Exploratory models of the blood pressure in HUNT2 and the dropout process, based on some covariates in HUNT1, are also formulated, and the resulting parameter estimates from these models are discussed. In Chapter 4, the method of this work is presented. This includes a detailed formulation of the models used, i.e. the naive blood pressure model and

the shared parameter model. Simulation studies, where the shared parameter model and the naive model are tested on data with known parameters, are introduced. A small analysis of prior sensitivity and an extension of the models with BMI added as a covariate are also considered. Further, in Chapter 5, the results are presented and interpreted, before further discussion and proposals for future work are given in Chapter 6, together with some conclusive remarks. A few supplementary remarks and R code are provided in the appendices.

2 | Background theory

2.1 Missing data

Data with missing values might cause difficulties when one wants to draw unbiased inferences from the data. The information hidden behind missing values could be of importance, and statistical conclusions might therefore be wrong when these values are not accounted for. How one should handle the missing values depends largely on how the missingness of values is related to the data. According to Little and Rubin (2002), such missing data mechanisms can be divided into three categories. These are respectively missing data completely at random (MCAR), missing data at random (MAR) and missing data not at random (MNAR). The data are MCAR if values are missing without any patterns, MAR if the probability that data are missing depends on other observable data and MNAR if the probability of missingness depends directly on the missing values themselves.

To exemplify these missing data mechanisms, consider a clinical study in which the blood pressure is the outcome of interest. If some participants accidentally forget to show up to get their blood pressure measured, then the data will be MCAR. Whether or not the blood pressure value is missing for a participant is in this case unrelated to both observed and unobserved data, and is instead entirely random. On the other hand, if men have a larger tendency of not showing up than women, then the probability that a blood pressure measurement is missing depends on sex, which is observed, and the data are MAR. Sex might in turn be correlated with the underlying blood pressure value, but conditionally on sex, the blood pressure values are missing at random. Alternatively, if those with high underlying blood pressure, even after conditioning on other observable data, do not show up for their blood pressure measurement, then the data are MNAR. Now, whether or not a blood pressure value is missing is directly related to the value that would have been observed if it was not missing.

The following notation used to present missing data theory is partly based on Little and Rubin (2002). Let $Y = (X, \boldsymbol{y})$ denote the matrix of the entire dataset, including the covariates X and the response variable $\boldsymbol{y}$. This is the underlying, complete data, which would have been observed in the absence of missing data. Assume that the covariates are fully observed, but that there are missing values in the response variable. Then, $\boldsymbol{y}$ can be divided into $\boldsymbol{y} = (\boldsymbol{y_o}, \boldsymbol{y_m})$, where $\boldsymbol{y_o}$ and $\boldsymbol{y_m}$ are the observed and unobserved part of $\boldsymbol{y}$, respectively. Further, let $\boldsymbol{m}$ be a missing indicator vector, which has the value $m_i = 1$ if

y_i is missing, and $m_i = 0$ if y_i is observed. The three missing data mechanisms are informally sketched in Figure 2.1. Z represents here the possible explanations of missingness that are completely unrelated to both X and $\boldsymbol{y}$.

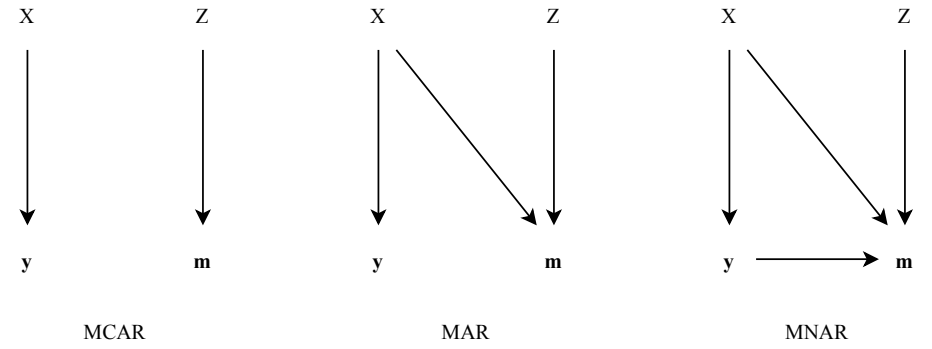

Figure 2.1: Illustration of causes of missingness $\boldsymbol{m}$ of variable $\boldsymbol{y}$ under the different missing data mechanisms. X represents completely observed covariates, $\boldsymbol{y}$ represents the response variable that is partly missing, and Z represents the component of causes of missingness not related to X or $\boldsymbol{y}$.

The difference between full data and observed data is highly important. The interest lies in the full data. However, inferences about the full data that are based on incomplete observations must rely on assumptions about the missing response distribution. The distribution of response measurements is characterized by $f(\boldsymbol{y}|X, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are some parameters describing this distribution. Similarly, the missing data mechanism is specified through the conditional distribution of the missing data indicator $\boldsymbol{m}$ given the data, $f(\boldsymbol{m}|X, \boldsymbol{y}, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ are unknown parameters describing the distribution. If the data are MCAR, then the missing data mechanism can be simplified to

$$f(\boldsymbol{m}|\boldsymbol{y}, X, \boldsymbol{\psi}) = f(\boldsymbol{m}|\boldsymbol{\psi}), \tag{2.1}$$

meaning that the distribution of missingness does not depend on any of the data, observed or not. Further, if the data are MAR, the missing data mechanism is instead given by

$$f(\boldsymbol{m}|\boldsymbol{y}, X, \boldsymbol{\psi}) = f(\boldsymbol{m}|\boldsymbol{y_o}, X, \boldsymbol{\psi}), \tag{2.2}$$

Missingness is now independent of the missing response values $\boldsymbol{y_m}$ conditionally on observed responses $\boldsymbol{y_o}$ and covariates X. If the data are MNAR, however, the missing data mechanism can not be simplified further, so the missingness is given by

$$f(\boldsymbol{m}|\boldsymbol{y}, X, \boldsymbol{\psi}) = f(\boldsymbol{m}|\boldsymbol{y_o}, \boldsymbol{y_m}, X, \boldsymbol{\psi}). \tag{2.3}$$

The distribution may still depend on $\boldsymbol{y_o}$ and X, as when data are MAR, but the crucial point is now the additional dependence on $\boldsymbol{y_m}$.

Usually, it is of interest to estimate the parameters $\boldsymbol{\theta}$ that describes the distribution of measurements. In a Bayesian setting, thoroughly presented in the upcoming sections of

the chapter, Bayes' formula can be used to obtain posterior parameter estimates, such that

$$f(\boldsymbol{\theta}|\boldsymbol{y}, X) = \frac{f(\boldsymbol{y}|X, \boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y}|X)} \propto L(\boldsymbol{\theta}|\boldsymbol{y}, x)f(\boldsymbol{\theta}). \tag{2.4}$$

When ignoring the missing data process, the likelihood of $\boldsymbol{\theta}$ can be written as

$$L_{ign}(\boldsymbol{\theta}|\boldsymbol{y}, X) \propto f(\boldsymbol{y_o}|X, \boldsymbol{\theta}) = \int f(\boldsymbol{y_o}, \boldsymbol{y_m}|X, \boldsymbol{\theta})d\boldsymbol{y_m}, \tag{2.5}$$

where the missing values $\boldsymbol{y_m}$ are integrated out. Maximizing this, a maximum likelihood estimate is obtained. Inserting (2.5) into (2.4) together with a prior distribution for $\boldsymbol{\theta}$, gives a posterior distribution for $\boldsymbol{\theta}$, and Bayesian inference is obtained. Inference based on this likelihood is valid if the missing data mechanism can be ignored.

Instead of ignoring the missing data mechanism, one would normally consider the full likelihood. This is a joint distribution of $\boldsymbol{y}$ and $\boldsymbol{m}$, and can be written as

$$f(\boldsymbol{y}, \boldsymbol{m}|X, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}|X, \boldsymbol{\theta})f(\boldsymbol{m}|\boldsymbol{y}, X, \boldsymbol{\psi}). \tag{2.6}$$

The actually observed data are $(\boldsymbol{y_o}, \boldsymbol{m})$, in addition to X. Therefore, similarly as with the likelihood (2.5) ignoring the missing data mechanism, the full likelihood becomes

$$\begin{aligned} &L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y_o}, X, \boldsymbol{m}) \propto f(\boldsymbol{y_o}, \boldsymbol{m}|X, \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \int f(\boldsymbol{y_o}, \boldsymbol{y_m}|X, \boldsymbol{\theta})f(\boldsymbol{m}|\boldsymbol{y_o}, \boldsymbol{y_m}, X, \boldsymbol{\psi})d\boldsymbol{y_m}. \end{aligned} \tag{2.7}$$

Bayesian inference under the full model for $\boldsymbol{y}$ and $\boldsymbol{m}$ is given by

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y_o}, X, \boldsymbol{m}) &= \frac{f(\boldsymbol{y_o}, \boldsymbol{m}|X, \boldsymbol{\theta}, \boldsymbol{\psi})f(\boldsymbol{\theta}, \boldsymbol{\psi})}{f(\boldsymbol{y_o}, \boldsymbol{m}|X)} \\ &\propto L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y_o}, X, \boldsymbol{m})f(\boldsymbol{\theta}, \boldsymbol{\psi}), \end{aligned} \tag{2.8}$$

where the full likelihood (2.7) is combined with a prior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$.

The question is now when inference about $\boldsymbol{\theta}$ should be based on the full likelihood (2.7) and when it can be based on the simplified likelihood (2.5) ignoring the missing data mechanism. In general, MCAR and MAR are said to be ignorable, meaning that the simpler likelihood can be used. If the data are MAR, (2.2) holds. Then, by using (2.5), (2.7) can be written as

$$\begin{aligned} &L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y_o}, X, \boldsymbol{m}) \propto f(\boldsymbol{y_o}, \boldsymbol{m}|X, \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= f(\boldsymbol{m}|\boldsymbol{y_o}, X, \boldsymbol{\psi}) \int f(\boldsymbol{y_o}, \boldsymbol{y_m}|X, \boldsymbol{\theta})d\boldsymbol{y_m} \\ &= f(\boldsymbol{m}|\boldsymbol{y_o}, X, \boldsymbol{\psi})f(\boldsymbol{y_o}|X, \boldsymbol{\theta}) \propto L_{ign}(\boldsymbol{\theta}|\boldsymbol{y}, X). \end{aligned} \tag{2.9}$$

As the likelihood can be written on the form (2.9) when data are MAR, then likelihood and Bayesian inference for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ can be done ignoring the missing data process when $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct and have independent priors (Little and Rubin, 2002). Therefore, as valid inference is obtained when missing data are MCAR or MAR without taking the missing process into account, these mechanisms are called ignorable. However, when the conditions for ignoring the missing process are not met, the data are non-ignorable. This is the reality when data are MNAR. It is not possible to tell explicitly whether or not data at hand are MNAR, but if this is the case, the missing process also needs to be modelled in order to obtain valid inference.

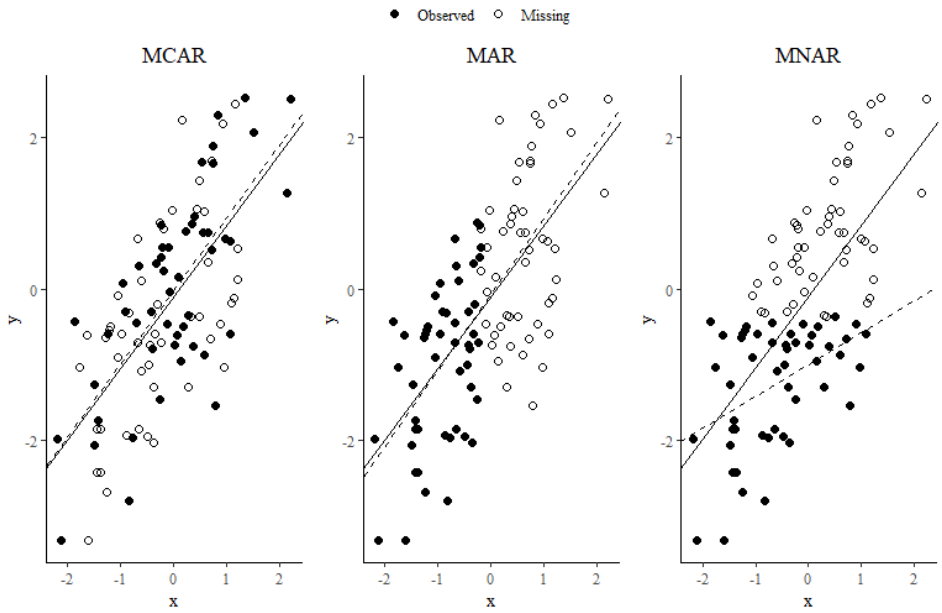

Figure 2.2: Illustration of the different missing data mechanisms with an underlying linear relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$. The true regression line fitted to all points are solid, whereas the regression line fitted to only the observed points under the respective missing data mechanism is dashed.

The missing data mechanisms are illustrated in a simple linear regression setting in Figure 2.2, where $\boldsymbol{y}$ is regressed onto the covariate $\boldsymbol{x}$. 100 correlated bivariate normal points (x_i, y_i) are generated, before half of them are removed. In the left plot, 50 observations are removed completely at random, making the data MCAR. In the middle plot, the 50 observations with the largest x-values are removed, making the data MAR. To the right, the 50 points with the largest y-values are removed, so the data are MNAR. In these plots, the true regression line, which would have been fitted in the absence of missing data, is plotted as a solid line, while the regression lines fitted to the observed observations only, are dashed. Both under MCAR and MAR, the regression line is unbiased, whereas the regression line is clearly downward biased under MNAR.

How one should handle the missing data depends strongly on the missing data mechanism and the objective of study. The most traditional method to handle missing data is a so-called complete-case analysis, in which those rows of the dataset with any missing values are deleted. If the data are MCAR, then a complete-case analysis will provide unbiased inferences. Otherwise, a complete-case analysis might lead to severely biased inferences. If the data are MAR, then a complete-case analysis might give biased results, depending on what the objective is. If the interest lies in the conditional distribution of a variable $\boldsymbol{y}$ with missing values, given completely observed X, then a complete-case analysis is satisfactory if the data are MAR. This is the case in a regression setting with missing values in the response variable. The incomplete cases give no information to the regression of $\boldsymbol{y_m}$ onto X, so unbiased maximum likelihood estimates are obtained after deleting the cases with missing values (Little, 1992). This can be seen from the middle plot of Figure 2.2. If the interest instead lies in the marginal distribution of the variable $\boldsymbol{y}$, then an analysis based on only the complete records will be biased under MAR. For example, the mean of the observed observations $\boldsymbol{y_o}$ in the middle plot of Figure 2.2 will differ from that of the full, underlying $\boldsymbol{y}$. Bias is also present in a regression analysis with missing values in X being MAR, but where $\boldsymbol{y}$ is known. In these cases, however, there exist good methods, such as multiple imputation, to handle the missing data and correct for the bias.

If the data are MNAR, fundamental identifiability issues are introduced, simply because the fact that the missing data are not observed means that there are no data with which to estimate the distribution of the missing values. In these cases, one also needs to model the missing data process in order to obtain valid inference. In other words, when ignorability is not assumed to be a suitable assumption, a more general class of models can be used, allowing missing data indicators to somehow depend on missing responses themselves. These models parametrize the conditional dependence between $\boldsymbol{m}$ and $\boldsymbol{y_m}$, given $\boldsymbol{y_o}$ and X. However, this association structure can not be identified from the observed data without the benefit of untestable assumptions. Therefore, inference depends on a combination of unverifiable parametric assumptions and possibly, in the Bayesian setting, informative prior distributions.

There are different approaches to formulating such models accounting for non-ignorable missing data, each differing in how the full model (2.6) for the data $\boldsymbol{y}$ and missingness $\boldsymbol{m}$ is factorized. Selection models use the factorization

$$f(\boldsymbol{y}, \boldsymbol{m}|X, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}|X, \boldsymbol{\theta}) f(\boldsymbol{m}|\boldsymbol{y}, X, \boldsymbol{\psi}), \tag{2.10}$$

whereas pattern-mixture models write the joint distribution

$$f(\boldsymbol{y}, \boldsymbol{m}|X, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}|X, \boldsymbol{m}, \boldsymbol{\theta}) f(\boldsymbol{m}|X, \boldsymbol{\psi}), \tag{2.11}$$

where the factorization is reversed compared to that of selection models.

Another way to handle non-ignorable missing data is alternatively to use a shared parameter model (Wu and Carroll, 1988), in which the missingness and data can be jointly modelled with a vector of shared random effects. This distribution is formulated as

$$f(\boldsymbol{y}, \boldsymbol{m}|\boldsymbol{\epsilon}, X, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}|\boldsymbol{\epsilon}, X, \boldsymbol{\theta}) f(\boldsymbol{m}|\boldsymbol{\epsilon}, X, \boldsymbol{\psi}), \tag{2.12}$$

where $\boldsymbol{\epsilon}$ are the shared random effects. The shared parameter approach appeals strongly to intuition, suggesting that $\boldsymbol{y}$ and $\boldsymbol{m}$ have a common, unobserved cause. The shared random effects framework is discussed further in Chapter 4, where a shared parameter model for non-ignorable missing blood pressure data is proposed.

For models dealing with non-ignorable missing data, one must be aware that identifiability of parameters is an important issue, as there may be too many nuisance parameters (Fitzmaurice et al., 1996; Wu, 2010). Due to the potential complexity of non-ignorable models, the models can be non-identifiable if there is not sufficient information in the data, meaning that two different sets of parameters may lead to the same observed likelihood. To show algebraically that all of the parameters in non-ignorable models are identifiable is not trivial (Fitzmaurice et al., 1996). In practice, one might often find out empirically that not all parameters are fully identifiable (Wu, 2010).

2.2 Bayesian inference

There exist two main paradigms for statistical analysis. With a classical, frequentist approach, the parameters are considered to be fixed and unknown, and the goal is to estimate the true parameters from the data. By contrast, in Bayesian statistics, the parameters are instead characterized by probability distributions. Given data $Y = (X, \boldsymbol{y})$, a parameter vector $\boldsymbol{\theta}$ and potentially a vector of hyperparameters $\boldsymbol{\phi}$, the goal in Bayesian inference is therefore to obtain the posterior distributions $f(\theta_i|X, \boldsymbol{y})$ and $f(\phi_j|X, \boldsymbol{y})$. The posterior distributions are obtained after specifying a model, specifying prior distributions for the parameters of the model, and then update the prior information about the parameters using the model and the data. In other words, the posterior distributions are obtained by combining prior information known beforehand and the information provided by the data, quantified by the likelihood. The likelihood summarizes the information the data have about the parameters. The maximum likelihood estimator, which maximizes the likelihood function, is typically used as point estimator in frequentist inference. However, in Bayesian inference, the likelihood and prior information are combined through Bayes' formula, which is on the form (2.4).

Prior probability distributions represent the knowledge one has in advance, a priori, about $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. If the prior knowledge about a parameter is specific, then this probability distribution is a so-called informative prior. On the other hand, vague prior beliefs can be represented by a diffuse probability distribution. In this case, most weight is put on the information provided by the data.

The posterior distributions obtained are often summarized by some quantities. For point estimates, the posterior mean, median or mode are typical choices. Uncertainty is typically represented by the standard deviation of the posterior distribution, or by forming credible intervals based on quantiles of the distribution. For example, the 2.5th and 97.5th quantiles can be used to make a 95% equal-tailed credible interval for a parameter.

Most often, it is complicated to express the posterior distribution in closed form (Daniels and Hogan, 2008). Therefore, the most common approach to Bayesian inference is to obtain a sample from the posterior distribution using techniques that do not need explicit evaluation of the denominator, which is a normalizing constant, of the posterior distribution.

Traditionally, Markov chain Monte Carlo (MCMC) sampling has been used to fit Bayesian models. The idea of MCMC algorithms is to obtain a sample from the posterior distribution by construction of a Markov chain having the posterior distribution of interest as stationary distribution. Marginal posteriors and the posterior of functions of the parameters are obtained by doing Monte Carlo integration using the sample from the Markov chain. A detailed presentation of MCMC methods can for example be found in Gelman et al. (2014). Although MCMC algorithms are extremely flexible, they turn often out to be slow and might even become computationally unfeasible (Blangiardo and Cameletti, 2015). Therefore, the newly developed INLA algorithm (Rue et al., 2009), which is a fast and accurate algorithm for approximate Bayesian inference, can be used as an alternative to MCMC algorithms for special cases of latent Gaussian models. The concepts behind this algorithm are discussed in the next sections.

2.3 Latent Gaussian models

The INLA methodology is designed for a subgroup of so-called latent Gaussian models, which a wide range of different Bayesian models are. In order to understand latent Gaussian models, it is useful to also know the concepts of generalized linear models and generalized additive models. These types of models are briefly introduced in this section.

2.3.1 Generalized linear models

In generalized linear models, first introduced by Nelder and Wedderburn (1972), a distribution is assumed for the observed response data $\boldsymbol{y} = (y_1, \cdots, y_n)$. In contrast to ordinary linear regression, the distribution does not have to be Gaussian, as long as it is part of the exponential family of distributions. The distribution of y_i is characterized through a linear predictor η_i which is defined as the mean μ_i through an appropriate link function $g(\cdot)$, such that $\eta_i = g(\mu_i)$. The linear predictor is given by

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi}, \tag{2.13}$$

where β_0 defines the intercept and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)$ quantifies the linear effects of the covariates $X = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_M})$ on the linear predictor $\boldsymbol{\eta}$.

If $g(\cdot)$ is the identity function, then this model is a linear model. However, for binary responses $\boldsymbol{y}$, meaning that they take on the values 0 and 1 only, the distribution of $\boldsymbol{y}$ is generally chosen to be the Bernoulli distribution, which is equivalent to a binomial distribution

with a single trial. In this case, μ_i is interpreted to be the probability p_i that y_i takes the value 1. A common link function is now the logit function, $\eta_i = \text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$, with $y_i \sim \text{Bernoulli}(p_i)$, making the model a logistic regression model. Logistic regression models are used to model the probabilities of certain events to happen. A general introduction to logistic regression can for example be found in James et al. (2013).

In the missing data setting, logistic regression is suitable to model the missing process. For instance, logistic regression is commonly used to model the probability of dropout in longitudinal studies, when missingness is assumed to be MNAR and a dropout model therefore is included as a part of a joint model (Diggle and Kenward, 1994; Molenberghs et al., 1997).

2.3.2 Generalized additive models

When making statistical models, one might often find that a model with only linear effects of covariates on the linear predictor as in (2.13) is insufficient to capture the true relationship. In generalized additive models, developed by Hastie and Tibshirani (1990), this restriction is relaxed, as the linear predictor now depends linearly on unknown smooth functions of covariates. The linear predictor is an additive predictor given by

$$\eta_i = \beta_0 + \sum_{m=1}^{M} f_m(x_{mi}), \tag{2.14}$$

where $\boldsymbol{f} = (f_1, \cdots, f_M)$ are functions defined in terms of the covariates $X = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_M})$. In such models allowing for non-linearities, the true shapes of the functions defining the non-linear relationships are typically not known and are usually estimated, either through semi-parametric or non-parametric methods (Gómez-Rubio, 2020). Such smoothing methods are useful to model complex relationships between the covariates and the additive predictor.

A random walk model can for instance be used to model such non-linearities (Fahrmeir and Tutz, 2001). In the Bayesian framework, in a random walk model of order 1, a prior is set on the function value at knots, such that

$$f(k_{i+1}) - f(k_i) \sim N(0, \sigma^2), \quad i = 1, \ldots, K-1. \tag{2.15}$$

Similarly, in a random walk model of order 2, a prior is set on the second-order differences

$$f(k_{i+1}) - 2f(k_i) + f(k_{i-1}) \sim N(0, \sigma^2), \quad i = 2, \ldots, K. \tag{2.16}$$

Fitting smooth functions to the data in this way allows for a large flexibility when modelling.

2.3.3 Latent Gaussian models

Now, latent Gaussian models are defined through a structured additive predictor given by

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}) + \epsilon_i, \tag{2.17}$$

where β_0 defines the intercept, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)$ quantifies the linear effect of the covariates $X = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_M})$ on the additive predictor, and where $\boldsymbol{f} = (f_1, \cdots, f_L)$ are a set of functions defined in terms of some covariates $Z = (\boldsymbol{z_1}, \cdots, \boldsymbol{z_L})$. In regression models, these functions can model non-linear covariate effects, such as random walk smoothing functions described above for generalized additive models, but also different random effects. Therefore, the linear predictor can include both fixed effects and random effects. The latent components of interest in (2.17) can be collected in a set of latent parameters $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$. Now, a latent Gaussian model is obtained if all elements of $\boldsymbol{\theta}$ have Gaussian priors assigned, making latent Gaussian models a subset of all Bayesian structured additive regression models on the form (2.17), which includes generalized linear models and generalized additive models. Further, $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_k\}$ denotes the vector of possible hyperparameters. Their prior distributions do not have to be Gaussian for the model to be a latent Gaussian model.

2.4 Integrated nested Laplace approximations

The INLA methodology, introduced by Rue et al. (2009), is designed for latent Gaussian models described above, which are models with a structured additive predictor on the form (2.17) and with Gaussian priors assigned to the latent parameters. Further, the INLA algorithm exploits properties of Gaussian Markov random fields and Laplace approximations for computationally efficient Bayesian inference. These concepts are briefly introduced here, following Blangiardo and Cameletti (2015) closely.

With observed data $\boldsymbol{y} = (y_1, \cdots, y_n)$, the likelihood function is given by

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^{N} f(y_i|\theta_i, \boldsymbol{\phi}), \tag{2.18}$$

where there is a connection between each observation y_i and only one element of $\boldsymbol{\theta}$.

A multivariate normal prior on $\boldsymbol{\theta}$ having mean $\mathbf{0}$ and precision matrix $\boldsymbol{Q}(\boldsymbol{\phi})$, which is the inverse of the covariance matrix, is assumed. Thus, the density function becomes

$$f(\boldsymbol{\theta}|\boldsymbol{\phi}) = (2\pi)^{-n/2} |\boldsymbol{Q}(\boldsymbol{\phi})|^{1/2} \exp(-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}(\boldsymbol{\phi})\boldsymbol{\theta}), \tag{2.19}$$

where $|\cdot|$ denotes the matrix determinant. $\boldsymbol{\theta}$ is now a latent Gaussian field. Due to the Markov property, $\boldsymbol{Q}(\boldsymbol{\phi})$ will be a sparse matrix if the components of $\boldsymbol{\theta}$ are assumed to

be conditionally independent. As a result, $\boldsymbol{\theta}$ is a Gaussian Markov random field (Rue and Held, 2005). In turn, this leads to computational benefit.

The joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be obtained through Bayes' formula and combining (2.18) and (2.19), such that

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) &\propto f(\boldsymbol{\phi}) \cdot f(\boldsymbol{\theta}|\boldsymbol{\phi}) \cdot f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) \\ &\propto f(\boldsymbol{\phi}) \cdot f(\boldsymbol{\theta}|\boldsymbol{\phi}) \cdot \prod_{i=1}^{N} f(y_i|\theta_i, \boldsymbol{\phi}) \\ &\propto f(\boldsymbol{\phi}) \cdot |\boldsymbol{Q}(\boldsymbol{\phi})|^{1/2} \exp\Big(-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}(\boldsymbol{\phi})\boldsymbol{\theta} + \sum_{i=1}^{n} \log\big(f(y_i|\theta_i, \boldsymbol{\phi})\big)\Big), \end{aligned} \tag{2.20}$$

where $f(\boldsymbol{\phi})$ is a prior distribution for the hyperparameters.

In Bayesian inference, the objectives are to obtain the marginal posterior distributions of each element of the parameter and hyperparameter vectors,

$$f(\theta_i|\boldsymbol{y}) = \int f(\theta_i, \boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi} = \int f(\theta_i|\boldsymbol{\phi}, \boldsymbol{y}) f(\boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi} \tag{2.21}$$

and

$$f(\phi_k|\boldsymbol{y}) = \int f(\boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi}_{-k}. \tag{2.22}$$

Therefore, $f(\boldsymbol{\phi}|\boldsymbol{y})$ and $f(\theta_i|\boldsymbol{\phi}, \boldsymbol{y})$ must be computed in order to obtain all the relevant marginals.

In the INLA algorithm, a Laplace approximation is used to produce numerical approximations to the posterior distributions. In order to understand Laplace approximations, consider the integral

$$\int f(x) dx = \int \exp\left(\log f(x)\right) dx, \tag{2.23}$$

where $f(x)$ is the density function of a random variable X. If $\log f(x)$ is represented by a Taylor series expansion evaluated in $x^* = \arg\max_x \log f(x)$, then the fact that $\frac{\partial \log f(x)}{\partial x}\Big|_{x=x^*} = 0$ yields an approximation to the integral given by

$$\int f(x) dx \approx \exp(\log f(x^*)) \int \exp\Big(-\frac{(x-x^*)^2}{2} \frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}\Big) dx. \tag{2.24}$$

The shape of the integrand is the same as the density of the normal distribution. By setting $\sigma^{2^*} = -1/\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}$,

$$\int f(x) dx \approx \exp(\log f(x^*)) \int \exp\Big(-\frac{(x-x^*)^2}{2\sigma^{2^*}}\Big) dx. \tag{2.25}$$

The integrand can now be seen as the kernel of a normal distribution with mean x^* and variance σ^{2^*}. Further, integrating over the interval (α, β) then gives the approximation

$$\int_{\alpha}^{\beta} f(x)dx \approx f(x^*)\sqrt{2\pi\sigma^{2^*}}(\phi(\beta) - \phi(\alpha)), \tag{2.26}$$

where $\phi(\cdot)$ is the cumulative density function of the normal distribution with mean x^* and variance σ^{2^*}. This is the Laplace approximation of the integral (2.23), used in the INLA algorithm when approximating the posterior distributions of the parameters and hyperparameters.

Further, the joint posterior distribution of the hyperparameters is approximated by

$$\begin{aligned} f(\boldsymbol{\phi}|\boldsymbol{y}) &= \frac{f(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})}{f(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})} \\ &\propto \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi})f(\boldsymbol{\theta}|\boldsymbol{\phi})f(\boldsymbol{\phi})}{f(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})} \\ &\approx \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi})f(\boldsymbol{\theta}|\boldsymbol{\phi})f(\boldsymbol{\phi})}{\tilde{f}(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\phi})} =: \tilde{f}(\boldsymbol{\phi}|\boldsymbol{y}), \end{aligned} \tag{2.27}$$

where $\tilde{f}(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})$ is the Gaussian approximation, obtained through the Laplace method, of $f(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})$, and $\boldsymbol{\theta}^*(\boldsymbol{\phi})$ is the mode for a given $\boldsymbol{\phi}$.

Further, there are three available options to estimate the posterior $f(\theta_i|\boldsymbol{\phi}, \boldsymbol{y})$ of the parameters. The fastest option is to use the marginals of the Gaussian approximation already computed. Another possibility is to perform a Laplace approximation one more time, or otherwise to use a so-called simplified Laplace approximation (Rue et al., 2009). Using the approximation $\tilde{f}(\theta_i|\boldsymbol{\phi}, \boldsymbol{y})$ obtained by one of these methods, together with the approximation $\tilde{f}(\boldsymbol{\phi}|\boldsymbol{y})$ from (2.27), the approximation to the marginal posterior of the parameters $f(\theta_i|\boldsymbol{y})$ in (2.21) is numerically solved through

$$\tilde{f}(\theta_i|\boldsymbol{y}) \approx \sum_j \tilde{f}(\theta_i|\boldsymbol{\phi}^{(j)}, \boldsymbol{y})\tilde{f}(\boldsymbol{\phi}^{(j)}|\boldsymbol{y})\Delta_j, \tag{2.28}$$

where $\boldsymbol{\phi}^{(j)}$ is a grid of relevant integration points with corresponding weights Δ_j. The grid of integration points are found by locating the mode $\boldsymbol{\phi}^*$ of $\tilde{f}(\boldsymbol{\phi}|\boldsymbol{y})$ and then exploring the distribution from there.

Although INLA is designed for latent Gaussian models, it is worth noticing that not all such models can be fitted efficiently by the algorithm. The INLA algorithm is most efficient when $\boldsymbol{\theta}$ is a Gaussian Markov random field, due to effective numerical methods for sparse matrices through a Cholesky decomposition. The number of hyperparameters should also be small, typically between 2 and 5, and not exceeding 20 (Rue et al., 2009; Wang et al., 2018). The reason for this restriction is to limit the dimensions in the numerical integration, as it can be expensive to integrate out a large number of hyperparameters.

A more comprehensive review of the theory behind INLA can be found in Rue et al. (2009), Blangiardo and Cameletti (2015) and Wang et al. (2018).

A package in R called R-INLA has been developed to perform approximate Bayesian inference with the INLA algorithm. This package is used in this work. R-INLA can for instance handle models with multiple likelihoods (Martins et al., 2013). Hence, it is well suited for joint modelling of blood pressure measurements and dropout, as is done in this work.

3 HUNT Study data and exploratory analysis

In this work, data from the Trøndelag Health Study, HUNT, are used for blood pressure modelling. More specifically, the two first health surveys, HUNT1 and HUNT2, are considered. In this chapter, the dataset and its variables are described and explored, with a special focus on the dropout process from HUNT1 to HUNT2.

3.1 Variables from HUNT1 and HUNT2

The full data from HUNT1 and HUNT2 contain large amounts of information. Numerous available variables provide information ranging from clinical measurements such as blood pressure and BMI, to whether or not the participants have or have had different diseases and use medication. Also, demographic data and questionnaire data covering lifestyle-related issues such as alcohol consumption, smoking and exercising are provided, to mention a few. In this work, however, the interest lies in those that participated in HUNT1 and their corresponding systolic blood pressure in HUNT2, which is missing if the participant dropped out prior to HUNT2. The dataset considered in this work consists of a few, selected variables from HUNT1, that are used as covariates in the models, to predict the blood pressure in HUNT2.

In order to reduce the complexity of the models, only a few variables from HUNT1 are selected. These variables are the systolic blood pressure, age, sex and BMI. From HUNT2, the systolic blood pressure, which is the outcome of interest, is included. However, also a missing indicator variable, which is created, is included from HUNT2. This is a binary variable,

$$\mathrm{m}_i = \begin{cases} 1, & \text{if participant } i \text{ is missing in HUNT2,} \\ 0, & \text{if participant } i \text{ is observed in HUNT2,} \end{cases} \tag{3.1}$$

i.e. stating whether or not a participant has missing blood pressure measurements in HUNT2, and hence are regarded as dropped out of the study.

The variables considered are summarized in Table 3.1.

Table 3.1: Summary of variables.

	Variable	Type	Description
Response (HUNT2)	BP_2	Numeric	Systolic blood pressure (mmHg)
	m	Binary	Observed (0) or missing (1) BP_2
Covariates (HUNT1)	age	Numeric[a]	Age (years)
	sex	Binary	Female (0) or Male (1)
	BP_1	Numeric[a]	Systolic blood pressure (mmHg)
	BMI	Numeric[a]	Body Mass Index (kg/m^2)

[a] Continuous covariates are standardized before the models are fitted.

The dataset consists originally of a large number of missing values, but only the participants with complete records from the variables of interest in HUNT1 are included in the dataset. The only variable with missing values is therefore the blood pressure in HUNT2, BP_2, corresponding to $\mathrm{m} = 1$. After limiting the number of included participants this way, 57351 participants are considered. These are the ones participating in HUNT1, with no missing values in any of the variables considered from HUNT1. Out of those, 37445 (65.3%) have a blood pressure measurement in HUNT2. 19906 (34.7%) do not have a blood pressure measurement in HUNT2 and are regarded as missing.

It is worth making a comment about the blood pressure variables from HUNT1 and HUNT2, BP_1 and BP_2. Some of the participants are using blood pressure medication at the time of HUNT1 or HUNT2. As it is of interest to look at the underlying systolic blood pressure, which would have been observed in the absence of any blood pressure medication, the blood pressure needs to be adjusted in these cases. Therefore, as suggested by Tobin et al. (2005), 15 mmHg is added to the measured systolic blood pressure of the participants using blood pressure medication in HUNT1 or HUNT2, in order to neutralize the effect of medication. However, a minor note by doing so is that in HUNT1, the question about current blood pressure medication only applied to a subset of the participants. Because of how the question was asked, 714 of the participants stated in HUNT1 that they had been using blood pressure medication previously, but they did not state whether they used it at the time of HUNT1. These participants did not get their blood pressure adjusted. Hence, not all of those actually using blood pressure medication in HUNT1 did get their blood pressure adjusted for treatment effects. Consequently, a small fraction of the participants considered should have had a higher blood pressure in HUNT1 than they had in this work.

3.2 Exploratory data analysis

A descriptive summary of the variables is given in Table 3.2. Each continuous variable is described in terms of mean and standard deviation. For the binary variable sex, the to-

tal number and fraction of participants belonging to each sex are stated. The summary is given for the participants overall, but also separately for those that showed up in HUNT2 and for those that did not.

Table 3.2: Descriptive statistics. Continuous variables are described with mean and standard deviation, while the binary variable sex is described in terms of the number and fraction females and males, respectively.

		Status in HUNT2	
Variable	Summary HUNT1 (n = 57351)	Observed (n = 37445, 65.3%)	Missing (n = 19906, 34.7%)
BP_2	-	143.7 ± 24.6	-
Age	48.8 ± 17.3	45.0 ± 14.2	55.8 ± 20.1
BP_1	139.3 ± 24.9	134.6 ± 21.3	148.3 ± 28.5
BMI	25.2 ± 3.9	24.9 ± 3.7	25.6 ± 4.4
Sex			
Female	28947 (50.5%)	19864 (68.6%)	9083 (31.4%)
Male	28404 (49.5%)	17581 (61.9%)	10823 (38.1%)

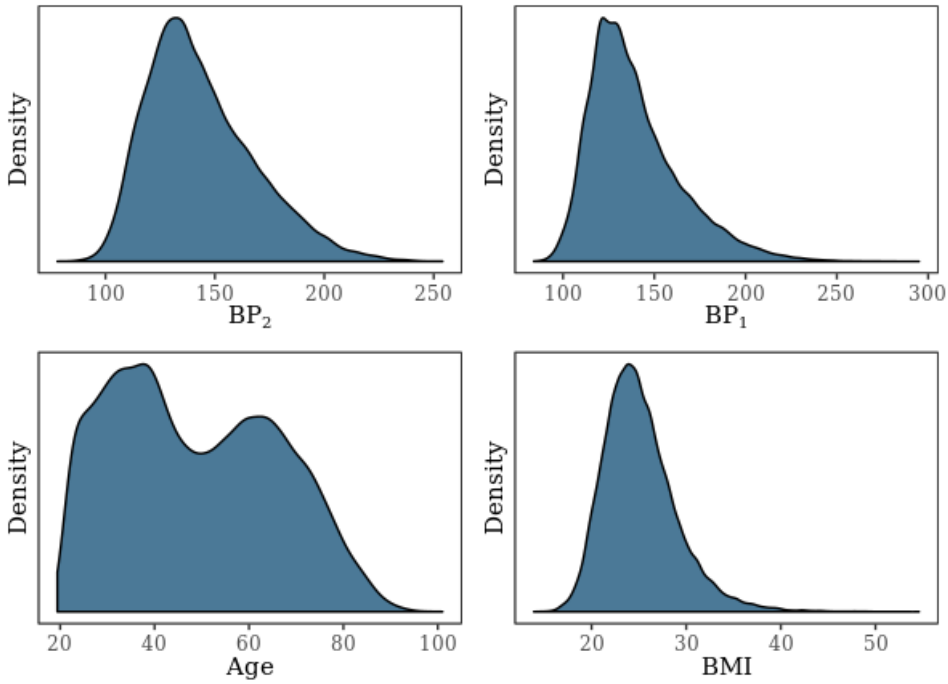

Figure 3.1: Density plots of the continuous variables.

To complement the information provided in Table 3.2, the distributions of the continuous variables are shown in Figure 3.1. Further, Figure 3.2 shows how the distributions of blood pressure, age and BMI in HUNT1 are for those that also participated in HUNT2 and for those that did not. From Table 3.2 and Figure 3.2, it can clearly be seen that the characteristics in HUNT1 differ between the participants that are observed and the participants that are missing in HUNT2. Naturally, a large portion of those that had a high age in HUNT1 did not show up in HUNT2. Also, the individuals that dropped out had on average a slightly higher systolic blood pressure in HUNT1. By contrast, the distribution of BMI for those dropping out is almost similar to the distribution of BMI for those showing up in HUNT2. In Table 3.2, it is shown that men are more likely than women to drop out, as 38.1% of the men drop out compared to 31.4% of the women. That there are clearly different characteristics in HUNT1 between those that show up in HUNT2 and those that do not motivates why it is vital to take missing values into account.

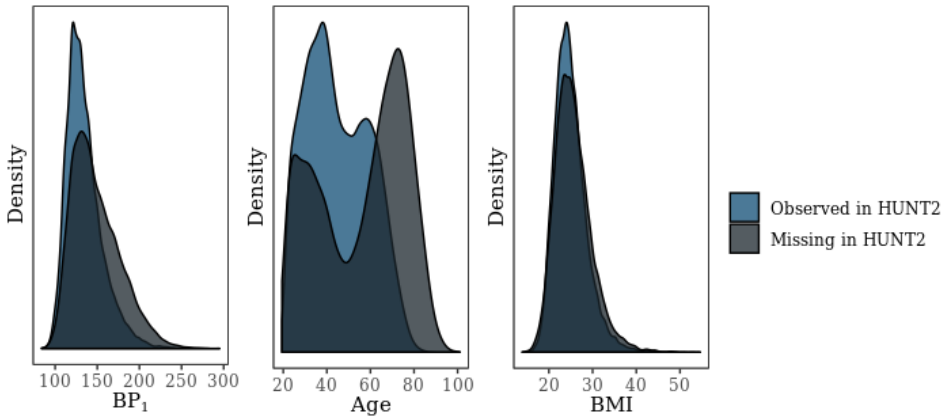

Figure 3.2: Distributions of blood pressure, age and BMI, respectively, in HUNT1, of the participants that also participated in HUNT2 (blue) and those that did not (grey).

3.3 Exploratory modelling

In order to explore how the different covariates influence the blood pressure in HUNT2, a simple frequentist blood pressure model is fitted to the data. This is a linear regression model with the blood pressure in HUNT2 as response variable, and the variables from HUNT1 as covariates. Only the participants with a measured blood pressure in HUNT2 are included, since the rest are missing. The model is therefore a complete-case model, in which the dropout process is ignored. The numerical covariates are all standardized in the following. The model is given by

$$\mathrm{BP}_{2,i} = \alpha_0 + \alpha_{\mathrm{BP}} \cdot \mathrm{BP}_{1,i} + \alpha_{\mathrm{age}} \cdot \mathrm{age}_i + \alpha_{\mathrm{sex}} \cdot \mathrm{sex}_i + \alpha_{\mathrm{BMI}} \cdot \mathrm{BMI}_i + \epsilon_i, \qquad (3.2)$$

where α_0, α_{BP}, α_{age} and α_{sex} are regression coefficients and ϵ is a normally distributed individual random effect.

In Table 3.3, resulting parameter estimates are presented. It can be seen that the blood pressure and age in HUNT1 have most impact on the blood pressure ten years later. For instance, if one has a blood pressure in HUNT1 that is one standard deviation higher than the mean, then the systolic blood pressure in HUNT2 is on average 16.2 mmHg higher, according to the model. Also sex and BMI play important roles, and all variables are highly significant.

Table 3.3: Summary of the simple linear blood pressure model (3.2) fitted to those who participated in both HUNT1 and HUNT2.

	Estimate	Std. Error	Pr(>|t|)
α_0	148.9416	0.1241	$< 2e{-}16$
α_{BP}	16.1638	0.1242	$< 2e{-}16$
α_{age}	6.5582	0.1256	$< 2e{-}16$
α_{sex}	-1.4112	0.1775	$1.9e{-}15$
α_{BMI}	1.5266	0.1018	$< 2e{-}16$

Also, a logistic model with the missing indicator variable in HUNT2 as response is fitted in order to model the dropout process. As whether or not a person shows up in HUNT2 is a binary variable, a logistic model is suitable. The same covariates as in the linear blood pressure model are included, such that

$$\begin{aligned}\text{logit}(p_i) = \log\big(\frac{p_i}{1-p_i}\big) = \beta_0 + \beta_{\text{BP}} \cdot \text{BP}_{1,i} + \beta_{\text{age}} \cdot \text{age}_i + \\ \beta_{\text{sex}} \cdot \text{sex}_i + \beta_{\text{BMI}} \cdot \text{BMI}_i,\end{aligned} \tag{3.3}$$

where β_0, β_{BP}, β_{age} and β_{sex} are regression coefficients and p_i is the probability that participant i of HUNT1 does not show up in HUNT2, i.e. $m_i = 1$, such that $m_i \sim \text{Bernoulli}(p_i)$.

Table 3.4 summarizes the dropout model. Age is the most important factor explaining whether or not participants show up in HUNT2, but also sex and the measured blood pressure in HUNT1 are important. A high age, high blood pressure in HUNT1 and being a male increase the dropout probability considerably, whereas high BMI slightly decreases the probability of not showing up.

In Figure 3.3, the coefficients of the dropout model are interpreted in terms of probability of dropout. To the left, the probability of dropout before HUNT2, according to this model, is plotted as a function of the blood pressure in HUNT1 for males and females, respectively, for a person being 50 years old and having a BMI of 22 in HUNT1. To the right, the probability of dropout is plotted as a function of age for a participant with systolic blood pressure in HUNT1 of 140 and a BMI of 22. As the dropout probability changes dramatically from low to high age, and from low to high BP_1, it is clear that these two factors have major impact on whether or not one shows up in HUNT2.

Table 3.4: Summary of the logistic dropout model (3.3).

	Estimate	Std. Error	Pr(>|z|)
β_0	-0.8637	0.0136	$< 2e{-}16$
β_{BP}	0.3059	0.0114	$< 2e{-}16$
β_{age}	0.5092	0.0114	$< 2e{-}16$
β_{sex}	0.3325	0.0187	$< 2e{-}16$
β_{BMI}	-0.0871	0.0100	$< 2e{-}16$

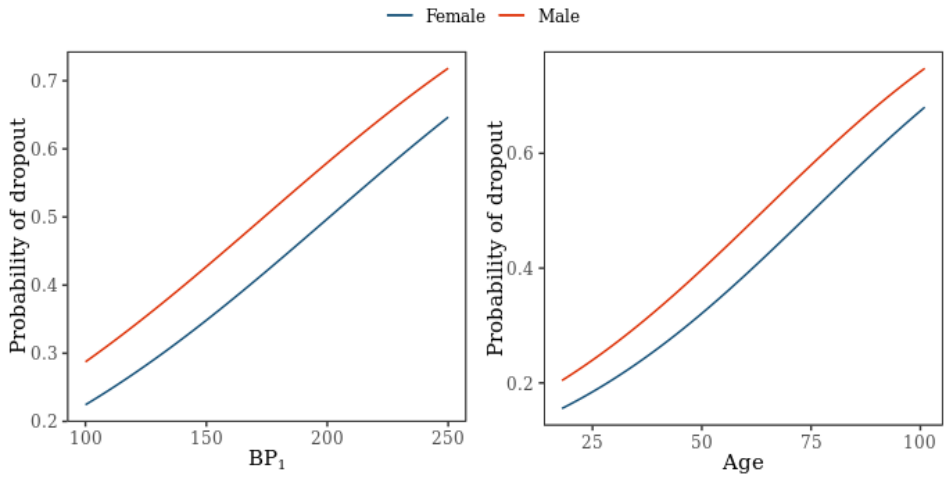

Figure 3.3: Probability of dropout given by a simple dropout model for females and males, as a function of BP_1 and age, respectively. To the left, the participant is 50 years old and has a BMI of 22 in HUNT1. To the right, the participant has a systolic blood pressure of 140 and a BMI of 22 in HUNT1.

Since the dropout regression itself is highly significant, meaning that some of the coefficient estimates other than the intercept with a high certainty deviate from zero, the data are at least MAR. For example, as increasing age gives a higher chance of being missing, the group of missing individuals is not a completely random subsample of the participants in HUNT1. If the data really are MAR, and not MNAR, the parameter estimates of the simple blood pressure model (3.2), in Table 3.3, are unbiased. However, the observed blood pressure distribution in HUNT2 is biased to that of the full population. Figure 3.4 serves as an example of how MAR provides biased distributional properties. Here, the blood pressure in HUNT2 is predicted for those missing, and the distribution of predicted blood pressures is compared to the observed distribution. The predicted distribution obtained from the linear model differs substantially from the distribution of the observed blood pressures. Therefore, if one for example wants to report the mean systolic blood pressure of the population based on the dataset used in this work, it is not enough to take the mean of the observed blood pressures in HUNT2. In order to obtain an unbiased marginal distribution of BP_2, the covariates need to be taken into account (Little and Rubin, 2002).

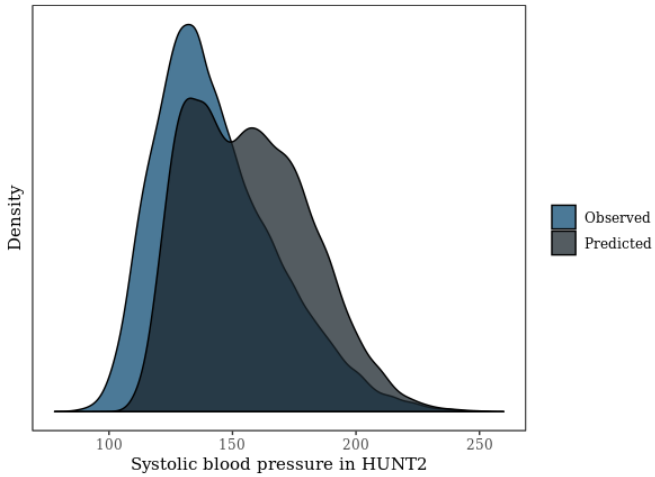

Figure 3.4: The distribution of observed blood pressures in HUNT2, together with the distribution of the predicted blood pressures in HUNT2 of those that dropped out prior to HUNT2, using the blood pressure model (3.2).

From this first exploration, it is clear that the variables considered from HUNT1 are influencing the blood pressure in HUNT2, based on the participants showing up in both of the surveys. Also, the variables influence the probability of dropout, so the data are clearly at least MAR. However, it is not possible to tell by inspection whether the data really are MNAR. In the next chapter, a Bayesian model accounting for MNAR dropout is formulated and compared to a Bayesian model assuming MAR dropout. In order to keep the models simple, BMI is kept out of the models to begin with, as BMI, according to the dropout model (3.3), is the least important covariate in explaining dropout. Later on, however, BMI is reintroduced.

4 | Models and method

The main goal of this work is to formulate a model that is able to make unbiased inferences about parameters describing the blood pressure in HUNT2 based on HUNT1. Two different Bayesian models are therefore proposed in this chapter. These are, respectively, a naive blood pressure model, introduced in Section 4.1, and a joint model, more specifically a shared parameter model, introduced in Section 4.3. The models presented are all fitted using the INLA framework through R-INLA, and the continuous covariates are standardized.

The naive model proposed in Section 4.1 ignores the dropout process and hence assumes that the data are MAR, similarly as the frequentist, exploring model (3.2) in Chapter 3.3. By contrast, the model proposed in Section 4.3 takes the dropout process into account together with the blood pressure model. The dropout process has to be well specified in order for such joint models to work (Mason et al., 2010). Therefore, a separate, naive dropout model is formulated in Section 4.2, in which possible non-linear effects between the covariates and the additive predictor are explored. The results from this modelling provide the basis for the dropout part of the joint model.

A major advantage of Bayesian modelling, in addition to the fact that one can incorporate data from external sources or prior knowledge, is that these models relatively easy can be adapted to allow for a possible non-ignorable missing data mechanism. This is done by adding a link between the missing values of the blood pressure model and the dropout model which models the probability of being missing. The blood pressure and dropout parts of the model are then jointly fitted. This is the case in the shared parameter model that is formulated in Section 4.3. The core of this model is that a random effect is shared between the blood pressure model and the dropout model.

4.1 Naive blood pressure model

With BP_1, age and sex as covariates, a simple Bayesian linear blood pressure model is given by

$$\mathrm{BP}_{2,i} = \alpha_0 + \alpha_{BP} \cdot \mathrm{BP}_{1,i} + \alpha_{age} \cdot \mathrm{age}_i + \alpha_{sex} \cdot \mathrm{sex}_i + \epsilon_i, \tag{4.1}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, with prior distributions assigned to the coefficients and σ_ϵ^2.

The coefficients have assigned the prior distributions

$$\alpha_0, \alpha_{\text{BP}}, \alpha_{\text{age}}, \alpha_{\text{sex}} \sim N(0, 10^6), \tag{4.2}$$

where the large variances reflect vague prior beliefs. Further, the prior used for σ_ϵ is

$$\sigma_\epsilon^2 \sim \text{invGamma}(1, 5 \cdot 10^{-5}), \tag{4.3}$$

which also contains little specific prior information about σ_ϵ.

This model is in the following referred to as the naive model. The reason for this is that the model assumes that the data are MAR, but not MNAR. The participants with missing blood pressure measurements in HUNT2 do not contribute to the likelihood, and the dropout process is ignored. Hence, if one were to assume that the data are MAR, then such a model is sufficient to provide unbiased inferences, but inferences are not valid if the data actually are MNAR.

4.2 Dropout model

Within the joint modelling approach that is proposed in the next section as an alternative to the naive model (4.1), a model for the dropout process also needs to be specified. Therefore, in this section, a naive logistic dropout model is proposed, in which the assumption of a linear relationship between the covariates and the additive predictor is relaxed. Thus, the continuous covariates BP_1 and age are modelled through smooth functions allowing for non-linear effects, such that

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + f(\text{BP}_{1,i}) + f(\text{age}_i) + \beta_{sex} \cdot \text{sex}_i, \\ m_i &\sim \text{Bernoulli}(p_i), \end{aligned} \tag{4.4}$$

where $f(\cdot)$ is a random walk model of order 2 as described in Chapter 2.3.

The goal of fitting this model is to investigate whether the covariates BP_1 and age are related non-linearly to the additive predictor in the dropout process, in order to be able to specify this correctly in the shared parameter model formulated in the next section. Mason et al. (2010) illustrate the importance of formulating a missing model which is a good approximation to the true missing process in order for joint models to reduce the bias from MAR models in the presence of MNAR, and to avoid convergence difficulties. Possible non-linearities could be discovered from (4.4), and hence be incorporated into the upcoming joint model to prevent such issues.

Priors are now set to the coefficients of the models and to the variance of the second order differences of the random walk models in (4.4), as given by (2.16). These priors are, respectively,

$$\beta_0, \beta_{\text{sex}} \sim N(0, 10^6), \tag{4.5}$$

and

$$\sigma^2_{\text{BP}}, \sigma^2_{\text{age}} \sim \text{invGamma}(1, 5 \cdot 10^{-5}). \tag{4.6}$$

These priors all incorporate non-specific prior information about the parameters into the model.

4.3 Shared parameter model

To account for data being MNAR, one generally needs to model the missing process together with the measurements. A naive blood pressure model as presented in Section 4.1 is then not satisfactory, as inferences must be made based on the full likelihood (2.7). In order to formulate a joint model for blood pressure values and the dropout process in the presence of data assumed to be MNAR, a shared parameter model inspired by Steinsland et al. (2014) is introduced. A shared random effect is included in the factorization of the full likelihood, such that the factorization takes the form (2.12). The random effect $\boldsymbol{\epsilon}$ characterizes the individual-specific blood pressure levels in HUNT2. By sharing this effect, a certain dependence between the blood pressure in HUNT2 and the dropout process is induced.

The shared parameter model (SPM) consists of two submodels, which is one model for the blood pressure, and one model for the dropout process, connected through $\boldsymbol{\epsilon}$. The blood pressure model is

$$\text{BP}_{2,i} = \alpha_0 + \alpha_{BP} \cdot \text{BP}_{1,i} + \alpha_{age} \cdot \text{age}_i + \alpha_{sex} \cdot \text{sex}_i + \epsilon_i, \tag{4.7}$$

with the same covariates as the naive model. $\boldsymbol{\epsilon}$ is a normally distributed random effect, $\epsilon_i \sim N(0, \sigma^2_\epsilon)$.

Further, $\boldsymbol{\epsilon}$ is shared with the missing data model, such that there is a relationship between the blood pressure model and the dropout model. The dropout process is modelled by logistic regression, given by

$$\text{logit}(p_i) = \beta_0 + \beta_{BP} \cdot \text{BP}_{1,i} + f(\text{age}_i) + \beta_{sex} \cdot \text{sex}_i + c \cdot \epsilon_i, \tag{4.8}$$

such that p_i is the probability that participant i will drop out before HUNT2. The missing data indicator is then $m_i \sim \text{Bernoulli}(p_i)$. Here $f(\cdot)$ is a random walk model of order 2. Further, c is an association parameter that, when different from zero, directly links the two submodels and makes missingness dependent on the underlying, potentially unobserved blood pressure in HUNT2 through the random effect ϵ_i. Due to this, the term *shared parameter model* is slightly misleading, since in fact $\boldsymbol{\epsilon}$ is treated as a covariate, and not a parameter, in the dropout submodel.

When fitting the separate dropout model (4.4) which includes smooth functions allowing for non-linear covariate effects, it turns out that age is clearly non-linear in the dropout process. This can be seen from Figure 5.1 in Chapter 5.1. This is the reason why age is fitted through a random walk function of order 2 in the dropout part (4.8) of the shared

parameter model. σ^2_{age} is the corresponding variance of the prior set to the second-order differences (2.16).

Further, the parameters of SPM have prior distributions assigned given by

$$\alpha_0, \alpha_{\text{BP}}, \alpha_{\text{age}}, \alpha_{\text{sex}}, \beta_0, \beta_{\text{BP}}, \beta_{\text{sex}} \sim N(0, 10^6), \tag{4.9}$$

where the variances are large in order to reflect vague prior knowledge.

The prior distributions of the hyperparameters must also be specified in order to obtain Bayesian inference from the shared parameter model. Prior distributions of the variances σ^2_ϵ and σ^2_{age} of the random effects are

$$\sigma^2_\epsilon, \sigma^2_{\text{age}} \sim \text{invGamma}(1, 5 \cdot 10^{-5}). \tag{4.10}$$

Further, the prior used for the association parameter c is

$$c \sim N(0, 1). \tag{4.11}$$

This prior reflects an initial belief that the parameter is not too far away from zero.

4.4 Inference from models

The objectives of inference for hierarchical Bayesian models are the posterior distributions of the latent variables and the hyperparameters. For the shared parameter model, a traditional approach would be to use Markov chain Monte Carlo methods to sample from the posterior distributions of the latent variables $\boldsymbol{\theta} = (\alpha_0, \alpha_{\text{BP}_1}, \alpha_{\text{age}}, \alpha_{\text{sex}}, \beta_0, \beta_{\text{BP}_1}, \beta_{\text{sex}}, f)$ and the hyperparameters $\boldsymbol{\phi} = (\sigma^2_{\text{age}}, \sigma^2_\epsilon, c)$. However, these methods are often not very efficient. With Gaussian priors assigned to the the latent variables, the shared parameter model can be shown to be a latent Gaussian Markov random field model, instead making it suitable for the INLA methodology.

The structured additive predictor of the shared parameter model can be written on the form (2.17),

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}) + \epsilon_i, \tag{4.12}$$

where f models the effects of age in the dropout process. In addition, the random iid effect $\boldsymbol{\epsilon}$ is copied into the linear predictor for the binomial observations, with the scaling parameter c. Since the latent variables have Gaussian priors, the shared parameter model is a latent Gaussian model.

Further, the latent variables are conditionally independent, so both the blood pressure model and the dropout model are Gaussian Markov random field models. Hence, the joint model is also a Gaussian Markov random field model. Additionally, the number of

hyperparameters ϕ is small. Therefore, the shared parameter model satisfies the conditions for models that can be fitted efficiently by INLA.

The joint model consists of both Gaussian observations, which are the blood pressures, and of binomial observations, which are the missing data indicator values. INLA supports models with multiple likelihoods, and it is possible to group the observations such that they have different additive predictors. Here, the data are divided into two groups such that the blood pressure data have a Gaussian likelihood assigned and the missing data indicators have a binomial likelihood assigned. The latent variables corresponding to each group are then linked by manipulating the structure of the data frame containing the two response variables and the covariates. If the number of individuals in the data is n, then the length of the additive predictor is $2n$. The latent variables in the Gaussian model are only defined for $i = 1, \ldots, n$, while the latent variables in the binomial model are defined for $i = n + 1, \ldots, 2n$. See the R code in Appendix B for details.

It should be noted that technically, when specifying the model in INLA, ϵ is split into ϵ_1 and ϵ_2, two separate normally distributed individual random effects with variances $\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_2}$, respectively. It is in fact ϵ_1 which is shared with the dropout part of the model. However, identifiability issues arise when σ_{ϵ_1}, σ_{ϵ_2} and c are simultaneously to be estimated. In order to overcome this, σ_{ϵ_2} is fixed. Here, it is fixed to a small value, 0.01. This means that almost all variability is concentrated to σ_{ϵ_1}, so $\sigma_\epsilon \approx \sigma_{\epsilon_1}$. ϵ_2 can thus be neglected. One can therefore without harm use ϵ and ϵ_1 interchangeably. Thus, only ϵ and σ_ϵ are referred to.

Now, inferences can be obtained based on the joint likelihood of the blood pressure observations and the missing indicator observations, together with the prior distributions defined. Similarly, the naive blood pressure model has Gaussian priors assigned with conditional independence between the latent variables, so the model is a Gaussian Markov random field, making it also suitable for the INLA framework.

Both the naive model and the shared parameter model are fitted to the HUNT data. The main interest lies now in the posterior estimates of the parameters of the blood pressure model, α_0, α_{BP}, α_{age} and α_{sex}, from the two different approaches. Also, interest lies in the variability σ_ϵ estimated from the two models, and, especially, the parameter c of the shared parameter model, as this is an indication of data being MNAR if it differs from zero.

4.5 Simulation studies

Two simulation studies are conducted in order to test the performance of the naive model and the shared parameter model on data with known, underlying dropout processes and parameters.

The simulation studies are based on the participants of the HUNT dataset. Further, the parameter estimates obtained by the shared parameter model are used to simulate blood

pressures in HUNT2 and simulate the dropout process for HUNT2. The parameters used in the first simulation study, given in Table 4.1, coincide with the posterior mean parameter estimates from SPM on the HUNT data. Since the true c is non-zero in this simulation study, the data are MNAR.

Table 4.1: Parameters used to simulate blood pressures in HUNT2 and to simulate the dropout process for HUNT2 in the simulation studies.

Parameter	True value in study 1	True value in study 2
α_0	152.49	152.49
α_{BP}	17.45	17.45
α_{age}	7.31	7.31
α_{sex}	-0.53	-0.53
β_0	0.13	0.13
β_{BP}	0.26	0.26
β_{sex}	0.44	0.44
σ_ϵ	18.10	18.10
c	0.046	0

In this simulation study, 100 new datasets are simulated. For each dataset, a shared parameter model and a naive model are fitted. The simulation procedure is given in Algorithm 1. The estimates obtained over the 100 simulations by the two approaches are evaluated by looking at the mean of posterior means, bias and coverage. The bias of a parameter estimate is here, in fact, a mean of biases over the 100 fitted models, and is given by

$$\text{Bias}(\hat{\theta}) = \frac{1}{100}\sum_{s=1}^{100}(\hat{\theta}_s - \theta), \tag{4.13}$$

where $\hat{\theta}_s$ is the posterior mean parameter estimate of a true parameter θ in simulation s. In addition, for each fitted model, a 95% equal-tailed credible interval A_s is obtained for each parameter. An indicator function,

$$I_s(\theta) = \begin{cases} 1, & \theta \in A_s, \\ 0, & \theta \notin A_s, \end{cases}$$

states whether or not the true, underlying parameter θ is covered by this interval in simulation s. Further, the coverage of a parameter is the fraction of simulations where the credible interval covers the true parameter. Thus, it is defined by

$$C = \frac{1}{100}\sum_{s=1}^{100} I_s(\theta). \tag{4.14}$$

Using these evaluation criteria, the shared parameter model can be compared with the naive model on the known MNAR data.

Algorithm 1: Simulation studies

Initialize true α_0, α_{BP_1}, α_{age}, α_{sex}, β_0, β_{BP_1}, $f(\mathrm{age})$, β_{sex}, c, σ_ϵ
for *s in 1:100* **do**
  Create $\epsilon_s \sim N(0, \sigma_\epsilon^2)$
  Give all participants new BP_2 using α_0, α_{BP_1}, α_{age}, α_{sex} and ϵ_s
  Make some participants missing in HUNT2 using β_0, β_{BP_1}, β_{sex}, $f(\mathrm{age})$, c and ϵ_s
  Fit SPM
  Fit naive model
end
return Parameter estimates obtained by each model in each simulation

In addition, a second simulation study is conducted, in which the underlying parameters are such that the data are MAR instead of MNAR. The posterior mean estimates from SPM on the HUNT data are still used to simulate new underlying blood pressures and to simulate the dropout process in HUNT2, but the parameter c is now set to 0 instead of the posterior mean provided by SPM. Except this, all parameters used are similar to those of the first part of the simulation study, as given in Table 4.1. Since the true value of c now is 0, there is no connection between the underlying blood pressure in HUNT2 and whether one shows up in HUNT2, after according for the observed covariates. Since the data are MAR, the naive model is expected to provide unbiased inferences. In order for the shared parameter model to be reliable under MNAR, it is vital that it also estimates the parameters unbiased under MAR, and that the association parameter c is estimated to be close to 0.

Similarly as the first simulation study, the second study also follows Algorithm 1. 100 new datasets are generated, and for each simulated dataset, both a shared parameter model and a naive model are fitted. The results are summarized through the mean of posterior means, coverage and bias.

4.6 Prior sensitivity

The association parameter c in the shared parameter model is of special interest, as this parameter identifies the dependence between the underlying blood pressure in HUNT2 and the dropout process, if there is such a dependency. It is therefore of interest to see whether inferences vary depending on the choice of prior for c. A small analysis of prior sensitivity is conducted, in which the shared parameter model is fitted to the HUNT data with three different Gaussian priors for c. The different priors tested are given in Table 4.2. SPM-P2 has a prior for c with higher variance than the original SPM, whereas the prior in SPM-P3 has a non-zero mean. Parameter estimates obtained from the three different models are still compared to those obtained from the naive model.

Table 4.2: Name of the different shared parameter models and their corresponding prior for c.

Model name	Prior for c
SPM	$N(0, 1^2)$
SPM-P2	$N(0, 10^2)$
SPM-P3	$N(1, 1^2)$

4.7 Extended models with BMI

The blood pressure model looked at so far is mostly for illustrative purposes, as it only includes a few, although very important, covariates. One would typically consider exploring more complex models, in which more covariates, or possibly interaction effects, are included as well. In the exploratory modelling section 3.3, BMI was shown to be an important factor in predicting the blood pressure in HUNT2, based on the participants that showed up. Therefore, a first natural extension is to add BMI to the shared parameter model, and then see how the parameter estimates change. BMI is added as a covariate, both in the blood pressure part and in the dropout part. Correspondingly, the naive model is also fitted with BMI added.

The naive model and the blood pressure part of the shared parameter model become

$$\mathrm{BP}_{2,i} = \alpha_0 + \alpha_{BP} \cdot \mathrm{BP}_{1,i} + \alpha_{age} \cdot \mathrm{age}_i + \alpha_{sex} \cdot \mathrm{sex}_i + \alpha_{\mathrm{BMI}} \cdot \mathrm{BMI}_i + \epsilon_i, \quad (4.15)$$

while the dropout part of the shared parameter model is

$$\mathrm{logit}(p_i) = \beta_0 + \beta_{BP} \cdot \mathrm{BP}_{1,i} + f(\mathrm{age}_i) + \beta_{sex} \cdot \mathrm{sex}_i + \beta_{\mathrm{BMI}} \cdot \mathrm{BMI}_i + c \cdot \epsilon_i. \quad (4.16)$$

Prior distributions are the same as defined in sections 4.1 and 4.3, together with the additional priors $\alpha_{\mathrm{BMI}}, \beta_{\mathrm{BMI}} \sim N(0, 10^6)$.

5 | Results

To begin with in this chapter, the resulting non-linear effects in the dropout model (4.4) are presented. These results provide the basis for the dropout model specification in SPM. Then, parameter estimates obtained from SPM and the naive model on the HUNT data are compared. Further, results from the simulation studies, from the prior sensitivity analysis and from the extended models with BMI are presented, before a few remarks about computational issues are provided.

5.1 Non-linear effects in dropout model

In the separate dropout model (4.4) formulated in Section 4.2, the covariates BP_1 and age are modelled through a random walk model of order 2. The resulting effects of these covariates on the additive predictor are plotted in Figure 5.1. Here, the effect of BP_1 is more or less linear, whereas the age effect is clearly non-linear. The youngest participants of HUNT1 are less likely to show up in HUNT2 than middle-aged participants. The probability of dropout then increases considerably the older one gets, after a certain age. Hence, this motivates the inclusion of the non-linear age effect in the dropout part of the shared parameter model, while it is justified to model BP_1 linearly.

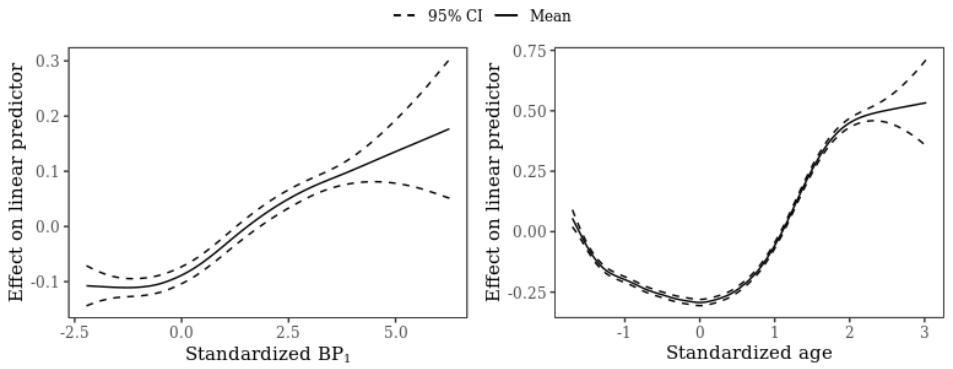

Figure 5.1: Effects of the covariates BP_1 and age on the additive predictor in the dropout model (4.4).

5.2 Parameter estimates

The parameter estimates obtained when fitting the naive model and the shared parameter model to the HUNT data are presented in Table 5.1. Further, the posterior densities of the blood pressure parameters obtained by the naive model and the shared parameter model are shown in Figure 5.2.

Table 5.1: Parameter estimates, in terms of posterior mean and 95% equal-tailed credible interval, obtained by the shared parameter model and the naive model.

	SPM		Naive	
	Posterior mean	95 % CI	Posterior mean	95 % CI
α_0	152.49	(152.23, 152.74)	148.97	(148.72, 149.22)
α_{BP}	17.45	(17.21, 17.69)	16.64	(16.40, 16.88)
α_{age}	7.31	(7.06, 7.56)	6.88	(6.63, 7.13)
α_{sex}	-0.53	(-0.89, -0.17)	-1.31	(-1.67, -0.95)
β_0	0.13	(0.04, 0.22)	-	-
β_{BP}	0.26	(0.23, 0.29)	-	-
β_{sex}	0.44	(0.40, 0.49)	-	-
σ_ϵ	18.10	(17.98, 18.18)	17.48	(17.39, 17.56)
c	0.046	(0.046, 0.046)	-	-

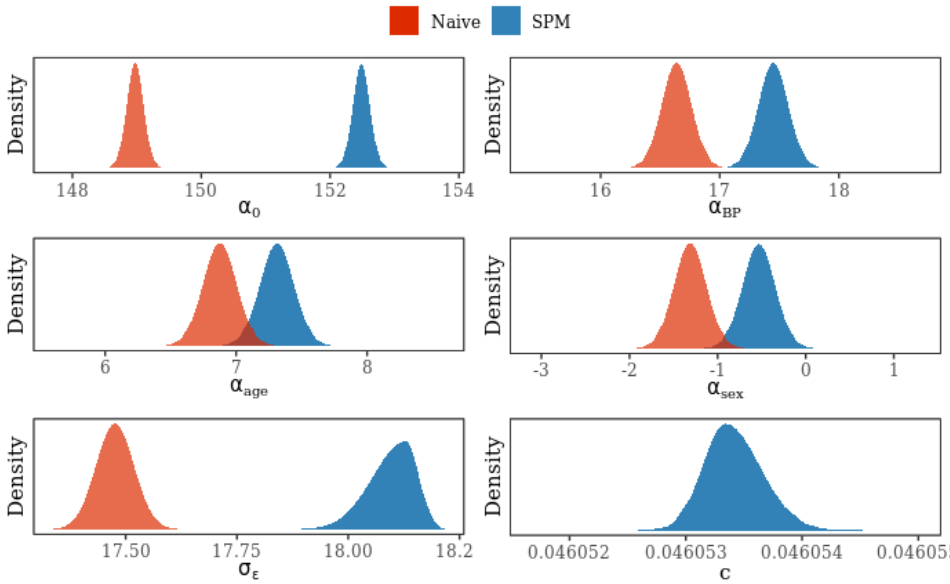

Figure 5.2: Posterior distributions of α_0, α_{BP}, α_{age}, α_{sex}, σ_ϵ and c from the shared parameter model and the naive model.

The parameter estimates obtained from the shared parameter model deviates from those obtained from the naive model. The blood pressure parameters α_0, α_{BP}, α_{age} and α_{sex} are all estimated to be clearly larger by the shared parameter model than by the naive model. As the association parameter c is estimated to be non-zero in the shared parameter model, the model indicates that the data are MNAR. Since this parameter is estimated to be larger than zero, participants with a large, positive residual ϵ in the underlying blood pressure model have a higher chance of not showing up in HUNT2 than those with smaller or negative residuals. Thus, more people with a high underlying blood pressure at the time of HUNT2, not explained by the covariates of the model, do not show up than people with a lower blood pressure, according to SPM.

The differences of the parameter estimates between SPM and the naive model are reasonable if the data are MNAR as SPM suggests. For instance, if many of the observations with largest residuals are missing, then it is natural that the naive model underestimates the true variability of the data and hence estimates σ_ϵ to be lower than SPM. Furthermore, the fact that the intercept is underestimated in the naive model is also reasonable if many of the largest blood pressure values in HUNT2 are missing. Similarly, since an increase of BP_1, age and sex means a higher probability of dropout, it is also explainable that the naive model underestimates the corresponding parameters α_{BP}, α_{age} and α_{sex} if the data actually are MNAR as SPM suggests. The reason for this is that participants with for example a high age, that were likely to be missing if the data only were MAR, might not be missing after all due to a negative residual and hence a low blood pressure. This results in an underestimation of the corresponding regression coefficient of the covariate. This is exactly what can be seen in Figure 2.2. In the plot to the right, the data are MNAR, but dropout is also covariate-dependent, since the covariate is positively correlated with the probability of dropout, exactly as is the case with BP_1, age and sex. Therefore, as the complete-case naive model in the figure then underestimates both the intercept and the regression coefficient, it is also natural that the naive blood pressure model does the same on the HUNT data if SPM provide valid inference. In any case, if the parameter estimates from SPM are valid, then the naive model clearly proves its insufficiency.

In Figure 5.3, the probability of dropout for females with average systolic blood pressure in HUNT1 are plotted as a function of the residuals of the underlying blood pressure model, according to the shared parameter model, for three different ages. The probability of dropout is plotted for ± 2 standard deviations of the residual ϵ. The figure shows the effect of the term $c\epsilon$ in the dropout part of the model, (4.8). Although an estimate of $c = 0.046$ at first sight might look small, Figure 5.3 shows the opposite. The probability of dropout increases drastically with increased residual, and hence increased blood pressure. The model therefore states that there is a clear connection between the underlying blood pressure one has at the time of HUNT2, and whether or not one shows up. Also, the different dropout probabilities for different age groups are clearly illustrated in this figure. For example, an 80-year old participant of HUNT1, with a large residual, and hence a high underlying blood pressure at the time of HUNT2, will almost certainly not show up in HUNT2. By contrast, a 50-year old participant in HUNT1 with a low underlying blood pressure at the time of HUNT2 will with great probability show up in HUNT2.

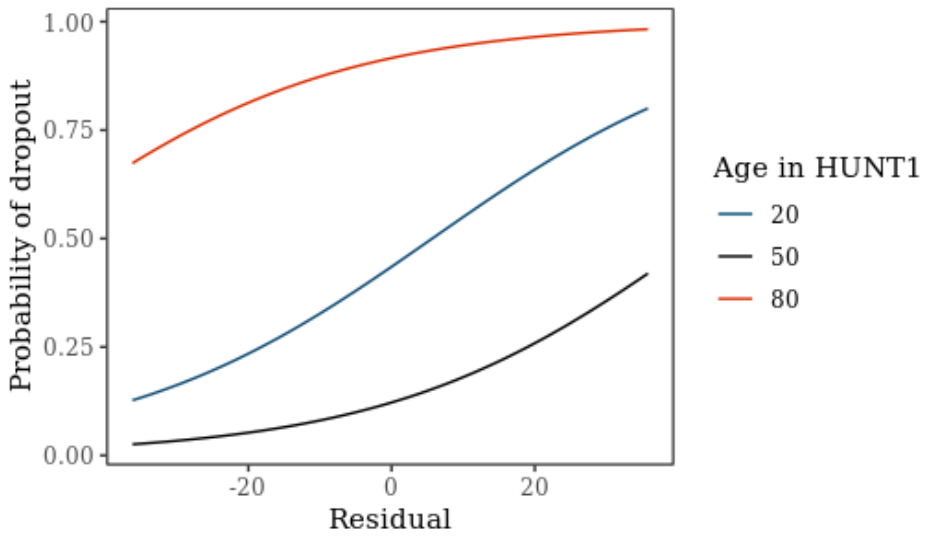

Figure 5.3: Probability of dropout for females with average systolic blood pressure in HUNT1 for three different age groups, as a function of their residual of the underlying blood pressure model, according to SPM.

5.3 Simulation studies

The results from the first simulation study, where the underlying data are MNAR, are presented in Table 5.2. Mean of the posterior means, bias and coverage of the parameter estimates obtained from the 100 simulated datasets are reported for the two models.

The shared parameter model clearly estimates the parameters better than the naive model. Biases are lower, and coverages are higher for all parameters. Having that said, the shared parameter model still underestimates the parameters somewhat, and coverages are relatively low, especially of the intercept and of σ_ϵ and c. In general, the shared parameter model states that there is a MNAR effect in the data, but that the effect is slightly lower than it really is.

Even though there are small biases in the estimates from SPM, it is interesting to see that the mean of the posterior means from the naive model in this simulation study, given in Table 5.2, almost coincide with the posterior means of the naive model on the true data, given in Table 5.1. The naive model therefore produces similar results on the known data simulated from the SPM parameters as on the true data. This could be an indication of the fact that the data really are MNAR as suggested by SPM.

Table 5.2: Summary of the parameter estimates obtained from the shared parameter model and the naive model in the first simulation study, when the underlying data are MNAR.

		SPM			Naive		
	True	Mean	Bias	Coverage	Mean	Bias	Coverage
α_0	152.49	151.50	-0.99	0.14	148.94	-3.55	0
α_{BP}	17.45	17.23	-0.22	0.49	16.65	-0.80	0
α_{age}	7.31	7.20	-0.11	0.82	6.94	-0.37	0.13
α_{sex}	-0.53	-0.75	-0.22	0.71	-1.35	-0.82	0
β_0	0.13	0.11	-0.02	0.96	-	-	-
β_{BP}	0.26	0.25	-0.01	0.79	-	-	-
β_{sex}	0.44	0.42	-0.02	0.75	-	-	-
σ_ϵ	18.10	17.70	-0.40	0.12	17.45	-0.65	0.03
c	0.046	0.033	-0.013	0.07	-	-	-

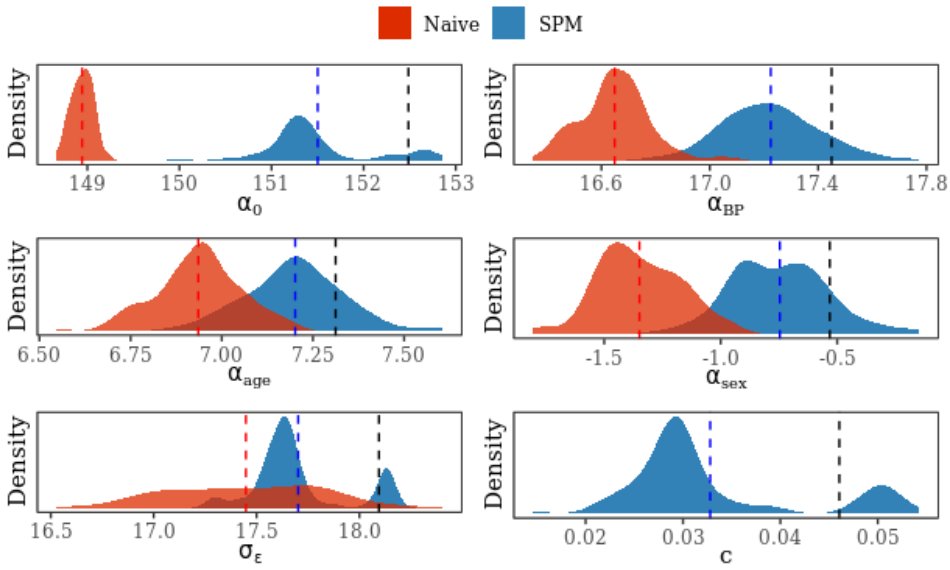

Figure 5.4: Density plots of the posterior means from the shared parameter model and the naive model obtained over the 100 simulated datasets in the first simulation study. The black line shows the true value of the parameter, whereas the blue and red lines indicate the mean of posterior means from SPM and the naive model, respectively.

In Figure 5.4, the distributions of the 100 posterior means obtained in the first simulation study from SPM and the naive model are plotted for α_0, α_{BP}, α_{age}, α_{sex}, σ_ϵ and c. Although it is clear that the posterior means of the shared parameter model estimates are much closer to the true parameter values than the posterior means from the naive model, it is evident that there are some identifiability issues involved. Thus, the INLA algorithm has problems exploring the posterior distributions correctly. To go deeper into this is outside

the scope of this thesis. In any case, the resulting posterior distributions are in most of the simulations slightly shifted from being around the true value. Similar behavior could possibly also be the case when fitting SPM to the true HUNT data, although this is difficult to state categorically. However, biases are unquestionably reduced going from the naive model to SPM on the MNAR data. Still, to investigate the identifiability issues further and possibly optimize SPM is of interest in future work.

In Table 5.3, the results from the second simulation study are presented. Here, the true value of c is 0, meaning that the data really are MAR and not MNAR. The naive model provides in this case more or less unbiased estimates. The shared parameter model also estimates the parameters with little bias and relatively high coverage, but performs at first sight not completely as well as the naive model. However, Figure 5.5 explains why there are marginally higher biases from SPM in this simulation study. In this figure, the distributions of the posterior means of α_0 and c, obtained from the 100 simulated datasets, are shown. It is clear that the posterior means are more or less concentrated around the true parameter values, but a few extreme outliers are present. It is not unlikely that these outliers also are the result of some identifiability issues. These outliers influence the mean of posterior means to become slightly biased. Due to this, also the coverages of the estimated parameters from SPM become marginally lower than from the naive model, except the coverage of α_0, which is considerably lower regardless of the outliers. However, if these few outliers are disregarded, the shared parameter model provides more or less unbiased estimates to the parameters when the data are MAR, and it performs as good as the naive model.

Table 5.3: Summary of the parameter estimates obtained from the shared parameter model and the naive model in the second simulation study, when the underlying data are MAR.

		SPM			Naive		
	True	Mean	Bias	Coverage	Mean	Bias	Coverage
α_0	152.49	152.84	0.35	0.58	152.49	0.00	0.96
α_{BP}	17.45	17.52	0.07	0.87	17.46	0.01	0.98
α_{age}	7.31	7.41	0.10	0.90	7.31	0.00	0.96
α_{sex}	-0.53	-0.73	-0.20	0.85	-0.54	-0.01	0.97
β_0	0.13	0.12	-0.01	0.92	-	-	-
β_{BP}	0.26	0.32	0.06	0.89	-	-	-
β_{sex}	0.44	0.48	0.04	0.89	-	-	-
σ_ϵ	18.10	18.30	0.20	0.49	18.20	0.10	0.21
c	0	0.010	0.010	0.33	-	-	-

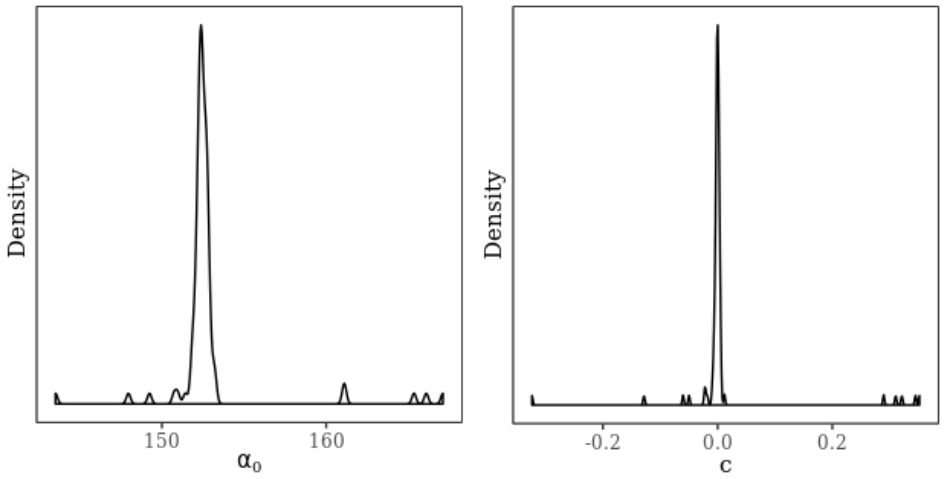

Figure 5.5: Distributions of the posterior means of α_0 and c from the 100 shared parameter models fitted in simulation study 2.

5.4 Prior sensitivity

The shared parameter model is in addition fitted with three different priors for c, given in Table 4.2, in order to see whether the model is sensitive to the choice of prior for c. 95% credible intervals for the parameters obtained from the three models are shown in Table 5.4. Further, 95% credible intervals for the parameters of the blood pressure model are plotted in Figure 5.6 for the three different shared parameter models and the naive model.

Table 5.4: 95% credible intervals obtained from the three different shared parameter models with the priors for c given in Table 4.2.

	Model		
	SPM	SPM-P2	SPM-P3
α_0	(152.23, 152.74)	(152.35, 152.86)	(152.26, 152.77)
α_{BP}	(17.21, 17.69)	(17.23, 17.21)	(17.21, 17.70)
α_{age}	(7.06, 7.56)	(7.08, 7.58)	(7.07, 7.57)
α_{sex}	(-0.89, -0.17)	(-0.87, -0.15)	(-0.89, -0.17)
β_0	(0.04, 0.22)	(0.03, 0.19)	(0.03, 0.19)
β_{BP}	(0.23, 0.29)	(0.24, 0.29)	(0.23, 0.29)
β_{sex}	(0.40, 0.49)	(0.40, 0.49)	(0.40, 0.49)
σ_ϵ	(17.98, 18.18)	(17.98, 18.19)	(17.98, 18.17)
c	(0.046, 0.046)	(0.048, 0.048)	(0.047, 0.047)

The resulting parameter estimates across these shared parameter models are consistent. Hence, the shared parameter model is apparently not highly prior sensitive with respect to the prior for the association parameter c which indicates how much MNAR the data

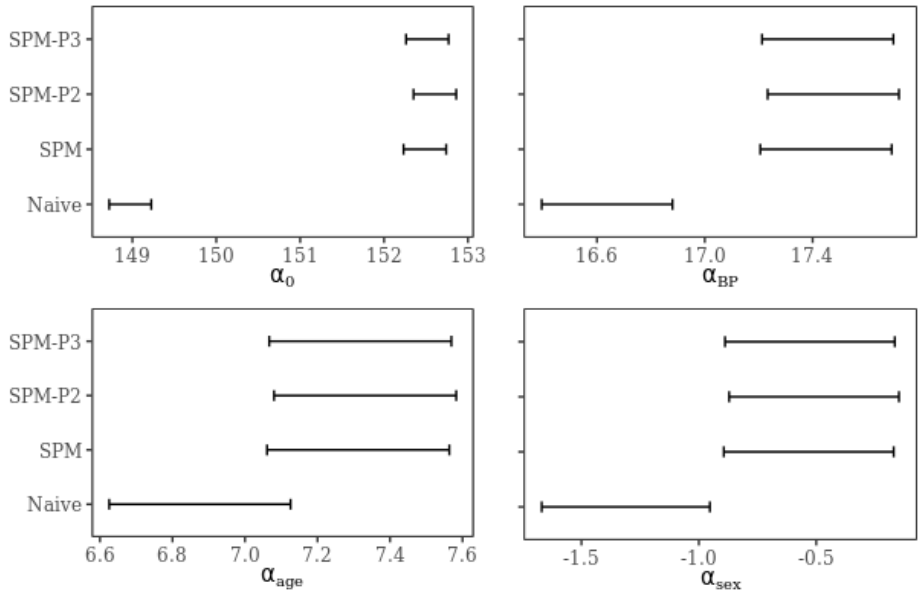

Figure 5.6: 95% credible intervals of the blood pressure model parameters from the naive model and the three shared parameter models with different priors for c

are. However, a further prior sensitivity analysis, in which several different kinds of priors for c, also other than normal distributions, are tested, could be conducted in further work. In addition, a prior sensitivity analysis where the priors of the other parameters also are varied is of interest.

5.5 Extended models with BMI

BMI is added as a covariate to the shared parameter model, both to the blood pressure part and the dropout part of the model, and also to the naive model. Resulting parameter estimates after the models are fitted to the HUNT data are given in Table 5.5. Furthermore, in Figure 5.7, the 95% equal-tailed credible intervals of α_0, α_{BP}, α_{age} and α_{sex} are plotted for the models with and without BMI as an additional covariate.

An interesting result is that the estimate of c increases noticeably. The posterior mean is now 0.059 compared to 0.046 without BMI, meaning that SPM suggests an increasing degree of MNAR after BMI is added. It is not necessarily intuitive that adding BMI to the model should lead to this. The case with BMI is that it is a relatively important covariate in the blood pressure model, as can be seen from the parameter estimates in Table 5.5, but in opposite to the other covariates included, it is not at all important in explaining whether one shows up in HUNT2, as β_{BMI} is estimated to be around 0. A possible explanation of why the model suggests a stronger degree of MNAR could therefore be that large residuals

Table 5.5: Parameter estimates, in terms of posterior mean and 95% equal-tailed credible interval, obtained by the shared parameter model and the naive model when BMI is included as a covariate.

	SPM		Naive	
	Posterior mean	95 % CI	Posterior mean	95 % CI
α_0	153.25	(152.99, 153.50)	148.95	(148.70, 149.20)
α_{BP}	17.20	(16.95, 17.45)	16.17	(15.92, 16.41)
α_{age}	7.16	(6.90, 7.41)	6.56	(6.31, 6.81)
α_{sex}	-0.48	(-0.84, -0.12)	-1.42	(-1.77, -1.07)
α_{BMI}	1.38	(1.17, 1.59)	1.53	(1.32, 1.73)
β_0	0.09	(0.01, 0.16)	-	-
β_{BP}	0.27	(0.24, 0.30)	-	-
β_{sex}	0.47	(0.43, 0.52)	-	-
β_{BMI}	0.02	(-0.01, 0.04)	-	-
σ_ϵ	18.03	(17.89, 18.18)	17.22	(17.21, 17.23)
c	0.059	(0.059, 0.059)	-	-

might be more correlated with dropout after BMI is added to the model than before. Participants with a large ϵ before BMI is added might get the residual explained to a certain degree simply by adding BMI into the blood pressure model. Then, since BMI does not contribute much to explain dropout, large residuals are even more correlated with dropout after adding BMI than before, and hence, c is larger.

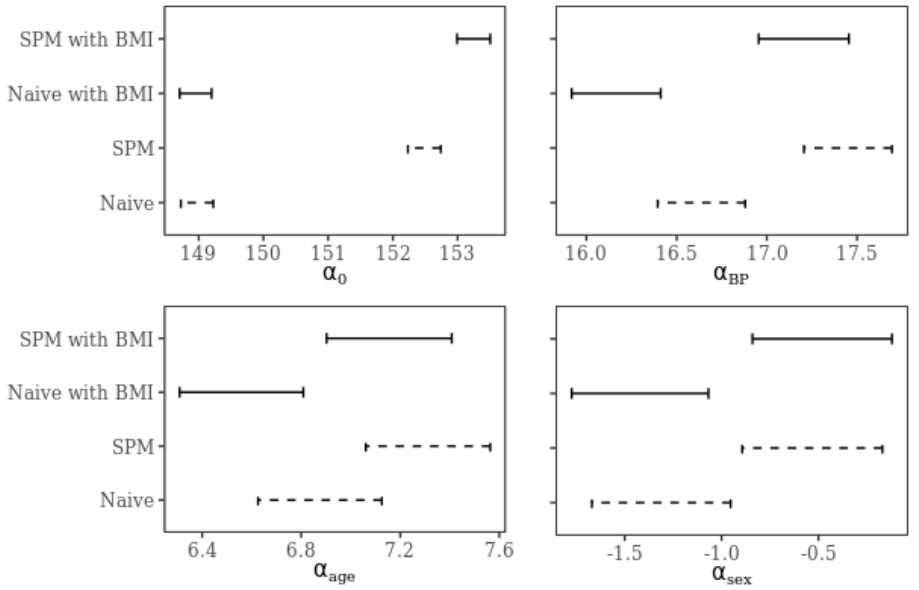

Figure 5.7: 95% credible intervals of the blood pressure model parameters of SPM and the naive model, with and without BMI as a covariate.

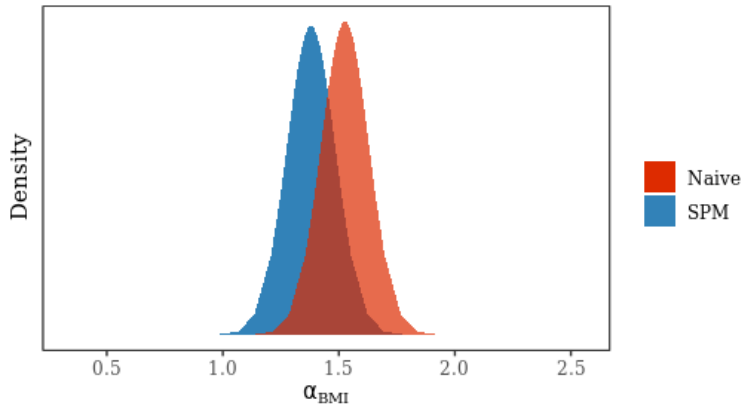

Figure 5.8: Posterior distribution of α_{BMI} obtained from SPM and the naive model after BMI is added as a covariate.

In the blood pressure model, the parameter estimates change slightly when adding BMI, as shown in Figure 5.7. However, the estimates from SPM and the naive model change almost synchronically with and without BMI, except the estimate for α_0, which increases in SPM but remains the same in the naive model. In Figure 5.8, the posterior distributions of α_{BMI} obtained from SPM and the naive model are plotted. The two distributions are to a certain degree overlapping. This is, in accordance with the discussion in Section 5.2, natural as BMI is not of much importance in the dropout process. If data are MNAR but dropout is not dependent on a covariate, then bias is added to the intercept, but the corresponding regression coefficient of that covariate remains unbiased.

5.6 Note on computational issues

As the shared parameter model is complex and as there are much data available, a few numerical and computational issues worth mentioning arise. Firstly, the resulting parameter estimates obtained by the shared parameter model vary slightly each time the model is fitted to the same data, also with the same priors. Secondly, when fitting the SPM, the posterior estimate of c has an extremely low standard deviation, meaning that the estimate is highly concentrated, as can be seen in Figure 5.2. The posterior mean of c varies in the interval $[0.046, 0.048]$ each time the model is fitted to the HUNT data, but the posterior is always extremely concentrated around the mean.

Additionally, when performing the simulation studies, there were some numerical issues. When fitting the shared parameter model to many of the simulated datasets, a warning showed up, stating that one of the eigenvalues of the Hessian matrix, which is used when exploring the distribution $\tilde{f}(\boldsymbol{\phi}|\boldsymbol{y})$ in the INLA algorithm, was negative, and that it was changed to become positive. This could possibly have influenced the precision of the estimates of the shared parameter model in the simulation studies. However, the estimates

were more or less consistent across the simulations, also when all Hessian eigenvalues were positive.

Also, in order for the shared parameter model to be able to converge, it is of huge importance to formulate a well-specified dropout process. For instance, a shared parameter model with a linear age effect in the dropout model does not converge within a reasonable time. Difficulties in achieving convergence when the dropout model is misspecified is in accordance with the findings by Mason et al. (2010). To allow for age to be modelled non-linearly is therefore a key to obtain meaningful results. The blood pressure model could also potentially include non-linear effects, but this is again more computationally expensive. Supplementary discussion of this is found in Appendix A.

6 | Discussion and conclusion

In this work, blood pressure data from HUNT1 and HUNT2 are considered, when some participants drop out of the follow-up study prior to HUNT2. Such missing values are commonly assumed to be MAR, but models that are based on this assumption might provide severely biased inferences if the data under study actually are MNAR. In the presence of data being MNAR, blood pressure data and the dropout process need to be jointly modelled. One approach for such modelling is proposed in this work, in which an individual random effect is shared between the blood pressure model and the dropout model. This shared parameter model is a Bayesian latent Gaussian Markov random field model. Approximate Bayesian inference is therefore obtained by fitting the model using the INLA methodology, through R-INLA. Inferences are compared to those obtained by a naive Bayesian model assuming MAR. Further, simulation studies are conducted in order to test the models on known, underlying parameters.

The parameter estimates obtained from the shared parameter model differ clearly from those obtained from the naive model. The association parameter c is estimated to be larger than zero, indicating that the data are MNAR. The results from the simulation studies partially support this.

Although the results seemingly show that the data are MNAR, one cannot categorically state that the parameter estimates obtained from the shared parameter model reflect the underlying truth. There are no possibilities to test empirically whether the data are MAR or MNAR, since that information is not available from the observed data. Every MNAR model has a MAR counterpart, meaning that they have exactly the same fit to the observed data, but differ in their predictions of the unobserved outcomes (Molenberghs et al., 2008). In general, in missing data problems, the conclusions about the dropout mechanism could therefore depend crucially on untestable distributional assumptions. If adding a non-random dropout component to a model leads to a noticeable change of the likelihood, then some real structure of the data is identified that the original model does not encompass (Kenward, 1998). Therefore, the fact that parameter estimates change remarkably from the naive model to the shared parameter model might tell more about the inadequacy of the naive model rather than the adequacy of the shared parameter model. Many authors underline the importance of being careful with interpreting evidence for or against data being MNAR by only using the data under study (Molenberghs and Verbeke, 2005). How well a model fits observed data can be assessed, but it is not possible to assess its fit to un-

observed data given the observed data. Therefore, a sensitivity analysis is recommended, in which, for example, different statistical models are considered simultaneously (Verbeke and Molenberghs, 2000). How stable inferences are across these models can be an indication of how much confidence one can place in them. A general strategy for sensitivity analysis is to consider different dependencies between the missing data process and the outcome or covariates (Verbeke and Molenberghs, 2000). A thorough approach to sensitivity analysis in the shared parameter framework is given by Creemers et al. (2010). Another approach to Bayesian sensitivity analysis can be found in Daniels and Hogan (2008).

It is also worth making a few remarks regarding some aspects of the data used from HUNT1 and HUNT2. For instance, as made clear in Chapter 3, all participants with missing values in the covariates from HUNT1 are removed from the analysis. In the dataset considered here, also training and educational level were included as variables, so only participants who also had records for these variables were selected. In total, 21851 individuals, who were only partly participating in HUNT1, were disregarded. In this work, only missing values in the response variable are considered. Still, missing values in the covariates could possibly affect the inferences drawn. Therefore, to also take missing covariate values into account in similar blood pressure studies would be interesting to look closer at in future work.

Another aspect worth some discussion is the fact that many of the oldest participants of HUNT1 do not show up in HUNT2 simply because they do not live anymore. Still, they are here regarded as having an underlying blood pressure at the time of HUNT2. This assumption is reasonable to question, as they in fact are dead. However, many of them might have died a short time before HUNT2, and their death could also be related to the blood pressure they had at the time of their death. The simplification done in this work is therefore not completely unreasonable. Ideally, one should perhaps include a binary variable stating whether the participants are dead or not in HUNT2 if that information is available, and separately model the dropout process for those that still are alive.

The models presented in this work are formulated with the purpose of illustrating how to account for data being MNAR. Typically, several other covariates and different effects are also included when modelling. The blood pressure model considered in this work could for instance include non-linear effects similarly as the dropout model of SPM. However, this is not done here. Further discussion around this matter is given in Appendix A.

In addition to extending the models to become more complex, other different approaches to further work are of interest. A thorough sensitivity analysis should be performed to check the robustness of the parameter estimates across different modelling assumptions. Also, as discussed previously, it is of interest to have a closer look at issues concerning parameter identifiability of SPM. Accordingly, possible adjustments to the shared parameter model should be considered. In addition, a detailed study of prior sensitivity is another suggestion for further work.

To summarize, it is evident from this work that wrongly assuming MAR when the data in reality are MNAR might lead to large biases. There is a clear indication that blood pressure data in HUNT are MNAR due to dropout. This indication is based on the parameter estimates obtained directly from the models, in addition to the results from the simulation studies. Even though the shared parameter model might not provide completely unbiased inferences, the naive model assuming MAR is by all accounts insufficient. Also, if the data actually are MAR, then the shared parameter model, according to the second simulation study, would provide more or less valid inferences. Hence, the shared parameter model is clearly preferable to the naive model. However, further analysis of sensitivity of different modelling assumptions should be conducted in order to be even more confident in the parameter estimates describing the blood pressure development.

Bibliography

Blangiardo, M., Cameletti, M., 2015. Spatial and Spatio-Temporal Bayesian Models with R-INLA. John Wiley & Sons, Inc.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., Kenward, M. G., 2010. A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. Biometrical Journal 52 (1), 111–125.

Daniels, M. J., Hogan, J. W., 2008. Missing Data In Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall/CRC.

Diggle, P., Kenward, M. G., 1994. Informative drop-out in longitudinal data analysis. Journal of the Royal Statistical Society. Series C (Applied Statistics) 43 (1), 49–93.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer.

Fitzmaurice, G., 2008. Missing data: implications for analysis. Nutrition 24, 200–202.

Fitzmaurice, G. M., Laird, N. M., Zahner, G. E. P., 1996. Multivariate logistic models for inclomplete binary responses. Journal of the American Statistical Association 91 (4), 99–108.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., 2014. Bayesian Data Analysis. Chapman & Hall/CRC.

Gómez-Rubio, V., 2020. Bayesian Inference with INLA. Chapman & Hall/CRC.

Hastie, T. J., Tibshirani, R. J., 1990. Generalized Additive Models. Chapman & Hall/CRC.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A. Y. M., Yang, C. W., 2013. Chronic kidney disease: global dimension and perspectives. The Lancet 382 (9888), 260–272.

Kearney, P. M., Whelton, M., Reynolds, K., 2005. Global burden of hypertension: analysis of worldwide data. The Lancet 365 (9455), 217–223.

Kenward, M. G., 1998. Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. Statistics in Medicine 17 (23), 2723–2732.

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, J., Stene, T. R., Bratberg, G., Heggland, J., Holmen, J., 2013. Cohort profile: The HUNT Study, Norway. International Journal of Epidemiology 42 (4), 968–977.

Lewington, S., Clarke, R., Qizilbash, N., Peto, R., Collins, A., 2002. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. The Lancet 360 (9349), 1903–1913.

Little, R. J. A., 1992. Regression with missing x's: A review. Journal of the American Statistical Association 87 (420), 1227–1237.

Little, R. J. A., 1995. Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association 90 (431), 1112–1121.

Little, R. J. A., Rubin, D. B., 2002. Statistical Analysis with Missing Data. John Wiley & Sons, Inc.

Martins, T. G., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with INLA: New features. Computational Statistics and Data Analysis 67, 68–83.

Mason, A. J., Best, N., Plewis, I., Richardson, S., 2010. Insights into the use of Bayesian models for informative missing data. `http://eprints.ncrm.ac.uk/1691/1/InsightsSubmitted.pdf`.

Molenberghs, G., Beunckens, C., Sotto, C., Kenward, M. G., 2008. Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 70 (2), 371–388.

Molenberghs, G., Kenward, M. G., Lesaffre, E., 1997. The analysis of longitudinal ordinal data with nonrandom drop-out. Biometrika 84 (1), 33–44.

Molenberghs, G., Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer.

Nelder, J. A., Wedderburn, R. W. M., 1972. Generalized linear models. Journal of the Royal Statistical Society: Series A (General) 135 (3), 370–384.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2), 319–392.

Schafer, J. L., Graham, J. W., 2002. Missing data: Our view of the state of the art. Psychological Methods 7 (2), 147–177.

Sparrow, D., Garvey, A. J., Rosner, B., Thomas Jr, H. E., 1982. Factors in predicting blood pressure change. Journal of the American Heart Association 65 (4), 789–794.

Steinsland, I., Larsen, C. T., Roulin, A., Jensen, H., 2014. Quantitative genetic modeling and inference in the presence of nonignorable missing data. Evolution 68 (6), 1735–1747.

Tobin, M. D., Sheehan, N. A., Scurrah, K. J., Burton, P. R., 2005. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Statistics in Medicine 24 (19), 2911–2935.

Van Buuren, S., 2018. Flexible Imputation of Missing Data. Chapman & Hall/CRC.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer.

Vonesh, E. F., Greene, T., Schluchter, M. D., 2006. Shared parameter models for the joint analysis of longitudinal data and event times. Statistics in Medicine 25 (1), 143–163.

Wang, X., Yue, Y. R., Faraway, J., 2018. Bayesian Regression Modeling with INLA. Chapman & Hall/CRC.

Whelton, P. K., 1994. Epidemiology of hypertension. The Lancet 344 (8915), 101–106.

Williams, B., Mancia, G., Spiering, W., et al., 2018. ESC/ESH Guidelines for the management of arterial hypertension. European Heart Journal 39 (33), 3021–3104.

Wilsgaard, T., Schirmer, H., Arnesen, E., 2000. Impact of body weight on blood pressure with a focus on sex differences. JAMA Internal Medicine 160 (18), 2847–2853.

World Health Organization, 2019. Hypertension. `https://www.who.int/news-room/fact-sheets/detail/hypertension`. Last accessed: 2020-03-18.

Wu, L., 2010. Mixed Effects Models for Complex Data. Chapman & Hall/CRC.

Wu, M. C., Carroll, R. J., 1988. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics 44 (1), 175–188.

A | Non-linear effects in blood pressure model

In the shared parameter model, age is modelled with a random walk-function of order 2 in the dropout part (4.8) after a separate dropout model (4.4) showed that age is not linear with respect to the log-odds of dropout. This specification is vital in order for the shared parameter model to converge. Having that said, in the naive blood pressure model, and in the blood pressure part of the shared parameter model, all covariates are modelled linearly. A blood pressure model allowing for non-linearities could instead have been proposed, such that

$$\mathrm{BP}_{2,i} = f(\mathrm{BP}_1) + f(\mathrm{age}) + \alpha_{\mathrm{sex}} \cdot \mathrm{sex} + \epsilon_i, \tag{A.1}$$

where $f(\cdot)$ is a random walk-model of order 2. When fitting this model, BP_1 turns out to be linear, but the effect of age on the additive predictor is actually non-linear, as can be seen in the left plot of Figure A.1. Therefore, a non-linear effect of age should perhaps be included in the blood pressure model as well as in the dropout model. However, there are several reasons for not doing so in this work. Firstly, introducing a random walk-function into the blood pressure models would reduce the interpretability of the models. Secondly, when specifying age non-linearly in the blood pressure part of the shared parameter model, it takes about twice as long time to fit the model, so it is not of computational benefit. Thirdly, and most importantly, the effect of age on the additive predictor in the blood pressure part of the shared parameter model turns actually out to be linear, as can be seen from the right plot of Figure A.1, even after allowing it to be non-linear. Thus, specifying the effect of age non-linearly in the dropout part is sufficient to have a linear effect of age in the blood pressure part. Therefore, it is reasonable to model all covariates completely linearly in the blood pressure models. Also, the most important in this work is to have the naive blood pressure model and the blood pressure part of the shared parameter model identically specified, such that they are comparable.

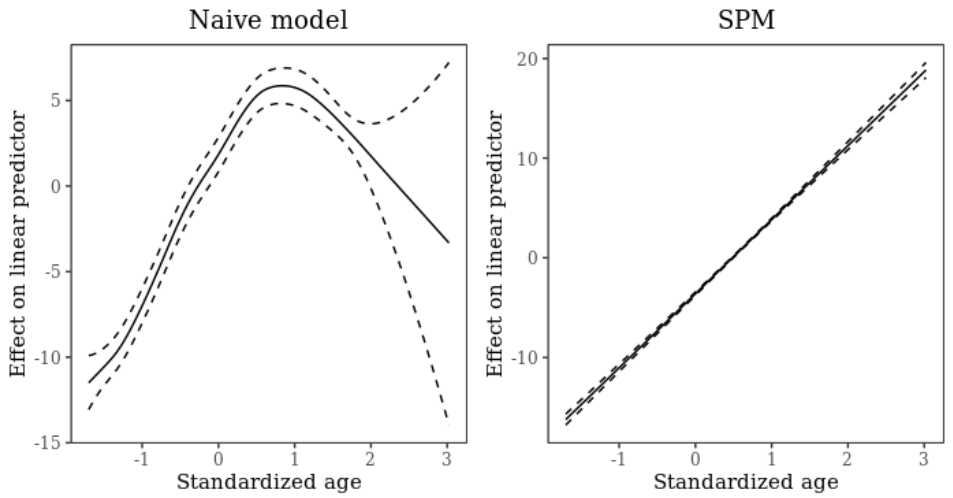

Figure A.1: Effect of age on the additive predictor in a naive blood pressure model when modelled with random walk function of order 2, and in the blood pressure part of a shared parameter model where age is modelled through a random walk function of order 2 both in the blood pressure and dropout submodels.

B | R code

B.1 Fitting models

```
library(INLA)
library(brinla)
INLA::inla.dynload.workaround()

# Data loaded into the data frame df

n <- nrow(df)

sigma2 <- 0.01 # Set to a neglectable value

# Prepare the response variables
y.gaussian <- c(df$bp2,rep(NA,n)
m.binomial <- c(rep(NA,n),df$missing)
joint.response <- list(y.gaussian,m.binomial)

# Specify number of trials for the binomial response variable
Ntrials <- c(rep(NA,n),rep(1,n))

linear.covariates <- data.frame(
        alpha.0 = c(rep(1,n),rep(NA,n)),
        beta.0 = c(rep(NA,n),rep(1,n)),
        y.BP1 = c(df$bp1,rep(NA,n)),
        y.AGE = c(df$age,rep(NA,n)),
        y.SEX = c(df$sex,rep(NA,n)),
        y.BMI = c(df$bmi,rep(NA,n)),
        m.BP1 = c(rep(NA,n),df$bp1),
        m.AGE = c(rep(NA,n),df$age),
        m.SEX = c(rep(NA,n),df$sex),
        m.BMI = c(rep(NA,n),df$bmi),
)
```

```
random.covariates <- data.frame(
        y.eps1 = c(1:n,rep(NA,n)),
        m.eps1 = c(rep(NA,n),1:n)
)

joint.data <- c(linear.covariates,random.covariates)
joint.data$Y = joint.response

formula.spm = Y ~ -1 + alpha.0 + y.BP1 + y.AGE + y.SEX +
        beta.0 + m.BP1 + m.SEX +
        f(m.AGE,model="rw2",constr=T) +
        f(y.eps1,model="iid") +
        f(m.eps1,copy="y.eps1",fixed=F,param=c(0,1))
# Copy eps1 into binomial dropout process
# Use option param to set mean and precision for Gaussian
# prior for association parameter

# Fit SPM:
spm <- inla(formula.spm,family=c("gaussian","binomial"),
        data=joint.data,Ntrials=Ntrials,verbose=T,
        control.family=list(list(initial=log(1/sigma2^2),
        fixed=T),list()))
# Use control.family to fix sigma2. Specified through
# log-precision

# Naive model:
formula.naive <- bp2 ~ bp1 + age + sex
naive <- inla(formula.naive,family="gaussian",
        data=df)
```

B.2 Simulation studies

```
# Load results obtained from SPM in B.1 in order to
# simulate new BP2 and dropout process

# True coefficients are posterior means from SPM:
a.0 <- spm$summary.fixed[[1]][1]
a.bp1 <- spm$summary.fixed[[1]][2]
a.age <- spm$summary.fixed[[1]][3]
a.sex <- spm$summary.fixed[[1]][4]
b.0 <- spm$summary.fixed[[1]][5]
b.bp1 <- spm$summary.fixed[[1]][6]
b.sex <- spm$summary.fixed[[1]][7]
# Obtain sigma as sd instead of precision:
sigma <- bri.hyperpar.summary(spm)[2]
c.sim1 <- bri.hyperpar.summary(spm)[3]
c.sim2 <- 0      # True c is 0 in simulation study 2

# Create function recreating the non-linear age effect
# in the dropout process
age.coeff <- function(age,age.spm,b.age.spm) {
    b.age <- b.age.spm[which(age.spm==age)]
}

age.spm <- spm$summary.random$m.AGE[[1]]
b.age.spm <- spm$summary.random$m.AGE[[2]]
b.age <- sapply(age,age.coeff,age.spm=age.spm,
    b.age.spm=b.age.spm)

# Create function simulating dropout process
make.missing.indicator <-
    function(n,b.0,b.bp1,b.age,b.sex,c,eps,df) {
        logit_p <- b.0 + b.bp1*df$bp1 + b.age +
            b.sex*df$sex + c*eps
        p <- exp(logit_p)/(1+exp(logit_p))
        p[logit_p>600] <- 1 # Avoid numerical error
        prob <- runif(n)
        missing <- p>prob
        return(as.numeric(missing))
    }

num.sim <- 100 # Number of simulations

# First simulation study:
for (i in 1:num.sim) {
```

```
    print(paste0(i,"/",num.sim))
    eps1 <- rnorm(n,mean=0,sd=sigma)
    eps2 <- rnorm(n,mean=0,sd=0.01)

    df$bp2 <- a.0 + a.bp1*df$bp1 + a.age*df$age +
        a.sex*df$sex + eps1 + eps2

    df$missing <- make.missing.indicator(n,b.0,b.bp1,
        b.age,b.sex,c.sim1,eps1,df)

    df$bp2[df$missing==1] <- NA

    # Use the code in B.1 to fit SPM and naive model
    # Save the results
}

# Repeat using c.sim2 instead of c.sim1 for the second
# simulation study
```