Bayesian Methods for Inference in Biostatistical Longitudinal Studies and Modelling of Missing Data

Hanadi Mohammed Alzahrani

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

July 2024

Abstract

Longitudinal studies repeatedly collect data from the same individuals over time to study long-term factors. A commonly used model in longitudinal studies is the linear mixed effects model, which considers the correlation between observations within individuals. There are two ways to fit the model in statistical fields: the Frequentist and Bayesian approaches. The Frequentist approach is widely used, while the Bayesian approach has become more common with computational advancements. The work in this thesis comprises a comparison study between the Frequentist linear mixed effects model and the Bayesian Hierarchical model, using simulated longitudinal data and data from a heart failure study (BIOSTAT-CHF). It was observed that inferences from both approaches were similar. However, the Bayesian approach offers an advantage by providing a probability distribution for the parameter estimates. This shows the probability of values falling within a certain range and incorporates prior information from previous studies into the inference.

In longitudinal studies, missing data is a common problem that can impact the statistical analysis estimates by producing biased estimates. A method that deals with non-ignorable missingness in the response using Correlated Random Effects (CRE) based on latent variables and Gibbs sampling has been proposed in the literature and has performed well in scenarios assuming semi-parametric modelling. However, when applied to linear mixed-effect modelling, the covariance matrix parameters had difficulty converging. To address this issue, the work in this thesis considers a weakly informative prior using the Inverse Wishart distribution. Additionally, this CRE method is unable to accommodate incomplete data in the analysis model explanatory variables. To address this problem, the work in this thesis proposed three methods to deal with missingness in the response and explanatory variables by adapting the CRE method.

Two proposed methods, the Two-Step and the GCRE-MAR methods, were designed to address non-ignorable missingness in the model response and ignorable missingness in the model explanatory variables. The GCRE-MNAR method was designed for non-ignorable missingness in both the model response and explanatory variables. In the Two-Step method, the CRE method was adapted by incorporating an additional step using the MICE algorithm, a common approach for handling MAR data and producing imputed datasets. The CRE method is then applied to the imputed MICE datasets.

The GCRE-MAR and GCRE-MNAR represent generalised versions of the CRE method. The GCRE-MAR method incorporates the incomplete explanatory variable model. The GCRE-MNAR method incorporates the incomplete explanatory variable model and the incomplete explanatory variable missingness process model. It considers correlated random effects between the incomplete explanatory variable model and the missingness process.

The proposed methods were compared with the CRE method and some baseline models using simulated longitudinal data for different numbers of repeated measures and missing proportion factors. The proposed methods perform similarly to the CRE method, given that the proposed methods consider missing data in both the response and explanatory variables. In contrast, the CRE method only has missing data in the response (no missing values are in the explanatory variables). Furthermore, the proposed methods outperform the available data method in out-of-sample predictive performance, and the parameter estimates closely match the parameters that generated the data.

Additionally, the proposed methods were applied to the BIOSTAT-CHF data, and the results were consistent regardless of the applied method. The correlated random effects indicated that the NT-proBNP missingness was MAR, and the eGFR missingness was MNAR. Finally, the sensitivity analysis for the misspecified missingness mechanism for the proposed methods had a small impact on the overall results, whereas the misspecified response missingness model resulted in biased parameter estimates for some of the analysis model coefficients.

Declaration

I declare that all the work presented in this thesis has been done by myself under the supervision of Dr. Benn Macdonald, Dr. Caroline E. Haig and Prof. John Cleland, except where otherwise stated. This thesis represents work completed between 2020 and 2024 in the School of Mathematics and Statistics at the University of Glasgow.

Acknowledgements

I would like to express my appreciation to my primary supervisor, Dr Benn Macdonald, whose guidance, consideration, and valuable feedback were crucial in completing this work. I am also profoundly grateful to my supervisors, Dr Caroline Haig and Prof John Cleland, for their kind support and encouragement throughout the stages of my PhD.

I thank the Ministry of Higher Education and King Saud bin Abdulaziz University for Health Sciences in Jeddah for sponsoring my PhD. I also want to thank the Saudi Arabian Cultural Bureau for their support during my study.

A warm thanks to my husband, Ahmed, for his support, patience, and belief in me. I appreciate that you left everything and travelled with me to achieve my goals. Your support has been crucial in enabling me to continue my work. I also want to extend a huge thanks to my wonderful kids, Yusef and Saif. Your love and laughter have been my greatest source of joy and motivation, inspiring me to continue my work. I apologise for being absent and busy, but everything I do is for you. You will always be the greatest achievement in my life; I love you.

I am deeply grateful to my parents, Dr Mohammed and Jawaher, for their spiritual support and encouragement in pursuing my dreams. Special thanks to my brother, Khaled, for being my unwavering companion on this journey. I also want to thank my siblings, Mazen, Hatoon, Hamzah, and Basil, for all their love.

Many thanks to my friends Antesar, Riham, Shaykhah, and Shuhrah for their endless support and understanding. Your presence in my life is indispensable.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

Missing Completely At Random	MCAR
Missing At Random	MAR
Missing Not At Random	MNAR
Linear Mixed Model	LMM
Bayesian Hierarchical Model	BHM
Multiple Imputation by Chained Equations	MICE
Markov Chain Monte Carlo	MCMC
Correlated Random Effects	CRE
Generalized Correlated Random Effects with MAR predictor	GCRE-MAR
Generalized Correlated Random Effects with MNAR predictor	GCRE-MNAR

Chapter 1

Introduction

1.1 Introduction

Longitudinal data analysis is the process of studying a dataset that contains a substantial number of participants' records taken over time. This data type provides researchers with a valuable forecasting and predictive modelling source. By monitoring subjects or entities over time, researchers can understand changes and trends related to disease evaluation, treatment effectiveness, individual development, economic trends, long-term effects, and more. Considering the time and money required for such studies before proceeding with the research is crucial (Caruana et al., 2015). Recently, there has been an enormous amount of research using longitudinal methods in health studies (Calman et al., 2013).

In longitudinal design studies, variables are repeatedly measured for each participant. Observations are recorded at baseline and follow-up times, and measures from the same participant are typically highly correlated compared with other participants (Murphy et al., 2022). Therefore, it's important to take into account the correlation between observations obtained from the same participant when analysing data obtained at multiple time points. The Linear Mixed Model (LMM) is a valuable method for handling this correlation while considering predictor variables. This model assumes a linear relationship between the response and predictor variables and takes into account both systematic variation that impacts the population as a whole and subject-specific variation.

There are two main conceptual frameworks for carrying out statistical inference: the Frequentist approach and the Bayesian approach. The common approach used in clinical research is the Frequentist approach (Perkins and Wang, 2004). The Frequentist statistical methods perhaps only use information from past studies at a design stage, not in the inference analysis. On the other hand, the Bayesian statistical approach uses prior knowledge and the observed data from the study and then combines them to update beliefs (Wong et al., 2010). This procedure continues whenever new data is available. Recently, Bayesian inference has been used in research related to health studies as well as a mixture of the Frequentist and Bayesian (Perkins and Wang, 2004). Both methods are widely employed, especially when addressing missing values, which is typically challenging using longitudinal data.

Longitudinal studies typically encounter individual dropout over time and commonly have missing data (Diggle et al., 2013). For example, patients may miss one or more scheduled appointments in medical studies, drop out of the study, refuse to answer sensitive questions, or have participants' data entered incorrectly. This problem makes the data analysis step complicated. Moreover, dealing with missing data inappropriately can lead to inefficient and biased results (Janssen et al., 2010; Little and Rubin, 2019; Mason et al., 2010; Molenberghs et al., 2014; Sterne et al., 2009). It can also influence the conclusions drawn from the study (Stavseth et al., 2019). One of the challenges when missing values exist is to rule out why data is missing. This is referred to as the missing data mechanism. Since the available data does not provide any information about the missing data and its relationship with other variables, it is crucial to understand the mechanisms that describe missing data, which will be explained in more detail in the following Section 1.2.

1.2 Missing Data

Missing data refers to unobtainable information; it is a probable and problematic complication in research fields, which should be considered in the statistical analysis step. There are two types of missing observation in clinical studies: lost follow-up when participants drop out of the study before the study ends, and intermittently missingness when a participant misses follow-up observation and attends later follow-ups. There can be various reasons why participants fail to respond or drop out of the study. Some of the reasons are benign. However, others greatly impact the statistical analysis, and it is essential to handle missing data correctly to ensure the validity of statistical analysis (Ma and Chen, 2018). The cause of missingness is necessary in determining the appropriate statistical method, and specific assumptions about the missingness mechanism need to be met. The suitability of a specific method is determined by the underlying mechanism causing the absence of data (Mason, 2010).

Missing data can be classified into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). This categorisation was introduced by Rubin (1976) and is still widely used today. The missing data mechanism is usually unknown, and relying simply on the data itself is insufficient to distinguish between MAR and MNAR (Molenberghs and Kenward, 2007). The crucial concern is whether the cause of missing data in a variable is associated with the variable itself; in this case, individuals with missing data have different characteristics than individuals with observed data. Consequently, the statistical analysis should be treated with considerable concern to avoid potential bias. Despite this, when the missing data is not associated with the variable itself, the influence of the missingness is benign and does not require a complex analysis (Molenberghs et al., 2014). Additionally, missing by design can occur when necessary information is misrecorded in data collection. For example, missing data can result from a survey question being phrased poorly (Chaudhuri and Agiwal, 2024).

To explore the missingness mechanisms, let's consider the outcome of interest as $Y_i = (y_{i1}, y_{i2}, \ldots, y_{im})'$, which is $m \times 1$ vector. The outcome of interest for the i^{th} subject at t^{th} observation or measurement is represented by y_{it}, where $t = 1, \ldots, m$. Assuming there are missing values in Y_i, therefore, we define a vector of missingness indicators as $R_i = (r_{i1}, r_{i2}, \ldots, r_{im})'$, represents the outcome missingness indicator of the same length as Y_i. Suppose $r_{ij} = 1$ if y_{it} is observed and $r_{it} = 0$ if y_{it} is not observed. The data contains information about R_i and Y_i, referred to as complete data, Y_i can be split into two sub vectors, Y_i^o and Y_i^m, according to the observed and missing outcome, respectively.

Complete data consists of observed data and missingness indicators (Y_i, R_i). There might be confusion between complete data and complete case analysis terminology. The former indicates a data set with a missing indicator variable, and the latter indicates an analysis based on omitting individuals with at least one missing observation on Y_i. The complete data density is expressed as $f(Y_i, R_i | X_i, \beta, \theta)$, where X_i is the design matrix in the Y_i model, β and θ refer to a model of interest and missingness process model parameters, respectively. The type of the missingness mechanism should be evaluated to come by adequate inference from incomplete data.

1.3 Missing Data Mechanism

The missing data mechanism describes the probability distribution of the missingness indicator of the outcome in the model R_i, given Y_i^o, Y_i^m and X_i. Generally, the study analyst does not control the missing data mechanism. However, the adequacy of the analysis and the assumption about the missingness type depends on whether these assumptions satisfy the data at hand (Molenberghs et al., 2014). Three types of missing data mechanisms determine how the R_i depends on Y_i and X_i. Based on these mechanisms, one can choose the appropriate analysis method. In this context, we will discuss each mechanism's definition in detail.

1.3.1 Missing Completely at Random (MCAR)

The data are considered missing completely at random (MCAR) if the probability of missingness is independent of both observed and unobserved outcome values. For example, the nurse was absent while collecting blood test samples. Therefore, the missing data is unrelated to the participants in the study. In MCAR, R_i is unrelated to Y_i^o and Y_i^m as follows:

$$f(R_i|Y_i^o, Y_i^m, X_i) = f(R_i). \tag{1.3.1}$$

An interesting property of MCAR is that the observed data is considered a random sample; therefore, moments and the joint distribution of the observed data do not change from moments and the joint distribution of the full data. Full data refers to fully observed data (no missing values). Therefore, any analysis method will yield identical results for both the observed and full data (Molenberghs et al., 2014).

Using complete case analysis will result in unbiased estimates. However, the efficiency may decrease (Ibrahim and Molenberghs, 2009) because of the reduced number of observations. When there is 5% or less missingness in the data, and the missingness is independent of observed and unobserved values, then it is acceptable to use the complete case analysis (Graham et al., 2009; Jakobsen et al., 2017). However, this situation is uncommon in real life data.

1.3.2 Missing at Random (MAR)

The data is said to be missing at random (MAR) if the probability of missing outcome depends on the observed data and is independent of unobserved data. For instance, younger participants did not show up to collect blood test samples. Therefore, the missing data is related to the age factor, and age needs to be fully observed. To be more specific, R_i is considered to be unrelated to Y_i^m given Y_i^o as follows:

$$f(R_i|Y_i^o, Y_i^m, X_i) = f(R_i|Y_i^o, X_i). \tag{1.3.2}$$

Since the missingness mechanism depends on the observed outcome Y_i^o, the distribution of Y_i is not the same distribution of Y_i in the target population. That is because the distribution of Y_i in each stratum has a different pattern of missingness. In cases of MAR, using complete case analysis or observed data can lead to biased inference, and it is not a sample from the target population (Molenberghs et al., 2014). This occurs when covariates related to the missing at random data mechanism are not included in the analysis.

Rubin (1976) proposed that inferences based on the likelihood can be made by ignoring the missing data mechanism. Therefore, MCAR and MAR are considered ignorable missing mechanism because inferences are only applied to the observed data, and there is no need to set a model for the missing data mechanism. The assumption of MAR implies that the likelihood for the i^{th} subject can be expressed as follows:

$$\begin{aligned} f\left(Y_i^o, R_i \mid X_i,\right) &= f\left(R_i \mid Y_i^o, X_i\right) \times \int f\left(Y_i^o, Y_i^m \mid X_i\right) dY_i^m \\ &= f\left(R_i \mid Y_i^o, X_i\right) f\left(Y_i^o \mid X_i\right). \end{aligned} \tag{1.3.3}$$

The Equation 1.3.3 is derived by ignoring the missing outcome from the joint distribution, where $f\left(R_i \mid Y_i^o, X_i\right)$ doesn't depend on Y_i^m as a results of integrator. Consequently, when the missing mechanism of the data is MAR, the missing value can be then predicted using the observed data and an adequate model for the joint distribution of Y_i, and there is no need to model $f\left(R_i \mid Y_i^o, X_i\right)$ (Ibrahim and Molenberghs, 2009; Molenberghs et al., 2014). Conceivably, MAR is the default assumption whenever there is missing data, except if there is a strong reason to prefer the MCAR assumption (Molenberghs et al., 2014).

1.3.3 Missing Not at Random (MNAR)

In contrast to MCAR and MAR, MNAR is called non-ignorable missingness. The missing data are considered MNAR if the probability of missingness depends on both the missing and observed values. For example, participants with specific health conditions refused to provide blood test samples. As a result, the missing data directly correlates with the participant's health status. This means that the conditional distribution of R_i depends on both Y_i^o and Y_i^m as follows:

$$f(R_i|Y_i,X_i) = f(R_i|Y_i^o,Y_i^m,X_i). \tag{1.3.4}$$

Another situation is that R_i can indirectly relate to Y_i^m through its association on unobserved random effect b_i as follows:

$$f(R_i|Y_i,b_i,X_i) = f(R_i|b_i,X_i). \tag{1.3.5}$$

The equation 1.3.5 is known as a Shared Random Effects (SRE), which will be discussed in Section 2.5.3. Since the MNAR mechanism is mentioned as non-ignorable missingness, that means the mechanism of the missing data cannot be ignored to make inferences about the complete data distribution (Molenberghs et al., 2014). Thus, the missing data mechanism model has to be specified to derive an adequate inference. The ordinary analysis methods are invalid when the missingness mechanism is MNAR; otherwise, it will produce bias estimates (Ibrahim and Molenberghs, 2009).

On the other hand, joint models for the outcome of interest and the missing data are applied to obtain adequate estimates. In Section 2.5.3, we will discuss three joint models that are commonly used: the selection model, pattern mixture model, and shared parameter models. MNAR is common in longitudinal studies (Ibrahim and Molenberghs, 2009).

1.4 Thesis Goals and Contributions

This thesis begins by comparing two commonly used approaches for analysing longitudinal study data: the Linear Mixed Effect Model from the Frequentist approach and the Bayesian Hierarchical Model from the Bayesian approach. The objective of this comparison is to identify the characteristics of each approach and determine their effectiveness. The aim of this study is to thoroughly examine each approach and identify their nuances of application. During the investigation, a typical problem arose, which is the issue of missing data, particularly in longitudinal studies. This is a commonly problematic issue and makes it challenging for researchers to obtain effective results. In this thesis, we propose procedures that deal with missing data, including a combination of non-ignorable and ignorable missingness.

The assumption of the MNAR mechanism can exist for the response and the predictors in the analysis model. There is existing literature on missing data in longitudinal studies. Daniels and Hogan (2008); Fitzmaurice et al. (2008) provides details about missing data methods for longitudinal studies. Molenberghs and Kenward (2007) describe missing data in clinical studies. Ibrahim and Molenberghs (2009) extensively reviews missing data methods in longitudinal studies, and Ma and Chen (2018) reviews the applications of Bayesian methods to handle missing data.

Some studies focus on the missingness of the response variable. For example, Gao (2004) used the shared random effect parameter models for missingness in longitudinal data. They proposed maximum likelihood estimation with Laplace approximation for parameter estimation to avoid high-dimensional integration over the random effect parameter distributions and found that their method was effective. Tsonaka et al. (2009) used a semi-parametric shared parameter model for the random effects distribution, which produced reliable parameter estimates regardless of the distributional assumption for the random effects.

Lin et al. (2010) proposed a method to handle non-ignorable missing data in longitudinal studies. Their approach, the Correlated Random Effects model, differs from the traditional Shared Random Effects model by allowing for different random effects in the outcome and missingness models. They obtained a closed-form expression for the likelihood function by transforming the integral in the likelihood function into a conditional expectation. The results of their simulations showed that their method produced reliable estimates. To avoid the complicated numerical integration that the log-likelihood function involves, Bhuyan (2019) came up with another way to model the "Correlated Random Effects model". They used a simple algorithm that utilized Gibbs sampling for estimation. The simulation results show that the estimates from their proposed method are similar to the estimates from full data analysis (with no missing values).

Other studies focus on the missingness of the predictors variables. Huang et al. (2005) addressed the challenge of nonignorable missing covariate data by ensuring the propriety of the joint posterior distribution under proper priors for regression coefficients in the missing data mechanism. Ibrahim et al. (2005) compare four commonly used methodologies for generalised linear models when dealing with missing covariate data. These include maximum likelihood (ML), multiple imputation (MI), fully Bayesian (FB), and weighted estimating equations (WEEs). The study aims to understand these methodologies' relationships, properties, advantages, disadvantages, and computational implementations.

Yu et al. (2013) developed an effective hierarchical Bayesian method with repeated binary responses and a joint model for time-dependent missing covariates. In a study by Erler et al. (2016), two methods for handling missing covariates in longitudinal models were compared: Multiple Imputations by Chained Equations (MICE) and the full Bayesian approach. The study found that the full Bayesian approach performed well without the need for the specificity of the longitudinal process in the imputation models. Enders et al. (2020) used a Bayesian imputation technique to handle missingness in the model's predictors with random coefficients, interaction, and nonlinear terms. Their technique yielded an accurate parameter estimation, similar to the full data results.

Nevertheless, some studies focused on handling missing data in both the model response and predictors. For example, Du et al. (2022), introduce a Bayesian Latent Variable Selection Model (BLVSM) to estimate model parameters and handle missing data when the response and predictors are MNAR or MAR in a linear model context. Stubbendick and Ibrahim (2003) use the selection model for nonignorable missing predictors and response in the normal random effect model. They employed maximum likelihood estimation with the Gibbs sampler and Monte Carlo EM algorithm, along with bootstrap. Their method performs well based on the simulation data results. This area can be developed by producing an approach for longitudinal models with missing values in the model response and predictors with different combinations of missingness characteristics using Bayesian inference.

This thesis proposes methods for addressing missing data in a longitudinal context using linear mixed models and Bayesian inference. These methods can account for missingness in both the model predictor and the response and can handle different types of missingness mechanisms, including MNAR and MAR. We extend the work outlined by Bhuyan (2019) in the class of Correlated Random Effects selection modelling, where the longitudinal response missingness is non-ignorable and the analysis model predictors are fully observed. This model-based estimation and imputation procedure will be extended to accommodate incomplete predictors. The proposed methods can address different missing data mechanisms, such as MNAR response with a completely observed predictor, MNAR response with MAR predictors and MNAR response and MNAR predictors.

Bayesian inference is carried out to simultaneously estimate the analysis model parameters and the missingness models. This is a useful alternative in such complex settings due to its ability to model incomplete responses and predictors jointly without solving the intractable log-likelihood function using approximation methods, such as in existing Frequentist methods (Lin et al., 2010). Furthermore, the Bayesian approach allows us to incorporate prior information from previous studies if available. We can also incorporate prior beliefs, such as eliciting information from experts in the field. This could be helpful for modelling the missingness process, as it provides insights into how missingness relates to unobserved data, which often cannot be inferred from the observed data alone. This provides an advantage of using Bayesian inference. Our goal is to estimate parameters of interest, provide reliable results while dealing with missing data, and offer decision-makers and clinicians a robust prognosis prediction method that allows them to forecast and analyse patients' health outcomes confidently.

Due to the extensive scope of the subject of missing data modelling, our study is focused on examining the effectiveness of Bayesian full probability modelling. Specifically, our focus is on analysing datasets with missing values in any variable within the model, particularly in a longitudinal context. Our approach is to extend joint models using the Correlated Random Effects model and compare it with the baseline models. The baseline models include the full data model with fully observed variables (no missing data) and the available model with observed data (missing values remain in the dataset, and no imputation is applied). The current Correlated Random Effects method addresses missing responses in longitudinal datasets. However, it is crucial to be aware that such datasets often contain missing predictors alongside missing responses. Therefore, our research investigates the incorporation of predictor missingness within the CRE method, widening the scope beyond addressing missing only in the response.

The key contributions made by this thesis can be outlined as follows:

- A comparison between the Frequentist approach's Linear Mixed Effects modelling and the Bayesian approach's Hierarchical Bayesian Modelling in the context of longitudinal data with continuous outcomes. This is evaluated both in a simulation and real data (from a clinical study investigating heart failure) setting.

- Implementing the Correlated Random Effects (CRE) method for nonignorable missingness in the response of longitudinal data, coupled with offering a solution to the non-convergence of the covariance matrix problem.

- Generate a Two-Step process using Multiple Imputations by Chained Equations (MICE) that adapt the CRE method to account for ignorable missingness in the predictor variables as well as the non-ignorable missingness in the response variable.

- Develop a generalisation of the CRE method to account for ignorable missingness in predictor variables instead of only non-ignorable missingness in the response variable via Gibbs sampling, named the Generalised Correlated Random Effects - Missing at Random (GCRE-MAR) method.

- Develop a further extension of the CRE method to account for non-ignorable missingness in predictor and response variables, named the Generalised Correlated Random Effects - Missing Not at Random (GCRE-MNAR) method

- Examine the proposed methods' performance compared to the baseline methods. The baseline models include the full data model with fully observed variables (no missing data) and the available model with observed data (missing values remain in the dataset, and no imputation is applied). This will be done using a simulated study with different scenarios that may be encountered in such studies and real-life data from a clinical study investigating heart failure.

- Evaluation of the proposed approaches through a sensitivity analysis.

1.5 Structure of the Thesis

In this thesis, we will focus on Bayesian inference under the longitudinal context and missing data. The objectives of this project are briefly described as follows: Chapter 1 presents the topic of the study by introducing some fundamental missing data principles and discusses the thesis goals. Chapter 2 explains the statistical background used throughout the thesis. Chapter 3 presents the real-world study data designed to investigate heart failure, as well as the synthetic data simulated to test the performance of the proposed methods. Chapter 4 features a comparative analysis of the two statistical paradigms (Frequentist and Bayesian inference) for longitudinal data and demonstrates the application of one of the recent approaches in handling non-ignorable missingness in longitudinal data, the CRE method.

Chapter 5 establishes the adapted CRE method, called the Two-Step method, for addressing ignorable missingness in the predictor and non-ignorable missingness in the response. The generalisation of the CRE method to allow for MAR in the predictors is introduced in Chapter 6. The CRE method is further extended to allow for MNAR in both the response and predictors. This extension involves considering Correlated Random Effects between the incomplete predictors and their missingness process and different Correlated Random Effects between the response and its missingness process, which is explained in Chapter 7. Afterwards, Chapter 8 explores a sensitivity analysis in order to evaluate the inference robustness of the proposed methods. A summary of the thesis, outline of the proposed methods’ strengths and limitations, and a discussion of possible directions for future work are provided in Chapter 9.

Chapter 2

Statistical Background

2.1 Introduction

In this chapter, we will discuss the statistical methodologies used in this thesis. In Section 2.2, we will begin by providing an overview of the primary methods used to analyse longitudinal data. Then, in Section 2.3, we will explain the Frequentist approach, including regression modelling. Section 2.4 will address Bayesian inference, including prior distributions, MCMC algorithms, and convergence assumptions.

We will briefly summarize the current methods used to handle missing data in Section 2.5 and describe the probit model used for the missingness indicator model in Section 2.6. To conclude this exploration, we will explain the criteria for assessing the model's performance in Section 2.7. This chapter serves as a fundamental primer, offering essential knowledge for understanding the technical statistical methodologies featured in the upcoming chapters.

2.2 Multilevel Linear Model

Multilevel linear models, also known as Linear Mixed Models or Hierarchical Linear Models, are a more complex form of the ordinary least squares model. It is used to analyse variance in the response variable when the data is hierarchically structured. This is an important method for generalising results from longitudinal data. Several methods are used to fit models for longitudinal data, including maximum likelihood (ML), restricted maximum likelihood (REML), and Bayesian inference (Boedeker, 2017; Raudenbush and Bryk, 2002). In longitudinal studies, responses within an individual are correlated since there are multiple observations of the same individual at different time points, violating the statistical independence assumption in the standard linear regression. Therefore, correlation must be considered when selecting an appropriate analysis method for the data (Carrière and Bouyer, 2002). Linear Mixed Effects Models (LMMs) can be used when adopting the Frequentist approach to inference, as the methodology supports repeated outcomes.

Similarly, the Hierarchical Bayesian model (HBM) is a natural alternative to LMMs. Bayesian modelling is a powerful approach to data modelling that has recently become widely used because it can handle quite complex data structures using Markov Chain Monte Carlo sampling. The advantages of Bayesian modelling are the ability to consider previous information using different priors, the ability to provide a sampling distribution of the parameters in the model, and the posterior distribution can be used as a prior distribution for future analysis. The inference of hierarchically structured data will be discussed in the upcoming sections.

2.3 Linear Mixed Effect Model

The Linear Mixed Effects Regression Model (LMM) is an extension of standard linear regression models that accounts for non-independent data by incorporating fixed and random effects in the fitted model. The fixed effect represents the systematic influence of the response variable, which is equivalent to the predictors in standard linear models. The correlation between observations for an individual (in a longitudinal context) is derived from sharing unobserved variables, which is the random effect. The inclusion of the random effect distinguishes mixed models from traditional standard models.

To capture the correlation structure between observations within the same individual in longitudinal data, we will analyse the data using a mixed-effects model with individual-specific random intercept. The linear mixed model described by Laird and Ware (1982) is expressed as follows:

$$\boldsymbol{Y}_i = \underbrace{\boldsymbol{X}_i\boldsymbol{\beta}}_{\text{fixed effects}} + \underbrace{\boldsymbol{Z}_i\boldsymbol{u}_i}_{\text{random effects}} + \boldsymbol{e}_i, \tag{2.3.1}$$

where $i = 1,\ldots,n$ represents individuals and $t = 1,\ldots,m$ represents repeated measures per individual. The model has the flexibility to handle data structures with both equal and unequal numbers of repeated measures per individual by incorporating the random effect, which accounts for the between and within individual variability (Pinheiro and Bates, 2006). In this thesis, we consider an equal number of repeated measures for each individual. Therefore, the total number of observations is $N = n \times m$. $\boldsymbol{Y}_i = (y_i(1), y_i(2), \ldots, y_i(m))'$ is a $m \times 1$ vector of the response variable, each $y_i(t)$ represents the response observation of t^{th} time point for the i^{th} individual, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ is a $J \times 1$ vector of fixed effect parameters, with $j = 1,\ldots,J$ number of predictors. Moreover, $\boldsymbol{X}_{ji} = (x_{1i}(t), x_{2i}(t), \ldots, x_{Ji}(t))$ is a $1 \times J$ vector of J predictors observed at time t for subject i. Thus, $\boldsymbol{X}_i$ is an $m \times J$ matrix of predictors associated with subject i and has the form:

$$\begin{pmatrix} x_{1i}(t) & \cdots & x_{Ji}(1) \\ \vdots & \ddots & \vdots \\ x_{1i}(m) & \cdots & x_{Ji}(m) \end{pmatrix}. \tag{2.3.2}$$

Furthermore, $\boldsymbol{u}_i$ is a vector of random effect parameters expressing the deviation from the population mean for each individual, and $\boldsymbol{Z}_i$ is a vector of the random intercept, which is constant across time within each individual. $\boldsymbol{e}_i = (e_i(1), e_i(2), \ldots, e_i(m))'$ is a vector of random errors, where, $e_i(t)$ shows additional deviation for the i^{th} individual at t^{th} time point. Moreover, the random effect u_i and the residuals e_i each independently and identically follow a normal distribution with a mean of zero and constant variance (Pinheiro and

Bates, 2006) as follows:

$$\begin{aligned} \boldsymbol{u}_i &\overset{\text{i.i.d}}{\sim} N(0, \sigma_B^2), \\ \boldsymbol{e}_i &\overset{\text{i.i.d}}{\sim} N(0, \sigma_A^2). \end{aligned} \tag{2.3.3}$$

Furthermore, they are uncorrelated, defined as follows:

$$E\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \text{Cov}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix},$$

where, $\mathbf{u} = [u_1, \ldots, u_n]'$, $\mathbf{e} = [e_1(1)', \ldots, e_n(m)']'$, $\mathbf{G} = \sigma_B^2\mathbf{I}$ and, $\mathbf{R} = \sigma_A^2\mathbf{I}$. Where, $\mathbf{I}$ is the identity matrix. Given that $\mathbf{e}^* = \mathbf{Zu} + \mathbf{e}$. The variance of $\mathbf{Y}$ calculated as follows:

$$\begin{aligned} \mathbf{Var}(\mathbf{Y}) &= \mathbf{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}^*) \\ &= \mathbf{Var}(\mathbf{X}\boldsymbol{\beta}) + \mathbf{Var}(\mathbf{e}^*) \\ &= 0 + \mathbf{ZGZ^T} + \mathbf{R}. \end{aligned} \tag{2.3.4}$$

Consequently, the response is assumed to be normally distributed with a mean equal to the fixed effects terms $(\mathbf{X}\boldsymbol{\beta})$ and the variance equals ($\mathbf{V} = \mathbf{ZGZ^T} + \mathbf{R}$). Accordingly,

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}). \tag{2.3.5}$$

To describe the variability in the data, we will assume a compound symmetry covariance structure for simplicity, where all observations have equal variances and a constant correlation between any two observations with different time points. However, there are other types of covariance structures to consider. For example, a first-order autoregressive covariance structure assumes that correlations between observations decrease exponentially as time increases. The Toeplitz covariance structure assumes that correlations depend only on the lag, with each lag having its own correlation parameter. Lastly, the unstructured covariance assumes no specific structure, meaning each pair of time points has a unique correlation parameter (Littell et al., 2000). We can rewrite Equation 2.3.5 in terms of the compound symmetry covariance matrix as follows:

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{I_n} \otimes \Sigma), \tag{2.3.6}$$

where $\mathbf{I_n}$ is the n-dimensional identity matrix, where n is the number of random intercepts/ participants in the model and Σ is $m \times m$ matrix, where m is the number of repeated measures. The diagonal elements of Σ are the summation of the two sources of variation in the LMM $(\sigma_E^2 + \sigma_B^2)$ representing each individual's variance at a different time point, and the off-diagonal is the between-individual variance (σ_B^2), which expresses the variation between measurements per individual. This is called compound symmetry structure, which is assumed for the variance-covariance matrix in the random intercept model of the longitudinal data as mentioned in (Donald and D., 2006). Suppose there are two repeated measures; the covariance matrix structure is expressed as follows:

$$\Sigma = \begin{bmatrix} \sigma_B^2 + \sigma_E^2 & \sigma_B^2 \\ \sigma_B^2 & \sigma_B^2 + \sigma_E^2 \end{bmatrix}. \tag{2.3.7}$$

The $\otimes$ indicates the Kronecker product, which is a multiplication of two matrices. For example, consider A and B as matrices. The Kronecker product is expressed as follows:

$$\boldsymbol{A}_{I\times P} \otimes \boldsymbol{B}_{K\times L} = \begin{bmatrix} A_{11}\boldsymbol{B} & A_{12}\boldsymbol{B} & \dots & A_{1P}\boldsymbol{B} \\ A_{21}\boldsymbol{B} & A_{22}\boldsymbol{B} & \dots & A_{2P}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ A_{I1}\boldsymbol{B} & A_{I2}\boldsymbol{B} & \dots & A_{IP}\boldsymbol{B} \end{bmatrix}_{IK\times PL}, \tag{2.3.8}$$

where each element from matrix $\boldsymbol{A}$ will be multiplied by the entire $\boldsymbol{B}$ matrix.

So, the product $\mathbf{I} \otimes \Sigma$ will be $N \times N$ block matrix as follows:

$$\mathbf{I}_n \otimes \Sigma_m = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma \end{bmatrix}_{nm \times nm} . \tag{2.3.9}$$

This indicates that between individuals are independent. However, observations within individuals are correlated, and this correlation is expressed through the variance-covariance matrix Σ. Therefore, this correlation must be considered to find an applicable analysis method to the data (Carrière and Bouyer, 2002). The correlation between observations for an individual is derived from sharing unobserved variables (the random effect). To obtain the marginal density for the data, the random effect in equation 2.3.1 must be integrated out, which is considered a nuisance parameter (Pinheiro and Bates, 2006). Since the random effect and the residuals are independent, we can represent this as:

$$\begin{aligned} L(\Theta \mid \boldsymbol{Y}) &= \prod_{i=1}^{n} p\left(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \sigma_A^2, \sigma_B^2\right) \\ &= \prod_{i=1}^{n} \int p\left(\boldsymbol{Y}_i \mid \boldsymbol{u}_i, \boldsymbol{\beta}, \sigma_A^2\right) p\left(\boldsymbol{u}_i \mid \sigma_B^2\right) d\boldsymbol{u}_i, \end{aligned} \tag{2.3.10}$$

where $\Theta = (\boldsymbol{\beta}, \sigma_B^2, \sigma_A^2)$ refers to the model parameters that need to be estimated. In most longitudinal studies, models are fitted to report marginal means and consider the covariance structure as nuisance parameters. This is named marginal / population average models, which explains the regression coefficients as a marginal response to changing predictors (Lee and Nelder, 2004). On the other hand, there are conditional models, where the regression coefficients explain each individual's response variable. Thus, the marginal model is expressed as follows:

$$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i \boldsymbol{\beta}, \tag{2.3.11}$$

and the conditional model is expressed as follows:

$$E(\boldsymbol{Y}_i|\boldsymbol{u}_i) = X_i\boldsymbol{\beta} + \boldsymbol{u}_i. \tag{2.3.12}$$

Choosing between marginal and conditional models in longitudinal studies depends on the study objectives. Lee and Nelder (2004) mentioned that in the case when the main interest is model predictions, both models can be used. Muff et al. (2016) state that there are no differences between the marginal and conditional formulas of the model when the output variable is assumed to be normally distributed. Therefore, we will consider the marginal model throughout the analysis. In particular, the parameters to be estimated in the mixed effects models are the fixed effects parameters in $\boldsymbol{\beta}$, the random effects parameters $\mathbf{u}$ and the variance-covariance matrix parameters $\mathbf{V}$.

2.3.1 Estimation Inference

To estimate the parameters in the Linear Mixed Effects Model (LMM), we will use the `lme` function from the `nlme` package in `R` developed by Pinheiro et al. (2007). The model parameters can be estimated using the Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) methods. The maximum likelihood method estimates the most likely parameters resulting from the observed data (Myung, 2003). In contrast, the REML method estimates the variance components by taking into consideration the degrees of freedom for the fixed effects in the model. The main difference between the REML and ML methods is the variance estimation (Peugh, 2010). When ML is used to estimate the variance, the fixed components are treated as known and measured without errors. In contrast, the fixed components are considered nuisance parameters when estimating the variance using the REML method. The maximum likelihood method tends to underestimate the variance components (Pinheiro and Bates, 2006), whereas REML results in less biased estimates than ML (Boedeker, 2017; Gałecki et al., 2013). In summary, the main feature of the REML method is that it is preferred for estimating the variance components as it takes into account the degrees of freedom for the fixed effects in the model (Boedeker, 2017). In the following sections, we will define each of these techniques.

Maximum Likelihood

The maximum likelihood estimation of the parameters in the normal model is reached via maximisation of the marginal likelihood, and the MLE of $(\boldsymbol{\beta}, \sigma_E^2, \sigma_B^2)$ is the one that maximises this expression:

$$L_{ML}(\Theta \mid \mathbf{Y}) = \prod_{l=1}^{N} \left\{ (2\pi)^{-\frac{m}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right) \right\}. \tag{2.3.13}$$

The Maximum Likelihood method is used to compare nested models with different fixed effects, providing an advantage over the Restricted Maximum Likelihood (REML) method. The REML will project data into two separate spaces, making it impossible to compare the likelihoods (George and Aban, 2015).

Restricted Maximum Likelihood

To derive an unbiased estimator of the variance component, we will assume that the fixed effect parameters are nuisance parameters and eliminate the fixed effects parameters from the likelihood function (Laird and Ware, 1982) as follows:

$$L_{REML}(\Theta \mid \mathbf{Y}) = \int L(\boldsymbol{\beta}, \mathbf{V} \mid \mathbf{Y})\, d\boldsymbol{\beta}. \tag{2.3.14}$$

REML starts by fitting the fixed effects parameters using generalised least squares (GLS) estimation, which is used as an alternative to the ordinary least squares (OLS) to account for the correlation structure in the data (Pinheiro and Bates, 2006). Then the residuals of the regression model are maximized to compute the estimates of the variance components. REML works by maximising the likelihood of the response model independent of β.

In a linear mixed effect model, closed-form estimations may not be available for ML and REML calculations, thus requiring the use of conventional numerical methods (Jiang and Nguyen, 2007).

Optimization Algorithms

Two common optimisation methods are often used to determine the parameter value that maximises the likelihood of the observed data given the statistical model. These methods are the Expectation Maximisation (EM) algorithm and the Newton-Raphson iterations (Stirrup, 2016). The EM algorithm was first introduced by Dempster et al. (1977) to compute the likelihood estimation for models with incomplete data. Since then, it has been widely used for likelihood estimation of linear mixed models, by treating the random effects as unobserved (latent) data (Laird and Ware, 1982). The EM algorithm works by first setting initial values for model parameters. Then, based on these initial/recent values, the expected values are computed (E-step). Next, it updates the parameters to maximise the likelihood using the computed expectations (M-step). Finally, these steps are repeated iteratively until convergence is reached. In contrast, the Newton-Raphson method is a widely used iterative optimisation method in linear mixed model estimation. Thisted (1988) defined this procedure, which begins with initial parameter estimates and is continually updated. In each iteration, the algorithm computes the gradient of the loglikelihood function around the recent parameter estimate to compute the next parameter estimate. It requires calculating the first and second derivatives of the likelihood function (Pinheiro and Bates, 2006; Stirrup, 2016). Based on this, the algorithm updates the parameter estimations, intending to maximise the likelihood. This iterative process continues until it reaches convergence to a roots of an equation.

These algorithms are useful for optimizing the complex likelihood often encountered in Linear Mixed Models (LMM). These algorithms estimate both fixed and random effects simultaneously. The `lme` function applies a hybrid optimization technique. It begins by computing initial estimates of the Θ parameters. After that, it implements the EM algorithm for multiple iterations to approach the optimum values, followed by a change to Newton-Raphson iterations for achieving convergence to the optimal values. The EM algorithm quickly approaches the optimal region for the parameters, but it becomes slow near the optimum. On the other hand, the Newton-Raphson algorithm requires a longer processing time than the EM algorithm, but it rapidly converges as it gets closer to the optimum (Pinheiro and Bates, 2006). To ensure robust statistical inference, it is essential to check the underlying assumptions. The primary assumptions of the linear mixed models are presented next.

2.3.2 Assumptions

It is essential to check the model assumptions in order to have the correct conclusion of the model analysis. The following are the assumptions of Linear Mixed Models.

1. Linearity:
   This means that the model's relationship between the response variable and predictors should follow a straight line. This assumption could be checked using graphical plots of response vs. predictors (Pinheiro and Bates, 2006).

2. Homogeneity of Variance:
   This assumes that the residual variances are equal within groups (Pinheiro and Bates, 2006). It can be checked using graphical plots (predicted values vs. residuals). There should be no pattern or trend for the homoscedasticity (constant variance) assumption to hold true.

3. Normally distributed residuals:
   This assumes that the residuals of the model are normally distributed. This can be checked using the Q-Q plots of residuals, where strong deviation points from the line of equality indicate that this assumption is

violated. Also, it is one way to identify the outliers (Pinheiro and Bates, 2006).

4. Independent residuals:
   The residuals are assumed to be independent within and between individuals.

2.4 Bayesian Hierarchical Linear Model

Lately, Bayesian methods have been widely used due to the latest improvements in computation capacity and the sharp growth of efficient algorithms in different disciplines. Furthermore, there has been a notable development in their ability to address missing data problems (Huang et al., 2005). Bayesian inference can incorporate additional information and often provide better outcomes despite small sample sizes (Cai et al., 2010). The major difference between the Frequentist approach and the Bayesian approach is the definition of probability. In the Frequentist approach, the probability is related to the frequency of events. On the other hand, in the Bayesian approach, the concept of probability is related to the degree of belief about statements. In a Frequentist's view, the underlying parameters are fixed throughout a repeatable process. Whereas in a Bayesian's view, underlying parameters are unknown and have a distribution (treated as a random variable). Therefore, under the Bayesian framework, all parameters are random, so there is no longer a need to distinguish between fixed and random effects. In this section, we will summarise some important concepts of the Bayesian approach.

Bayes' theorem, introduced by Bayes (1763), is fundamental in Bayesian inference, which calculates the conditional probability, for example, when there are two events, event X and event Y. The Bayes' rule is expressed as follows:

$$p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}, \tag{2.4.1}$$

where $p(X)$ is the probability event X occurs, $p(Y)$ is the probability event Y occurs and $p(X \mid Y)$ is the conditional probability of event X occurring given that event Y occurred. The Bayes' rule can be rewritten in the form of Bayesian statistics as follows:

$$p(\Theta \mid Y) = \frac{p(Y \mid \Theta)p(\Theta)}{p(Y)}, \tag{2.4.2}$$

where Θ is an unknown parameter which is considered a random variable, Y is the observed data, $p(\Theta)$ is the prior distribution, $p(Y \mid \Theta)$ is the distribution of the observed data or likelihood, either, $p(Y) = \int p(\Theta)p(Y \mid \Theta)d\Theta$ in case of a continuous variable or $p(Y) = \sum_{\Theta} p(\Theta)p(Y \mid \Theta)$ in case of a discrete variable. As a consequence that $p(Y)$ does not depend on Θ, this can be considered as a constant, and thus the Equation (2.4.2) can be rewritten as follows:

$$p(\Theta \mid Y) \propto p(Y \mid \Theta) \times p(\Theta), \tag{2.4.3}$$

where $p(\Theta \mid y)$ is the posterior distribution, which is the basis of any inference in Bayesian statistics; it summarises the knowledge of uncertain quantities. The marginal likelihood for the data is expressed in Equation 2.3.5.

2.4.1 Prior Distribution

The prior distribution plays a crucial role in Bayesian statistics as it represents the uncertainty of the parameter Θ before observing the data. There are two primary types of prior distribution: informative and non-informative. An informative prior assigns more probability density to some locations than others, providing additional knowledge about a variable. In contrast, a non-informative prior expresses vague information about a variable with a wide probability distribution, which has a minimal impact on the estimated parameters. In such cases, most of the information is obtained from the data, and the inferences would be similar to those from the Frequentist inference (Lesaffre and Lawson, 2012). Other names of noninformative priors that appeared in literature are: "nonsubjective", "objective", "default", "weak", "diffuse", "flat", and "minimally informative" (Lesaffre and Lawson, 2012).

The non-informative prior can be an improper prior if the integral of this distribution does not integrate into one. One example of an improper prior is the Jeffreys priors (Robert et al., 2009). However, if an improper prior leads to an improper posterior, the inference would be considered invalid (Gelman, 2006). To make Bayes theorem easier to work with, we can use a conjugate prior where the prior and posterior distributions come from the same distribution family. The posterior distribution will be the same as the prior distribution but with updated hyperparameters. When we lack information about the parameter of interest, we can use non-informative prior. However, informative priors can be obtained from previous studies or expert knowledge through prior elicitation. The concept of prior elicitation has been addressed in numerous studies, such as Agiashvili et al. (2021); Azzolina et al. (2021); Choy et al. (2009); James et al. (2010); Kinnersley and Day (2013); Martin et al. (2012); Zapata-Vázquez et al. (2014).

2.4.2 Bayesian Inference

Bayesian inference relies on the posterior distribution, which can sometimes be challenging to calculate in a closed form. In such cases, the Markov Chain Monte Carlo (MCMC) algorithm is used to sample from it. This algorithm is widely used in modern Bayesian computing, particularly when the posterior distribution is difficult to work out analytically. MCMC is a numerical simulation method used to generate a random set of points from the parameter space that is drawn from the posterior distribution. It's then used to estimate the distribution of the parameters and, finally, compute summary statistics from it. Basically, MCMC is used when it is unachievable to sample Θ directly from the posterior distribution $p(\Theta \mid y)$, it samples iteratively so that at each step of the process, the draws from the distribution become closer to $p(\Theta \mid y)$ as the number of iterations increases.

The MCMC algorithm is a numerical simulation of a chain of a series of random variables with the transition probability property that the chain depends only on the present state of the chain rather than the entire past of the chain. Therefore, the draws are not independent samples. More explanation about MCMC can be found in Gamerman and Lopes (2006). The most popular MCMC algorithms are the Gibbs Sampler, the Metropolis-Hastings Algorithm and the Hamiltonian Monte Carlo (HMC) Algorithm. In this study, we intend to use the `brms` function in `R` (Bürkner, 2017) to implement a Bayesian hierarchical model for the data analysis, which utilises the HMC algorithm. Each MCMC algorithm will be explained in the following sections.

2.4.3 Gibbs Sampler

Geman and Geman (1984) introduced the Gibbs Sampler, which is used when it is possible to sample from the conditional posterior distribution, often requiring each parameter's conditional distribution to have a standard distribution form by using conjugate priors to the likelihood of parameters (Gelman et al., 2013). Thus, each parameter is sampled from its conditional distribution depending on the remaining parameters' most recent values. To illustrate this, suppose each parameter Θ_p is sampled from the conditional distribution given all previous components of Θ_{-p} as:

$$p(\Theta_p \mid \Theta_{-p}, Y), \tag{2.4.4}$$

where p represent the total number of parameters in the model and Θ_{-p} contains all the components of Θ, excluding Θ_p. The process of the Gibbs sampling algorithm for S samples is implemented as follows:

Algorithm 1 Gibbs Sampling Algorithm

```
Choose initial values of $\Theta^{(0)}$
for $1,\ldots,S$ iterations do
1- sample $\Theta_1^{(s)}$ from its conditional distribution $p\left(\Theta_1 \mid \Theta_2^{(s-1)},\ldots,\theta_p^{(s-1)}\right)$
2- sample $\Theta_2^{(s)}$ from its conditional distribution $p\left(\Theta_2 \mid \Theta_1^{(s)},\Theta_3^{(s-1)},\ldots,\Theta_p^{(s-1)}\right)$
⋮
$p$ - sample $\Theta_p^{(s)}$ from its conditional distribution $p\left(\Theta_p \mid \Theta_1^{(s)},\ldots,\Theta_{p-1}^{(s)}\right)$
```

Repeat steps 1 through p for a long enough number of S draws until convergence. In the case when the conditional distribution cannot be found in a standard distribution form, one can use the Metropolis-Hastings Algorithm.

2.4.4 Metropolis-Hastings Algorithm

The Metropolis-Hastings Algorithm (M-HA) (Hastings, 1970) is a random walk with acceptance and rejection rates to converge to the identified posterior distribution. The algorithm assumes a proposal distribution $q(\Theta^*|\Theta^{s-1})$ to generate a candidate sample Θ^*, where the proposal distribution depends on the recent value of Θ^{s-1}. The decision to accept or reject the proposed sample Θ^* is based on the acceptance probability $\alpha(\Theta^*,\Theta^{s-1})$, which is then compared with a random value η drawn from the uniform distribution as $\eta \sim Unif(0,1)$. If $\alpha(\Theta^*,\Theta^{s-1}) \geq \eta$ then the proposed value Θ^* will be accepted, otherwise the value of Θ^s will remain unchanged $\Theta^{s-1} = \Theta^s$. The process Metropolis-Hastings Algorithm is implemented as follows:

Algorithm 2 Metropolis-Hastings Algorithm

```
Choose initial values of $\Theta^{(0)}$
for $1, \ldots, S$ iterations do
1- Sample $\Theta^{(*)}$ from proposal distribution at time $s$, $q\left(\Theta^{*} \mid \Theta^{(s-1)}\right)$
2- Calculate the acceptance rate as follows:
    $$\alpha = \frac{p\left(\Theta^{*} \mid Y\right) q\left(\Theta^{(s-1)} \mid \Theta^{*}\right)}{p\left(\Theta^{(s-1)} \mid Y\right) q\left(\Theta^{*} \mid \Theta^{(s-1)}\right)}$$
3 - Set
    $$\Theta^{s} = \begin{cases} \Theta^{*} & \text{with probability } \min(\alpha, 1) \\ \Theta^{s-1} & \text{otherwise .} \end{cases}$$
```

where $p\left(\Theta \mid Y\right)$ is the posterior distribution. Repeat the steps outlined in Algorithm 2 for a sufficient number of iterations, S, until convergence is achieved. Gibbs sampling is a special case of the M-HA, where each random variable is updated successfully using the conditional distribution.

The Metropolis Algorithm (MA) introduced by Metropolis et al. (1953) is a special case of Metropolis-Hastings Algorithm, when the proposal distribution is symmetric and meet the following condition: $p\left(\theta^{*} \mid \Theta^{s}\right) = p\left(\Theta^{s} \mid \Theta^{*}\right)$ for all Θ^{*}, Θ^{s} and s. The steps for Metropolis Algorithm MA are similar to Metropolis-Hastings Algorithm, except for the calculation of the acceptance rate, which will be calculated as follows:

$$\alpha = \frac{p\left(\Theta^{*} \mid Y\right)}{p\left(\Theta^{(s-1)} \mid Y\right)}. \tag{2.4.5}$$

The proposal distribution can highly affect the performance of the M-HA and MA (Rosenthal et al., 2011). The commonly used proposal distribution is the Normal distribution as follows:

$$\Theta^{s+1} \sim N(\Theta^{s}, \sigma^{2}), \tag{2.4.6}$$

where the mean (Θ^s) is the value of the parameter at the current step of the chain and the variance (σ^2), which is referred to as the step size. The adjustment of the step size is necessary to control the performance of the algorithm, and it is associated with the acceptance probability $\alpha(\Theta^*, \Theta^{s-1})$. If σ^2 is too large, then most proposed values will be rejected. Contrarily, if σ^2 is too small, then most proposed values will be accepted. In both cases, the chain will take a long time to converge. The recommended acceptance rate for high dimensions is around 0.23 and 0.44 for one dimension (Gelman et al., 2013).

To avoid proposing negative values for strictly positive parameters, for instance, the variance parameters, one can use reflection. To explain the idea of reflection, let $\Theta \in (a,b)$ and assume a symmetric proposal as explained previously in Equation 2.4.6 then, if the proposed values are outside the interval, the excess is reflected into the interval as follows:

$$\begin{aligned} \text{if } \ \Theta' &\leq a, \text{then} \\ \Theta' \ \text{ is reset to } \ &2a - \Theta', \end{aligned} \tag{2.4.7}$$

where a is zero because we assume positive random variables; this process will be repeated until Θ' is positive. Moreover, this method will not break the symmetric assumption when using Metropolis Algorithm. That is because if Θ can reach Θ' through a number of reflections, it is also possible for Θ' to reach Θ through the same number of reflections, so that the Hastings ratio $\frac{q(\Theta'|\Theta)}{q(\Theta|\Theta')} = 1$ (Yang and Rodríguez, 2013).

2.4.5 Hamiltonian Monte Carlo (HMC) Algorithm

The main drawbacks of the previously mentioned MCMC algorithms (Gibbs and M-H algorithms) lie in their slow convergence when applied to correlated parameters in high dimensional models (Hoffman et al., 2014; Neal et al., 2011) and the random walk behaviour (Gelman et al., 2013). Neal et al. (2011) introduced an MCMC algorithm employing Hamiltonian dynamics, which is a physical system in order to avoid the issues of correlated parameters and random walk behaviour. Furthermore, Hoffman et al. (2014) proposed a No-U-Turn Sampler as an extension to the HMC algorithm that provides efficient and accelerated convergence.

We will use the `brms` (Bürkner, 2017) function from `brm` package in `R` to conduct the Bayesian Hierarchical model. The `brms` function is a versatile function that utilises Hamiltonian Monte Carlo (HMC) with the No-U-Turn Sampler (NUTS) algorithm to estimate Bayesian Hierarchical model parameters. Next, we will briefly overview the HMC and NUTS algorithms.

Hamiltonian Monte Carlo (HMC) Algorithm

The Hamiltonian Monte Carlo (HMC) Algorithm is a modified version of the Metropolis Algorithm (Nishio and Arakawa, 2019), which considers the parameters to be a physical system component and then suggests a new proposal by letting it alter through Hamiltonian dynamics. The Hamiltonian function $H(\Theta, p)$ is defined as follows:

$$H(\Theta,p) = U(\Theta) + K(p), \tag{2.4.8}$$

where $U(\Theta)$ is the potential energy, $K(p)$ is the kinetic energy, and Θ and p are the position and momentum vectors, respectively, describing the particle's motion. The dynamics have the characteristic to preserve the H invariant (Nishio and Arakawa, 2019). This means Θ and p will remain constant over time. The potential energy $U(\Theta)$ defined as:

$$U(\Theta) = -\log f(\Theta), \tag{2.4.9}$$

where Θ is the variable to be estimated and $f(\Theta)$ is the probability density function of Θ (The posterior distribution of Θ). Every variable Θ has an auxiliary momentum variable p, where $K(p)$ commonly follows a normal distribution with mean zero and M covariance matrix, to generate a proposal distribution that enables the use of the posterior distribution's gradient information. The kinetic energy $K(p)$ helps the algorithm move faster around the parameter space (Gelman et al., 2013). The joint density function of $f(\Theta,p)$ has the following form:

$$f(\Theta,p) = exp[-H(\Theta,p)]. \tag{2.4.10}$$

HMC produce samples of Θ and p from the aforementioned joint distribution. Thus, we can select only Θ as samples from the target distribution. According to the Hamiltonian dynamics concepts, the samples proceed with respect to the total energy, as expressed by the following sequence of differential equations, known as Hamilton's equations:

$$\frac{d\Theta}{dt} = \frac{dH}{dp} \tag{2.4.11}$$

$$\frac{dp}{dt} = \frac{dH}{d\Theta}, \tag{2.4.12}$$

determining how the system moves over time t. Moreover, these equations cannot be solved analytically; therefore, a numerical method is needed to break down the time t interval into a series of shorter time intervals ϑ. The common numerical method HMC uses is the leapfrog, which sequentially updates Θ and p. It starts with a half step for p and the full step for Θ by incorporating the updated values for p. The leapfrog method achieves the proposal points (Θ^*, p^*) via L steps of step size ϑ. The proposed values of Θ^* and p^* are accepted or rejected by applying the Metropolis acceptance probability as follows:

$$\alpha = min\{1, exp[H(\Theta,p) - H(\Theta^*,p^*)]\}, \tag{2.4.13}$$

if the proposed values are accepted, then the next state equals (Θ^*, p^*). If rejected, the current state will stay the same. The challenging point of HMC is that the sampling efficiency is sensitive to the values of the step size ϑ and the number of steps L. For example, if the value of ϑ is chosen to be very large, this results in a low acceptance rate. On the other hand, if the value of ϑ is chosen to be very small, it will take a long time to explore the target distribution. Moreover, a small value of L will lead to high autocorrelation, and a large value of L will take longer computational time (Neal et al., 2011). Additionally, it may revert the parameters back to their initial values (Nishio and Arakawa, 2019).

No-U-Turn Sampler

Hoffman et al. (2014) introduced the No-U-Turn Sampler (NUTS) to overcome the limitation of specifying the value of L. This is used as an extension to the HMC to prevent the Markov chain from moving backwards by stopping the simulation when the distance between the current and proposed position starts decreasing (Varsi, 2021). For every iteration, NUTS automatically chooses a suitable value for L to increase the distance at each leapfrog step and prevent random walk behaviour (Nishio and Arakawa, 2019). This is achieved by checking the following:

$$\frac{\partial Q}{\partial \tau} = \frac{\partial}{\partial \tau} \frac{(\Theta^* - \Theta)'(\Theta^* - \Theta)}{2} = (\Theta^* - \Theta)' p < 0, \tag{2.4.14}$$

where Q is half the squared distance between the current and proposed position, τ is the time $1 \leq \tau \leq L$. The aforementioned equation means that leapfrog steps will keep running until the derivative of Q with respect to time is less than zero (Nishio and Arakawa, 2019). However, this condition does not guarantee reversibility or correct convergence (Nishio and Arakawa, 2019). By using the doubling method (start with one leapfrog step, then two, four, and so on) for slice sampling, NUTS avoids this problem.

The slice sampling Neal (2003) is an MCMC algorithm that samples from a target distribution by selecting points uniformly beneath the distribution curve $f(\Theta)$. This approach involves using an auxiliary variable ν and a joint distribution that is distributed uniformly over the region as follows: $D = \{(\Theta, \nu) : 0 < \nu < f(\Theta)\}$ beneath the surface $f(\Theta)$. The joint distribution is expressed as follows:

$$f(v,\Theta)=\begin{cases}\frac{1}{z} & \text{if } \ 0\leq v\leq \pi(\Theta)\\ 0 & \text{otherwise}\end{cases}, \tag{2.4.15}$$

where $\pi(\Theta)$ is a kernel of $f(\Theta)$, $z=\int \pi(\Theta)d\Theta$ and the marginal distribution of $f(v,\Theta)$ is $f(\Theta)$. Thus, by taking a sample from $f(v,\Theta)$ and discarding v, the Θ can be extracted from the target distribution. These steps are performed via slice sampling by sampling v and Θ in turn (Nishio and Arakawa, 2019). First, fix Θ and sample v from:

$$p(v\mid\theta)\sim \text{Uniform}(0,\pi(\Theta)), \tag{2.4.16}$$

to satisfy $v\leq\pi(\theta)$. Next, find the slice by fixing v and sample θ uniformly from the horizontal sliced region S defined by:

$$S=\{\theta: v\leq\pi(\Theta)\}. \tag{2.4.17}$$

Then, sample a new value of Θ uniformly from S. The issue in the slice sampling algorithm is to identify S boundaries. Neal (2003) proposed the doubling method, where the size of an initial segment which contains the current value of Θ is randomly chosen and then expanded by doubling its size until the endpoints are outside set S.

The procedure employs Hamiltonian dynamics to move the Markov chain forward or backwards. Every step has a randomly selected direction. This method involves iteratively repeating the process, where the number of steps is doubled, and the directions are changed. Consequently, Leapfrog steps are monitored by NUTS. When the following U-Turn condition is satisfied, the algorithm stops.

$$\left(\Theta^{+}-\Theta^{-}\right)'p^{-}<0 \text{ or } \left(\Theta^{-}-\Theta^{+}\right)'p^{+}<0, \tag{2.4.18}$$

where Θ^{+} and p^{+} are the forward direction after a Leapfrog step of Θ and

p and Θ^- and p^- are the backward direction. For more information and a pseudo-code of the HMC with the NUTS algorithm, see Hoffman et al. (2014); Nishio and Arakawa (2019); Varsi (2021).

The length of the chain is a critical concern when running the MCMC algorithm. Therefore, we need criteria to evaluate the chain's convergence, which will be explained next.

2.4.6 Convergence

Assessing the convergence is crucial to ensure that MCMC chains converge to the posterior distribution in the long run. There are several ways to visually and statistically inspect convergence for each parameter. We will discuss these methods next.

Visual Inspection

The trace plots can assess the convergence visually, which charts the parameter value at each iteration against the iteration number. These plots help show how the MCMC chain moves within the parameter space. We draw multiple traces from various starting points within the parameter space to generate these plots. We then discard the initial iterations, known as the warm-up period, to reduce their influence. The smooth movement in the trace plots indicates that the MCMC chain has successfully converged. However, if the movement is stuck in some parameter space or has seasonal trends, this suggests a lack of convergence.

Geweke

The Geweke test is a tool used to diagnose the convergence of Markov Chain Monte Carlo (MCMC) chains. It was proposed by Geweke (1992). The test is designed to divide a single chain into two non-overlapping segments, usually the first 10% and last 50% of iterations after discarding the burn-in period (Best et al., 1995). The means and standard deviations for each segment are then calculated, and the Z-test is applied to determine if the means of the beginning and end of the chain are equal. The null hypothesis is that the means are equal. The test statistic is compared to the standard normal distribution. If the absolute value of the test statistic is less than or equal to 2, the test is not significantly different from zero, and the chain is assumed to have converged (Ntzoufras, 2011).

Gelman-Rubin Diagnostic

The Gelman-Rubin diagnostic (Gelman and Rubin, 1992) can be used to check the convergence by running multiple chains with different starting points and then splitting those chains into halves to calculate the within-chain and between-chain variance. Suppose we have m chains each of length n as ϕ_{ij}, where $i = 1,...,n$ and $j = 1,...,m$. Then the between chain (B) and the within chain (W) variance is calculated as follows:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\phi}_{.j} - \bar{\phi}_{..}\right)^2, \text{ where } \quad \bar{\phi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\phi_{ij}, \quad \bar{\phi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\phi}_{.j},$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \text{ where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\phi_{ij} - \bar{\phi}_{.j}\right)^2,$$

$$\widehat{\text{var}}(\phi) = \frac{n-1}{n}W + \frac{1}{n}B,$$

$$\widehat{R} = \sqrt{\frac{\sqrt{\widehat{\text{var}}}(\phi)}{W}}, \tag{2.4.19}$$

where $\widehat{R}$ is the estimated potential reduction. If the chains converged, the between-chains variability values would be small. Consequently, $\widehat{R}$ gets closer to one. Whenever $\widehat{R}$ is greater than 1.1, this indicates non-convergence, thus it might be necessary to run the chains for longer.

Autocorrelation

Using the Markov Chain Monte Carlo (MCMC) algorithm in Bayesian inference produces correlated samples. This is due to the structure of drawing the samples where the new samples are drawn depending on the previous sample only (the Markov property). As the sampler moves wider around the parameter space, the autocorrelation decreases, indicating the samples' high efficiency and leading to a better estimate, while high autocorrelation results in slower convergence. Applying thinning, that is, discarding every s^{th} iteration from the chain, can help to reduce the autocorrelation. Thinning is a useful option when there is limited computer memory and storage space, but the reduction in the number of samples via thinned Markov chains can produce less precise results (Link and Eaton, 2012). The autocorrelation is also affected by the step size of the proposal distribution when using the Metropolis-Hastings (MH) algorithm, where a smaller step size leads to higher autocorrelation. The autocorrelation can be assessed by using an autocorrelation function plot.

The `coda` package (Plummer et al., 2006) in `R` is used to implement the convergence diagnostics discussed. When determining the length of a run, we take into account various factors, including the running time, running multiple MCMC chains and the number of parameters needed to store complete posterior samples, rather than just summary statistics. This is particularly important in our research, which involves multiple runs. We aim to achieve effective sample sizes in the hundreds or thousands. This level of adequacy is sufficient for the exploratory nature of our work, as stated in Mason (2010).

2.5 Methods for Handling Missing Data

In different fields of research, missing data occurs when the relevant variables' values are not measured or recorded. Addressing this issue is essential to ensuring the reliability and validity of statistical analysis. There are two general categories of approaches to dealing with missing data: the Ad-Hoc methods and the Model-Based methods. This section provides a general overview of common approaches to deal with missing data.

2.5.1 Ad-Hoc Methods

Ad-Hoc methods are considered a less complicated and straightforward way to handle missing data. It works by offering prompt resolutions for the presence of missing data without applying complex imputation methods. Ad-Hoc methods typically involve producing a single "complete" dataset, which is analysed using the standard complete-data techniques (Mason, 2010). Although this strategy is straightforward, it tends to be inadequate as it reduces precision and creates bias if the missing data mechanism is not Missing Completely at Random (MCAR) (Little and Rubin, 2019; Ma and Chen, 2018; Van Buuren, 2018). Complete-case data analysis is a common technique in which rows with incomplete records are excluded from the analysis (Little and Rubin, 2019). This method is implemented as a default technique in many statistical software programs by automatically excluding rows with a missing value for any variable. For example, widely used functions for longitudinal data in `R` are: `lme`, `nlme` and `brm` which uses all available data (Ibrahim and Molenberghs, 2009). For instance, rows containing missing values are omitted when data is in long format, meaning only missing observations are excluded, not the entire patient data (Pinheiro et al., 2017). By using all of the available data in longitudinal data, these tools remove the complete-case bias (Ibrahim and Molenberghs, 2009). Complete case analysis showed biased estimates and noticeably reduced power as expressed by Janssen et al. (2010); Knol et al. (2010); Van Buuren (2018). It is suggested to use the complete case if the proportion of missingness is less than 5% since the possible effect of the missing data is minimal (Jakobsen et al., 2017).

Additionally, there are single imputation techniques in which the missing data is filled with a single substituted value. The mean, median or mode values can be used to replace the missing value, which is one technique. Also, regression modelling can be used to predict the value of the missing data derived from the relationships between variables. In longitudinal data, the Last Observation Carried Forward (LOCF) can be applied to fill in the missing data, which assumes the observed value will continue until a new observation is recorded. Meanwhile, these techniques can potentially induce bias and improperly account for statistical uncertainty. As a result, they are not suggested for accurate analysis (Schafer and Graham, 2002). Detailed information on these techniques can be found in Little and Rubin (2019).

2.5.2 Model-Based Methods

Alternative to Ad-Hoc methods are Model-Based methods that use the information from observed data along with specific assumptions about the cause of the missing value (Mason, 2010) and effectively address the uncertainty caused by missing data. There are several approaches, such as Bayesian full probability modelling, weighting methods, multiple imputations, and maximum likelihood methods, which employ the Expectation maximization (EM) algorithm (Mason, 2010).

Expectation Maximization (EM)

A well-known approach to handling missing values is the Expectation Maximization (EM) approach, which uses likelihood inference (Dempster et al., 1977). EM alternates between an expectation step and a maximization step. In the expectation step, the expected value of the missing data is derived based on the observed data and the latest parameter estimates. In the maximization step, the parameter values are estimated by maximizing the likelihood of the parameters given the latest values of the missing data (Mason, 2010). The EM algorithm considers missing data as random variables that could be excluded from the likelihood function or integrated out of the likelihood as if they were not sampled (Schafer and Graham, 2002). The EM estimator is unbiased and efficient when the missing mechanism is MAR (Missing At Random) (Graham, 2003). The limitation of the EM algorithm is that the M step has no closed form. Although it requires numerical optimization techniques to overcome this constraint, which may become more computationally difficult, and when the proportion of missingness is large, the convergence rate might be very slow (Little and Rubin, 2019). The steps of the EM algorithm are expressed as follows:

Algorithm 3 Expectation Maximization Algorithm

```
Choose initial values of $\Theta^{(0)}$
for $1, \ldots, S$ iterations do
E-step: Compute the expectation of log-likelihood $l(\Theta|Y)$ as:
    $$Q(\Theta|\Theta^s) = E[l(\Theta|Y)|\Theta^s]$$
M-step: the next proposed estimate of $\Theta$ is calculated by maximizing the expectation from step E:
    $$\Theta^{s+1} = argmax_{\Theta}\, Q(\Theta|\Theta^s)$$
```

The algorithm repeats the process until it reaches a state of convergence. The convergence is satisfied whenever the difference between the new Θ^{s+1} and the current Θ^s estimates is small. EM methods maximise the log-likelihood function of the observed data to estimate the parameters directly, rather than "filling in" the missing data such as the Multiple Imputation method (Dong and Peng, 2013).

Multiple Imputation (MI)

Unlike the single imputation, where missing data is filled with a single plausible value, multiple imputation fills the missing data with multiple plausible values introduced by Rubin (1987), and it is one of the popular ways to deal with missing data. It has been recently widely used. There is extensive literature such as: Hayati Rezvan et al. (2015); Lee and Simpson (2014); O'Kelly and Ratitch (2014); Rubin (2004); Van Buuren (2018). Multiple Imputations will result in multiple complete datasets, where the values that are missing have been filled in/ imputed with plausible values. Usually, MI generates more than two datasets. Knowing that the imputed datasets are different from one another, the produced estimates of each dataset will not be identical, which allows the capture of additional uncertainty that arises by the missing values to the estimates (Rubin, 2004).

Generally, MI works by deriving the full conditional distribution for each incomplete variable, and the imputed values can be sampled from these distributions to create multiple complete datasets with imputed missing data. Multiple Imputation by Chained Equations (MICE) is a popular MI technique, which will be discussed in Section 5.2. MI and MICE methods consist of three steps: First, generate multiple datasets that have imputed missing data using the appropriate imputation model. Next, fit the appropriate statistical analysis model on each imputed dataset. Then, calculate the pooled estimates to gain the overall results and take into consideration the uncertainty produced by the missing values.

The methods proposed for pooling (combining) the results, known as Rubin's Guidelines (Erler, 2019), are explained as follows: The average estimates obtained from the analyses of K imputed datasets can be used to calculate the pooled (combined) estimate for a parameter vector Θ as:

$$\overline{\Theta} = \frac{1}{K} \sum_{\ell=1}^{K} \widehat{\Theta}_{\ell}, \tag{2.5.1}$$

where $\widehat{\Theta}_{\ell}$ represents the estimate derived from the ℓ^{th} imputed dataset. The total variance of Θ is composed of the within and between imputation variances as $T = \overline{V_W} + V_B + \frac{V_B}{K}$. The within-imputation variance is calculated as follows:

$$\overline{V_W} = \frac{1}{K} \sum_{\ell=1}^{K} \widehat{V_W}_{\ell}, \tag{2.5.2}$$

where $\widehat{V_{w\ell}}$ denotes the estimated variances of the Θ_ℓ from each imputed dataset. The between imputation variance B is calculated as:

$$V_B = \frac{1}{K-1} \sum_{\ell=1}^{K} \left(\widehat{\Theta}_\ell - \overline{\Theta}\right) \left(\widehat{\Theta}_\ell - \overline{\Theta}\right)^\top . \tag{2.5.3}$$

The pooled estimates can be implemented in the `MICE` package in `R` (Van Buuren and Groothuis-Oudshoorn, 2011). MI can decrease bias and improve precision in situations where the variable with missing data is related to other observed variables in comparison with complete case analysis (Lee and Simpson, 2014). Generally, it is important to mention that multiple imputation and EM algorithms produce valid statistical estimates in the context of the MAR condition (Little and Rubin, 2019). The estimates of the parameters may be biased whenever the used analysis procedure assumes MAR, and the true underlying mechanism of missing values is MNAR (Yang and Maxwell, 2014).

2.5.3 Joint Models for MNAR

Joint modelling is a statistical technique that combines the missing data process and the analysis of the model of interest into a single, simultaneous process. This approach is useful when dealing with non-ignorable missing data, which requires modelling the missingness process. Estimating parameters with non-ignorable missing data can be challenging since it requires the specification of the joint distribution of the data and the missing data mechanism, which is necessary to produce unbiased estimates (Ibrahim and Molenberghs, 2009). In such cases, a missingness model is required to address the non-ignorable missing data mechanism (Ma and Chen, 2018). To address non-ignorable missing data, three common frameworks based on the factorisation forms of the joint models for (Y_i, R_i) are available (Little and Rubin, 2019): the selection model (SM), the Pattern mixture model (PMM), and the Shared parameter model (SPM).

Selection Model (SM)

The selection model factors the joint distribution into the complete data model for the response and the probability of missingness on the response. Following the notation in Section 1.3.3, the selection model is expressed as:

$$f(Y_i, R_i|X_i) = f(Y_i|X_i)f(R_i|Y_i, X_i), \tag{2.5.4}$$

where $f(Y_i|X_i)$, the marginal distribution for Y_i, represents the response model and $f(R_i|Y_i, X_i)$, is conditional distribution of the missing indicator R_i given Y_i, represents the missingness model. Typically, the response model has a multivariate normal distribution and a logistic or probit model for the missingness model (Mason, 2010). Selection models are useful when the main interest is the marginal distribution of the outcome and understanding the overall distribution (Ibrahim and Molenberghs, 2009; Ma and Chen, 2018). However, it requires complex computational methods to fit the model, and the reliability of the parameter estimates could be affected by model assumptions (Ibrahim and Molenberghs, 2009).

Pattern Mixture Model (PMM)

The pattern mixture model groups the data according to different missing patterns where the joint distribution is factored as the conditional distribution of the response given the missing mechanism $f(Y_i|R_i, X_i)$ and the marginal model for the missing mechanism $f(R_i|X_i)$ as follows:

$$f(Y_i, R_i|X_i) = f(Y_i|R_i, X_i)f(R_i|X_i), \tag{2.5.5}$$

where the marginal distribution of the response would be a mixture of normal distributions since the distribution of the response depends on the missingness process. This is because the response distribution depends on the missingness process, meaning the response models will have different coefficients for different missing patterns. It can directly address identifiability issues by assuming a different distribution of the outcomes for each pattern, which makes it easier to explore the sensitivity to model parameters (Ibrahim and Molenberghs, 2009). However, the PMM has a disadvantage in that the parameters of interest are not available directly. Averaging over patterns is required to obtain the marginal distribution, which means that the regression coefficients cannot be used to analyse how each predictor affects the overall distribution directly (Ibrahim and Molenberghs, 2009).

In the context of longitudinal data, the SM and PMM have been reviewed by Fitzmaurice (2003); Michiels et al. (2002).

Shared Parameter Model (SPM)

In the shared parameter model, both Y_i and R_i are assumed to depend on a shared random effect. This means that the model for complete data Y_i and the model for the missingness mechanism R_i are connected through random effects. The shared parameter model can be thought of as a special case of the mixed effect model that uses the selection model. In this case, the probability of missingness is linked to random effects as follows:

$$f\left(Y_i, R_i, b_i \mid X_i\right) = f\left(Y_i \mid b_i, X_i\right) f\left(R_i \mid X_i, Y_i, b_i\right) f\left(b_i \mid X_i\right), \qquad (2.5.6)$$

where the random effects and the missingness process parameters can be viewed as nuisance parameters (Ibrahim and Molenberghs, 2009). A probit link function and the predictors were used to determine the missing probability, where the missing indicator follows a Bernoulli distribution. SPM has the ability to specify response models and missingness models more easily and is able to deal with structured data with multiple levels (Ma and Chen, 2018). On the other hand, it requires integration over the random effects, which makes it challenging to have a closed form (Daniels and Hogan, 2008). The SRE is generally used as an alternative to the selection model in a longitudinal context, in which the response model and the missing response indicator model have exactly the same random component b_i. In many circumstances, the underlying latent factors affecting the missingness might differ from those affecting the response, although they are correlated because of common risk factors. To model this case Lin et al. (2010) incorporate a correlation between the random effects in SRE, named as Correlated Random Effects (CRE) Model. This factorises the joint distribution as $p(Y_i,R_i,b_i,c_i \mid X_i) = p(Y_i \mid b_i,X_i)\ p(R_i|Y_i,c_i,X_i)\ p(b_i,c_i)$, where b_i and c_i are random effects in the response model and missing indicator model, respectively.

2.6 Probit Model

The traditional linear regression model assumes a continuous dependent variable that is used to model and predict the response based on a set of explanatory variables, also known as predictors. However, the response variable may not always be available on a continuous scale. It could be a binary variable that takes only two values. For example, an event of interest occurs or not, yes or no, pass or fail, and it can also be coded as 1 and 0. Probit regression is a statistical method that can be used to model a binary response given a set of predictors. This model is used in different studies to estimate the probability that an event can occur. The conditional probability that the response variable Y is equal to 1, given a set of predictors $X = [X_1,X_2,\ldots,X_J]$ is defined as:

$$E(Y|X) = p(Y = 1|X) = \Phi(X\beta), \tag{2.6.1}$$

where $\Phi(.)$ is the cumulative standard normal distribution function. Consequently, the error term in the Probit model has a zero mean and variance equal to 1, which is the variance of the standard normal distribution (De Leeuw et al., 2008). Probit regression assumes that the probability of the event of interest (1) occurring is determined by a normally distributed latent variable produced by the linear combination of predictor variables. Thus, the latent response formulation using the Probit link function is defined as (Gelman and

Hill, 2006):

$$\begin{aligned}
y_i &= \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* < 0 \end{cases} \\
y_i^* &= X_i\beta + \varepsilon_i \\
\varepsilon_i &\overset{\text{i.i.d}}{\sim} \mathrm{N}(0,1),
\end{aligned} \tag{2.6.2}$$

where y_i^* is the continuous latent variable. We can interpret parameters in terms of a latent variable y_i^* as follows: the latent variable y_i^* change associated with a one-unit change in the individual predictor variable, with the other predictor variables held constant (Agresti, 2010). The variance of the residuals is set to be equal to one to assure that the effects in the probit and the latent variable models are the same (Agresti, 2018).

To conduct Bayesian inference and obtain the posterior distribution of the probit model coefficients, we need to add a prior distribution to the β in Equation 2.6.2, denoted as $\beta \sim p(\beta)$. Albert and Chib (1993) used the latent variable to make the conditional distributions of the model parameters equivalent to those in a Bayesian normal linear regression model with Gaussian noise. They achieved this by simulating the latent continuous data from truncated normal distributions and then calculated the posterior distributions of parameters using the Gibbs sampling algorithm. Therefore, if the latent variable y_i^* is larger than 0, then event 1 is assumed to occur. Otherwise, event 0 is assumed (Johnson and Albert, 2006).

2.7 Model Evaluation

It is important to assess the fit of a model in any statistical analysis. Evaluating the performance of a method includes measuring the bias, accuracy, and coverage by using the parameter values generated from simulated data results (Burton et al., 2006). We can better understand a method's performance by evaluating multiple performance criteria since results may vary across these criteria. We will define the formulations of commonly used performance criteria, where β is the data generating parameter value and $\hat{\beta}$ is the estimated parameter value. The Root Mean Square Error (RMSE) is a measure of overall accuracy in the same units as the original data, incorporating bias and variability (Collins et al., 2001). The RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{\beta}_i - \beta\right)^2}{n}}, \tag{2.7.1}$$

where $\hat{\beta}_i$ represents the i^{th} estimated parameter value, with $i = 1, \ldots, n$ representing the estimated parameter samples. The RMSE is also used to measure the out-of-sample performance, where $\hat{\beta}_i$ is the predicted value and β is the observed value. As part of evaluating our models, we examine the bias and efficiency of the parameter estimates. The Relative Bias (RB) of a parameter estimate is used to assess the deviation of the parameter estimates from the data-generating parameter (Burton et al., 2006). The RB is defined as:

$$\text{Relative bias } = \frac{(\hat{\beta} - \beta)}{\beta}, \tag{2.7.2}$$

To evaluate how well a statistical method captures the data-generating parameter values, we use the Coverage Rate (CR), which is the percentage of times the data-generating parameter values appear between the 2.5 and 97.5 percentiles of the posterior distribution of the parameter (Mason, 2010). The CR is defined as:

$$\text{Coverage Rate} = \Pr(\hat{\beta}_{LCL} \leq \beta \leq \hat{\beta}_{UCL}), \tag{2.7.3}$$

where $\hat{\beta}_{LCL}$ and $\hat{\beta}_{UCL}$ are the lower and upper confidence limits for β (Nevalainen et al., 2009). It is recommended that the coverage rate be greater than 90 (Collins et al., 2001).

The Kolmogorov-Smirnov (KS) test is a non-parametric statistical test applied to continuous data. It is used to determine the statistical significance between two independent sample distributions or a sample distribution to a specific probability distribution. The Null hypothesis is that the two samples have the same distribution, whereas the alternative hypothesis is that the two samples have different distributions. The test statistic for the Kolmogorov-Smirnov test is expressed as:

$$D_{KS} = \max |F_1(o) - F_2(o)| , \tag{2.7.4}$$

where $F_1(o)$ is the empirical distribution function of sample data 1 and $F_2(o)$ is the empirical distribution function of sample data 2. The KS test statistic is represented by D_{KS}, the maximum absolute difference between the two empirical distribution functions of sample data 1 and 2. The test statistic is then compared with the critical value obtained from the KS table. If D_{KS} is large, it implies that the two distributions are coming from different distributions, while a small D_{KS} implies that the two distributions are similar (Bonnini et al., 2014). The KS test will be executed using the `ks.test()` function in `R` (R Core Team et al., 2013) and we will use the p-value with significance level 0.05. The null hypothesis is rejected if the p-value is less than 0.05.

Chapter 3

Data

3.1 Introduction

This chapter will present the main datasets used in this thesis. The datasets consist of real-world data collected from a heart failure study, named BIOSTAT-CHF (Voors et al., 2016), and a synthetic dataset that we created to support the methodological exploration. We created the synthetic dataset because we know the actual parameter values that we need to evaluate the proposed methods in the upcoming chapters. We will provide a detailed description of both datasets in the following sections.

The BIOSTAT-CHF dataset is based on a study investigating patients with heart failure. Therefore, we will begin this chapter by introducing heart failure in Section 3.2. Next, in Section 3.3, we will describe the BIOSTAT-CHF dataset, including defining the medical terms used in the dataset in Subsection 3.3.1, and explore the data in Subsection 3.3.2 to gain insight into the data. After that, we will explain the synthetic data setup in Section 3.4.

3.2 Heart Failure

Chronic Heart Failure (CHF) is a major public health problem where the heart fails to pump blood sufficiently to meet the body's needs. This can lead to morbidity and mortality. In developed countries, mortality from Heart Failure

(HF) remains high, with over one million people in Europe and the United States and 26 million people worldwide developing heart failure annually (Ambrosy et al., 2014). According to James et al. (2018), approximately 64 million people worldwide suffer from heart failure. Furthermore, there has been an increase in the total number of patients diagnosed with heart failure (Groenewegen et al., 2020). Patients with CHF are often hospitalised at least once every two years (Malmberg and Persson, 2000), and it is estimated that heart failure costs $108 billion worldwide annually (Cook et al., 2014).

Additionally, HF has become more common as people age, and survival rates for HF patients remain low worldwide (Ponikowski et al., 2014). As people get older, they are more likely to develop cardiovascular risk factors (Smeets et al., 2019). However, heart failure has recently become more common in younger people, not just older individuals (Groenewegen et al., 2020). Heart failure can be considered the chronic stage of another disease that leads to heart dysfunction, making it difficult to identify an exact cause for an individual (Groenewegen et al., 2020). Symptoms of heart failure can be similar to those of other diseases, such as chronic obstructive pulmonary disease or obesity, which can lead to misdiagnosis. Additionally, echocardiography is not commonly performed in primary care units, which may contribute to undiagnosed cases of heart failure (Caruana et al., 2000).

As Ponikowski et al. (2014) suggested, a global methodology is needed to recognise the best methods of tending to the issue of heart failure to fuse the fundamental measures into regular practice. It is of interest to fit a robust prognosis prediction model in order to predict medical outcomes, such as worsening of heart failure. With the availability of data from longitudinal studies, we can predict early signs of heart failure by analysing patients' records. This can help provide patients with medication at an early stage to slow down HF progression and prevent hospitalisation. Given that each patient's journey with HF is different, robust statistical methods are needed to assist cardiologists and decision-makers in predicting the patients' HF. In the next section, we will present the BIOSTAT-CHF dataset, which was collected for a HF study.

3.3 BIOSTAT-CHF Data

A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure (BIOSTAT-CHF) dataset was derived from 11 European countries and funded by a grant from the European Commission. It was designed to improve and support risk prediction models in individuals with HF (Voors et al., 2016). The study was conducted between 2010 and 2015, and the recruitment period lasted 25 months, with a median and Inter Quartile Range (IQR) follow-up of 21(15-27) months. It consists of an index cohort and a validation cohort. The index cohort comprises 2516 European heart failure patients, and the validation cohort comprises 1738 from Scotland, UK. The enrolment of patients in this study was from inpatient or outpatient clinics. Moreover, the age of patients from the index cohort was more than 18 years old with symptoms of heart failure worsening or new onset of heart failure. The provided dataset (BIOSTAT-CHF) has two time points, the measurements were recorded at baseline and at nine months (second visit). The BIOSTAT-CHF dataset is the main source of real data to develop and test the performance of the proposed methods, which was provided by the Robertson Centre for Biostatistics, University of Glasgow.

The BIOSTAT-CHF dataset has been valuable in numerous studies, including developing risk models to predict mortality and hospitalizations due to heart failure. They found that specific clinical routines can provide essential prognostic information for patients with HF, and the mortality predictors differed from those of hospitalization due to HF (Voors et al., 2017). Another study compared the characteristics and outcomes of HF patients treated as inpatients versus outpatients. The inpatients were sicker, but some outpatients also had poor prognoses, suggesting an overlap in conditions (Ferreira et al., 2019). Since the BIOSTAT-CHF study is multicenter, researchers have also compared participants' characteristics, treatment, and outcomes based on their European geographical locations, and the study showed significant differences in the clinical treatment of heart failure across different regions (Lombardi et al., 2020).

3.3.1 Definitions of Medical Phrases

In this section, we will define the BIOSTAT-CHF dataset biomarkers' medical phrases used in the study and how they are related to heart failure. Table 3.3.1 expresses a very brief summary of some medical definitions related to heart failure. Next, we will further define the medical phrases and how they are related to the HF.

Table 3.3.1: Definitions of some phrases related to heart failure.

Phrases	Definition
Left Ventricular Ejection Fraction **(LVEF)**	It is a measure of how much blood is pumping out of left ventricle of the heart.
Beta-blocker	Is a medication used for chronic heart failure.
Heart Rhythms	Are the patterns of the heartbeats.
N-terminal pro-B-type natriuretic peptide **(NT-proBNP)**	Is one of the most powerful prognostic biomarkers in cardiovascular diseases.
Estimated Glomerular Filtration Rate **(eGFR)**	Is considered the best general indicator of how well kidneys are working.

- Left Ventricular Ejection Fraction **(LVEF)** is useful for diagnosing heart failure. It measures how much blood is pumping out of the left ventricle of the heart. A 'weak' heart muscle has a low LVEF. The now used categorization for HF patients based on LVEF is provided in the study by Ponikowski et al. (2014), as expressed in Table 3.3.2 below. An LVEF of less than 40% indicates heart failure.

Table 3.3.2: Categories of heart failure based on LVEF values.

LVEF ranges	Type of HF	Definition
LVEF $\geq$ 50%	HFpEF	Heart failure with preserved ejection fraction
40% $\leq$LVEF$<$ 49%	HFmrEF	Heart failure with mid-range ejection fraction
LVEF $<$ 40%	HFrEF	Heart failure with reduced ejection fraction

- **Beta-blocker** is a medication used to treat chronic heart failure. It has been shown to improve chronic heart failure patients' survival (Jost et al., 2005). In patients with heart failure, the heart beats too fast, which can exhaust the heart muscle. Beta-blockers work by slowing down the heart rate and allowing the heart to recover.

- **Heart Rhythms** refer to the patterns of heartbeats, and they can be classified into three levels: sinus rhythm, atrial fibrillation/flutter, and pacemaker. A healthy heart usually beats in sinus rhythm, which is the normal rhythm. Atrial fibrillation/flutter, on the other hand, is characterized by an irregular and often fast heart rate. A pacemaker is a small device implanted into the body, used to regulate heart rhythm in case of serious arrhythmia. Atrial fibrillation is a commonly occurring condition in patients who have heart failure (Savarese et al., 2022), and according to a study by Shahid and Lip (2016), more than half of the patients with atrial fibrillation are diagnosed with heart failure.

- N-terminal pro-B-type natriuretic peptide (**NT-proBNP**) is a hormone produced by the heart when it is under stress, which usually happens when the body is overloaded with fluids. This hormone activates kidneys to dispose of more salt and water. NT-proBNP is one of the most powerful prognostic biomarkers for detecting HF (Oremus et al., 2014; Wilson Tang, 2007) and evaluating its severity (Tsutsui et al., 2023; Werhahn et al., 2022). It provides strong and independent prognostic information in patients with heart failure (Weber and Hamm, 2006). Additionally, the levels of NT-proBNP increase in patients with atrial fibrillation (AF)

(Patton et al., 2009).

- Estimated Glomerular Filtration Rate **(eGFR)**, is a measurement of the clearance of exogenous filtration markers and is considered the best general indicator of how well kidneys are working (Levey et al., 2003). The value of eGFR gives information about how much kidney function a patient has, and as eGFR decreases, the kidney disease gets worse. Worsening renal function is associated with poor outcomes in HF patients with chronic kidney disease (Metra et al., 2012). Levey et al. (2009) defined the CKD-EPI equation that can be used to calculate the eGFR value using serum creatinine level, race, sex and age. The CKD-EPI equation is expressed as:

$$\begin{aligned} eGFR = 141 \times min\left[\frac{S_{Cr}}{k},1\right]^{\alpha} \times max\left[\frac{S_{Cr}}{k},1\right]^{-1.209} \times \\ = 0.993^{Age} \times 1.018[Sex] \times 1.159[Race] \end{aligned} \tag{3.3.1}$$

where,

$$\text{Sex} = \begin{cases} 1, \text{ if female} \\ 0, \text{ if male} \end{cases}, \quad \text{Race} = \begin{cases} 1, \text{ if black} \\ 0, \text{ otherwise} \end{cases}$$

S_{Cr} is serum creatinine in mg/dl , k is equal to 0.7 when sex is female and 0.9 when male, and α is equal to -0.329 when sex is female and -0.411 when male and age expressed in years. These values were obtained by modelling serum creatinine using two slope linear splines with six knots, sex, race, and age. which can help to overcome the problem of underestimating eGFR at higher values, especially eGFR $> 60\, mL/min/1.73m^2$. This was expressed by Levey et al. (2009).

The New York Heart Association **(NYHA)** classification consists of four categories depending on the patient's restriction during physical exercise, e.g. walking for a certain distance. It is a way of classification to detect heart failure. There are four categories in total, with Class III and

IV being the most severe in terms of symptoms (Raphael et al., 2007). The classes are outlined in Table 3.3.3 below:

Table 3.3.3: New York Heart Association classification based on the severity of symptoms and physical activities.

NYHA Class	Degree of symptoms with physical activity
Class I	No limitation of physical activity.
Class II	Slight limitation of physical activity.
Class III	Marked limitation of physical activity.
Class IV	Unable to carry on any physical activity without discomfort.

3.3.2 Exploratory Data Analysis

Descriptive statistics will be presented to provide an overview of patients characteristics and study variables in the BIOSTAT-CHF dataset.

The majority of the patients in BIOSTAT-CHF were male (73%), and the average age was 69 years with a 12 year standard deviation (SD). The average Body Mass Index (BMI), similar regardless of the sex, is 28 (SD 5). About 14% of patients were current smokers at baseline, and 27.8% were alcohol consumers at baseline.

The majority of patients in the BIOSTAT-CHF dataset (63%) are male with the LVEF$< 40\%$ (HFrEF), which is an indication of heart failure with reduced ejection fraction. There are about 49% of patients who were categorised in Class III in regards to the NYHA. The mean (SD) of the baseline eGFR is 63.5(23.5) $ml/min/1.73m^2$, which, based on the National Kidney Foundation (2002) classification, is "Kidney damage with mild decrease in eGFR".

The NT-proBNP is highly right-skewed as shown in Figure 3.3.1, with median (IQR) equal to 3386 ng/L (IQR =1613 ng/L - 88955 ng/L). Furthermore, about half (45%) of the participants in the study have "Sinus Rhythm", 24% have "Atrium fibrillation/ flutter", 13% have "Pacemaker" and 16% are characterised as Other as presented in Figure 3.3.2.

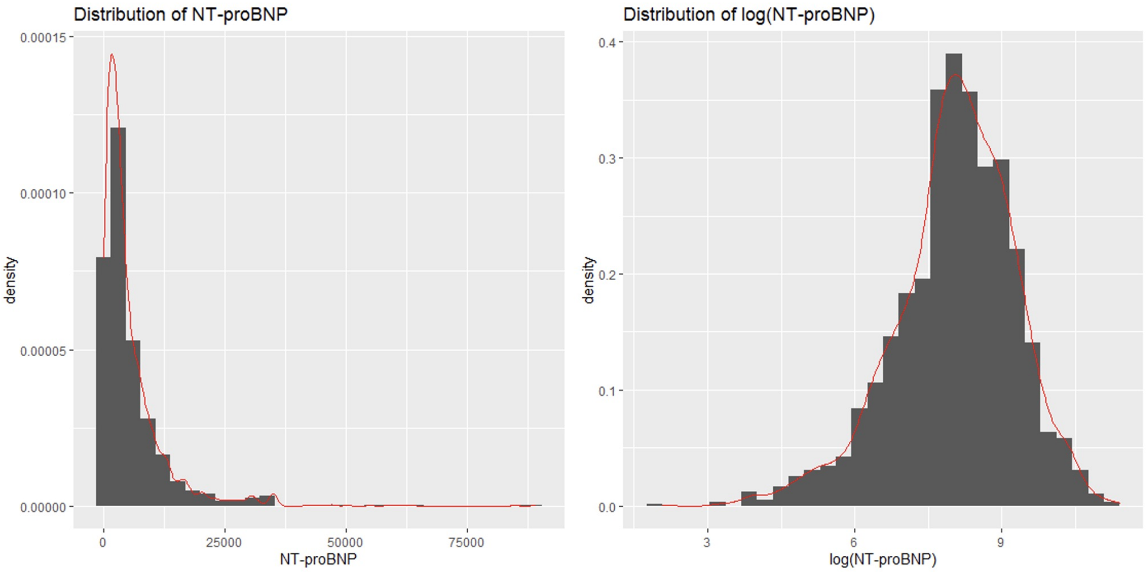


Figure 3.3.1: A histogram and density curve in red are shown for NT-proBNP on the left-hand side and log(NT-proBNP) on the right side. The log transformation made the distribution of NT-proBNP more symmetric.

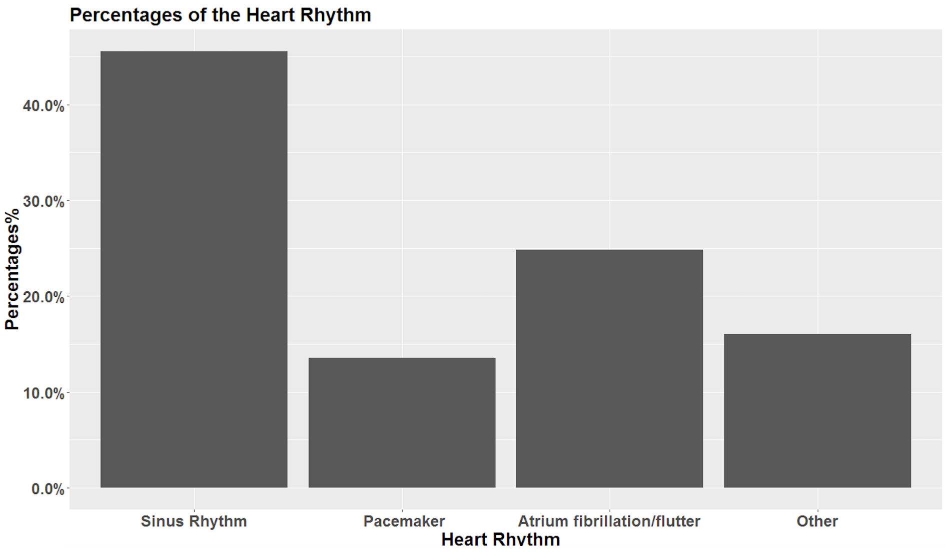


Figure 3.3.2: A bar plot of the Heart Rhythm categories of participants in the BIOSTAT-CHF dataset. The majority have Sinus Rhythm.

The BIOSTAT-CHF dataset contains missing data. A descriptive analysis was conducted to explore the type of missingness in the variable of interest. Figure 3.3.3 shows that there is a large proportion of missing values in the NT-proBNP variable. The number of missing observations in NT-proBNP and eGFR increased in the second visit. However, there are no missing values in the other variables used in this study. From Figure 3.3.4, we can gain insights into the possible relationships between variables and the missingness. It is noticeable that the largest proportion of missing values in NT-proBNP and eGFR occurs when the heart rhythm takes on the category "others" and in the second visit.

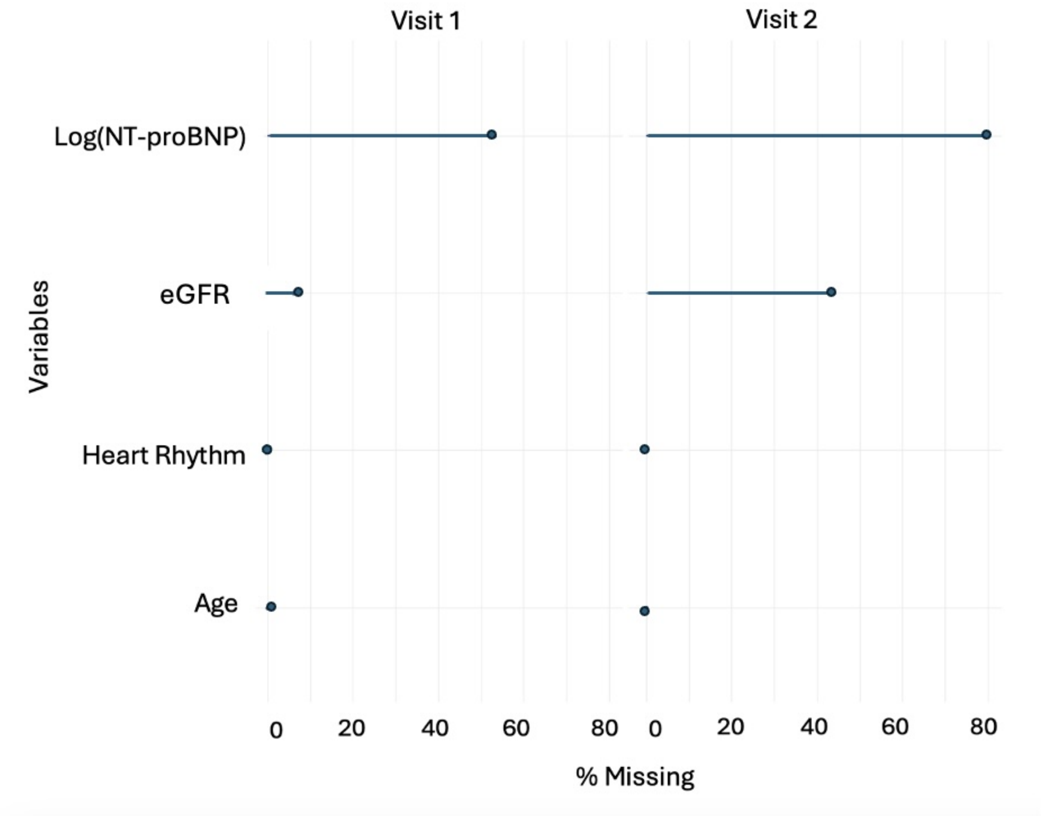


Figure 3.3.3: A graph displaying the percentage of missing values for each variable of interest, separated by the visit time. The variable log(NT-proBNP) has a higher percentage of missing data.

Furthermore, the missing values in the NT-proBNP occur with a high percentage, regardless of the category of the heart rhythm. The missing values in NT-proBNP are unrelated to Age and eGFR. On the other hand, the missing values of eGFR are associated with lower values of NT-proBNP and older participants. Understanding this association provides valuable insights into the missing data pattern and can guide further investigations.

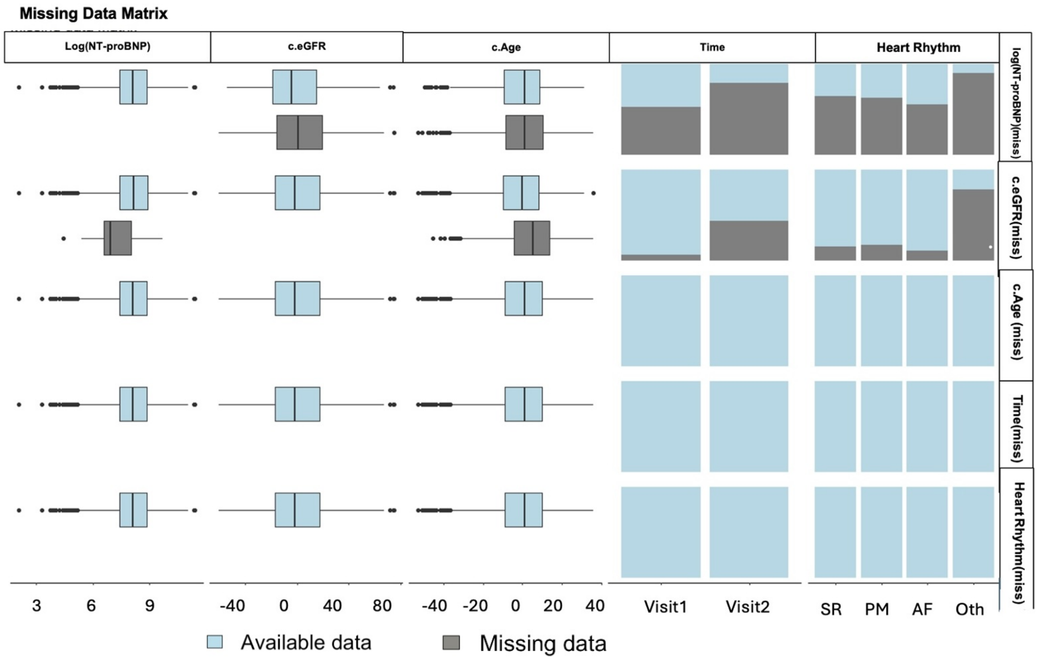


Figure 3.3.4: A missing data matrix plot is a graphical representation that shows the association between each variable of interest and the missingness. The x-axis represents the characteristics of the variable of interest, while the y-axis represents the missing values of each variable of interest. The continuous variable characteristics are described using a box plot, while categorical variable characteristics are described using a bar plot.

3.4 Synthetic Data

In this section, we will describe our process for generating simulated data. This will allow us to compare our results to known truth and evaluate different statistical methods in upcoming chapters. Additionally, the simulated dataset with known real values of missing observations will enable us to evaluate the effectiveness of missing data modelling methods. Therefore, we generate synthetic data scenarios by intentionally introducing missingness into complete simulated datasets.

As this research focuses on repeated measures, we will generate longitudinal data using the linear mixed effects model with a subject-specific random intercept. The simulated data set-up was inspired by Bhuyan (2019) and Ibrahim et al. (2002) and has been chosen to reflect the data found in similar studies.

The analysis model for simulation considers a continuous response and two continuous predictors, which are time-varying, one binary time-invariant predictor and subject-specific random intercept. This is the widely used linear mixed effect regression model as follows:

$$Y_i(t) = \beta_0 + \beta_1 X_{1i}(t) + \beta_2 X_{2i}(t) + \beta_3 X_{3i} + u_i \tilde{Z}_i(t) + e_i(t), \tag{3.4.1}$$

where $Y_i(t)$ is the response variable for the i^{th} subject at t^{th} time point, $i = 1, \ldots, n$ and $t = 1, \ldots, m$. $\beta_0 = 10$ is the overall intercept, $\beta_1 = 2$ and $\beta_2 = 5$ are regression coefficients associated with the time-varying fixed effect, and $\beta_3 = 15$ is the regression coefficient associated with the time-invariant fixed effect. The random components are subject-specific random intercepts $u_i \overset{\text{i.i.d}}{\sim} N(0,2)$ and $e_i \overset{\text{i.i.d}}{\sim} N(0,9)$ are the model's residuals.

The time-varying predictor X_2 is generated from a uniform distribution as:

$$X_{i2}(t) \sim U(0,2) \tag{3.4.2}$$

The time-invariant predictor X_3 is generated from a Bernoulli distribution as:

$$X_{i3} \sim Bern(0.6) \tag{3.4.3}$$

To assume possible missingness in the analysis model predictor, we consider the incomplete predictor X_1 to be a function of the other two complete predictors and accommodates variations at different time points within each subject, generated from the linear mixed model as follows:

$$X_{i1}(t) = \alpha_0 + \alpha_1 X_{2i}(t) + \alpha_2 X_{3i} + w_i \tilde{Z}_i(t) + r_i(t), \tag{3.4.4}$$

where $\alpha_0 = 1$ is the overall intercept, $\alpha_1 = 3$ and $\alpha_2 = 0.4$ are regression coefficients associated with the fully observed predictors. The individual's random intercept is assumed $w_i \overset{\text{i.i.d}}{\sim} N(0,2)$, and the residuals are assumed $r_i \overset{\text{i.i.d}}{\sim} N(0,0.5)$.

Multiple independent simulations of the same setup are run as part of a parallel simulation technique (Burton et al., 2006). The number of generated simulated datasets was constrained by factors such as the proportion of missingness and sample sizes, statistical methods, computer availability, and time limitations. Therefore, we generated 100 datasets with a fixed number of subjects ($n = 100$), varying number of repeated measures ($m = 2, 4$ & 8) and proportion of missingness to approximate a range of missing values and repeated measures that may occur in real-life data. Specifically, we introduced 20%, 40%, and 60% missing data in the response variable of the analysis model, along with a fixed 20% missing data in the incomplete predictor. Additionally, we generate 20%, 40%, and 60% missing data in the incomplete predictor of the analysis model, with a fixed 20% missingness in the response variable of the analysis model. We initially set up to 60% missingness in either the model response or the incomplete predictors. However, we technically exceeded this 60% proportion of missingness. For instance, when there is 60% missingness in the model response, there is also 20% missingness in the incomplete predictor. This indicates that the overall missingness is likely greater than 60% missingness.

Then, some of the observations will be deleted based on various models of missingness settings, which will be explained in detail in each upcoming chapters. By ensuring that the missing values are known, this straightforward setup allows us to emphasize the essential features of the performance of our proposed methods. In the case of simulated datasets where the missing values are known, we also implement the models of the same structure but with a complete set of values, referred to as the "Full data" model. This Full model serves as a benchmark for evaluating the performance of the proposed methods on simulated datasets. Furthermore, we apply the model of interest to complete cases, denoted as the "Available data" model. This allows us to compare the fit of the model of interest with a commonly employed approach, that of complete case analysis.

Typically, simulations aim to replicate results that could have arisen from a single study. The credibility of the model parameter's posterior distribution and prediction can be verified using different data that has not been explicitly incorporated into the model. Therefore, we will create test data with identical settings but with no missingness. Comparing the performance of several methods comes next. It is essential to use more than one performance criterion, such as Root Mean Square Error, Relative Bias, and Coverage Rate, since findings vary between several criteria (Burton et al., 2006). A combination of posterior samples over 100 simulations will be presented as a measure for the true estimate of interest. Criteria equations and definitions were explained in Section 2.7.

Chapter 4

Exploring Longitudinal Data Inferences and the CRE Method

4.1 Introduction

After presenting our data in Chapter 3, we will start this chapter by fitting a model to the BIOSTAT-CHF data. Next, we will examine the performance of linear mixed effect models using the two approaches in the statistics framework to model longitudinal data, the Frequentist and Bayesian approaches. The aim is to determine which model provides a more robust prognosis prediction of the response. Since we are dealing with longitudinal data, we consider an issue that influences the model performance, which is missing data. To address this, we employ a recent approach that introduces a Correlated Random Effects (CRE) method using a Gibbs sampler for longitudinal data in situations where the data contains missing responses generated by an informative missingness mechanism. We will explain the CRE method and identify the non-convergence factors that affect the model's performance. This chapter will use simulated and real-data (BIOSTAT-CHF) datasets.

This chapter is structured as follows: Section 4.2 defines the model for fitting the BIOSTAT-CHF data in order to predict heart failure. Section 4.3 performs a comparative study between the Bayesian and Frequentist approaches and presents the corresponding results for comparison using simulated and real data. Section 4.4 highlights the methodology adopted for non-ignorable missingness in the response and explains a possible solution to overcome non-convergence issues, we will present this using simulated data and then apply the CRE method to the real data. Finally, Section 4.5 discusses the main findings.

4.2 BIOSTAT-CHF Model

The BIOSTAT-CHF dataset introduced in Section 3.3 contains multiple records for each participant to detect changes, risk factors, or long-term predictor variables for patients with heart failure. To analyze this dataset, we need to capture the dependency of observations within participants by using the Linear Mixed Effect Model (LMM). Creating a single model using all variables in the BIOSTAT-CHF dataset can be challenging. Therefore, it is necessary to select a set of representative variables as a foundational step to apply the method for predicting heart failure in individuals.

NT-proBNP is a strong prognostic factor in heart disease, as mentioned in Section 3.3, and it has a considerable amount of missing values, as shown in Figure 3.3.3, which made it desired to estimate it more effectively. Therefore, it is considered as a response variable in the proposed LMM, with visit numbers creating repeated measures. The patient's age, eGFR, Heart Rhythm, and visit time are predictive factors in the model. The individuals (patients) are considered as the random effect. This particular set of variables has been chosen as benchmarks to start the model and recommended by an expert cardiologist with extensive expertise in the field, Professor of Cardiology John Cleland, in the School of Cardiovascular & Metabolic Health at the University of Glasgow. Moreover, the use of this particular set of predictors is also supported by existing research that predicts an individual's risk of heart failure (Voors et al., 2017).

The model proposed for the BIOSTAT-CHF dataset is expressed as follows:

$$Y_i(t) = \underbrace{\beta_0 + \beta_1 \text{Age}_i(t) + \beta_2 \text{eGFR}_i(t) + \beta_3 \text{HR}_i(t) + \beta_1 \text{Time}_i(t)}_{\text{fixed effects}} + \underbrace{u_i + e_i(t)}_{\text{random terms}} \;;$$

$$u_i \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_B^2\right) \;\; \& \;\; e_i(t) \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_A^2\right). \tag{4.2.1}$$

Linear Mixed Models (LMM) in Equation 4.2.1 were conducted to investigate how the following fixed effects: age, eGFR, Heart Rhythm (HR) and index time of visit can affect the patient's NT-proBNP ($Y_i(t)$). The characteristics of these predictors were discussed in Section 3.3. The model takes into consideration the within-subject variation via the random effect u_i and within individual variation via the model's residuals ($e_i(t)$), where both represent the random terms in the model.

Log transformation

The use of NT-proBNP violated the normality assumption in the LMM. A log transformation was applied to address this issue, which is a suggested transformation technique (Gelman and Hill, 2006) to stabilize variance (Jiang and Nguyen, 2007) and improve normality in skewed error distributions. The Q-Q plot is a tool for checking the normality of error assumptions by matching the residual quantiles with a normal distribution. The Q-Q plot with a curvature or departure from the line of equality indicates non-normality distributed residuals in the model. Therefore, the normality was re-evaluated after the log-transformed the response, which resulted in a closer alignment with the normal distribution as shown in Figure 4.2.1. As a result, the assumption of normality is satisfied in the model with the log-transformed response variable. Moreover, the residual vs. fitted values plot showed that the model with NT-proBNP as the response variable violates the assumption of constant variance of residuals. The log-transformed model validated the assumption of homoscedasticity with evenly spread data points around fitted values. The plot is located in Section A in the Appendix.

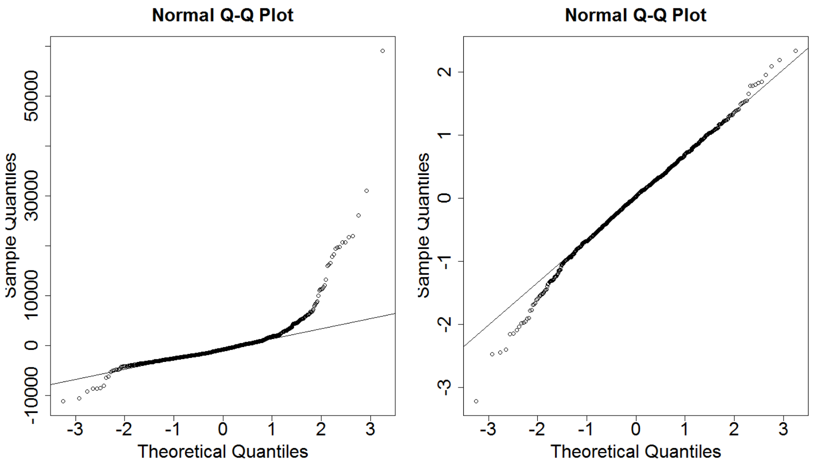


Figure 4.2.1: Two different BIOSTAT-CHF models are compared by displaying their Q-Q plots side by side. The Q-Q plot of the model fitted using NT-proBNP on the left-hand side, as the response variable shows curvature, indicating that the residuals are not normally distributed. On the other hand, the Q-Q plot of the model fitted using the logarithmic transformed response variable log(NT-proBNP) on the right-hand side, shows that the residuals are reasonably normally distributed, as the residual points fall mainly along a line of equality. Therefore, the model with a log transformation NT-proBNP is considered to be a better fit for the data.

Centering

Centering is a statistical transforming strategy that subtracts a variable's mean from its observed values. This approach is useful because it eliminates the correlation between parameters in a regression model (Lynch, 2007), leads to a meaningful interpretation of the regression parameters and makes the model more stable (Paccagnella, 2006). Additionally, centred models may converge faster than models that do not use centred variables (Paccagnella, 2006). While centring a variable will change the values of the regression parameters, it will not affect the association between the response and predictors (Enders and Tofighi, 2007). Therefore, interpreting a centred model is simply the expected response value when the predictors are equal to the mean value. Since we have only two visits per subject in the BIOSTAT-CHF dataset, we will centre the continuous predictor variables "age" and "eGFR" to the overall mean. The distribution of these variables before and after applying centring transformation is shown in Figure 4.2.2. The original distribution of variables is centred around their mean values, while the transformed variables are centred around zero.

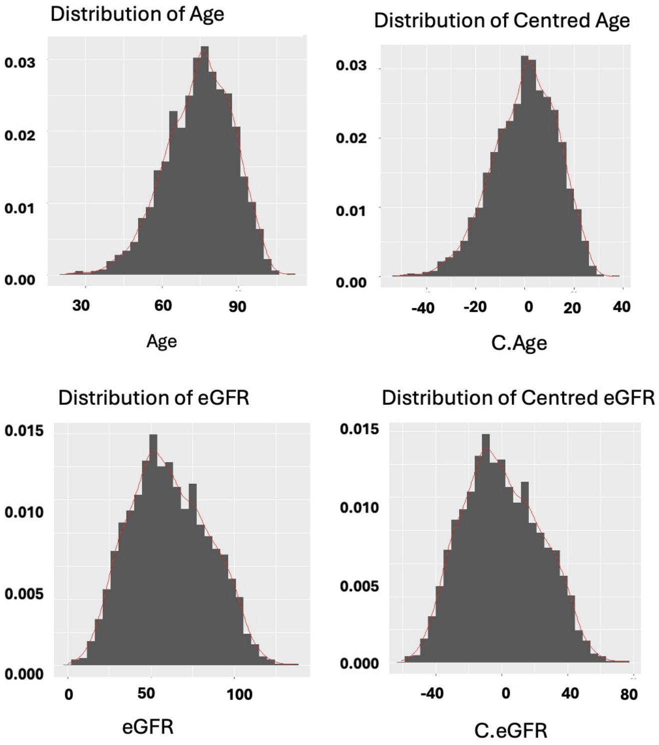


Figure 4.2.2: The histogram and density curve (in red) of the continuous predictors in the BIOSTAT-CHF data model are represented. The top panel shows the distribution of age, while the bottom panel shows the distribution of eGFR. The distribution of each variable is displayed on the left-hand side, while on the right-hand side, the distribution of centred variables is presented. These predictors display a symmetrical distribution and have been centred around their mean to enhance the interpretation of the results and reduce potential correlation issues.

We will rewrite the model in Equation 4.2.1, taking into account the transformations discussed previously as follows:

$$\begin{gathered} log(NT-proBNP)_i(t) = \beta_0 + \beta_1 \text{C.Age}_i(t) + \beta_2 \text{C.eGFR}_i(t) + \beta_3 \text{HR}_i(t) \\ + \beta_1 \text{Time}_i(t) + u_i + e_i(t); \\ u_i \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_B^2\right) \;\; \& \;\; e_i(t) \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_A^2\right). \end{gathered} \tag{4.2.2}$$

4.3 Comparison Study Between the Frequentist and the Bayesian Approaches in the Context of Longitudinal Data

We aim to compare two statistical approaches, the Frequentist Linear Mixed Effects Model and the Hierarchical Bayesian Model (HBM), to determine which method is more reliable in predicting longitudinal data. We will use the Root Mean Square Error (RMSE) to evaluate the model prediction of unobserved outcomes. To compare the estimates from the Bayesian approach and the Frequentist approach, we use the average of the posterior distribution obtained from the Bayesian approach and compare it with the point estimates from the Frequentist approach. We will compare the prediction errors between the two approaches by re-fitting the model 50 times, splitting the data into 30% test and 70% training dataset with replacement. This process will produce comparable RMSEs distribution from the two approaches, which can help us to quantify the uncertainty between the two approaches. We can compare the RMSE using the two approaches visually by plotting the density plot of the RMSE. The calculation of the RMSE is shown in Equation 2.7.1 in Section 2.7 Additionally, we will test whether the two distributions of the RMSEs are similar by using the Kolmogorov-Smirnov test (defined in Section 2.7). We will report the results of these tests in Subsection 4.3.1. Our ultimate goal is to determine which method provides a more reliable longitudinal data prediction.

4.3.1 Results

In this section, we will compare the results obtained from the Bayesian posterior distribution and the point estimates from the Frequentist approach. This will help us understand the relative location of the point estimates from the

Frequentist approach compared to the posterior distribution of the Bayesian approach. We used the `brm` function from the `brms` package (Bürkner, 2017) to fit the Bayesian hierarchical model using the Hamiltonian Monte Carlo (HMC) algorithm. Section 2.4.5 explains the HMC algorithm in detail.

We used the default prior in the `brm` function, which is considered non-informative. The default prior for the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)$ is a flat prior over the real numbers, meaning there are no restrictions on the values for the coefficients. Furthermore, the default prior for the standard deviations of the random effect and the residual is the Student-t distribution with 3 degrees of freedom, location 0, and scale 12.3. This is expressed as:

$$\begin{aligned} p(\boldsymbol{\beta}) &\propto 1, \\ p(\sigma_A^2) &\sim \text{Student}-\text{t}(3, 0, 12.3), \\ p(\sigma_B^2) &\sim \text{Student}-\text{t}(3, 0, 12.3). \end{aligned} \tag{4.3.1}$$

On the other hand, we used the `lme` function from the `nlme` package (Pinheiro et al., 2007) to fit the Frequentist Linear Mixed Model (LMM) as mentioned in Section 2.3.1. Both functions were applied using the `R` program (R Core Team et al., 2013). In the refitting process, we will compute the average of the posterior distribution of each model parameter to calculate the Root Mean Squared Error (RMSE). In the following sections, we will present the results using simulated and real data.

Simulated Data Results

In this section, we will present the results obtained from simulated data described in Section 3.4 with four repeated measures. However, the results of other repeated measures are in Section A in the Appendix. Using the simulated data, we can identify the location of the data-generated parameter values, as well as the Frequentist point estimates, in the posterior distribution. Figure 4.3.1 shows that the Frequentist point estimates are located in the centre of the Bayesian posterior distribution of the analysis model parameters. Additionally, the 95% Confidence Interval (CI) falls within the range of the posterior distribution. This alignment indicates that both approaches precisely capture the underlying relationships within the data. Furthermore, the data-generated parameter value is closely aligned with the posterior distribution, further strengthening the accuracy of both methodologies.

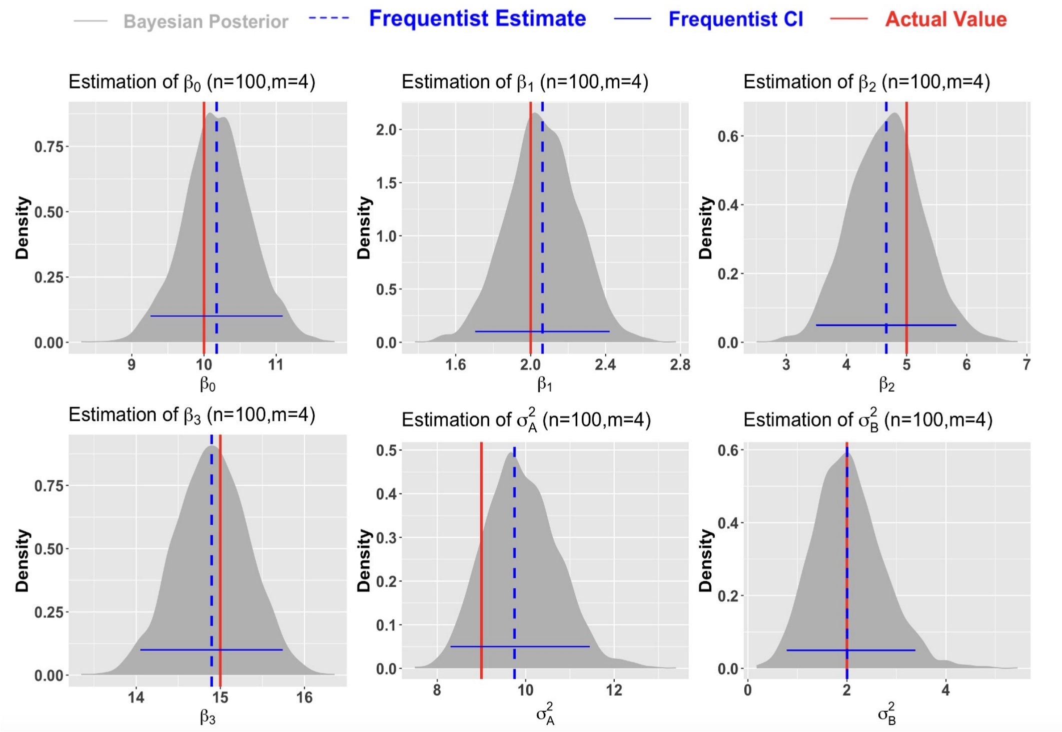


Figure 4.3.1: The grey curve represents the posterior density obtained using the Bayesian approach. There are three lines; a vertical blue dashed line representing the point estimate using the Frequentist approach, the blue horizontal line representing the 95% CI and a vertical red solid line representing the data generated value. Each plot represents one of the analysis model parameters that generated the simulated data with four repeated measures. The data-generated parameter value and the point estimates are located in the centre of the posterior distribution, indicating a good level of agreement between the Bayesian and Frequentist approaches.

As both approaches have produced similar estimated values, we would investigate which model performs better on unseen data. This will help us understand how these approaches will work with real data when we do not know the actual values of the parameter estimates. Furthermore, we aim to determine the level of uncertainty associated with each approach. To do this, we used the resampling process discussed in Section 4.3, which involves splitting the dataset into training and test data 50 times with replacement. In each iteration, each individual could be selected for either the training or testing set, allowing for varied inclusion of individuals across the iterations. Figure 4.3.2 shows the distribution of the RMSEs from the two approaches obtained using this process. It is evident from the plot that both approaches produce identical out-of-sample RMSE values, which indicates that there is no significant difference in terms of out-of-sample performance between the two approaches.

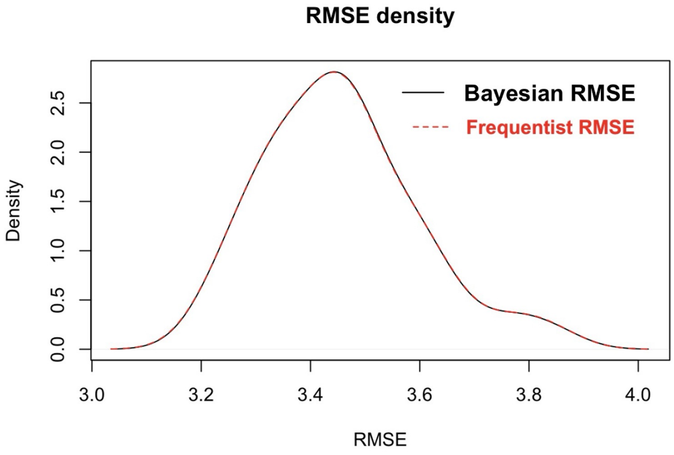


Figure 4.3.2: The plot displays two curves by applying the refitting process. The black curve represents the out-of-sample RMSE calculated using Bayesian inference, while the red dashed curve shows the RMSE calculated using Frequentist inference. Simulated data with four repeated measures were used to generate the curves. Interestingly, both approaches showed identical out-of-sample performance, indicating that there is no preference between them.

A full posterior distribution in Bayesian analysis shows a complete view of uncertainty compared with point estimates. Point estimates provide a single value for a parameter, whereas the posterior distribution covers a range of plausible values and the associated probabilities. This gives us a better understanding of the parameter's possible values and provides a range of values within which the parameter estimates value is likely to lie. For example, Figure 4.3.1 provides an overview of the plausible values. These values have a high density around the average value, which is close to the data-generating value and Frequentist parameter estimate, representing values close to the "truth". With values far away from the data-generating parameters, there is a relatively sharp drop in density. This approach not only provides a range of values like the Frequentist approach but also indicates which values within that range are more plausible. Credible intervals may be helpful for decision-making using Bayesian inference.

To calculate the out-of-sample performance using the RMSE, one can take advantage of applying Bayesian inference, which allows us to produce a distribution of the RMSE, while Frequentist inference only provides a single value for the RMSE for one dataset. In Figure 4.3.3, we can see the RMSE distribution obtained through Bayesian inference and a calculated RMSE using inference for one of the resampling data with four repeated measures.

It has been found that the RMSE calculated using Frequentist inference is located at the centre of the RMSE distribution obtained through Bayesian inference. This suggests that the average of the RMSE distribution obtained from Bayesian inference matches the RMSE calculated using Frequentist inference, which is consistent with our previous observations. This conclusion applies to all 49 datasets and 2 and 8 repeated measures. Therefore, we can use Bayesian inference to produce a distribution of possible ranges of values for the RMSE, which provides a more comprehensive evaluation of the out-of-sample performance.

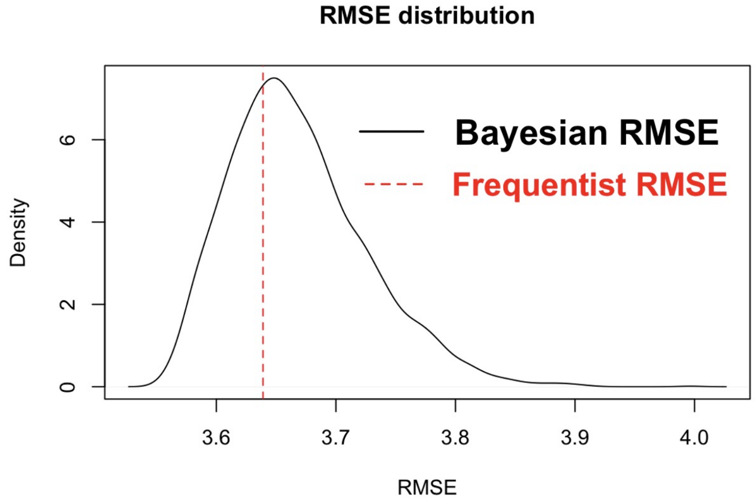


Figure 4.3.3: The plot shows a black curve representing the out-of-sample RMSE calculated using Bayesian inference, while the red dashed line displays the RMSE calculated using Frequentist inference. The RMSE value using Frequentist inference is located at the centre of the RMSE distribution, using one out of fifty of the resampling simulated data with four repeated measures.

Sensitivity Analysis of Informative Priors

The advantage of using the Bayesian approach is that it incorporates prior information into the inference. In this section, we aim to explore the impact of different informative priors on the posterior estimates of the analysis model compared to the Frequentist estimates. We want to assess the sensitivity of the posterior estimates to different scenarios of prior information. Since we use synthetic data, we know the data generating parameter values. Therefore, we will set three different scenarios of informative priors.

The first scenario involves a strong informative prior, with a normal distribution centered at the data-generating parameter value with a standard deviation of 0.2, reflecting strong prior information with less uncertainty around the data generating parameter value. The second scenario is a moderate informative prior, which is also centred around the generating parameter value but with a standard deviation of 3, allowing for larger uncertainty compared to the strong informative prior. Additionally, we will use an informative prior that is centred around a misleading value with a standard deviation of 3, similar to the moderately informative prior. The mean of the misleading values is randomly selected within ±2 standard deviations of the data generating parameter value. This will help us assess the robustness of Bayesian inference by examining the impact of using an informative prior based on an inaccurate belief about the data generating parameter value with considerable uncertainty.

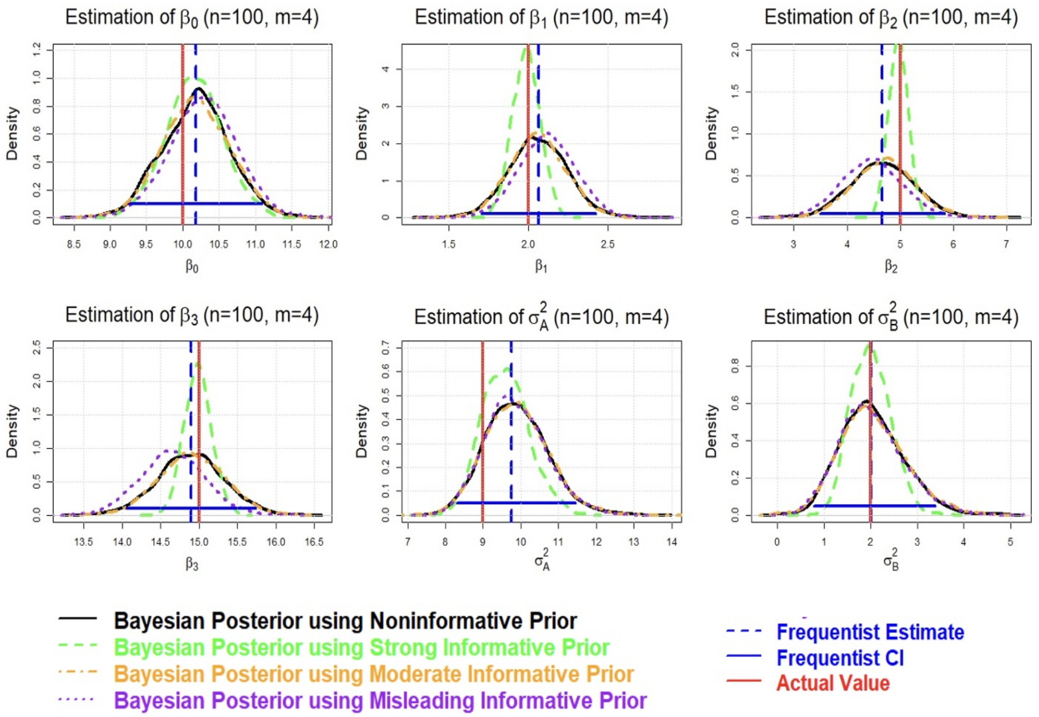


Figure 4.3.4: The black curve represents the posterior density obtained using the Bayesian approach with noninformative prior, the green curve represents the posterior density obtained using the Bayesian approach with strong informative prior, the orange curve represents the posterior density obtained using the Bayesian approach with moderate informative prior, and the purple curve represents the posterior density obtained using the Bayesian approach with misleading informative prior. There are three lines; a vertical blue dashed line representing the point estimate using the Frequentist approach, the blue horizontal line representing the 95% CI and a vertical red solid line representing the data generated value. Each plot represents one of the analysis model parameters that generated the simulated data with four repeated measures. The posterior distribution shifted to the data-generated parameter value with higher density when using a strong informative prior.

The results from Figure 4.3.4 shows that when using noninformative, moderately informative, and misleading informative priors, both the data generating parameter values and the Frequentist approach parameter estimates are located at the centre of the posterior distributions. In contrast, when a strongly informative prior is used, the centre of the posterior distribution is more skewed towards the data generating parameter values with higher density. This suggests that the Bayesian approach benefits from using a strong and correct informative prior; it positively affects the results by making the posterior noticeably more concentrated around the true value, leading to more accurate and confident parameter estimates. On the other hand, using weak or misleading informative priors may not yield the same precision as strong and correct informative prior, but the results still remain good and similar to those obtained through the Frequentist approach.

Including informative prior distribution had a negligible impact on the parameter estimates of the analysis model. We will now assess the performance on unseen data under different informative prior distribution scenarios. We will use the resampling process discussed in Section 4.3 to evaluate the level of uncertainty associated with each scenario.

Figure 4.3.5 displays the distribution of the RMSEs of the Bayesian approach across different informative prior scenarios and the Frequentist approach. The plot shows that the out-of-sample RMSE values are very similar, indicating that there is no significant difference in terms of out-of-sample performance between the Frequentist approach and the Bayesian approach with different informative prior distributions. This was evaluated using the Kolmogorov-Smirnov test (described in Section 2.7), which failed to reject the null hypothesis with a p-value greater than 0.05. This means there is insufficient evidence to conclude that there are differences between the RMSE distributions. These results are consistent across different numbers of repeated measures (2 & 8), leading to the same conclusion.

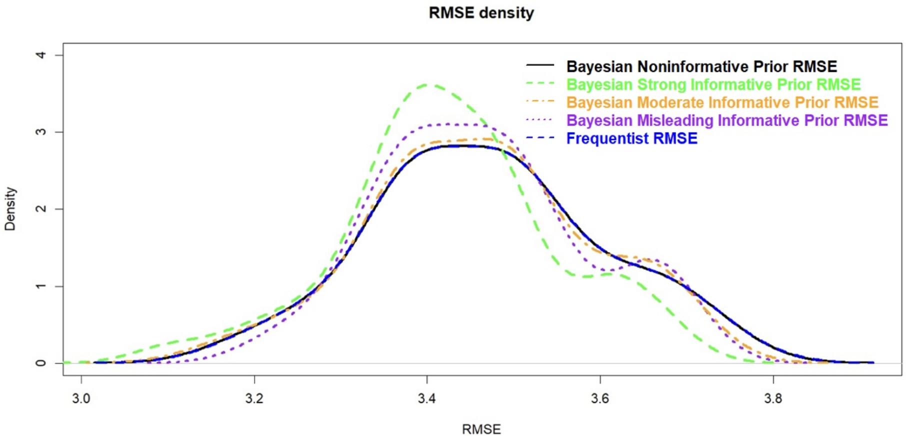


Figure 4.3.5: The plot displays different out-of-sample RMSE curves by applying the refitting process. The black curve represents the RMSE obtained using the Bayesian approach with a noninformative prior, the green curve represents the RMSE obtained using the Bayesian approach with a strong informative prior, the orange curve represents the RMSE obtained using the Bayesian approach with a moderate informative prior, and the purple curve represents the RMSE obtained using the Bayesian approach with a misleading informative prior. The blue curve shows the RMSE calculated using the Frequentist inference. Simulated data with four repeated measures were used to generate the curves. Interestingly, all approaches showed similar out-of-sample performance, indicating that there is no preference between them.

Real Data Results

We will use the BIOSTAT-CHF dataset to evaluate the performance of fitting a longitudinal data model using Bayesian and Frequentist approaches. This real-world dataset will give us an indication of how these two approaches behave in practice. We will use the model expressed in Equation 4.2.2 to fit the model. However, there is missing data in the BIOSTAT-CHF dataset, so we will use complete case analysis (where participants with missing data are eliminated, which means that only participants with fully observed data remain) to compare the two approaches at this stage. In the upcoming chapters, we will further analyse the BIOSTAT-CHF dataset to address the missing data using proposed methods. The results of these analyses will be presented in Section 4.4.7 of this chapter and in the results sections of Chapters 5, 6 and 7.

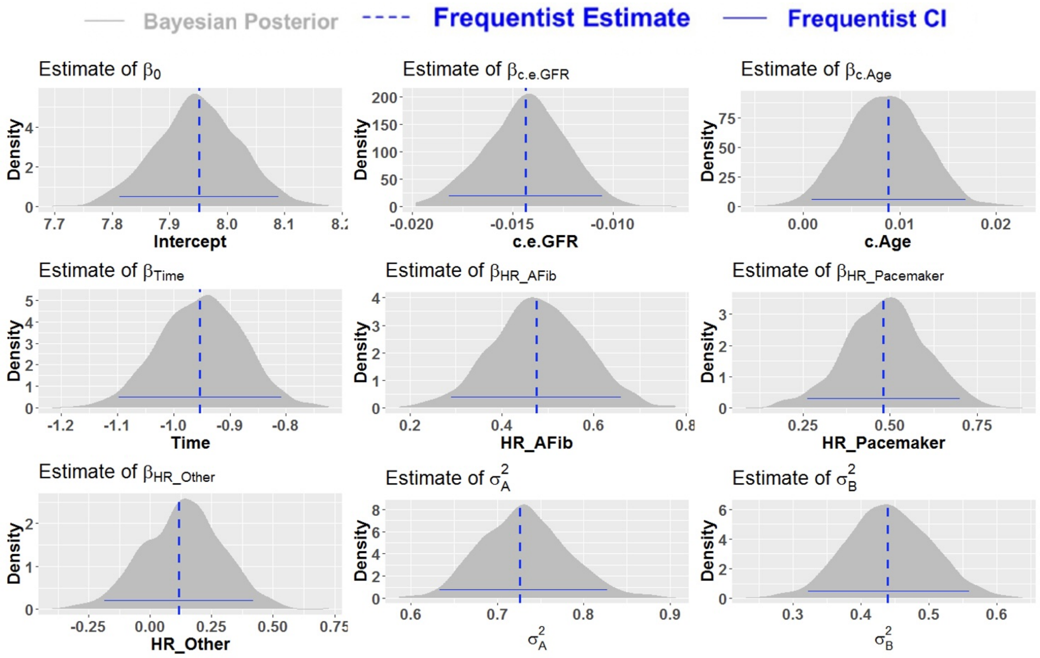


Figure 4.3.6: The plot shows two approaches of statistical inference for the BIOSTAT-CHF model parameters. The grey curve represents the posterior density curve based on the Bayesian approach, while the vertical blue dashed line represents the point estimate using the Frequentist approach, and the blue horizontal line represents the 95% CI. Each plot corresponds to a specific analysis model's parameter. The point estimates are located in the centre of the posterior distribution, indicating a good level of agreement between the two approaches.

Figure 4.3.6 presents the estimates obtained from fitting the BIOSTAT-CHF data model using Bayesian and Frequentist approaches. The figure shows a posterior density and Frequentist point estimates, with the point estimates located at the centre of the posterior distribution. Additionally, the 95% CI is located within the bulk of the posterior distribution. This indicates that the average of the posterior estimates is similar to the Frequentist point estimates for all BIOSTAT-CHF analysis model parameters.

These estimates reveal practical insights. For instance, we observe a negative relationship between the centred eGFR and log(NT-proBNP). This implies that for every one-unit ng/L increase in eGFR from its mean value, the log(NT-proBNP) will decrease on average by 0.01 units, holding all other variables constant. Similarly, the centred age shows a positive correlation with log(NT-proBNP). This suggests that for every one-year increase in age from its mean value, the log(NT-proBNP) will increase on average by 0.01 units, again holding all other variables constant.

In addition, patients with Sinus heart rhythm have the lowest rate of log(NT-proBNP) compared to those with other heart rhythm categories. Furthermore, patients' log(NT-proBNP) during their second visit will decrease on average by 0.95 units while all other variables are kept constant. The Interclass Correlation Coefficient (ICC) is approximately 0.64, which is the estimated correlation of two measurements on the same patient.

The resampling process described in Section 4.3 will be employed to assess the performance of two approaches using the BIOSTAT-CHF dataset. This process involves splitting the dataset into training and test data 50 times with replacement. Figure 4.3.7 displays the sampling distribution of the RMSEs obtained from both approaches using this process. The plots reveal that both approaches yield comparable out-of-sample RMSE densities. However, the Bayesian approach produces somewhat lower RMSE values than the Frequentist approach. Nevertheless, the KS test indicates no significant difference in out-of-sample performance between the two approaches.

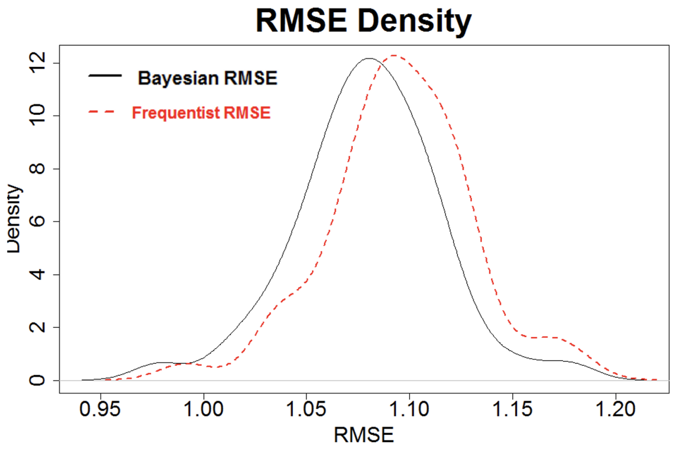


Figure 4.3.7: The plot displays two curves by applying the refitting process. The black curve represents the out-of-sample RMSE calculated using Bayesian inference, while the red dashed curve shows the RMSE calculated using Frequentist inference. The data used is the BIOSTAT-CHF dataset. Both approaches showed comparable out-of-sample performance, but the Bayesian RMSE values are shifted towards lower values, which is indicative of better out-of-sample prediction performance.

4.4 Application of the CRE Method

In the previous section of this chapter, we discussed fitting a linear mixed model to estimate the mean response based on a collection of predictor variables. This approach is commonly used in longitudinal studies to account for dependencies in the data. However, non-ignorable missing values can be a problem in this type of study for various reasons, and inference based on observed data may be biased. This section will discuss a recent approach introduced by Bhuyan (2019), which uses Correlated Random Effects (CRE) through a Gibbs sampler to handle non-ignorable missingness in the response.

The CRE is a generalisation of the SRE model framework mentioned in Section 2.5.3, mainly when the correlation value between the random effects in the CRE model is one. On the other hand, when the correlation between the random effects is zero (independent), this indicates that the missingness is ignorable. In the Correlated Random Effects model, the computational challenges arise due to the intractable numerical integration in the log-likelihood function as presented in Lin et al. (2010). To avoid approximation methods, Bhuyan (2019) suggested an alternative modelling approach using the Gibbs sampler, where the model parameters and latent variables are estimated at each iteration.

The following subsections are associated with the expression of the CRE model, the form of the joint distribution, the prior distribution, and the sampler algorithm. These elements represent the CRE method proposed by Bhuyan (2019). However, we have applied this method to a linear mixed model with fixed effects, including subject-specific random effects, using the linear regression parametric form. In contrast to the original study, they assumed a semi-parametric regression form using Legendre polynomials (LP) as basis functions. Finally, we will present the simulated and real-world data results and discuss the potential MCMC non-convergence issue using weakly informative prior distributions.

4.4.1 CRE Model

Mixed effects regression is a standard framework for studying the relationship between longitudinal response and predictor variables. For a continuous response measured over m different time points from n subjects and a set of predictors. The response for the i^{th} subject at the t^{th} time point, which we denote by $Y_i(t)$, can thus be modelled as the following:

$$\begin{gathered} Y_i(t) = \mu + \sum_{j'=1}^{J'} \lambda_{j'} X_{j'i}(t) + u_i \tilde{Z}_i(t) + e_i(t), \\ u_i \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_B^2\right) \;\; \& \;\; e_i(t) \overset{\text{i.i.d}}{\sim} N\left(0, \sigma_A^2\right). \end{gathered} \tag{4.4.1}$$

where J' express the number of predictors of fixed effects, μ is the fixed intercept represents the mean of the overall population, $\lambda_{j'}$ is the regression coefficient associated with the j'^{th} fixed effects, $X_{j'i}(t)$ is the value of the j'^{th} fixed effect for subject i at time t. Subject-specific random effects u_i capture the longitudinal dependence, and $\tilde{Z}_i(t)$ is the value of the random effect for subject i at time t. The model's residuals are expressed as $e_i(t)$. Next we will define binary missing data indicator $U_i(t)$, where $U_i(t) = 0$ if $Y_i(t)$ is missing and $U_i(t) = 1$ if $Y_i(t)$ is observed. The latent response variable can be written as:

$$Y_i(t) = \begin{cases} Y_i^*(t), & \text{if } U_i(t) = 1, \\ missing, & \text{if } U_i(t) = 0. \end{cases} \tag{4.4.2}$$

The regression model given in Equation 4.4.1 can be rewritten as follows:

$$Y_i^*(t) = \mu + \sum_{j'=1}^{J'} \lambda_{j'} X_{j'i}(t) + u_i \tilde{Z}_i(t) + e_i(t). \tag{4.4.3}$$

In addition, we consider a probit regression model (discussed in Section 2.6) that defines the missingness as a normal distribution latent variable $U_i^*(t)$ as follows:

$$U_i(t) = \begin{cases} 1, & \text{if } U_i^*(t) > 0, \\ 0, & \text{if } U_i^*(t) \leq 0. \end{cases} \tag{4.4.4}$$

In Equation 4.4.4, $U_i^*(t)$ is a continuous latent missingness indicator variable for subject i at time t that represents the latent proclivity for missing data. Then we consider the following model for missing response mechanism with the same set of predictors in the response model as follows:

$$U_i^*(t) = \tau + \sum_{j'=1}^{J'} \theta_{j'} X_{j'i}(t) + v_i \tilde{Z}_i(t) + \varepsilon_i(t), \tag{4.4.5}$$

where J' express the number of fixed effects, τ is the fixed intercept represents the mean of the overall population, $\theta_{j'}$ is the regression coefficient associated with the j'^{th} fixed effects, which expresses the systematic influence of missingness due to the unobserved response variables. Subject-specific random effects v_i capture the longitudinal dependence and are assumed to be i.i.d following normal distribution as $N(0, \sigma_C^2)$. The residuals $\varepsilon_i(t)$ are assumed to be i.i.d following normal distribution as $N(0,1)$. Therefore, the probit regression can be written as:

$$p(U_i(t)=0|Y_i(t)) = 1-\Phi\left(\tau+\sum_{j'=1}^{J'}\theta_{j'}X_{j'i}(t)+v_i\tilde{Z}_i(t)\right), \qquad (4.4.6)$$

where $\Phi()$ is the standard normal cumulative distribution function, and the predicted z-score of missing propensity is $\left(\tau+\sum_{j'=1}^{J'}\theta_{j'}(t)X_{ji}(t)+v_i\tilde{Z}_i(t)\right)$. The Φ will return the proportion of the area under that z-score in a standard normal density. Furthermore, the probit model includes a zero value as a threshold, which divides the standard normal into two parts. So that if $U^*(t)$ greater than zero then $U_i(t)=1$ and if $U^*(t)$ less than zero then $U_i(t)=0$.

To incorporate the possible correlation between the response variable $Y_i^*(t)$ and the response missing indicator variable $U_i^*(t)$, we consider u_i and v_i are correlated random vectors following a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = \begin{pmatrix}\sigma_B^2 & \sigma_D^2\\ \sigma_D^2 & \sigma_C^2\end{pmatrix}$, where σ_D^2 represents the covariance between u_i and v_i random effects.

4.4.2 CRE Joint Distribution

The Bayesian approach is used to estimate the model parameters in Equation 4.4.3 and in Equation 4.4.5 using an iterative MCMC algorithm, Gibbs sampling, to simultaneously impute missing values and produce analysis estimates. For Gibbs sampling to be carried out, one needs to sample from the joint posterior of the model parameters and latent variables. Let $\boldsymbol{Y}=(Y_{11}(t),\ldots,Y_{nm}(t))$, $\boldsymbol{Y}^*=\left(Y_{11}^*(t),\ldots,Y_{nm}^*(t)\right)$, $\boldsymbol{U}=(U_{11}(t),\ldots,U_{nm}(t))$ and $\boldsymbol{U}^*=\left(U_{11}^*(t),\ldots,U_{nm}^*(t)\right)$, along with the joint posterior is expressed as follows:

$$\begin{aligned} p\left(\Theta_{Y,U},\boldsymbol{Y}^*,\boldsymbol{U}^* \mid \boldsymbol{Y},\boldsymbol{U}\right) \propto p\left(\Theta_{Y,U}\right)\times\prod_{i=1}^{n}\int\prod_{t=1}^{m} f\left(Y_i^*(t),U_i^*(t)\mid u_i,v_i\right)\times \\ \left\{I\left(U_i^*(t)>0\right)I\left(U_i(t)=1\right)+I\left(U_i^*(t)\le 0\right)I\left(U_i(t)=0\right)\right\}\times g\left(u_i,v_i\right)du_idv_i, \end{aligned} \qquad (4.4.7)$$

where $\Theta_{Y,U} = \{\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\theta}}, \sigma_A^2, \Sigma\}$ denote a set of all analysis model parameters involved in Equation 4.4.3 and Equation 4.4.5. $\tilde{\boldsymbol{\lambda}} = [\mu, \boldsymbol{\lambda}]$ denotes a vector of the overall mean and regression coefficients of fixed effects in the response model and $\tilde{\boldsymbol{\theta}} = [\tau, \boldsymbol{\theta}]$ denotes a vector of regression coefficients of fixed effects and overall mean in the missingness indicator response model. The joint prior distribution for Θ is represented by $p\left(\Theta_{Y,U}\right)$, $g\left(u_i, v_i\right)$ is the joint distribution of u_i and v_i follow $N(0,\Sigma)$, $f\left(Y_i^*(t), U_i^*(t)\right)$ is the joint distribution, and $I(A)$ is an indicator variable which takes the value 1 if A occurs and zero otherwise.

To derive the posterior distributions of the joint model parameter $\Theta_{Y,U}$, we need to define the prior distribution, which explains the information of each parameter uncertainty before seeing the data $p\left(\Theta_{Y,U}\right)$. It is an important part of Bayesian statistics to derive the posterior distribution.

4.4.3 CRE Prior Distribution

The following non-informative prior is considered for $\Theta_{Y,U}$ as specified by Bhuyan (2019).

$$p\left(\tilde{\boldsymbol{\lambda}}, \sigma_A^2\right) \propto \frac{1}{\sigma_A^2}; \;\; p\left(\tilde{\boldsymbol{\theta}}\right) \propto 1; \;\; and \;\; p(\Sigma) \propto \frac{1}{|\,\Sigma\,|}. \tag{4.4.8}$$

The non-informative priors shown in Equation 4.4.8 are used because there is no prior information about these parameters and to minimise the influence of the prior on the parameters estimate. Furthermore, these priors are conjugate priors with normally distributed data.

The prior distribution is multiplied by the probability distribution of the data to produce a posterior distribution. Bayesian inference is based on the posterior distribution, which sometimes is not found in a closed form. In such cases, Markov Chain Monte Carlo (MCMC) can be implemented as a numerical simulation method to generate a random set of points from the parameter space drawn from the posterior distribution to estimate the distribution of the parameters. Then, summary statistics can be computed from it. Using conjugate priors, the full conditional distribution for each parameter can be found in closed form, allowing the Gibbs sampler to sample from the joint posterior distribution. This process is explained in Section 4.4.4.

4.4.4 CRE Gibbs Sampler

The Gibbs sampler is an iterative procedure that estimates the variables sequentially one by one. For example, estimate one random variable while holding all other variables on their current values (constant). In the Bayesian approach, the response model parameters, the response missing indicator model parameters, the response, and the missing response indicator are variables to be estimated. The Gibbs sampler invokes the following steps:

1. Estimate the latent response model's random intercept u_i.
2. Estimate the latent response variable $Y_i^*(t)$.
3. Estimate the latent response model's regression coefficients and residual variance $\tilde{\boldsymbol{\lambda}}$ & σ_A^2.
4. Estimate the latent response missingness model's random intercept v_i.
5. Estimate the latent response missingness indicator $U_i^*(t)$.
6. Estimate the latent response missingness model's regression coefficients $\tilde{\boldsymbol{\theta}}$.
7. Estimate the covariance matrix that represents the correlation between the random intercept in the response and missing model Σ.

Steps 1 to 7 are repeated until the MCMC chains converge and produce enough posterior samples. The convergence is assessed visually using the trace plot and statistically using the Gelman-Rubin diagnostic, as mentioned in Section 2.4.6. The Gibbs sampler produces a posterior distribution for each variable, which is used to conduct Bayesian inference. The Gibbs sampler algorithm is described in Algorithm 4, where $\boldsymbol{X}_{J'}$ is the design matrix consisting of fixed effect variables. The full conditional densities of the model's parameters are from standard densities. Accordingly, the Gibbs sample can be directly applied to estimate the model parameters. It is essential to assess the convergence to make sure that the MCMC chain will converge to the stationary distribution, which is the posterior distribution.

Algorithm 4 CRE Method Gibbs Sampling Algorithm

```
Choose initial $\{\Theta^0_{Y,U} = \tilde{\boldsymbol{\lambda}}^0, \tilde{\boldsymbol{\theta}}^0, \sigma_A^{2^0}, \Sigma^0\}$ and $\{u^0, v^0, Y^{*0}, U^{*0}\}$.

for $1, \ldots, S$ iterations do

    -Sample $u_i^{S+1} \sim p\left(u_i \mid Y_i^{*^S}(t), \tilde{\boldsymbol{\lambda}}^S, \sigma_A^{2^S}, \Sigma^S, v_i^S, \boldsymbol{X}_{J'}\right)$

    -Sample $Y_i^{*^{S+1}}(t) = \begin{cases} Y_i^{*^S}(t), & \text{if } Y_i(t) \textit{ is observed} \\ p\left(Y_i^*(t) \mid u_i^{S+1}, \tilde{\boldsymbol{\lambda}}^S, \sigma_A^{2^S}, \boldsymbol{X}_{J'}\right), & \text{if } Y_i(t) \textit{ is missing.} \end{cases}$

    -Sample $\sigma_A^{2^{S+1}} \sim p\left(\sigma_A^2 \mid Y_i^{*^{S+1}}(t), \tilde{\boldsymbol{\lambda}}^S, u_i^{S+1}, \boldsymbol{X}_{J'}\right)$.

    -Sample$\tilde{\lambda}^{S+1} \sim p\left(\tilde{\boldsymbol{\lambda}} \mid Y_i^{*^{S+1}}(t), u_i^{S+1}, \sigma_A^{2^{S+1}}, \boldsymbol{X}_{J'}\right)$.

    -Sample $v_i^{S+1} \sim p\left(v_i \mid U_i^{*^S}(t), \tilde{\boldsymbol{\theta}}^S, \Sigma^S, u_i^{S+1}, \boldsymbol{X}_{J'}\right)$.

    -Sample $U_i^{*^S}(t) = \begin{cases} p\left(U_i^*(t) \mid \tilde{\boldsymbol{\theta}}^S, v_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ left truncated}^* \text{ at } 0, & \text{if } Y_i(t) \textit{ is observed}, \\ p\left(U_i^*(t) \mid \tilde{\boldsymbol{\theta}}^S, v_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ right truncated}^* \text{ at } 0, & \text{if } Y_i(t) \textit{ is missing}. \end{cases}$

    -Sample $\tilde{\theta}^{S+1} \sim p\left(\tilde{\boldsymbol{\theta}} \mid U_i^{*^{S+1}}(t), v_i^{S+1}, \boldsymbol{X}_{J'}\right)$

    -Sample $\Sigma^{S+1} \sim p\left(\Sigma \mid u_i^{S+1}, v_i^{S+1}\right)$

* truncated normal distribution.
```

The full conditional distribution for $(u_i, v_i, Y_i^*(t), U_i^*(t), \tilde{\lambda}, \tilde{\theta})$ are the normal distribution and for σ_A^2 and Σ are the inverse gamma distribution and the inverse-Wishart distribution, respectively.

4.4.5 Non-Convergence in CRE Method

A drawback of the CRE method is that the covariance matrix parameters may struggle to converge. After further investigation, it was found that this could be due to a vague prior (we will refer to the $\frac{1}{|\Sigma|}$ as a vague prior) was set for the covariance matrix as mentioned in Section 4.4.3. We fixed this problem by considering a weakly informative prior. By "weakly informative prior", we mean that we can incorporate information into the prior distribution, such as the values of the hyperparameters, that gives little guidance about the expected values of the parameters. For example, a normal distribution prior with a large standard deviation (such as 100) allows for a wide range of expected values. This differs from a "vague prior", which provides minimal (non-informative) guidance about the possible values of the parameters and allows the data to guide the inference process. For instance, a uniform distribution prior with the interval $(0,1)$ implies that all possible values are equally likely. We will use the Inverse Wishart distribution as a prior distribution for the covariance matrix, it has the advantage of simplifying the posterior distribution because it's a conjugate prior with normally distributed data (Alvarez et al., 2014; Huang and Wand, 2013). The density function of the Inverse Wishart (IW) distribution $p(\Sigma) \sim IW(\nu, \Lambda)$ is:

$$p(\Sigma) = \frac{|\Lambda|^{\frac{\nu}{n}}\ |\Sigma|^{\frac{-(\nu+p+1)}{2}}\ e^{\frac{-tr(V\Sigma^{-1})}{2}}}{2^{\frac{\nu p}{2}}\Gamma\frac{\nu}{2}}, \tag{4.4.9}$$

where p is the dimension of the covariance matrix, Λ is $p \times p$ scale matrix, which, in practice, is an identity matrix for non-informative priors and ν is the degrees of freedom and meets the constraint that $\nu > p+1$ so that the mean exists. Because the mean of the IW is defined as:

$$E[\Sigma] = \frac{\Lambda}{\nu - p - 1}. \tag{4.4.10}$$

As the degrees of freedom ν increase, the posterior mean shifts towards the prior mean. This is because the prior mean will have a larger weight, as the posterior mean is a weighted mean of the prior and the sample mean (Zhang, 2021), leading to greater certainty about the information in Λ and making the prior more informative. Here, Λ represents the position of the inverse Wishart distribution used in the parameter space. As the element value of Λ decreases, the variance of IW distribution decreases and might result in overestimating or underestimating the variances (Schuurman et al., 2016). On the other hand, setting larger values of elements in ν can affect the position of the parameter space. So, the specification of Λ and ν should be balanced (Schuurman et al., 2016). The commonly used less informative IW prior is with small ν and the identity matrix Λ (Schuurman et al., 2016). Since the prior mean is affected by the value of the degrees of freedom, it is suggestible to fix the prior mean by setting the scale matrix $\Lambda = (\nu - p - 1)I$. Accordingly, the prior mean would be fixed and equal to I regardless of the degrees of freedom value choice.

$$E(\Sigma) = \frac{(\nu - p - 1)I}{\nu - p - 1} = I. \tag{4.4.11}$$

In our study, we will set $\nu = 4$, which satisfies the constrain $\nu > p + 1$ and $\Lambda = (\nu - p - 1)I$, which solved the non-convergence and reduces the influence of the prior on the posterior distribution.

4.4.6 Creating Synthetic Data for Simulation

The performance of the CRE method using a non-informative prior and weakly informative prior over the covariance matrix parameters, as discussed in the previous sections, is examined using a simulated study. We will use the simulation data mentioned in Section 3.4. In order to generate missing values in the model response, assuming its missingness mechanism is MNAR, we will generate the missing values on the response Y based on the response missingness model as follows:

$$U_i^*(t) = \theta_0 + \theta_1 X_{1i}(t) + \theta_2 X_{2i}(t) + \theta_3 X_{3i}(t) + v_i \tilde{Z}_i(t) + \varepsilon_i(t), \tag{4.4.12}$$

where $\boldsymbol{\theta} = \{\theta_0, \theta_1, \theta_2, \theta_3\}$ is a vector of the regression coefficients associated with the fixed effects. The values of $\boldsymbol{\theta} = \{-0.8, -0.4, 3, 4\}$ were chosen to produce 20% missing data proportion in the response, using a probit regression equation to connect missingness probabilities of the response Y to values of Y through the latent missingness indicator regression model U^* for non-ignorable missingness. This was discussed in Section 2.6. The missing data indicator U^* for each observation is sampled from the binomial distribution with a success rate equal to the observation's missingness probability from the probit model, where the value is one if the corresponding Y is observed and zero if missing. Moreover, $v_i \overset{\text{i.i.d}}{\sim} N(0,2)$ and the residuals $\varepsilon_i(t) \overset{\text{i.i.d}}{\sim} N(0,1)$. The covariance matrix associated with the random effects is $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$. The data is simulated for 100 participants and four repeated measures.

4.4.7 Results

This section is divided into two parts. In the first part, we will evaluate the performance of adopting the weakly informative prior over the covariance matrix parameter against the non-informative prior in the CRE method using the simulated data. This enables us to test how the MCMC converge using these two priors. In the second part, we will apply the CRE method to the BIOSTAT-CHF dataset.

Simulated Data Results

For the purpose of illustration, we will show the results of one generated dataset with 20% missingness in the model response with four repeated measures. However, the results are consistent across other missingness percentages as well. In the following chapters, the performance of the CRE method will be assessed using simulated data with varying repeated measures and the proportion of missingness.

To demonstrate the convergence performance of the CRE method, we show the trace and density plots of the response model parameters, missingness response model parameters, and the covariance matrix parameters in Figure 4.4.1 using the vague prior as $p(\Sigma) \propto \frac{1}{|\Sigma|}$. We ran three chains, each with different starting values and ran the MCMC for $50,000$ iterations; the first half of the samples are discarded to burn-in, and thinning is carried out by taking every 10^{th} sample. The Gelman-Rubin diagnostic factor, which is explained in Section 2.4.6, was greater than 1.1, indicating that the chain did not achieve convergence. Even with a larger number of iterations of $100,000$, the MCMC chain did not show any convergence; the corresponding plots are shown in Section A in the Appendix. Thus, samples are not from the posterior distributions.

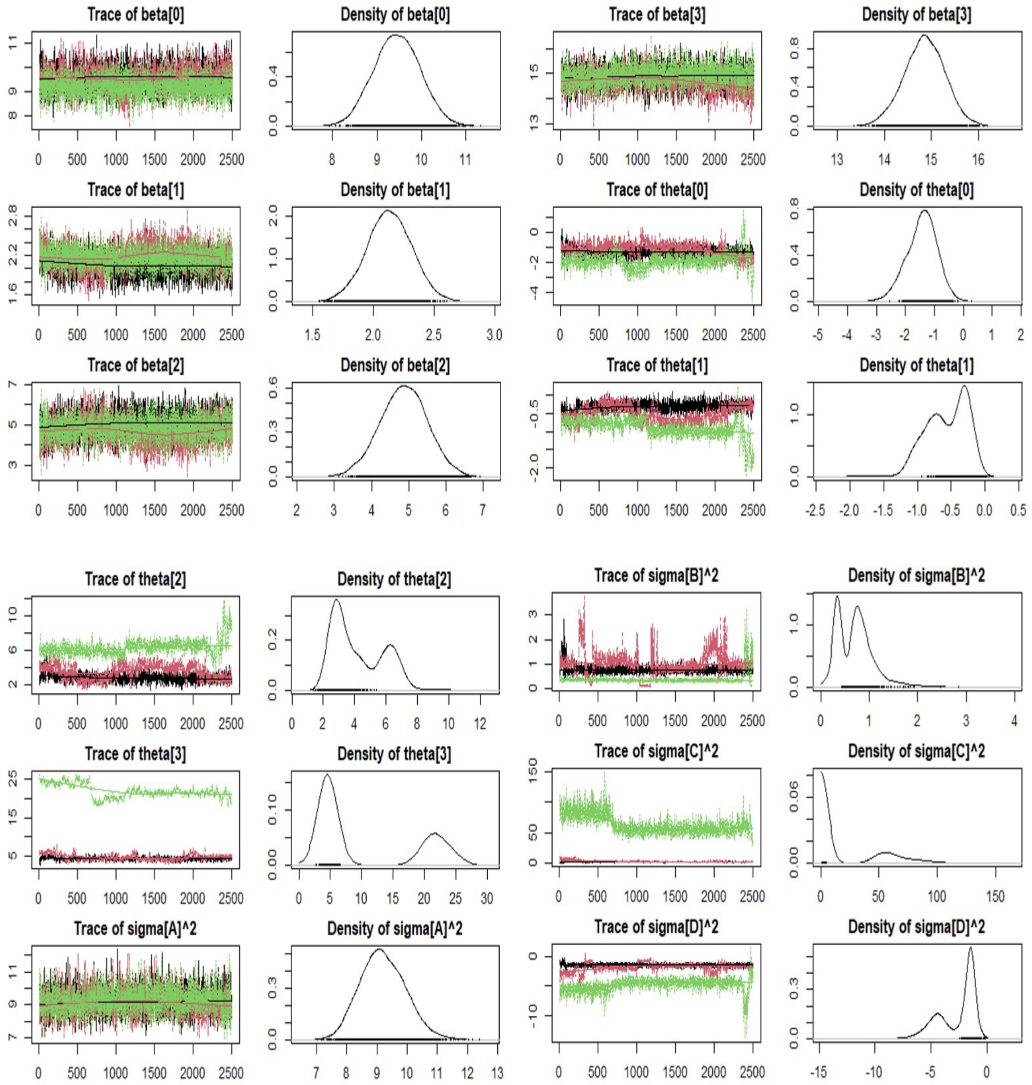


Figure 4.4.1: Trace and density plots of the response model, missingness response model and covariance matrix parameters posterior distributions, using the CRE method for MNAR response with the vague prior for the covariance matrix as $p(\Sigma) \propto \frac{1}{|\Sigma|}$. The parameters for the missingness response model ($theta[0]$, $theta[1]$, $theta[2]$ and $theta[3]$) and the covariance matrix parameters ($sigma[B]^2$, $sigma[C]^2$ and $sigma[D]^2$) did not converge.

To overcome this problem, a weakly informative prior as $p(\Sigma) \propto IW(\nu, \Lambda)$ was used for the covariance matrix in the model as explained in Section 4.4.5. The posterior distribution of the response model parameters, missingness response model parameters, and the covariance matrix parameters are expressed in Figure 4.4.2. Three chains were created using different initial values and run for 50,000 iterations. The first half of each chain was discarded to burn in, and every 10^{th} sample was taken. The Gelman-Rubin diagnostic factor was found to be less than or equal to 1.1. This indicates that the chains have converged, and samples are drawn from the posterior distributions.

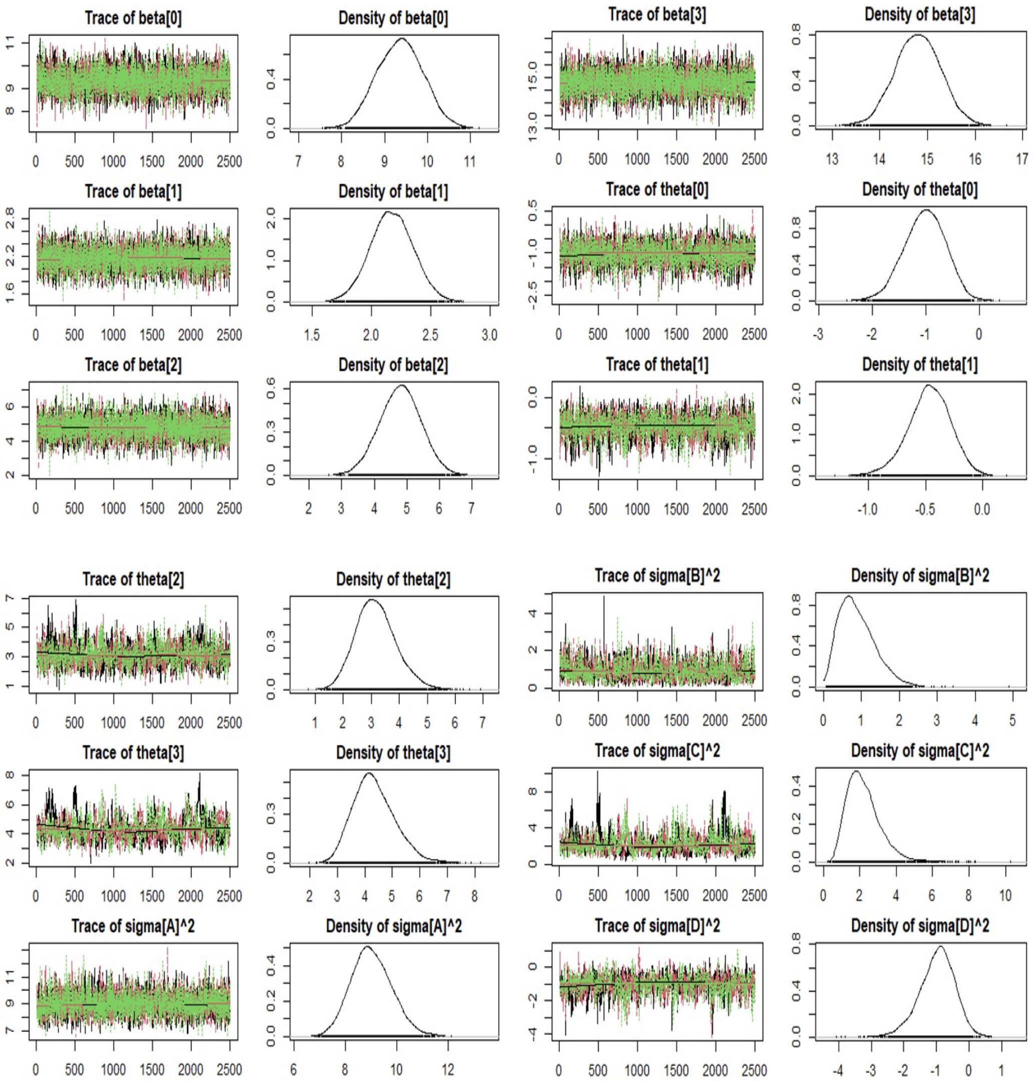


Figure 4.4.2: Trace and density plots of the response model, missingness response model and covariance matrix parameters posterior distributions, using the CRE method for MNAR response with the IW distribution as a weakly prior for the covariance matrix as $p(\Sigma) \propto IW(\nu, \Lambda)$. The trace plots stabilize around a central value without trends indicating convergence.

Real Data Results

In this section, we will apply the CRE method to the BIOSTAT-CHF dataset. The CRE method is specifically designed to deal with non-ignorable missingness in the response variable in longitudinal data. Since the CRE method can only handle missingness in the response variable, we will first filter the data to exclude missing observations in the incomplete predictor variable, which is the eGFR, before applying this method. This step ensures that the data aligns with the assumptions of the CRE method. We will use the model described in Equation 4.2.2.

Additionally, we will compare the results obtained from the CRE method with the baseline method, which uses the observed data and does not handle the missing values. This will be applied using the Bayesian hierarchical model with HMC, explained in Section 2.4.5 through the `brm` function in the `brms` package (Bürkner, 2017) in `R` (R Core Team et al., 2013).

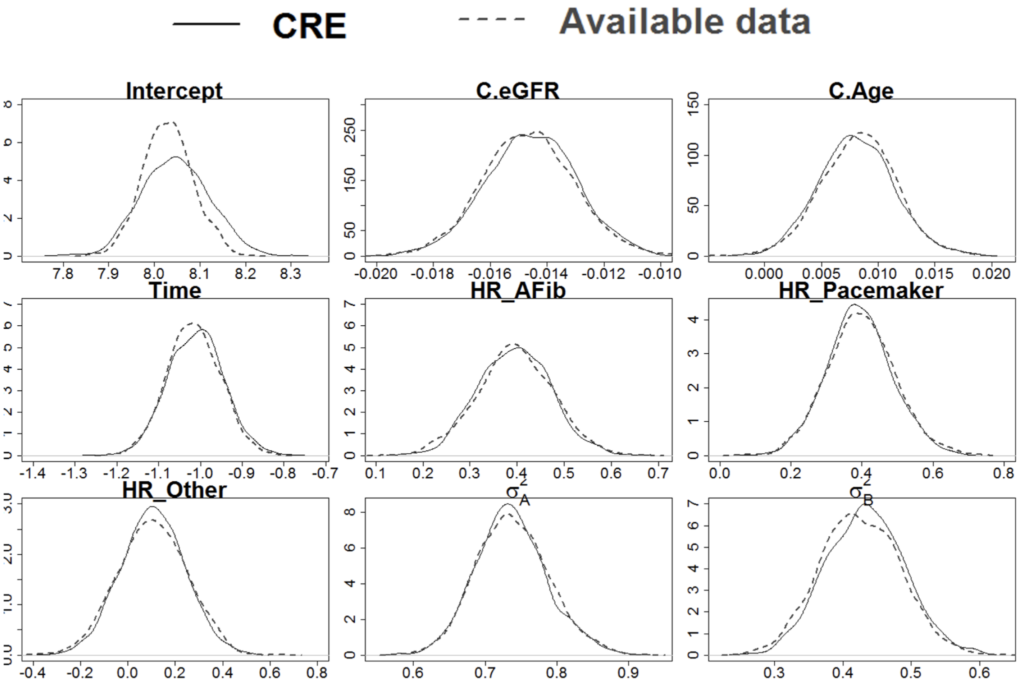


Figure 4.4.3: The posterior distribution using the available data is represented by a grey dashed curve, while the posterior distribution using the CRE method is represented by a black curve using the BIOSTAT-CHF dataset. Both posterior distributions are overlaid.

The BIOSTAT-CHF analysis model parameters posterior distribution estimates of the CRE method, and the available data exhibit a similar distribution, as shown in Figure 4.4.3. This similarity suggests that both methods have comparable estimates. However, the posterior density curve of the intercept using the available data model has a higher density, and this difference is significant based on the KS test. Additionally, there was a significant difference between the CRE method and the available data for time, between individual variance σ_B^2, centred eGFR and age coefficients based on KS test (explained in 2.7). The BIOSTAT-CHF model estimates are similar to those in section 4.3.1. The centred eGFR and the second time visit have a negative relationship with the log(NT-proBNP), while the centred age positively correlates with the log(NT-proBNP). Sinus Rhythm has a lower log(NT-proBNP) than other heart rhythm types, and the ICC is the same.

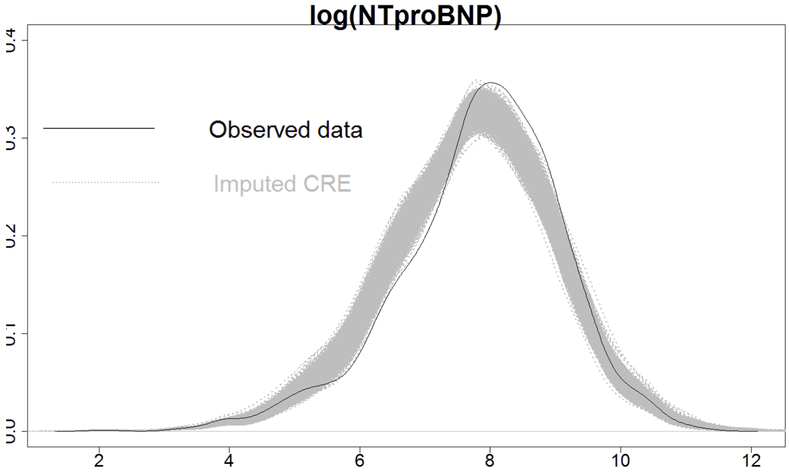


Figure 4.4.4: Density plot show the response observed values in the black solid curve and the CRE imputed response values in the grey dashed curves, where each curve is a different draw of the latent variable from the posterior using the Gibbs sampler. The density of the imputed values using the CRE method at each Gibbs sampler iteration is similar to the density of the observed values.

The CRE method has the advantage of estimating the probability of missingness in the model response being MNAR via σ_D^2. In the BIOSTAT-CHF dataset, this value equals -0.38, indicating a weak/moderate indication of MNAR in the log(NT-proBNP). Furthermore, the negative estimate of σ_D^2

suggests that patients with higher log(NT-proBNP) are more likely to have missing outcomes. The second visit and the centred eGFR are positively associated with the missingness of the log(NT-proBNP).

Another advantage of using the CRE method is that it can impute missing data in the response variable. In Figure 4.4.4, we can observe the log(NT-proBNP) in the BIOSTAT-CHF dataset with the use of draws of the latent variable from the posterior using the Gibbs sampler. The imputation process was successful as the observed density and density from each iteration were similar.

4.5 Discussion

This chapter acts as a foundation for motivating investigations in upcoming chapters. We began the chapter by setting up a model for the provided real data, the BIOSTAT-CHF data, which we will use throughout this thesis. Our focus in this chapter was on fitting longitudinal data and finding challenges researchers might face. The longitudinal data can be fit using two popular statistics approaches: the Frequentist and the Bayesian approaches. Our first goal was to determine which approach would yield better parameter estimates and how each approach behaves when used to predict unseen data to reflect real-world scenarios. We utilised resampling techniques to ensure a fair comparison between the two approaches. Our second goal was to handle the missingness that is common in such studies. In order to achieve this, we presented the CRE method from the literature to handle MNAR in the response variable. We also discussed the non-convergence issue that we encountered. We used simulated data, where we already knew the actual data generating parameter values, as well as real-world data using the BIOSTAT-CHF dataset.

In the Bayesian approach, we used the HMC utilized by the built-in function in `R`, the `brm` function Bürkner (2017). The HMC tends to converge to locations of higher posterior density faster than the Metropolis Algorithm because of the posterior gradient functions (Porter and Carré, 2014; Thomas and Tu, 2021). Hamiltonian dynamics allows a chain to move through energy trajectories with minimal computational cost, thereby decreasing correlations in the chain. Additionally, the HMC efficiency remains unchanged despite dealing with large dimensions; shorter chains are needed than MCMC (Hajian, 2007). Using the same dataset and number of iterations, the HMC produced the results within 2.1 minutes, whereas the Metropolis Algorithm required 7.5 hours in our application. Additionally, the Metropolis algorithm requires regular checks of the acceptance rate and adjusts the step size accordingly (Porter and Carré, 2014). Thus, in this chapter and the upcoming chapters, we will use the HMC algorithm for the baseline models to perform Bayesian inference.

To build a model for the BIOSTAT-CHF dataset, which is a study designed for heart failure, we used specific predictors that are known to affect the NT-proBNP and are commonly used in heart failure literature. The study conducted by Voors et al. (2017) also used this dataset to test the risk of mortality, HF hospitalisation, or a combination of both for patients with HF using Cox regression. We replicated their results (these Cox regression results are omitted from the thesis for brevity) and also found that NT-proBNP has a large effect on the event of interest. Additionally, this variable had a lot of missingness in the BIOSTAT-CHF study. This makes it a valuable variable to investigate for improving methods of handling missing values in a biostatistical context. We considered this finding while proposing the model, given that we are dealing with a continuous response. Although other variables in the dataset could be used as predictors, for the purpose of testing the proposed approaches in this chapter and upcoming chapters, we decided to use a simple model based on data exploration in Chapter 3 (e.g., Figure 3.3.4). This resulted in a suitable model to test the upcoming methods and gain a better understanding of their behaviour. However, in the future, it will be possible to scale up to larger models.

We used a linear mixed effect model in this chapter and throughout the thesis. We evaluated the performance of the Bayesian and Frequentist approaches using simulated data. We obtained similar results using both approaches, regardless of the incorporation of informative prior distribution in Bayesian inference. This implies that the data has a substantial impact on the inference. The estimates were close to the data-generating parameter values. Furthermore, using the BIOSTAT-CHF dataset, the Frequentist estimates aligned with the average of the posterior distribution and comparable out-of-sample prediction. However, the Bayesian approach resulted in lower RMSE values than the Frequentist approach, using the resampling technique for comparison. The advantage of Bayesian inference is that it can produce a distribution of parameter estimates. This can help decision-makers or healthcare researchers understand the possible values of a parameter by identifying the values with a high density and providing information about a more plausible range of values anywhere within the interval. Another advantage of Bayesian modelling is that it can include data from other sources and incorporate informative prior information (White et al., 2007).

For each simulation scenario, we conducted 100 repetitions, which were constrained by computational time. To assess the simulation uncertainty in our results, we used the Monte Carlo standard errors (MCSE) for the bias estimate to evaluate the precision of the parameter estimates from our analysis model (Morris et al., 2019). A smaller MCSE indicates that the number of simulated data repetitions is sufficient to obtain reasonably precise estimates. Overall, the Monte Carlo standard errors ranged from a minimum of 0.01 to a maximum of 0.07 for simulated data with four and eight repeated measures. When only two repeated measures were used, this range increased to a minimum of 0.02 and a maximum of 0.11. These values are not relatively low compared with the desired values in the literature (Cro et al., 2024; Mokkink et al., 2023; Seide et al., 2020). Nevertheless, they balanced the level of precision and computational time. The results showed unbiased estimates and good out-of-sample performance with 100 repetitions. This analysis serves as the foundation for the remainder of the thesis, where we address similar trade-offs to ensure our analyses are both accurate and practical.

There were missing values in the BIOSTAT-CHF variables, so we used a complete case analysis. We then applied a recent CRE method designed to handle non-ignorable missingness in longitudinal model responses. This method was introduced by Bhuyan (2019). However, our application differed from theirs in that we used a linear parametric mixed model, whereas they used a semiparametric model. We tested this approach using simulated data and introduced a weakly informative prior to overcome non-convergence issues. We applied the method to the BIOSTAT-CHF data with missingness only in the response, as the method was designed to handle. We also compared the CRE method with the baseline model that uses only available data. The posterior estimates were similar, but the CRE method had the advantage of imputing missing values of the response through Gibbs sampler iterations, which showed similar density to the observed values.

Additionally, the CRE method indicated weak to moderate MNAR in the log(NT-proBNP). As the log(NT-proBNP) increased, it was more likely to be missing, which is obvious since high values of log(NT-proBNP) indicate unhealthy patients who are more likely to have HF. Observations from the second visit were also more likely to be missing, which is common in repeated measures studies where patients are lost to follow-ups. As eGFR decreased, the outcome was more likely to be missing, which makes sense because low eGFR indicates unhealthy kidney function and may affect the follow-ups in the study.

The results of the BIOSTAT-CHF model show consistent findings using the CRE method, available data (with only missing values in the response), and complete case analysis, where participants with missing values are excluded, leaving only those with fully observed data. This leads to the conclusion that the centred eGFR and the second time visit have a negative correlation with the log(NT-proBNP), while the centred age has a positive correlation with the log(NT-proBNP). Additionally, patients in sinus rhythm have lower log(NT-proBNP) levels compared to those with other heart rhythm types.

The BIOSTAT-CHF model has missing values in the model response and predictor eGFR, thus requiring the CRE method to handle missingness in both. In this chapter, when the CRE method was applied, any missing value in the predictors caused that row of data to be omitted from the model. This is undesirable since it reduces the dataset size, causing a loss of information. The upcoming chapters will address this issue and explain how the CRE method can be adapted accordingly. Furthermore, the CRE method will be applied to various simulated data settings and compared with the generalised version of the method that is developed in Chapter 5, Chapter 6 and Chapter 7. These chapters will present the results of the CRE method.

Chapter 5

Proposed Two-Step Method

Some parts of this chapter have been submitted to and accepted for publication in the 6th International Conference on Statistics: Theory and Applications 2024 (ICSTA'24) (Alzahrani et al., 2024).

5.1 Introduction

The Correlated Random Effects (CRE) method has been discussed in Chapter 4, and one of the method's drawbacks is that it cannot accommodate missingness in the model predictors in addition to the MNAR in the response. In longitudinal studies, it's common to have missing data in both the predictor and response variables due to participant dropout. The missingness in the model predictors can appear in different studies, for example, clinical trials, epidemiological studies and surveys (Tang and Zhao, 2014), it's common to have missing predictors in medical record data (Hsu et al., 2023) specifically when the predictor variable is measured over time. To overcome this issue, we propose a Two-Step method to deal with missing values in the response and predictors, assuming the model response has non-ignorable missingness and ignorable missingness in predictors.

Multiple imputation (Rubin and Schenker, 1986) is a popular method for handling missing data. It has become widely used recently due to its availability in development software, making it accessible to analysts (Hayati Rezvan et al., 2015). There is extensive literature on multiple imputation methods. Rubin (2004) demonstrates how multiple imputations can be used to handle nonresponse in sample surveys and censuses. Graham et al. (2009) discuss different MI methods for a normal model. Lee and Carlin (2010) compared two multiple imputation methods: fully conditional specification (FCS) and multivariate normal imputation (MVNI), which were found to produce unbiased results compared to complete-case analysis. Lee and Simpson (2014) discusses the advantages and disadvantages of using multiple imputations in observational and experimental studies. Van Buuren and Groothuis-Oudshoorn (2011) introduced and documented the `"mice"` package in `R`, which uses the MICE Algorithm. The Multiple Imputation by Chained Equations (MICE) algorithm is a technique to impute missing values under the MAR assumption (Van Buuren, 2018) due to its overall performance and ease of use (Erler et al., 2016).

So far, our focus has been on modelling non-ignorable missing responses using the existing CRE method while assuming completely observed predictors. This chapter aims to enhance the CRE approach by allowing missing values in the model's predictor. We propose the Two-Step method, where each step is based on existing techniques, but the formulation and application of the Two-Step method in this context represent a novel approach. The process of the Two-Step method is illustrated in Figure 5.1.1. The process starts by imputing the missing observations in the model's predictor using the MICE Algorithm, which is commonly used to impute incomplete predictors (Erler et al., 2016) as a first step. Next, we apply the CRE method to impute missingness in the model response and produce model estimates simultaneously as a second step. The MICE method works by creating multiple copies of the data and replacing the missing data values in each copy. Then, a statistical method is used for each imputed dataset, which is, in this case, the CRE method. Finally, calculate pooled estimates to get overall results and to allow consideration of the uncertainty produced by the missing values (Van Buuren, 2018).

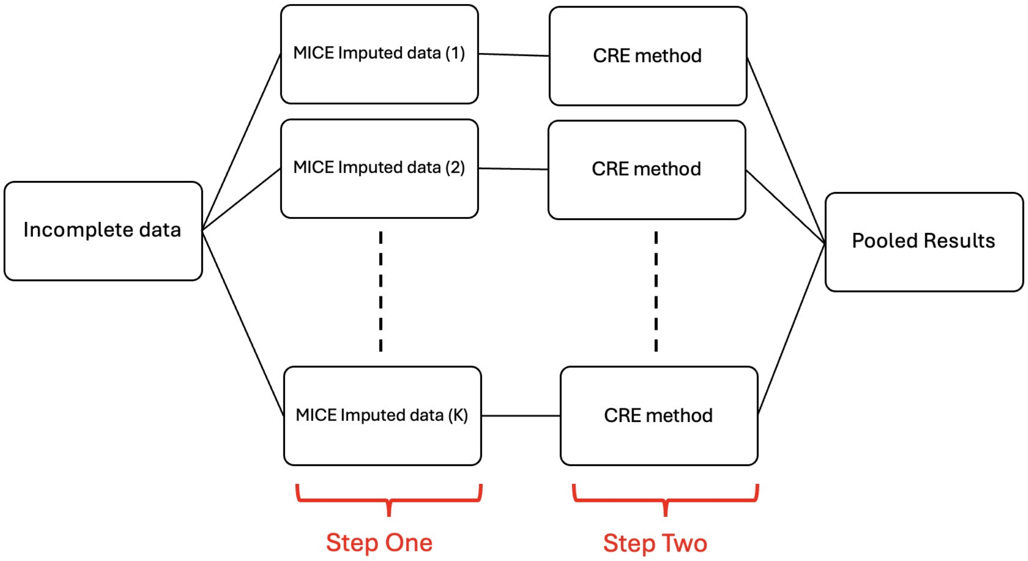


Figure 5.1.1: The Two-Step method is a statistical technique that involves the following steps. First, incomplete predictors are imputed using the MICE Algorithm, which results in K-imputed datasets. Second, the CRE method is applied to the imputed predictors data for each of these K datasets. Finally, the posterior distributions from each dataset are combined to produce overall estimates. This is done by pooling the posterior distributions together.

In this chapter, we will discuss scenarios where the predictor variable has ignorable missing values, and the response variable has non-ignorable missing values using the proposed Two-Step method. The chapter is divided into several sections, each of which will cover different aspects. First, we will define the MICE Algorithm in Section 5.2. Then, we will introduce the model in Section 5.3 and the Two-Step method in Section 5.4. Section 5.5 will illustrate the creation of simulated data, and Section 5.6 will present the results, which will include the findings associated with the simulated data in Subsection 5.6.1 and the application of the proposed method to the real-world dataset (BIOSTAT-CHF) in Subsection 5.6.2. Finally, in Section 5.7, we will discuss the main findings of the study.

5.2 MICE Algorithm

Multiple Imputation by Chained Equations (MICE) is a method that is used to impute missing data on each variable in an iterative manner. It is also known as sequential regression multiple imputation (Huque et al., 2018). MICE requires an imputation model for each variable with missing data. It works by deriving the full conditional distribution for each incomplete variable, where the incomplete variable is imputed conditional on all other variables. Therefore, the imputed values are sampled from these distributions (Austin et al., 2021; Van Buuren, 2018; White et al., 2011). Imputation under conditionally specified models can be implemented in available software (Van Buuren and Groothuis-Oudshoorn, 2011), such as `R` (R Core Team, 2020). Various research studies discuss the imputing longitudinal data using LMM for the imputation model (Resche-Rigon and White, 2018; Schafer and Yucel, 2002; Van Buuren et al., 2011). Huque et al. (2018) carried out a comparison study between Multiple Imputation (MI) approaches for repeated measures. The study found that MI methods provided less biased estimates. However, some approaches based on generalised linear mixed-effects model performed well in imputing missing data in longitudinal studies.

The MICE Algorithm begins by choosing a random sample of the incomplete variable's observed values and setting up the incomplete variable imputation model. In each iteration, the process goes through all the incomplete variables and samples the model's parameters from its conditional distribution based on the observed part of the current variable values and the latest completed data values of other variables. Then, it draws imputed values from the predictive distribution of missing values given the other variables and parameters. Finally, it fills in the incomplete variable with the imputed values from the last iteration. The algorithm creates multiple copies (K) of the data and replaces the missing values in each copy with predicted values from observed data. Then, a standard statistical method for each imputed dataset is applied.

Finally, the pooled estimates are computed to get general results and to consider the uncertainty produced by the missing values (Van Buuren, 2018). The process of the MICE Algorithm for a dataset consists of vectors of variables $\boldsymbol{X}$ is outlined in Algorithm 5, based on Van Buuren (2018).

Algorithm 5 MICE Algorithm

```
-Specify an imputation model: $p(\boldsymbol{X}_j^{mis} \mid \boldsymbol{X}_j^{obs}, \boldsymbol{X}_{-j}, \boldsymbol{X}_{j'})$.
- For each $j$ fill in initial values $\boldsymbol{X}_j^0$ by random drawing from $\boldsymbol{X}_j^{obs}$
-Repeat $w = 1, \dots, W$ iterations
-Repeat $j = 1, \dots, J$ incomplete variables
-Define $\boldsymbol{D}^w = [\boldsymbol{X}_1^w, \dots, \boldsymbol{X}_{j-1}^w, \boldsymbol{X}_{j+1}^{w-1}, \dots, \boldsymbol{X}_{J'}^{w-1}]$ as the currently complete data not including the variable $\boldsymbol{X}_j$.
-Draw $\boldsymbol{\phi}_j^w \sim p(\boldsymbol{\phi}_j^w \mid \boldsymbol{X}_j^{obs}, \boldsymbol{D}^w, \boldsymbol{X}_{j'})$.
-Draw imputations $\boldsymbol{X}_j^w \sim p(\boldsymbol{X}_j^{mis} \mid \boldsymbol{X}_j^{obs}, \boldsymbol{D}^w, \boldsymbol{X}_{j'}, \boldsymbol{\phi}_j^w)$.
-end of $J$.
-end of $W$.
```

where $\boldsymbol{X}_j$ is the j^{th} incomplete variable, $j = 1, \dots, J$, and $\boldsymbol{X}_{j'}$ is the j'^{th} complete variable, $j' = 1, \dots, J'$ and X_{-j} is all other incomplete variable except $\boldsymbol{X}_j$. $\boldsymbol{X}_j^{mis}$ and $\boldsymbol{X}_j^{obs}$ are the missing and observed observations in the j^{th} variable, respectively. The imputation model imputes missing values for the incomplete variable by using the incomplete variable as the outcome variable and the other variables as predictors, as expressed in Equation 3.4.4. The vector of the imputation model parameters for variable j is represented by $\boldsymbol{\phi}_j$. The MICE Algorithm is a useful tool for producing multiple imputations by executing it in parallel K times. Typically, it requires only a few iterations, generally between $W = 5$ and 10 (Van Buuren, 2018). This algorithm works as a Gibbs sampler, which is a Bayesian simulation technique. It takes samples from the conditional distributions to obtain samples from the joint distribution. Conditional distributions in MICE represent the distributions of missing data variables given the observed data variables (Van Buuren, 2018).

5.3 Proposed Model

Consider a continuous response measured over m different time points from n subjects and a set of predictors, some of which may have partially observed values. The response for the i^{th} subject at the t^{th} time point, denoted as $Y_i(t)$, can be modelled in a way similar to the CRE model presented in Chapter 4 as follows:

$$Y_i(t) = \mu + \sum_{j=1}^{J} \beta_j X_{ji}(t) + \sum_{j'=1}^{J'} \lambda_{j'} X_{j'i}(t) + u_i \tilde{Z}_i(t) + e_i(t), \tag{5.3.1}$$

where J' and J represent the number of predictors of fixed effects that are fully observed and partially observed, respectively. μ is the fixed intercept that represents the mean of the overall population. The j^{th} partially observed fixed effects coefficient is denoted by β_j, while the j'^{th} fully observed fixed effects coefficient is denoted by λ_j, $X_{ji}(t)$ is the value of the j^{th} partially observed fixed effect for subject i at time t and $X_{j'i}(t)$ is the value of the j'^{th} fully observed fixed effect for subject i at time t. Additionally, subject-specific random effects u_i are included to capture longitudinal dependence and are assumed to be i.i.d from $N(0, \sigma_B^2)$ and $\tilde{Z}_i(t)$ is the value of the random effect for subject i at time t. The residuals $e_i(t)$ are assumed to be i.i.d from $N(0, \sigma_A^2)$. Similar to Section 4.4.1, we can represent the regression model defined in Equation 5.3.1 as a latent variable:

$$Y_i^*(t) = \mu + \sum_{j=1}^{J} \beta_j X_{ji}(t) + \sum_{j'=1}^{J'} \lambda_{j'} X_{j'i}(t) + u_i \tilde{Z}_i(t) + e_i(t), \tag{5.3.2}$$

where $Y_i(t) = Y_i^*(t)$ if it's observed and $Y_i(t)$ is missing otherwise. Consider a probit regression model that defines the missing response mechanism with predictors as a normal distribution latent variable as follows:

$$p(U_i(t) = 0|Y_i(t)) = 1 - \Phi\left(\tau + \sum_{l=1}^{L} \theta_l X_{li}(t) + v_i \tilde{Z}_i(t)\right), \tag{5.3.3}$$

where, L represents the total number of predictors, which includes both observed and partially observed predictors $(L = J^{'} + J)$, the function $\Phi()$ is the standard normal cumulative distribution function, and the predicted z-score of missing propensity is $\left(\tau + \sum_{l=1}^{L} \theta_l X_{li}(t) + v_i \tilde{Z}_i(t)\right)$, which represent the missing response mechanism model. The function Φ returns the proportion of the area under that z-score in a standard normal density. Furthermore, the probit model includes a zero value as a threshold, which divides the standard normal distribution into two parts. If $U^*(t)$ is greater than zero then $U_i(t) = 1$ and if $U^*(t)$ less than zero then $U_i(t) = 0$, where, $U_i^*(t)$ is a continuous latent missingness indicator variable for subject i at time t modelled as:

$$U_i^*(t) = \tau + \sum_{l=1}^{L} \theta_l X_{li}(t) + v_i \tilde{Z}_i(t) + \varepsilon_i(t), \tag{5.3.4}$$

where θ_l denotes the regression coefficients of the l^{th} fixed effects expresses the systematic influence of missingness due to the unobserved response variable, τ is the fixed intercept represents the mean of the overall population. The subject-specific random effects v_i capture the longitudinal dependence and assumed to be i.i.d $N(0, \sigma_C^2)$, and the residuals $\varepsilon_i(t)$ are assumed to be i.i.d from $N(0,1)$. Moreover, the random effects u_i and v_i are correlated random vectors following a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_B^2 & \sigma_D^2 \\ \sigma_D^2 & \sigma_C^2 \end{pmatrix}$, where σ_D^2 represent the covariance between the random effects u_i and v_i.

5.4 Two-Step Method

The Two-Step method is a helpful method that involves using two different techniques to handle missing data. Each technique has its own strength. The CRE method has been shown in Bhuyan (2019) to effectively handle non-ignorable missingness in the model response, while the MICE Algorithm effectively handles ignorable missingness in the model predictors. Combining these approaches could extend the CRE method to also handle missingness in the model predictors, thereby obtaining more accurate overall results in practice.

In the Two-Step approach, we will use the response and predictor variables to impute the predictor variables that have missingness. This is because the imputation and analysis steps are performed separately, and the imputation model must include the response variable and other predictors (Erler et al., 2016; Moons et al., 2006) to identify and capture any relationships present in the data, which can improve the imputation accuracy. The recently recommended number of imputations ranges from five to ten for more accurate estimates of the regression coefficients and standard errors (Austin et al., 2021). We will produce ten imputed MICE datasets using the Two-Step method, which may be computationally expensive but can increase the accuracy of the results. Estimates based on a low number of imputations can be acceptable. However, it will increase the variability across repeated analyses, leading to less precise inferences. In contrast, using a large number of imputations produces more consistent estimates across repeated analyses, which reduces the Monte Carlo error. Furthermore, a large number of imputations may be necessary for studies that compare different methods (White et al., 2011).

The Two-Step method starts by using the MICE Algorithm to produce multiple imputed datasets of the missing observations in the analysis model's predictors. Next, the CRE method is applied to each imputed data set to handle missingness in the model response, and the analysis model is estimated simultaneously as a second step using Gibbs sampling as mentioned in Section 4.4.4. Then, the overall parameter estimates are obtained by combining the posterior distributions. This is done by aggregating the posterior distributions from each dataset into a single distribution.

Let $\boldsymbol{X}_J$ be the incomplete predictor, $\boldsymbol{X}_{-J}$ be the other incomplete predictors except for $\boldsymbol{X}_J$ and $\boldsymbol{X}_{J'}$ be all fully observed predictors, $\boldsymbol{\phi}_J$ is the vector of regression coefficients for the imputation model of $\boldsymbol{X}_J$, and K is the number of imputed datasets. The Two-Step method presented in Algorithm 6.

Algorithm 6 Two-Step Algorithm

```
for $1,\dots,K$ generated dataset do
    1-Specify an imputation model: $p(\boldsymbol{X}_j^{mis} \mid \boldsymbol{X}_j^{obs}, \boldsymbol{X}_{-j}, \boldsymbol{X}_{j'})$.
    - For each $j$ fill in initial values $\boldsymbol{X}_j^0$ by random drawing from $\boldsymbol{X}_j^{obs}$
    for $w = 1,\dots,W$ iterations do
        for $j = 1,\dots,J$ incomplete variables do
            -Define $\boldsymbol{D}^w = [\boldsymbol{X}_1^w,\dots,\boldsymbol{X}_{j-1}^w,\boldsymbol{X}_{j+1}^{w-1},\dots,\boldsymbol{X}_{J'}^{w-1}]$ as the currently complete data not including the variable $\boldsymbol{X}_j$.
            -Draw $\boldsymbol{\phi}_j^w \sim p(\boldsymbol{\phi}_j^w \mid \boldsymbol{X}_j^{obs}, \boldsymbol{D}^w, \boldsymbol{X}_{j'})$.
            -Draw imputations $\boldsymbol{X}_j^w \sim p(\boldsymbol{X}_j^{mis} \mid \boldsymbol{X}_j^{obs}, \boldsymbol{D}^w, \boldsymbol{X}_{j'}, \boldsymbol{\phi}_j^w)$.
        end of J.
    end of W.
    end of step one.
    2- Apply the CRE method for the joint distribution $p\left(\Theta_{Y,U}, \boldsymbol{Y}^*, \boldsymbol{U}^* \mid \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{X}_{\boldsymbol{J}}^{\boldsymbol{k}}, X_{J'}^k\right)$ using Gibbs sampler in Algorithm 4.4.4 for each K generated datasets.
end of K.
end of step two.
-Combine the K posterior distributions by pooling them together into one posterior distribution.
```

As a consequence of using this approach, there will be K posterior distributions for each parameter estimate. Therefore, these posteriors will be combined by mixing them into a single posterior for each analysis model parameter to calculate the parameter inference. The posterior distribution is approximated by these (K) mixed drawings, aligning with the principles of Bayesian statistical inference described by Gelman et al. (2013) and recommended by

Zhou and Reiter (2010). However, since the chains are from different data sets, the combined chain might exhibit non-convergence; for example, the Gelman-Rubin statistic could be greater than 1.1. Hence, the convergence should be inspected for each chain (Bürkner, 2017).

The Two-Step method involves two main steps: the MICE Algorithm and the CRE method. During the first step, the MICE procedure implements a Linear Mixed Effect Model (LMM) to impute missing values for partially observed time-varying predictors, given all other variables.

In the Two-Step method, we used the `2l.pan` function from the `MICE` package (Van Buuren and Groothuis-Oudshoorn, 2011) in `R` (R Core Team, 2020) to impute missing data. The function uses Gibbs sampler and hierarchical models, assuming a conditional LMM for the partially observed time-varying predictor (Huque et al., 2018; Schafer and Yucel, 2002).

5.5 Creating Synthetic Data for Simulation

In this section, we aim to evaluate the performance of the proposed method in dealing with non-ignorable missingness in the response and ignorable missingness in the predictor. To achieve this, we generate simulated data that demonstrates MNAR in the model response and MAR in the model predictor in a longitudinal context. We will use the simulation data mentioned in Section 3.4. The missing values on the response $Y_i(t)$ were generated based on the following model:

$$U_i^*(t) = \theta_0 + \theta_1 X_{1i}(t) + \theta_2 X_{2i}(t) + \theta_3 X_{3i}(t) + v_i \tilde{Z}_i(t) + \varepsilon_i(t), \qquad (5.5.1)$$

where θ_0 is the overall intercept, θ_1, θ_2 & θ_3 are regression coefficients associated with the fixed effects. The missing data indicator U for each observation is determined by sampling from the binomial distribution, with a success rate equal to the observation's missingness probability from the probit model, which is defined in Section 2.6. The probit regression equation connects missingness probabilities of the response Y to values of Y through the latent missingness indicator regression model U^* for non-ignorable missingness. If the corresponding Y is observed, the value of U is one; if it is missing, the value is zero.

The values of $\boldsymbol{\theta} = [\theta_0, \theta_1, \theta_2, \theta_3]$ were derived to produce the desired missing data proportion. Table 5.5.1 shows these values for each missingness percentage and number of repeated measures. Moreover, v_i are assumed to be $N(0,2)$ and the residuals $\varepsilon_i(t)$ follows $N(0,1)$. The covariance matrix associated with the random effects is $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.

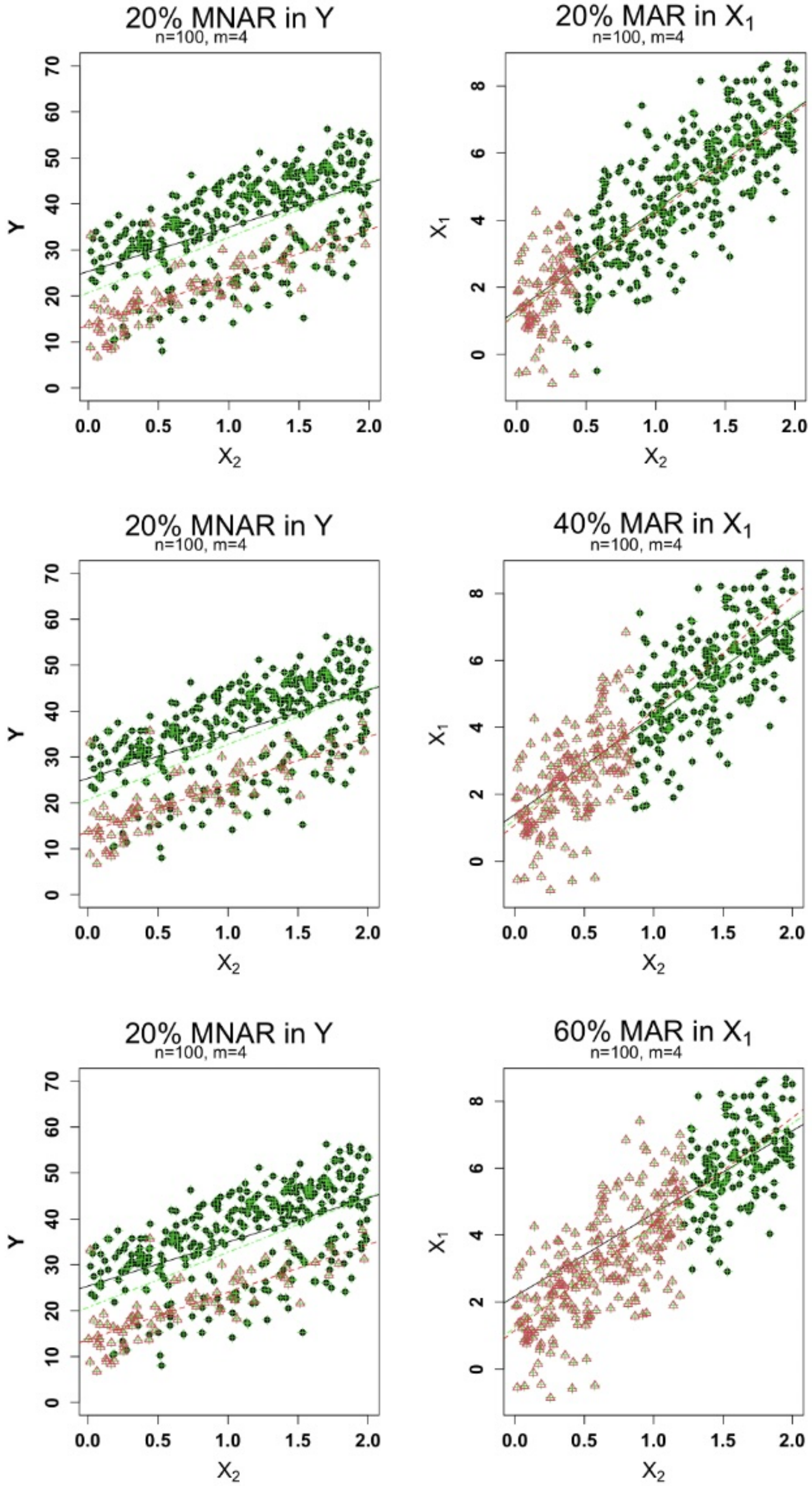
	$\mathbf{p_y = 20\%}$			$\mathbf{p_y = 40\%}$			$\mathbf{p_y = 60\%}$		
Parameter	m=2	m=4	m=8	m=2	m=4	m=8	m=2	m=4	m=8
θ_0	-2	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8
θ_1	-0.3	-0.4	-0.4	-0.4	-0.4	-0.4	-0.7	-0.7	-0.7
θ_2	3	3	2	1	1	1	1	1	1
θ_3	7	4	4	4	4	4	4	4	4

Table 5.5.1: The table presents the values of the coefficients $\boldsymbol{\theta}$ for the missing response indicator regression model, for various proportions of missingness and a number of repeated measures. This information is used to generate missing response values for the Two-Step method, where $\mathbf{p_y}$ represents the percentage of missing values in the model response, and m indicates the number of repeated measures.

The missing values for predictor $X_{i1}(t)$ were generated using the function `deleteMARcensoring()` in the `missMethod` package (Rockel, 2020) in `R`. This assumes that the missingness in $X_{i1}(t)$ is related to the values of $X_{i2}(t)$. Missing values are assumed in $X_{i1}(t)$ whenever the corresponding $X_{i2}(t)$ value is within the p^{th} quantile, where p is the proportion of missingness. This approach enforces the MAR mechanism, where the missing values of $X_{i1}(t)$ depend on the observed values of another variable, $X_{i2}(t)$.

We produced simulation data with 100 subjects, and we tested the impact of the following factors by varying their values. The first factor is the number of repeated measures per subject, which we set to $m = 2$, $m = 4$, or $m = 8$. The second factor was the proportion of missing data, which we varied in different ways. We considered cases where the response variable had 20%, 40%, or 60% missing values, while the incomplete predictor variables had a fixed missing proportion of 20%. We also explored scenarios where the incomplete predictor variables had 20%, 40% or 60% missing values, and in such cases, the response variable correspondingly had 20% missing values. By doing so, we were able to examine the effectiveness of the proposed method under different proportions of missing data. To improve generalisability, we produced 100 replications for each condition to understand how the method behaves under different conditions.

The missingness of the generated data is shown across different percentages of incomplete predictors and responses in Figure 5.5.1. The observed and missing values of the simulated datasets are illustrated for four repeated measures. Missing values in the response occur when the response variable values themselves are lower, confirming that MNAR is the missing mechanism. Additionally, the missing values in the incomplete predictor X_1 are associated with small values in the continuous complete predictor X_2, indicating that MAR is the missing mechanism. Comparable patterns are found in different values of the repeated measures $m = 2$ & 8.


20% MNAR in Y
n=100, m=4
Y
0
10
20
30
40
50
60
70
0.0
0.5
1.0
1.5
2.0
X_2
20% MAR in X_1
n=100, m=4
X_1
0
2
4
6
8
0.0
0.5
1.0
1.5
2.0
X_2
20% MNAR in Y
n=100, m=4
40% MAR in X_1
n=100, m=4
20% MNAR in Y
n=100, m=4
60% MAR in X_1
n=100, m=4

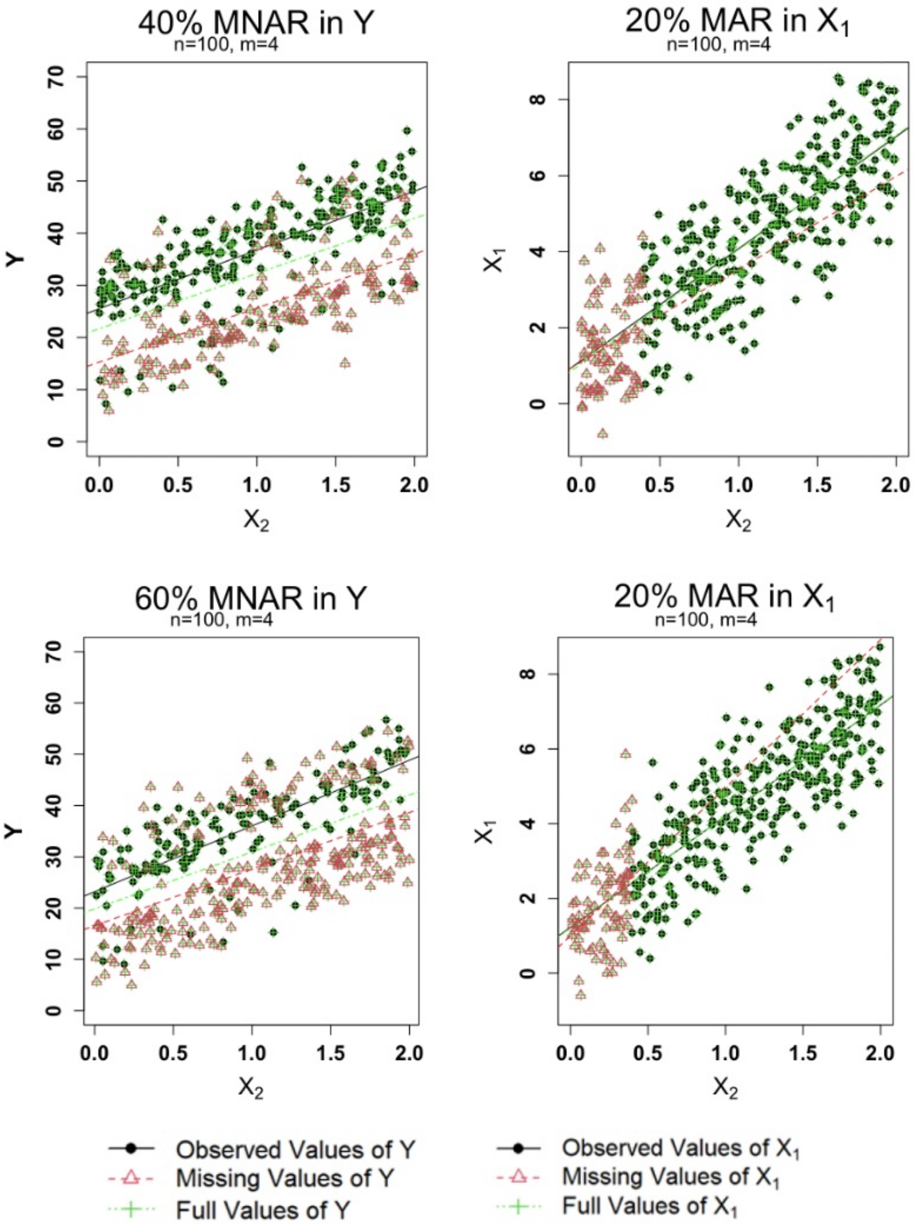


Figure 5.5.1: Scatter plot showing the association between the response (y-axis) and the complete predictor (x-axis) on the left-hand side plots and between the incomplete predictor (y-axis) and the complete predictor (x-axis) on the right-hand side plots. The black dots represent observed values (i.e. the available data), while the red triangles represent missing values. The green crosses show the full data, including both observed and missing values. The plot also displays three regression lines, representing different observed and missing values and illustrating the varying trends for MNAR in the response and MAR in the incomplete predictor. The missingness in the response is dependent on the small values of the response itself, while the incomplete predictor's missingness is dependent on the small values of the complete predictor. Both represent the desired missingness mechanism.

We will test the performance of the proposed Two-Step method using the simulated data. This enables us to evaluate the proposed method compared to some baseline approaches, including the full data that contains all observed data before excluding any missing values to reflect the best-case circumstance and the available data, where any missing values (in the response or predictor variables) cause that entire row to be removed from the dataset. We will also compare the findings of the Two-Step method with the results obtained from other methods. These methods include using the MICE Algorithm for imputing the missing values in the model response and predictor and the CRE method (discussed in Chapter 4) when there are fully observed predictors.

In this chapter, all models will be fitted using R (R Core Team et al., 2013), which employs MICE, Two-Step, CRE and baseline methods. The MCMC simulations will performed for 50,000 iterations, with a thinning rate of 10 applied and half of the iterations designated as a burn-in phase. A single chain will be produced due to the computational time and storage of the Two-Step method being cost-effective, and to assess convergence, we examine the Geweke convergence statistic (defined in Section 2.4.6) for individual parameters and consider convergence achieved if all absolute values of the test statistic are ≤ 2. We also visually examine the trace plots for each parameter. Based on these criteria, all the runs discussed in this chapter are deemed to have converged. The `brm` function is used to fit the full and available data methods, and the `brm_multiple` function is used to fit and pool the results of the MICE imputed data by combining the posterior distributions (Bürkner, 2017). Both functions belong to the `brms` package (Bürkner, 2017), which uses the HMC algorithm defined in Section 2.4.5, we will produce three chains, and each one was executed for 2000 iterations.

We will compare the Two-Step method performance based on model parameter estimates accuracy via Root Mean Square Errors (RMSE), Relative Bias (RB), and Coverage Rate (CR). To ensure equal scale, Weighted Root Mean Square Errors (WRMSE) are used to assess the overall method accuracy across all models of interest parameters, which were weighted by the data-generating parameters. We discussed these criteria in Section 2.7. We will investigate how these criteria are distributed across 100 replications.

5.6 Results

This section presents the results of the proposed Two-Step method and the comparative methods, applied to simulated data and the BIOSTAT-CHF dataset.

5.6.1 Simulated Data Results

The results of the Two-Step method using the CRE method demonstrate that imputed response values match the true generated values. Figure 5.6.1 shows the average of imputed values at each iteration of the Gibbs sampler of the response, falling within the ± 2 Standard Deviation (SD) range of the average imputed values across different proportions of missingness in the response and for four repeated measures. This conclusion applies to other repeated measures and across different proportions of missingness. These conclusions demonstrate how well and consistently the Two-Step method captures true values that were not seen.

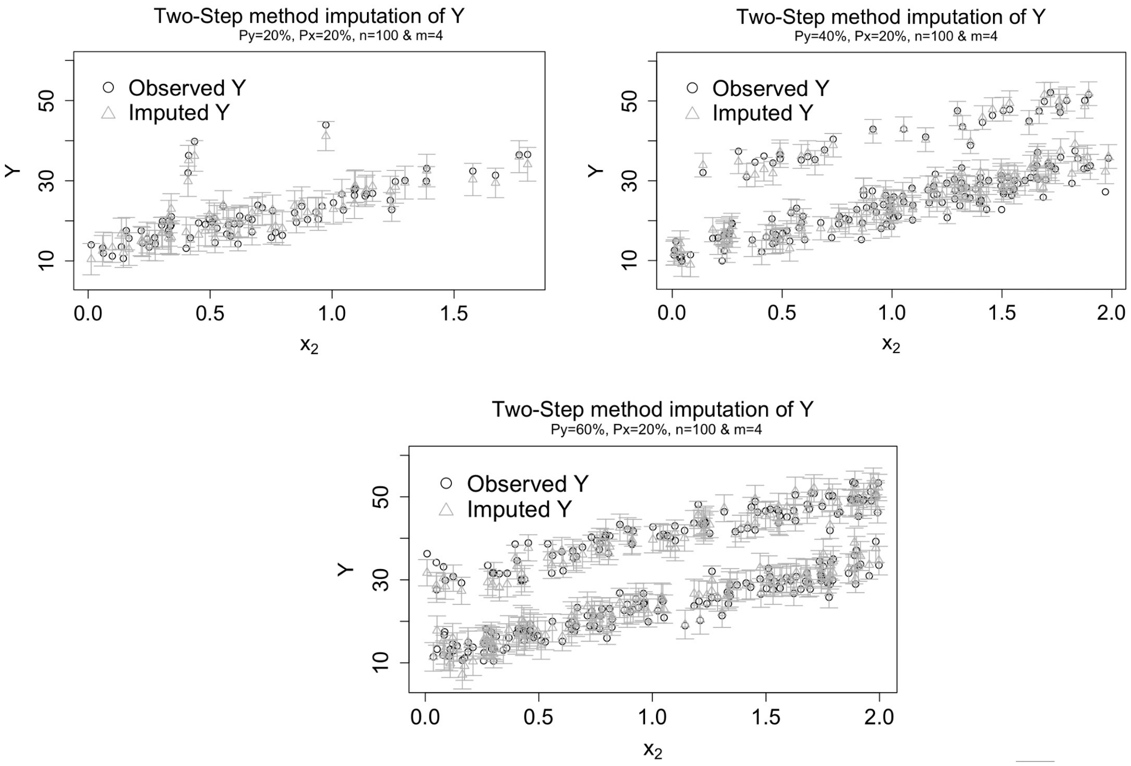


Figure 5.6.1: Scatter plots represent the model response Y on the y-axis and the fully observed predictor X_2 on the x-axis to assess the true values (depicted as black circles) against the average of imputed values and the ± 2 SD (represented by grey triangles and vertical lines) using the Two-Step method for various missingness in the model response with four repeated measures. The true values are mostly enclosed within the imputed values region.

In order to assess how well the Two-Step method performs in comparison to baseline models, CRE method and MICE Algorithm method of the response and predictor, we calculated the Weighted Root Mean Squared Error (WRMSE) for each generated dataset from the simulated study. We then visualized the distribution of the WRMSE for different applied models, proportions of missingness, and repeated measures in Figure 5.6.2. The results show that the Two-Step method has similar WRMSE values to the CRE method and MICE Algorithm method, except with slightly larger values and uncertainty of the WRMSE when using the MICE Algorithm with 60% missing values in the incomplete predictor model and with $m = 2$ and 4. The available data method, on the other hand, has larger WRMSE values overall and even larger WRMSE uncertainty with 60% missing values in the incomplete predictor model.

Considering the general inequality among methods, we will analyse the particular differences for every analysis model parameter next. The RMSE analysis displays the degree of variation between the estimated and data-generating parameters, giving us insight into how accurate our estimations are with the data-generating parameters.

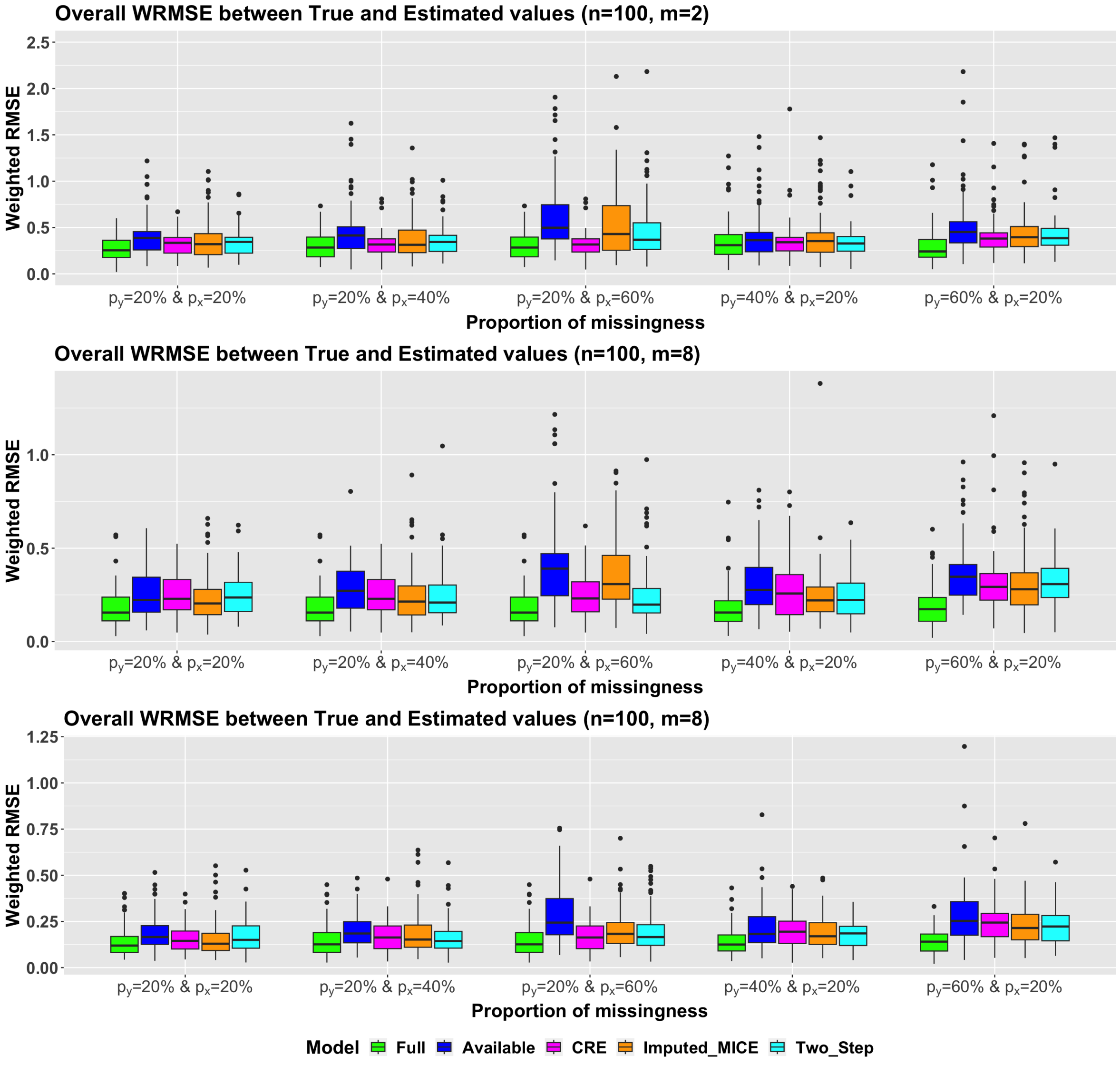


Figure 5.6.2: Boxplots represent the overall WRMSE for each method across different proportions of missing data and repeated measures. The y-axis represents the WRMSE values, while the x-axis represents one of the proportions of missing data in the model response and incomplete predictor. Each boxplot corresponds to one of the applied methods, and each plot represents different repeated measures. The results show that all methods have similar WRMSE, with a slight increase in the available data method.

The Two-Step method has been shown to produce results similar to the CRE and MICE Algorithm methods in terms of the RMSE of the analysis model fixed coefficients and variance components. However, the CRE method produces less RMSE than the Two-Step and MICE Algorithm methods in cases where a large proportion of missing data (60%) is in the model's incomplete predictor.

When analysing residual variance σ_A^2, the MICE Algorithm method produces slightly larger RMSE values than the Two-Step method in situations where 60% of data is missing in the model incomplete predictor. Additionally, the MICE Algorithm method produces larger RMSE values and higher levels of uncertainty for the between-individual variances σ_B^2 when $m = 2$, across different proportions of missing data, and when 60% of data is missing in the model incomplete predictor with $m = 4$ and 8.

On the other hand, the available data method produces larger RMSE values overall and more substantial RMSE uncertainty when there is a large proportion of missing data (60%) in the model incomplete predictor, specifically for the fixed effect coefficients. The RMSE distributions representing β_0 and β_1 can be found in Figure 5.6.3, while the RMSE distribution for variance components can be found in Figure 5.6.4 across different repeated measures and proportions of missingness. The results of RMSE for β_2 and β_3 can be found in Section B in the Appendix.

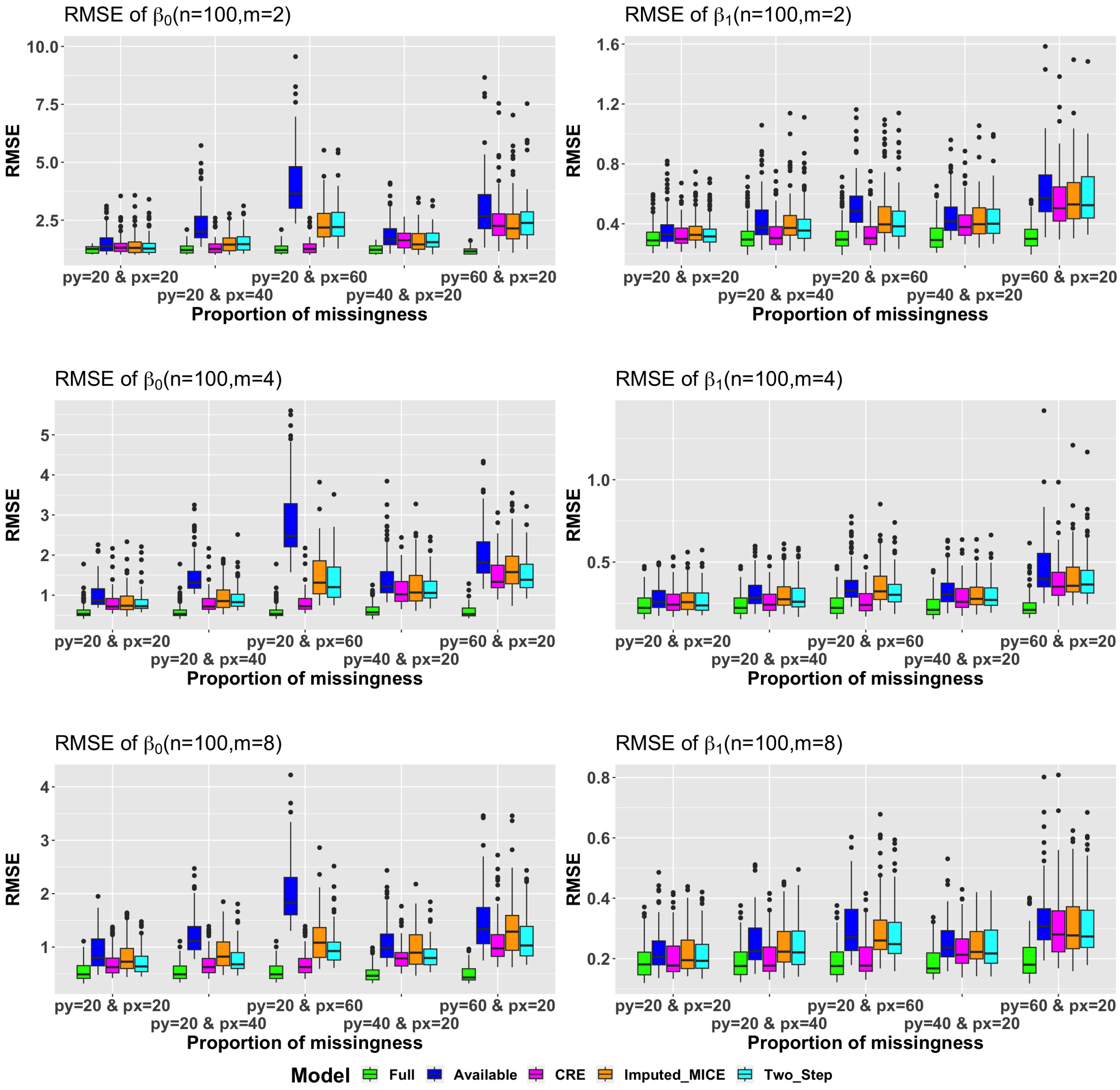
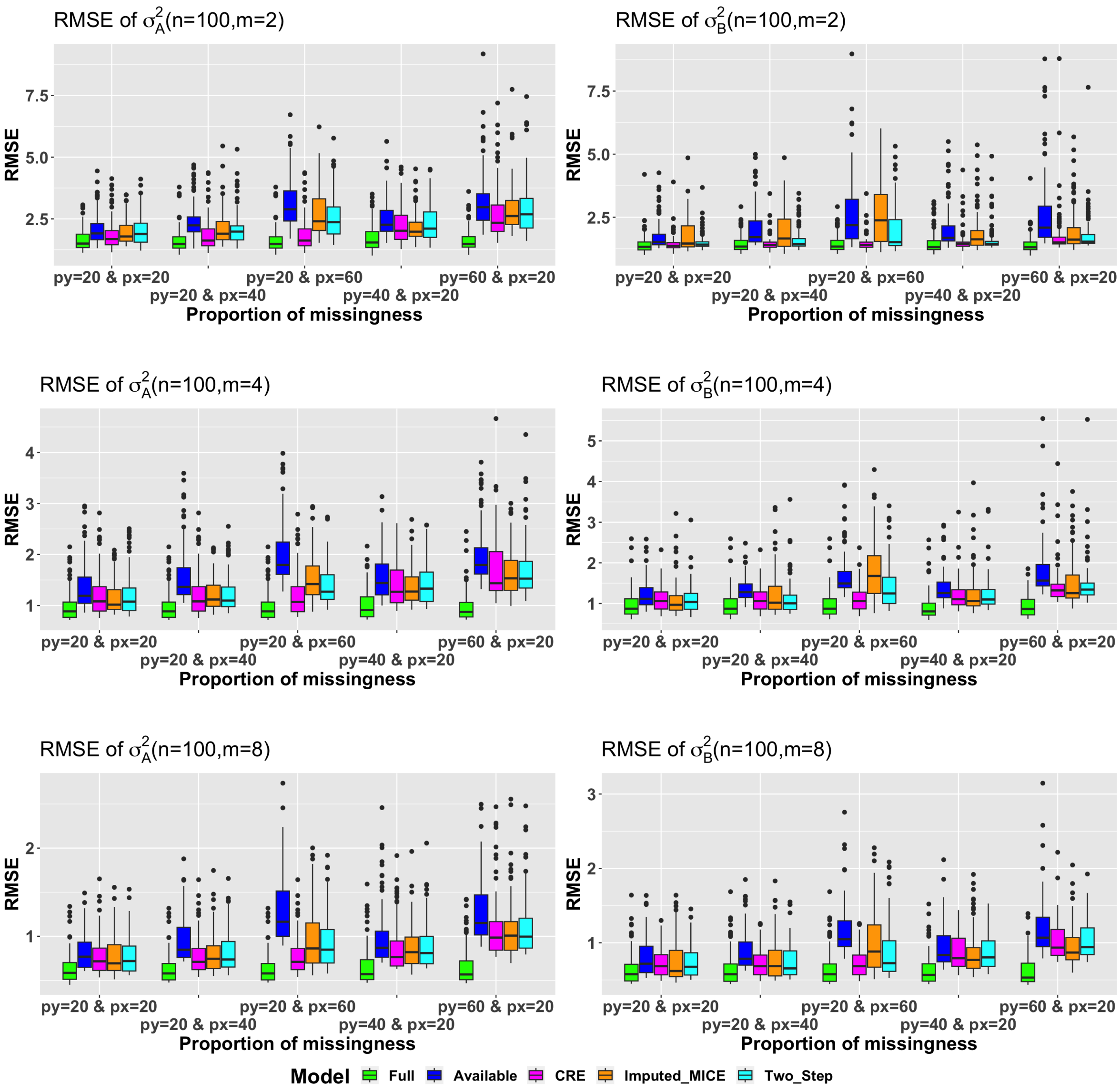


Figure 5.6.3: Boxplots illustrate the RMSE of β_0 in the left-hand side plots, and β_1 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RMSE values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. It's obvious that the available data method has quite larger RMSE values than other methods.



RMSE of σ^2_A(n=100,m=2)

RMSE of σ^2_B(n=100,m=2)

RMSE of σ^2_A(n=100,m=4)

RMSE of σ^2_B(n=100,m=4)

RMSE of σ^2_A(n=100,m=8)

RMSE of σ^2_B(n=100,m=8)

Figure 5.6.4: Boxplots illustrate the RMSE of σ_A^2 in the left-hand side plots, and σ_B^2 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RMSE values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. It is evident that the available data method produces larger RMSE values compared to other methods. Also, the available data and MICE Algorithm methods have larger RMSE values and uncertainty for σ_B^2 with m=2.

The RB evaluates the model estimates to identify patterns of overestimation or underestimation, pointing out differences in the precision of data-generating parameters across all applied methods. Results of simulated data show that the Two-Step method has a similar RB to the CRE method, MICE Algorithm method, and available data method for β_1 and β_2, with larger RB uncertainty when there is 60% missingness in the model's incomplete predictor for β_2 and β_0 using the available data method. The Two-Step method has a similar RB to the CRE method for β_0 and β_3. However, the Two-Step method has larger RB uncertainty when there is a 60% missingness in the model's incomplete predictor.

The available data method and the MICE Algorithm method tend to underestimate the intercept parameter, and this increases as the number of repeated measures increases, and with a larger proportion of missingness in the model response (40% and 60%). On the other hand, these methods tend to overestimate β_3, which increases as the number of repeated measures increases, with a larger proportion of missingness in the model response (40% and 60%). The results for β_0 and β_1 are presented in Figure 5.6.5, while β_2 and β_3 results can be found in Section B in the Appendix.

From Figure 5.6.6, the MICE Algorithm method underestimates the residual variance and overestimates the between-individual variance as the number of repeated measures decreases. Additionally, the MICE Algorithm method and the Two-Step method tend to underestimate the residual variance and overestimate the between-individual variance with 60% missingness in the model's incomplete predictor.

We have assessed the accuracy of parameter estimation for analysis model parameters and found some valuable insights into the performance of various methods. The results show that the Two-Step method outperforms the available data method, which is typically used for longitudinal data in terms of RMSE. This indicates that the Two-Step method has superior precision in estimating parameters. Additionally, it's important to note that the available data and MICE Algorithm methods demonstrate biased estimation when large proportions of missing data are present in the model's incomplete predictor, while the proposed Two-Step method only showed bias for the variance components with 60% missingness in the incomplete predictor. On average, the Two-Step estimates are closer to the data-generating parameters than the available data method.

This suggests that the Two-Step method is more effective in reducing estimation errors. Furthermore, the relative bias analysis reveals that the Two-Step method provides unbiased estimations, which implies that it consistently provides estimates that are centred around the true values. The MICE Algorithm and Two-Step methods produced biased estimates for the variance components. This was due to applying the MICE Algorithm in the first step of the Two-Step method, which produced biased estimates for the variance components. This occurred when there were 60% missing values in the incomplete predictor.

The CR assesses how effectively each approach covers the data-generating parameters of the analysis model parameters. For most analysis model parameters, the CR varies between 0.9 and 0.99, with some exceptions. For instance, the model intercept β_0 and slope β_3 using the MICE Algorithm method has low CR for all proportions of missingness with $m = 8$ and high proportions of missingness in the model response (40% and 60%) with $m = 2$ and 4. Additionally, for 60% missingness in the incomplete predictor with $m = 2$ for the model intercept β_0 and slope β_1 for 60% missingness in the model response with $m = 2$ and 8. The available data method has low CR with 20% missingness in the model response and incomplete predictor with $m = 8$ for β_0 and all proportions of missingness except 60% missingness in the model response with $m = 8$ for β_3.

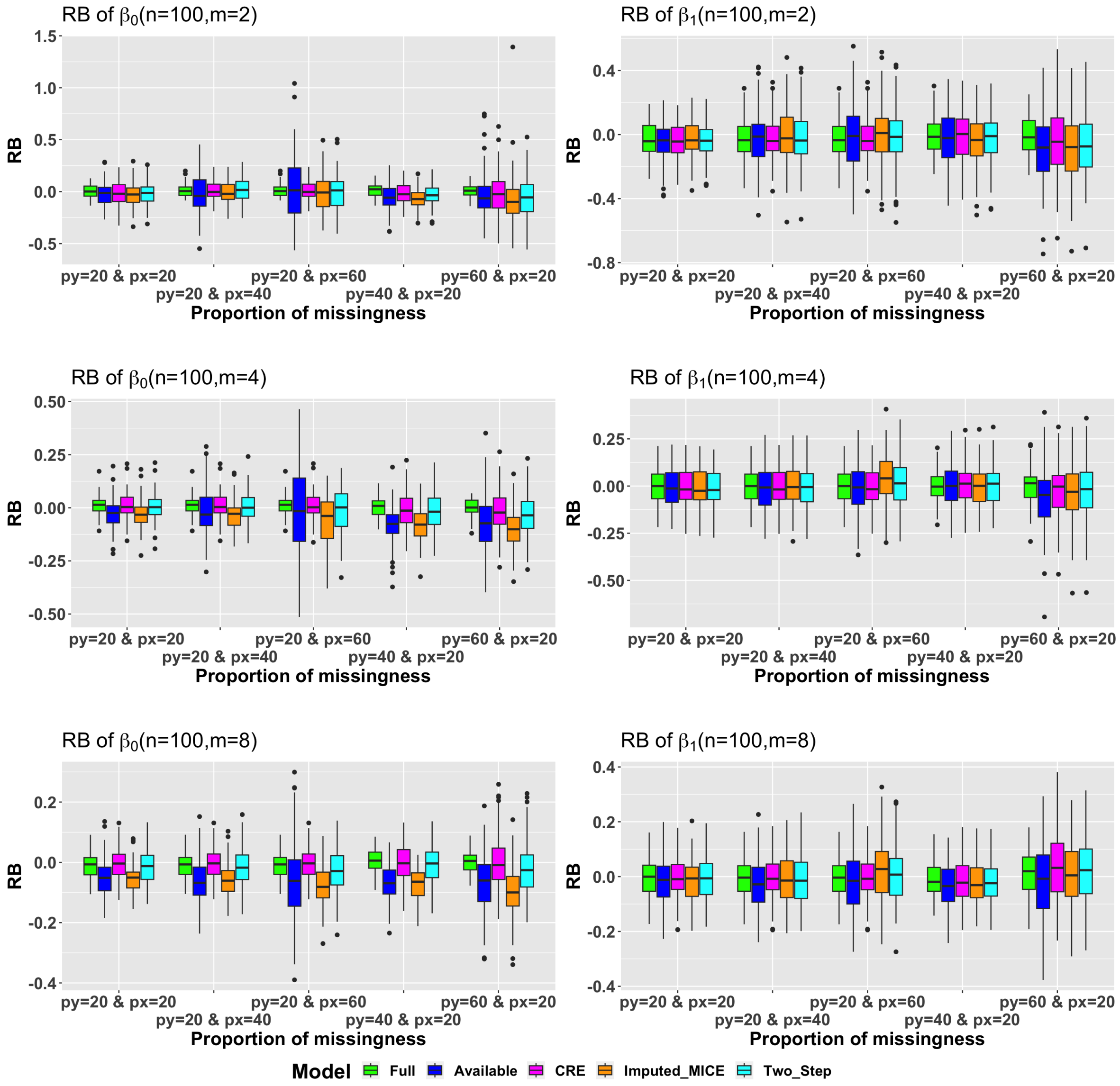


Figure 5.6.5: Boxplots illustrate the RB of β_0 in the left-hand side plots, and β_1 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RB values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. It's obvious that the available data method RB uncertainty increases when there is 60% missingness in the incomplete predictor.

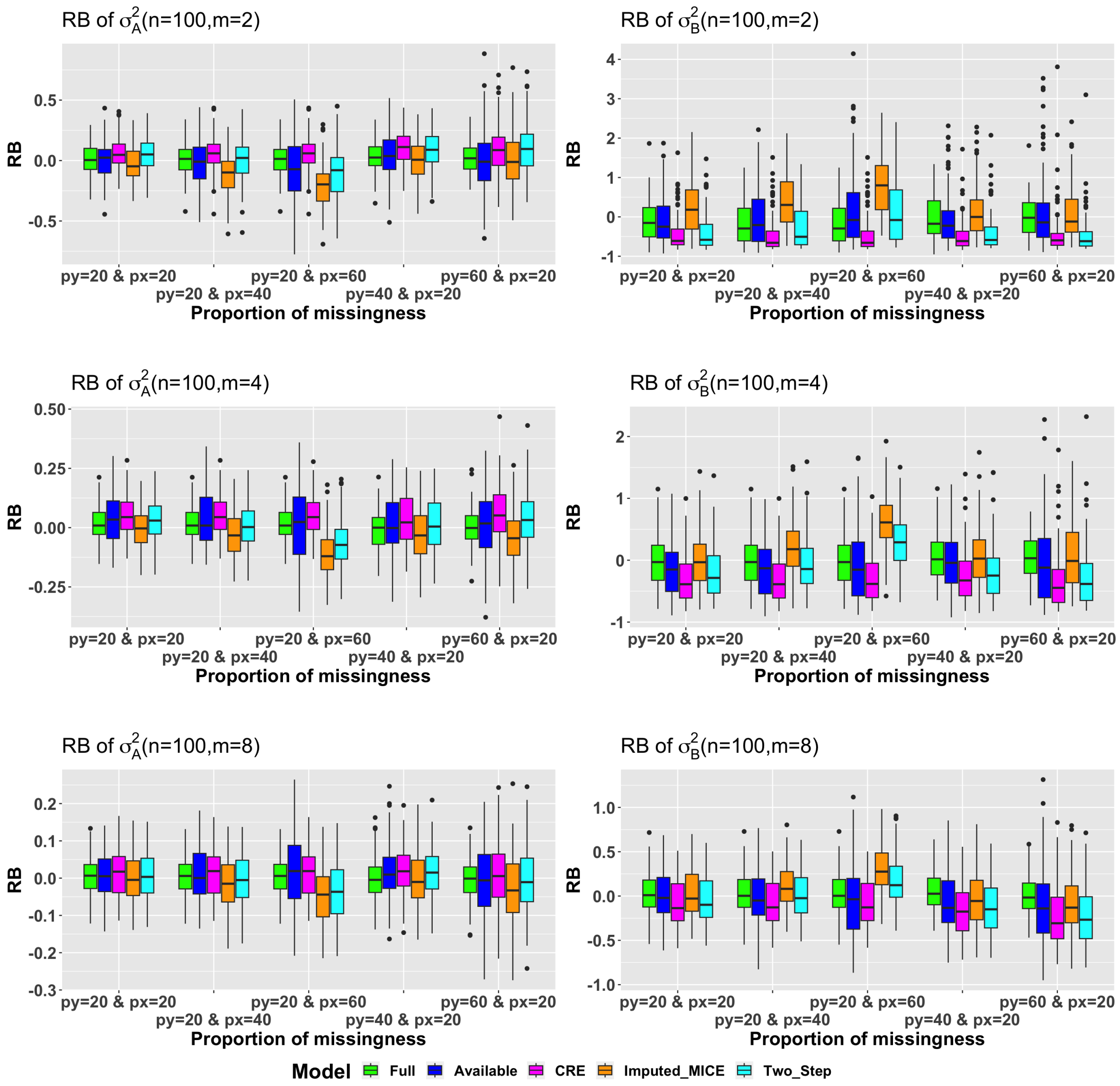


Figure 5.6.6: Boxplots illustrate the RB of σ_A^2 in the left-hand side plots, and σ_B^2 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RB values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. The Two-Step and MICE Algorithm methods produce biased results when there is 60% missingness in the incomplete predictor.

Regarding the variance components, the MICE Algorithm method CR is low for 60% missingness in the model incomplete predictor for different repeated measures. The CRE method has a low CR for the between-individual variance σ_B^2 with $m = 4$ for all missingness proportions and with $m = 2$ for 20% missingness in the model response regardless of the proportion of missingness in the model incomplete predictor. The Two-Step method has a low CR for residual variance σ_A^2 with $m = 8$ and 60% missingness in the model incomplete predictor. Section B in the Appendix contains plots related to the CR for each analysis model parameter.

Regarding out-of-sample prediction performance, it has been observed that the Two-Step method and the CRE method perform better than the available data method and the MICE Algorithm method. The Two-Step method performs better than the available data method as the proportion of missingness increases. On the other hand, the MICE Algorithm method outperforms the available data method. The results for one scenario of a proportion of missingness are shown in Figure 5.6.7, while the rest can be found in Section B in the appendix.

As the number of repeated measures increases, the RMSE of the missingness response model parameters using the Two-Step method decreases. The slope θ_1 has the lowest RMSE value and uncertainty among all the parameters. As the number of repeated measures increases, the model's parameters become unbiased. However, the covariance parameter between the random effects σ_D^2 has larger RB uncertainty as the repeated measure increases. The covariance parameter between the random effects σ_D^2 has a low CR with $m = 2$ and 4 for 20% missingness in the model response and incomplete predictor. The between-individual variance parameter in the missing response model has a low CR with $m = 2$ for 20% missingness in the model response, regardless of the missingness in the incomplete predictor. The fixed effect coefficient shows a low CR with $m = 8$ for 60% missingness in the incomplete predictor, except the model's intercept θ_0. In all other cases, the CR of the missingness response model parameters using the Two-Step method varies between 0.9 and 0.99. The CR plots can be found in Section B in the appendix.

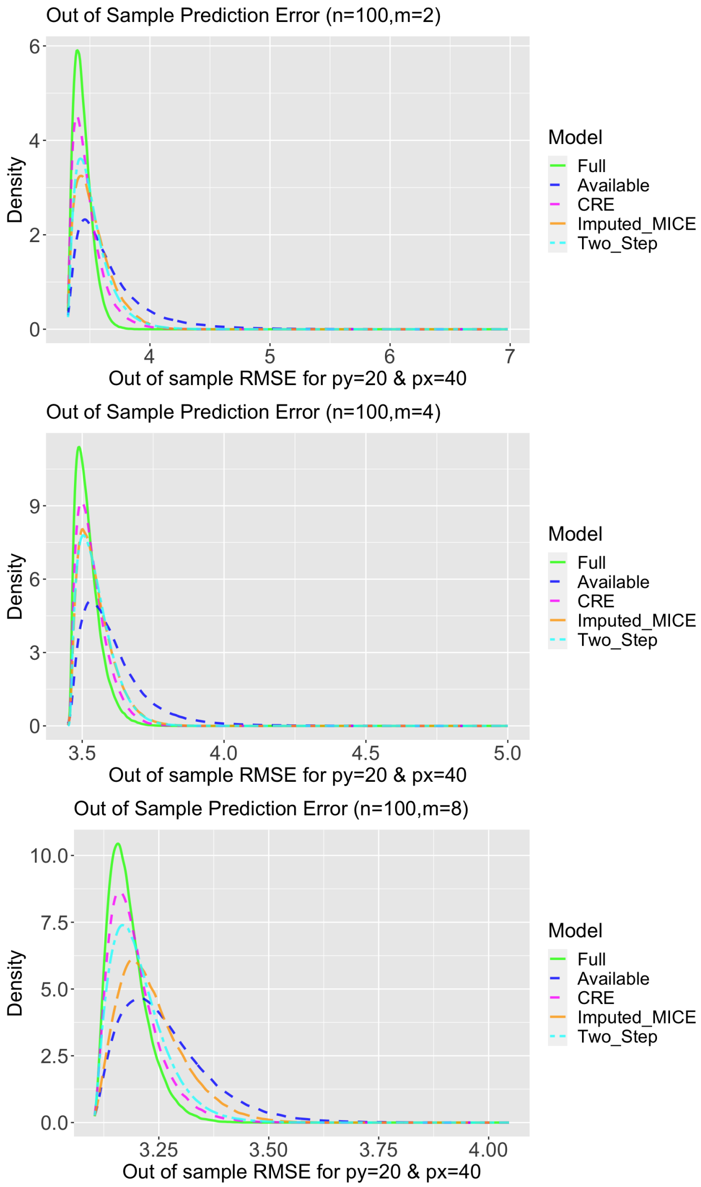


Figure 5.6.7: The density plots of the out-of-sample RMSE for different methods across various repeated measures with 20% missingness in the response and 40% in the incomplete predictor. Each density curve corresponds to one of the methods used, and each plot corresponds to a different value of repeated measures. The Two-Step method performs better than the available data method, which appears to have less density and slightly shifted to the right.

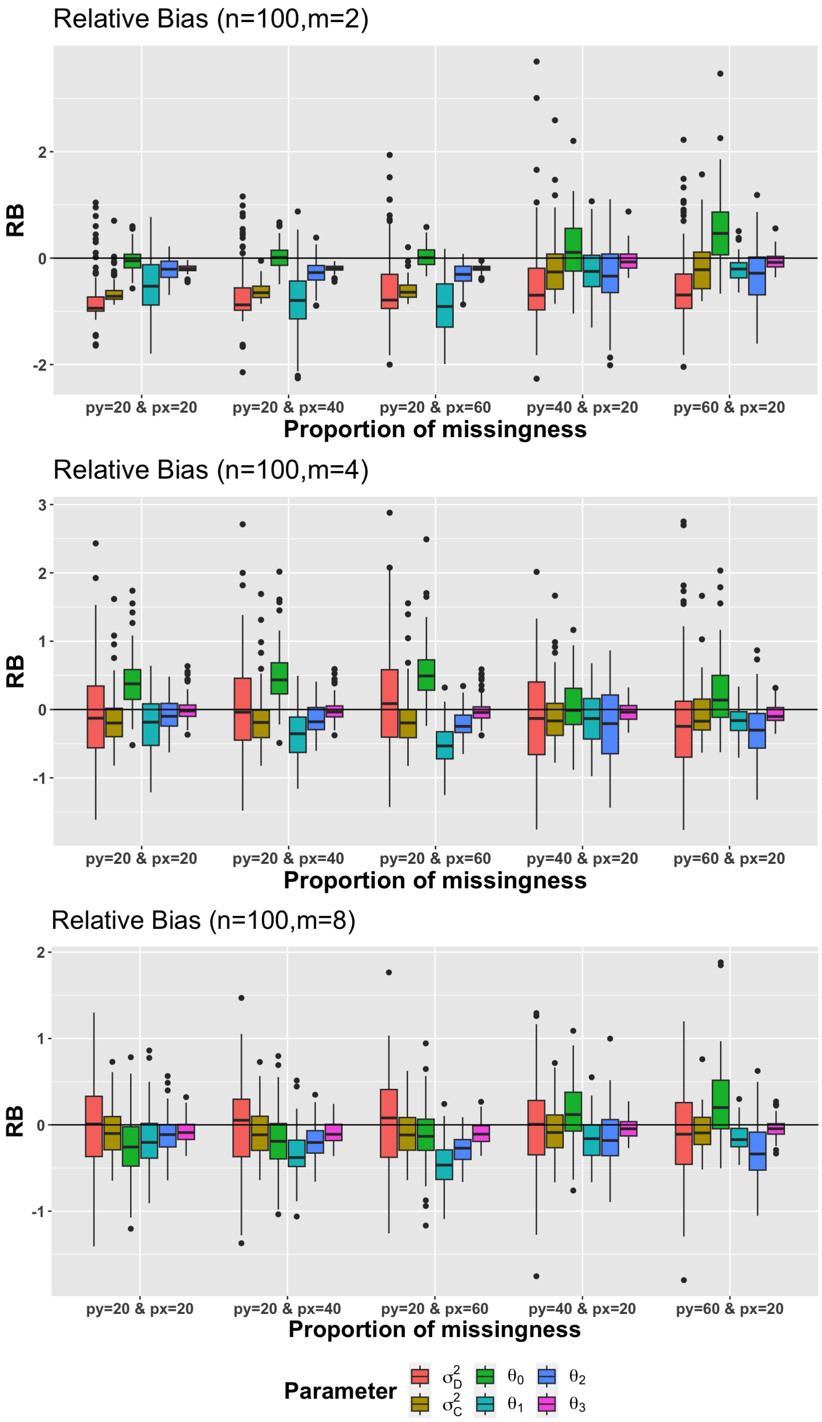


Figure 5.6.8: Boxplots illustrate the RB of the missingness response model parameters using the Two-Step method across various proportions of missingness and repeated measures. The y-axis represents the RB values, while the x-axis shows the proportion of missingness, and each boxplot represents a specific missingness response model parameter. Each plot corresponds to a different repeated measure value. As the repeated measure increases, the model parameter becomes more unbiased.

5.6.2 Real Data Results

Many longitudinal real-world datasets are subject to missing values, which should be handled carefully in the statistical analysis. In order to handle missingness in the real-world BIOSTAT-CHF dataset, we applied the Two-Step method to deal with missing values in the model response and predictor. We have also applied the MICE Algorithm method and available data as baseline approaches. We have used the Bayesian Hierarchical Model (BHM) using HMC expressed in Section 2.4.5 as a default method for dealing with repeated measures (mixed models) for the available data approach. The implementation of the Two-Step and baseline approaches in the BIOSTAT-CHF dataset is based on the model described in Equation 4.2.2 in Section 4.2.

The MICE approach was applied to the response variable "log(NT-proBNP)" and to the incomplete predictor variable "c.eGFR", which have missing values, to generate ten imputed datasets. The density plot in Figure 5.6.9 shows the blue curves for the observed values of log(NT-proBNP) and c.eGFR, overlain by the ten generated imputed datasets in red curves. The plot indicates that the imputation process effectively captures the underlying distribution of the variables, resulting in imputed values that are very similar to the observed values. However, there is a slight leftward shift in the MICE-generated data for log(NT-proBNP) in comparison to the original data.

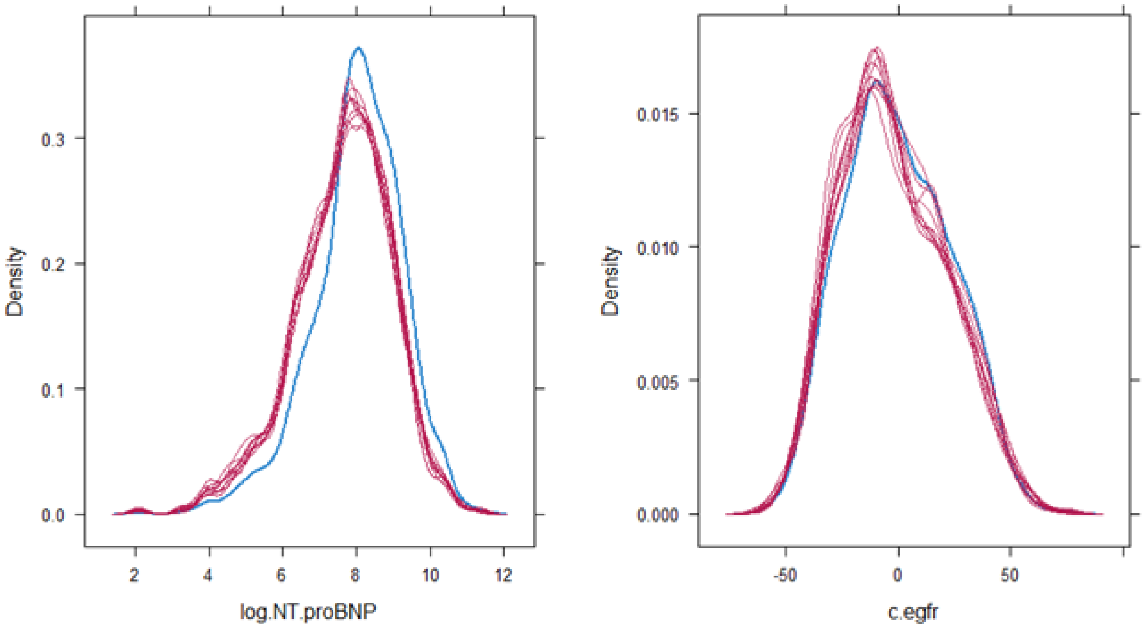


Figure 5.6.9: The density plot illustrating the distribution of the "log(NT.proBNP)" is on the left, and the incomplete predictor "eGFR" (which has been centred) is on the right in the BIOSTAT-CHF dataset. The plot displays a blue curve for the observed dataset and red curves for ten MICE-imputed datasets. The observed and imputed data have similar densities, except the imputed data for the response shifted leftward from the observed data distribution.

Figure 5.6.10 shows the posterior distributions obtained from three different methods: the Two-Step method, the available data method, and the pooled estimates from the MICE Algorithm. The posterior distribution of the model's parameters indicates that the Two-Step and available data methods have similar posterior distributions. However, the pooled posterior distribution from the MICE Algorithm differs noticeably, particularly in the variance component. The Kolmogorov-Smirnov test indicates that eGFR, Pacemaker, and the residual variance σ_A^2 have p-values greater than 0.05. This suggests no significant difference between the Two-Step and available data methods in these parameters. Additionally, the covariance between the random effects is $\sigma_D^2 = -0.25$, which suggests that there is weak/moderate MNAR evidence of the response variable. The result is similar to what we found when applying the CRE method in Section 4.4.7. It is worth mentioning that the pooled posterior distribution from the MICE Algorithm appears less smooth than the other methods despite using 10,000 iterations with HMC, with the Gelman-Rubin diagnostic being less than 1.1.

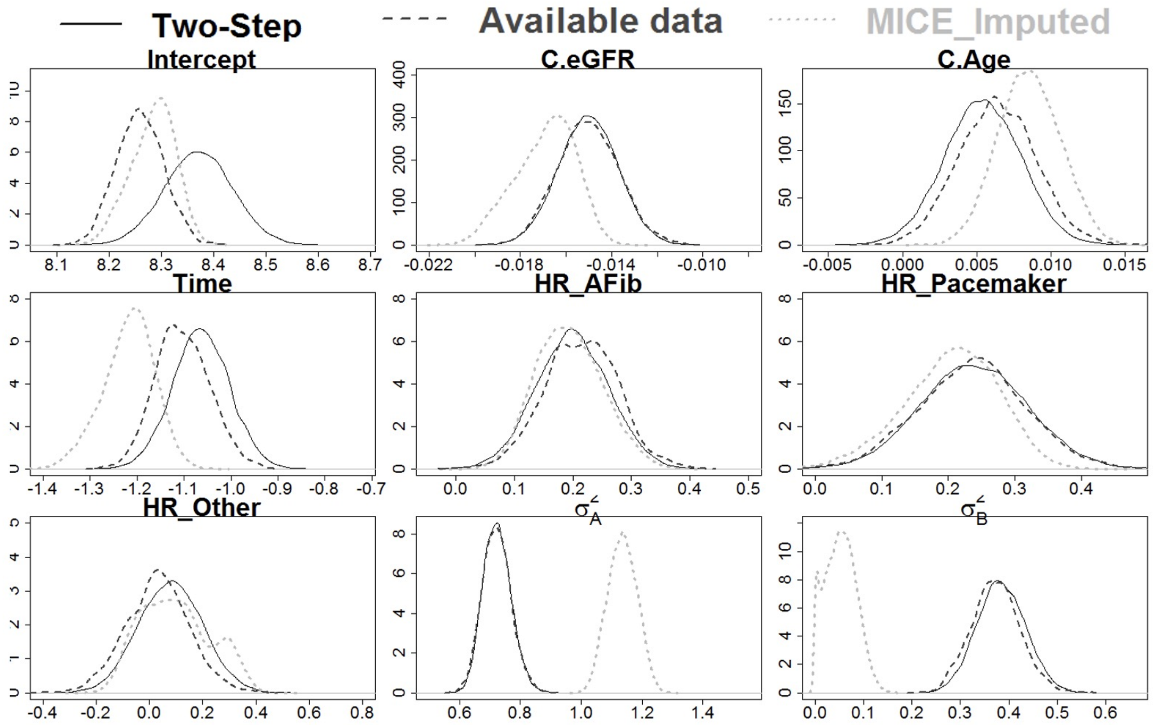


Figure 5.6.10: The plot consists of three curves representing different methods used in the BIOSTAT-CHF model. The black solid curve represents the posterior distribution of the Two-Step method, while the dark grey dashed curve represents the available data method. The light grey dashed curve represents the MICE imputed response and predictor. Each plot represents one of the model parameters. The MICE Algorithm converged to variance parameter estimates different from the Two-Step and available data methods.

The density of the observed and Two-Step imputed response variable values are shown in Figure 5.6.11. Each grey curve is a different draw of the latent variable from the posterior, whereas the black curve represents the observed data. The similarities between the observed and imputed data density curves indicate that the imputation process successfully preserved the overall data distribution.

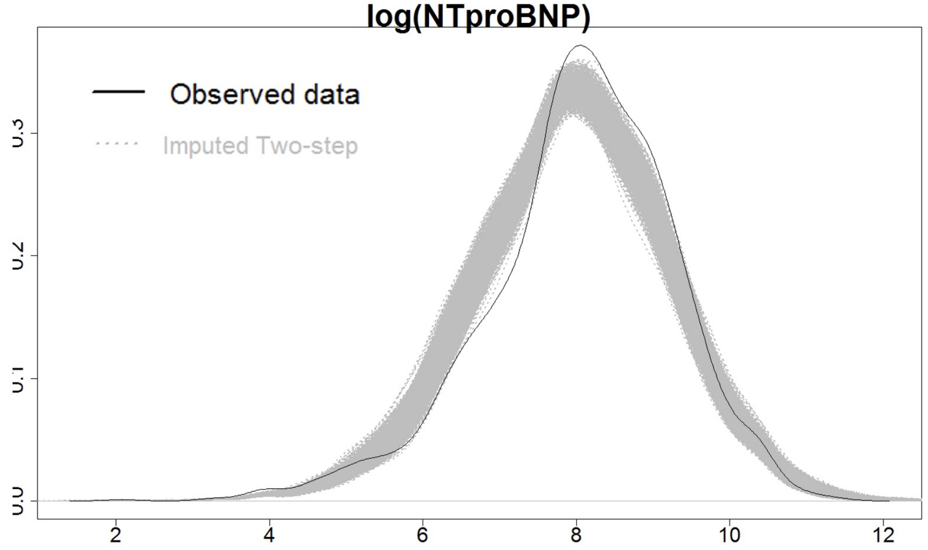


Figure 5.6.11: Density plot show the response observed values in the black solid curve and Two-Step imputed response values in the grey dashed curves, where each curve is a different draw of the latent variable from the posterior using Gibbs sampler. The density of the imputed values using the Two-step method at each Gibbs sampler iteration is similar to the density of the observed values.

5.7 Discussion

We have proposed a Two-Step method to address the issue of missing data when it occurs in both the analysis model response and predictor in the longitudinal study. This is a common occurrence in longitudinal datasets, as the variables are measured repeatedly over time to analyze their effects over time and draw sufficient conclusions. Our proposed method aims to handle such missingness effectively and draw accurate conclusions from the analysis. The Two-Step method is used to handle missing data in two stages. Firstly, it employs the MICE Algorithm to handle missing values in the analysis model predictor and generates multiple complete versions of the data. Subsequently, it applies the CRE method, which is a recent approach for dealing with non-ignorable missingness in the model response for longitudinal data. Bhuyan (2019) introduced this method to be efficient using Gibbs sampling. Since the MICE Algorithm assumes the MAR mechanism, which deals with ignorable missingness, and the CRE method was proposed to handle non-ignorable missingness in the response, the Two-Step method is designed to handle non-ignorable missingness in the analysis model response and ignorable missingness in the analysis model incomplete predictor.

The proposed Two-Step method was tested using simulated data. The results showed that the method can accurately estimate the true generated values of the response through Gibbs sampling. The analyst can then use this estimation to impute data that resembles the unobserved data. Additionally, the Two-Step method outperforms the available data and MICE Algorithm methods in terms of the overall analysis model's parameters WRMSE. When it comes to out-of-sample performance, both the Two-Step and CRE methods perform similarly (Given that the CRE method has fully observed explanatory variables, whereas the Two-Step method can handle up to 60% missingness in the explanatory variables) and better than the MICE Algorithm and available data methods. These results are quite promising, especially for longitudinal studies and healthcare, as they suggest that the Two-Step method produces reliable conclusions even with unseen or untrained data. This indicates that the proposed method handles missing data more effectively, leading to more reliable clinical decisions.

In terms of estimating the individual analysis model's parameters using the RMSE criteria, the Two-Step and MICE Algorithm methods show similar performance to the CRE method. However, the Two-Step and MICE Algorithm methods tend to have higher RMSE values and uncertainty when there is 60% missingness in the incomplete predictor, with the Two-Step method having a lower RMSE than the MICE Algorithm method for the fixed effects coefficient. The Two-Step method has a smaller RMSE than the MICE Algorithm method for estimating σ_A^2 with 60% missing values in the incomplete predictor. Similarly, the MICE Algorithm method shows a larger RMSE in estimating σ_B^2 when $m = 2$, and with 60% missingness in the incomplete predictor for $m = 4$ and 8. Additionally, the available data shows a noticeably larger RMSE compared to other methods, specifically for analysis model fixed effects coefficients with 60% missing values in the incomplete predictor.

Generally, the applied methods result in unbiased estimates for the analysis model parameters. Except for the available data, which tend to have larger RB uncertainty with 60% missingness in the incomplete predictor in some fixed effects coefficients. The Two-Step method and the CRE method are similar in their performance. However, the Two-Step method has larger RB uncertainty if there is 60% missingness in the incomplete predictor. This is due to the fact that the Two-Step method is designed to handle extra missingness compared to the CRE method, where the latter only considers fully observed predictors. As a result, the proposed method introduces extra error into the model. The available data method and the MICE Algorithm method produce biased results for β_0 and β_3 as the repeated measures increase and when there is a larger proportion of missingness in the response. The MICE Algorithm method produces biased estimates of variance parameters with low repeated measures. The MICE and Two-Step methods produce biased variance estimates with 60% missingness in the incomplete predictor. The MICE Algorithm method struggles to estimate the data-generating parameters accurately when the model response has a high proportion of missing values for β_0 and β_3. In some scenarios, the CRE method also struggles to capture the data-generating of the between-individual variance σ^2_B parameter.

Regarding the response missingness model parameters using the Two-Step method, the estimates become more accurate and unbiased as the repeated measures increase. Notably, the slope θ_1 has less uncertainty regarding RMSE and is therefore inferred more consistently. This slope parameter is associated with the incomplete predictor that has been imputed. The covariance between the random effects of the response models, σ^2_D, showed unbiased results with repeated measures of more than two. However, it has higher uncertainty, which indicates that the proposed Two-Step method can estimate the possibility of missingness in the model response characteristic due to MNAR across various data samples. Nevertheless, the proposed method was unable to capture the actual value of σ_D^2 with low repeated measures and low missing values in the response and incomplete predictor.

The Two-Step method was applied to the BIOSTAT-CHF dataset to determine how well it can handle missing data in a real-world scenario. The results showed that the proposed Two-Step method produced comparable estimates to the available data, but the Two-Step method has the advantage of producing imputed response values that closely resembled the observed ones, making predictions for people that have missing values and generate parameter estimates that indicate that the missingness in the response is less likely to be MNAR. In contrast, the MICE Algorithm method produced different parameter estimates in terms of variance parameters.

It's worth noting that the proposed Two-Step method is computationally expensive because it involves applying the CRE method to multiple imputed datasets. As a result, it requires K computational time of the CRE method, and the computational time and storage space may be limited. However, this problem can be addressed by introducing joint modelling, which simultaneously handles ignorable missingness in the incomplete predictor and non-ignorable missingness in the model response. This can reduce the amount of time needed to run the method and eliminate the need for a pooling results step, as the joint modelling can consider uncertainty. This will be further explored in the upcoming Chapter 6. It's also important to test the proposed Two-Step method with misspecified missingness assumptions to determine how sensitive the conclusion is when the missingness mechanism is not assumed. This will be addressed in Chapter 8.

Chapter 6

Extension to the CRE Method with Ignorable Predictors (GCRE-MAR)

6.1 Introduction

This chapter will extend the Correlated Random Effects (CRE) method (Bhuyan, 2019) by incorporating missingness in the model predictor variables into the joint modelling framework. The CRE method has proven to be a valuable tool for analysing data with non-ignorable missingness in the model response (Bhuyan, 2019). However, its original version needs to be improved in its ability to accommodate scenarios where the model's predictors can have missing values, too. The proposed extension of the CRE method is specifically designed to handle ignorable missingness in the model predictor variable as well as the non-ignorable missingness in the model response by using Gibbs sampling. This approach simplifies the incorporation of the incomplete predictor model's conditional distribution into the CRE method, allowing us to simultaneously impute the missing predictor and response and analyse the main model.

It is not uncommon in longitudinal studies to have missing values in the model response as well as the model predictors due to various factors since the data collection involves repeatedly measuring model predictors across multiple time points, which can increase the possibility that the predictors contain missing data points. Thus analysing and imputing the time-varying predictors require further attention to account for the relationship between the time-variant predictors and the response. We explore this by developing an extended CRE joint model incorporating more realistic model predictor missingness assumptions. Since the missing predictor distributions are required in cases with incomplete predictors, in addition to a response model (Ma and Chen, 2018), incorporating the partially observed predictor variables into the joint modelling process enables us to capture intricate relationships and interdependencies that might exist among variables in the dataset.

There are two main approaches to handling missing values in the predictor variable in joint modelling. The first method uses multivariate distributions to model all missing predictors as linearly associated (Ma and Chen, 2018). The second method involves dividing the joint distribution into a sequence of one-dimensional conditional distributions for each missing predictor. This approach iteratively estimates missing predictor values based on the available observed data and the conditional distributions of other variables (Erler et al., 2016; Ibrahim et al., 2002), which is preferred when there are mixed types (Ibrahim et al., 2002) and a large number of missing predictors (Ma and Chen, 2018). This will be explained in the upcoming Section 6.2.

We propose a new method called Generalised Correlated Random Effects with a Missing at Random predictor (GCRE-MAR). This method does not distinguish between dropout or intermittent missingness, which provides flexibility in terms of missing data model specification. Additionally, it can handle both non-monotone and monotone missing data patterns. Our goal is to increase the applicability of the CRE method by creating the GCRE-MAR method and offering a solution to real-world data problems.

This chapter explores various aspects of the proposed method. The chapter begins with an explanation of the model itself in Section 6.2, followed by an explanation of the joint distribution in Section 6.3. Section 6.4 describes the prior distribution setup, and Section 6.5 illustrates the sampler algorithm, while Section 6.6 details the simulated data setup. The proposed method results are presented in Section 6.7, which includes simulated data results in Subsection 6.7.1 and results for an application to real data (the BIOSTAT-CHF data) are shown in Subsection 6.7.2. Finally, Section 6.8 compares the results of the GCRE-MAR method to the baseline methods, highlighting its advantages and drawbacks.

6.2 Proposed Model

In order to account for missingness in the model predictors, consider a continuous response measured over m different time points from n subjects and a set of predictors, some of which possibly have partially observed values. The response variable for the i^{th} subject at the t^{th} time point, which we denote by $Y_i(t)$ can thus be modelled as follows:

$$Y_i(t) = \mu + \sum_{j=1}^{J} \beta_j X_{ji}(t) + \sum_{j'=1}^{J'} \lambda_{j'} X_{j'i}(t) + u_i \tilde{Z}_i(t) + e_i(t), \qquad (6.2.1)$$

where J' and J represent the number of predictors of fixed effects that are fully observed and partially observed, respectively. μ is the fixed intercept representing the mean of the overall population. The j^{th} partially observed fixed effects coefficient is denoted by β_j, while the j'^{th} fully observed fixed effects coefficient is denoted by λ_j, $X_{ji}(t)$ is the value of the j^{th} partially observed fixed effect for subject i at time t and $X_{j'i}(t)$ is the value of the j'^{th} fully observed fixed effect for subject i at time t. Additionally, subject-specific random effects u_i are included to capture longitudinal dependence and are assumed to be i.i.d $N(0, \sigma_B^2)$ and $\tilde{Z}_i(t)$ is the value of the random effect for subject i at time t. The residuals $e_i(t)$ are also assumed to be i.i.d, following a Normal distribution as $N(0,\sigma_A^2)$. Similar to the CRE model discussed in Section 4.4.1, the binary missing response indicator $U_i(t)$ is defined as:

$$U_i(t) = \begin{cases} 1, & \text{if } Y_i(t) \ \ \textit{is obseverd}, \\ 0, & \text{if } Y_i(t) \ \ \textit{is missing}. \end{cases} \tag{6.2.2}$$

Consider a probit regression model (as described in Section 2.6) that defines the missingness as a normal distribution latent variable set to $U_i^*(t) > 0$ if $U_i(t) = 1$ and $U_i^*(t) \leq 0$ if $U_i(t) = 0$, which models the response missingness mechanism as follows:

$$U_i^*(t) = \tau + \sum_{j'=1}^{J'} \theta_{j'} X_{j'i}(t) + v_i \tilde{Z}_i(t) + \varepsilon_i(t), \tag{6.2.3}$$

where J' represents the number of fully observed predictors of fixed effects, $\theta_{j'}$ denotes the regression coefficients of the j'^{th} fully observed fixed effect expressing the systematic influence of missingness due to the unobserved response variable and τ is the fixed intercept representing the mean of the overall population. The subject-specific random effects, denoted by v_i, capture the longitudinal dependence and are assumed to be i.i.d following Normal distribution as $N(0,\sigma_C^2)$. The residuals, denoted by $\varepsilon_i(t)$, are also assumed to be i.i.d following Normal distribution as $N(0,1)$. Moreover, the random effects u_i and v_i are considered as correlated random vectors following a Multivariate Normal distribution with a mean vector of zeros and covariance matrix as $\Sigma = \begin{pmatrix} \sigma_B^2 & \sigma_D^2 \\ \sigma_D^2 & \sigma_C^2 \end{pmatrix}$, where σ_D^2 refers to the covariance between the random effects u_i and v_i.

The regression model given in Equation 6.2.1 can be rewritten as follows:

$$Y_i^*(t) = \mu + \sum_{j=1}^{J} \beta_j X_{ji}(t) + \sum_{j'=1}^{J'} \lambda_{j'} X_{j'i}(t) + u_i \tilde{Z}_i(t) + e_i(t), \tag{6.2.4}$$

where $Y_i^*(t)$ is a latent random variable set to $Y_i(t) = Y_i^*(t)$ if $U_i(t) = 1$ and $Y_i(t)$ is missing if $U_i(t) = 0$.

Suppose there are J partially observed predictors and $L - J$ (where $L = J + J'$) fully observed predictors. The joint distribution of all partially observed predictors factors as:

$$\begin{aligned}
&p(Y_i(t), X_{1i}(t), \ldots, X_{Ji}(t) \mid X_{(J+1)i}(t), \ldots, X_{Li}(t)) = \\
&p(Y_i(t) \mid X_{1i}(t), \ldots, X_{Ji}(t), X_{(J+1)i}(t), \ldots, X_{Li}(t)) \times \\
&p(X_{1i}(t), \ldots, X_{Ji}(t) \mid X_{(J+1)i}(t), \ldots, X_{Li}(t)),
\end{aligned} \tag{6.2.5}$$

where $p(Y_i(t), X_{1i}(t), \ldots, X_{Ji}(t) \mid X_{(J+1)i}(t), \ldots, X_{Li}(t))$ is the joint conditional distribution for all incomplete predictors and the response, $p(Y_i(t) \mid X_{1i}(t), \ldots, X_{Li}(t))$ is the distribution of $Y_i(t)$ and $p(X_{1i}(t), \ldots, X_{Ji}(t) \mid X_{(J+1)i}(t), \ldots, X_{Li}(t))$ is the conditional distribution of the incomplete predictors and represents the predictors model.

There are two suggestions to specify the conditional distribution of the incomplete predictor model. The first approach is called *separate specification* (Enders, 2022; Enders et al., 2020), which refers to independently specifying a univariate predictor model for each incomplete predictor. It assumes the conditional distribution $p(X_{1i}(t), \ldots, X_{Ji}(t) \mid X_{(J+1)i}(t), \ldots, X_{iL}(t))$ is a multivariate normal distribution, such that the incomplete predictors are linearly related. For multiple continuous incomplete predictors, we can use a multivariate normal distribution or a multivariate probit regression for binary incomplete predictors (Ma and Chen, 2018). Based on this assumption, we can specify the full conditional distribution for each incomplete predictor, given all other incomplete and complete predictors as a univariate normal distribution. The second approach is called *sequential specification* (Erler et al., 2016; Ibrahim et al., 2002; Lüdtke et al., 2020), and this can accommodate a non-linear relationship between the incomplete predictors. It factors the joint distribution of all incomplete predictors in a sequential manner of univariate distributions for each predictor, where each incomplete variable is modelled conditionally on other variables (Ibrahim et al., 2002) as follows:

$$
\begin{aligned}
&p(X_{1i}(t),\ldots,X_{Ji}(t) \mid X_{(J+1)i}(t),\ldots,X_{Li}(t)) = \\
&p(X_{Ji}(t) \mid X_{(J+1)i}(t),\ldots,X_{Li}(t)) \times p(X_{(J-1)i}(t) \mid X_{iJ},X_{(J+1)i}(t),\ldots,X_{Li}(t)) \times \\
&p(X_{(J-2)i}(t) \mid X_{Ji},X_{(J-1)i},X_{(J+1)i}(t),\ldots,X_{Li}(t)) \times \ldots\ldots\ldots\ldots \times \\
&p(X_{(1)i}(t) \mid X_{Ji},X_{(J-1)i},X_{(J-2)i}(t),X_{(J+1)i}(t),\ldots,X_{Li}(t)).
\end{aligned}
\tag{6.2.6}
$$

This approach is equivalent to *separate specification* when assuming a multivariate normal distribution for incomplete predictors, and it is preferred when dealing with mixed data types or high dimensions of incomplete predictors (Ma and Chen, 2018). Implementing *sequential specification* is generally complicated because it requires working out how to factorize the joint distribution to achieve the desired model, where the order of the conditional distributions will lead to different joint distributions (Ma and Chen, 2018).

The recommendation of the order of the predictors is to condition the categorical variables on the continuous variables and according to the percentages of the missing value so that it starts with the variable that has the lowest percentage of missing values (Erler et al., 2016). Nevertheless, the results would be unbiased regardless of the order of the conditional distribution of missing predictors whenever the models fit the data well (Zhu and Raghunathan, 2015). On the other hand, *separate specification* only needs to specify the required univariate predictor model and nothing else.

In this study, we will use *separate specification* for the time-varying partially observed (incomplete) predictor model for the reason that we are dealing with a linear relationship between variables and only have a few predictors to deal with. We are specifically working with one incomplete continuous predictor. This approach allows us to handle time-varying predictors that are measured on the same time scale as the outcome. The model can be specified similarly to a Linear Mixed Model (LMM) as follows:

$$X_{ji}(t) = \delta + \sum_{j'=1}^{J'} \alpha_{j'} X_{j'i}(t) + w_i \tilde{Z}_i(t) + r_i(t), \tag{6.2.7}$$

where $X_{ji}(t)$ is the j^{th} partially observed predictor for subject i at time t. $\alpha_{j'}$ denotes the regression coefficients of the j'^{th} fully observed fixed effects, δ is the fixed intercept representing the mean of the overall population. The random intercept $w_i \overset{\text{i.i.d}}{\sim} N(0, \sigma_E^2)$ and the residuals $r_i(t) \overset{\text{i.i.d}}{\sim} N(0, \sigma_F^2)$. This is known as a *separate specification* because each incomplete predictor requires a unique regression model. In our study, we will assume one partially observed predictor and two fully observed predictors.

In order to estimate the values of partially observed predictors $X_{ji}(t)$, we need to derive its conditional distribution given all other predictor variables as discussed. Suppose the missing response indicator $U_i^*(t)$ in Equation 6.2.3 is conditional on the incomplete predictors. In that case, the posterior distribution of the incomplete predictors $X_{ji}(t)$ should consider the influence of the missing indicator variable $U_i^*(t)$. However, this assumption may cause a collinearity problem (Du et al., 2022). Thus, we consider independence between incomplete predictors $X_{ji}(t)$ and the missing response indicator $U_i^*(t)$.

6.2.1 GCRE-MAR Assumptions

We assume that some predictors in the analysis model in Equation 6.2.1 are partially observed, and their missing values depend on other observed predictors under the MAR missingness mechanisms assumption. Furthermore, the model in Equation 6.2.3 expresses the systematic influence of missingness due to the unobserved response variables, and we posit the missingness of the response does not depend on the partially observed predictors for simplicity in deriving the complex conditional distribution of the partially observed predictors.

6.3 Joint Distribution

We propose a Bayesian estimation method that can simultaneously estimate the parameters associated with the joint model using Gibbs sampling. To distinguish between the observed and partially observed predictor, as mentioned previously, we use $X_{ji}(t)$ for a predictor that contains missing values (partially observed predictor) and $X_{j'i}(t)$ for the fully observed predictor (predictor that doesn't contain any missing values). Similar to Section 4.4.2, let $\boldsymbol{X}_j = (X_{j11}(t),\ldots,X_{jnm}(t))$ and $\boldsymbol{X}_{J'} = (X_1^T,\ldots,X_{J'}^T)$. $\boldsymbol{X}_{J'}$ is a matrix containing each fully observed predictors as a vector. The joint posterior density for the latent variables and the parameters associated with the proposed model is:

$$p\left(\Theta_{Y,U},\Theta_{x_{mis}},\boldsymbol{X}_j,\boldsymbol{Y}^*,\boldsymbol{U}^* \mid \boldsymbol{Y},\boldsymbol{U},\boldsymbol{X}_{J'}\right) \propto p\left(\Theta_{Y,U}\right)\times p\left(\Theta_{x_{mis}}\right)\times$$
$$\prod_{i=1}^{n}\int_{-\infty}^{\infty}\prod_{t=1}^{m} f\left(Y_i^*(t),U_i^*(t) \mid u_i,v_i,\boldsymbol{X}_j,\boldsymbol{X}_{J'}\right)\times f\left(\boldsymbol{X}_j \mid w_i,\boldsymbol{X}_{J'}\right)\times f(w_i)\times$$
$$\{I\left(U_i^*(t)>0\right)I\left(U_i(t)=1\right)+I\left(U_i^*(t)\leq 0\right)I\left(U_i(t)=0\right)\}\times g\left(u_i,v_i\right)du_idv_idw_id\boldsymbol{X}_j, \tag{6.3.1}$$

where $\Theta_{Y,U} = \{\mu,\boldsymbol{\beta},\boldsymbol{\lambda},\tau,\boldsymbol{\theta},\sigma_A^2,\Sigma\}$ and $\Theta_{x_{mis}} = \{\delta,\boldsymbol{\alpha},\sigma_E^2,\sigma_F^2\}$. The joint prior is denoted by $p(\Theta)$, the joint distribution of Y^* and U^* is denoted by $f(Y^*,U^*)$, the joint distribution of u_i and v_i is represented by $g(u_i,v_i)$, and is a multivariate normal distribution $N(0,\Sigma)$, the conditional distribution of partially observed predictor $\boldsymbol{X}_j$ is represented by the predictor model $f\left(\boldsymbol{X}_j\right)$, $f(w_i)$ is the incomplete predictor model random intercept distribution and $I(A)$ is an indicator variable which takes value 1 if A occurs and zero otherwise.

6.4 Prior Distribution

In this section, we will consider the following priors for $\Theta_{Y,U}$ and $\Theta_{x_{mis}}$:

$$\begin{aligned} &p\left(\tilde{\boldsymbol{\beta}},\sigma_A^2\right) \propto \frac{1}{\sigma_A^2};\ p\left(\tilde{\boldsymbol{\theta}}\right) \propto N(0,b);\ p\left(\Sigma\right) \propto IW(\nu,\Lambda); \\ &p\left(\tilde{\boldsymbol{\alpha}},\sigma_F^2\right) \propto \frac{1}{\sigma_F^2};\ p(\sigma_E^2) \propto IG(b_0,b_0), \end{aligned} \tag{6.4.1}$$

where $\tilde{\boldsymbol{\beta}} = [\mu, \boldsymbol{\beta}, \boldsymbol{\lambda}]$ is a vector of the overall intercept and regression coefficients of fully and partially observed predictors in the response model, $\tilde{\boldsymbol{\theta}} = [\tau, \boldsymbol{\theta}]$ is a vector of the overall intercept and regression coefficients of fully observed predictors in the missingness model and $\tilde{\boldsymbol{\alpha}} = [\delta, \boldsymbol{\alpha}]$ is a vector of the overall intercept and regression coefficients of fully observed predictors in the incomplete predictor model. To ensure convergence, we use weakly informative priors for the matrix variable $p\left(\Sigma\right)$ that we discovered and resolved in Chapter 4 and the missingness model coefficients $p\left(\tilde{\boldsymbol{\theta}}\right)$.

Prior information is required to support convergence in the missingness model coefficients as suggested by Du et al. (2022) and in practice using the GCREMAR method, so we set a relatively large prior variance value for this purpose $b = 10$. When we tested values of $b < 5$, the inference became more sensitive to the choice of the prior, which had a stronger influence on the results. However, for values of $b > 5$, the inference was less affected by the prior choice, allowing the data to have a greater influence on the posterior estimates. Therefore, we selected $b = 10$ to ensure that the prior remained non-informative. In Section 4.4.5, we discussed that the Inverse Wishart (IW) distribution solved the non-convergence of the matrix variable and the hyperparameters are set to $\Lambda = (\nu - p - 1)I$, with $\nu = 4$.

For other priors, we use non-informative priors. We use the Inverse-Gamma (IG) distribution as a conjugate prior for the variance parameter of normally distributed data $p(\sigma_E^2)$. Lower values of b_0 lead to non-informativeness, and as b_0 gets closer to zero, it may produce an improper posterior density (Gelman, 2006). Therefore, we assume a reasonable value of $b_0 = 1$. Overall, we chose conjugate priors, which allow us to derive the full conditional distributions of each variable in a closed form and implement the Gibbs sampler.

6.5 Gibbs Sampler

The advantage of factorising the joint distribution of data is that the estimation of the parameters of interest is computed within each iteration of the imputation procedure and is conditional on the imputed variables' current value. The simultaneity of the analysis and the imputation produce a posterior distribution of the parameters, which automatically considers the uncertainty due to the missing values, and no pooling and further analysis are required, which is often required for other multiple imputation approaches. Moreover, the response variable is not included in the incomplete predictor model because the relationship between the incomplete predictor and the response is considered in the joint likelihood of the data.

The first steps are the 1-7 steps outlined in Section 4.4.4, followed by generating estimates for the predictor model's random components, coefficients, and missing predictor values. Additional steps to the Gibbs sampler in Section 4.4.4 include:

8. Estimate the predictor model's random intercept, w_i.
9. Estimate the predictor model's random intercept variance, σ_E^2.
10. Estimate the predictor model's regression coefficients, $\tilde{\boldsymbol{\alpha}}$.
11. Estimate the predictor model's residual variance, σ_F^2.
12. Estimate the missing values of the incomplete predictor, $X_{ji}(t)$.

The entire structure of the Gibbs sampler is given in Algorithm 7 overleaf. In each iteration, the missing response and predictor values will be simulated simultaneously. This allows us to obtain direct and joint estimates from the posterior distributions of the parameters and of the missing values using the GCRE-MAR method.

Algorithm 7 GCRE-MAR Method Gibbs Sampling Algorithm

```
Choose initial values of $\{\Theta^0_{Y,U} = \tilde{\boldsymbol{\beta}}^0, \tilde{\boldsymbol{\theta}}^{\mathbf{0}}, \sigma_A^{2^0}, \Sigma^0\}$, $\{\Theta^0_{mis} = \tilde{\boldsymbol{\alpha}}^{\mathbf{0}}, \sigma_F^{2^0}, \sigma_E^{2^0}\}$ and $\{X_j^0, w^0, u^0, v^0, Y^{*0}, U^{*0}\}$.
for $1,\ldots,S$ iterations do
  -Sample $u_i^{S+1} \sim p\left(u_i \mid Y_i^{*^S}(t), \tilde{\boldsymbol{\beta}}^S, \sigma_A^{2^S}, \Sigma^S, v_i^S, \boldsymbol{X}_{J'}\right)$.
  -Sample $Y_i^{*^{S+1}}(t) = \begin{cases} Y_i^{*^S}(t), & \text{if } Y_i(t) \textit{ is observed} \\ p\left(Y_i^*(t) \mid \tilde{\boldsymbol{\beta}}^S, X_{ji}^S(t), u_i^{S+1}, \sigma_A^{2^S}, \boldsymbol{X}_{J'}\right), & \text{if } Y_i(t) \textit{ is missing.} \end{cases}$
  -Sample $\sigma_A^{2^{S+1}} \sim p\left(\sigma_A^2 \mid \tilde{\boldsymbol{\beta}}^S, X_{ji}^S(t), Y_i^{*^{S+1}}(t), u_i^{S+1}, \boldsymbol{X}_{J'}\right)$.
  -Sample $\tilde{\boldsymbol{\beta}}^{S+1} \sim p\left(\tilde{\boldsymbol{\beta}} \mid Y_i^{*^{S+1}}(t), u_i^{S+1}, X_{ji}^S(t), \sigma_A^{2^{S+1}}, \boldsymbol{X}_{J'}\right)$.
  -Sample $v_i^{S+1} \sim p\left(v_i \mid U_i^{*^S}(t), \tilde{\boldsymbol{\theta}}^S, \Sigma^S, u_i^{S+1}, \boldsymbol{X}_{J'}\right)$.
  -Sample $U_i^{*^S}(t) = \begin{cases} p\left(U_i^*(t) \mid \tilde{\boldsymbol{\theta}}^S, v_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ left truncated}^* \text{ at } 0, & \text{if } Y_i(t) \textit{ is observed}, \\ p\left(U_i^*(t) \mid \tilde{\boldsymbol{\theta}}^S, v_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ right truncated}^* \text{ at } 0, & \text{if } Y_i(t) \textit{ is missing.} \end{cases}$
  -Sample $\tilde{\boldsymbol{\theta}}^{S+1} \sim p\left(\tilde{\boldsymbol{\theta}} \mid U_i^{*^{S+1}}(t), v_i^{S+1}, \boldsymbol{X}_{J'}\right)$.
  -Sample $\Sigma^{S+1} \sim p\left(\Sigma \mid u_i^{S+1}, v_i^{S+1}\right)$.
  -Sample $w_i^{S+1} \sim p\left(w_i \mid X_{ji}^S(t), \sigma_F^{2^S}, \sigma_E^{2^S}, \boldsymbol{X}_{J'}\right)$.
  -Sample $\sigma_E^{2^{S+1}} \sim p\left(\sigma_E^2 \mid w_i^{S+1}, \boldsymbol{X}_{J'}\right)$.
  -Sample $\sigma_F^{2^{S+1}} \sim p\left(\sigma_F^2 \mid X_{ji}^S(t), \tilde{\boldsymbol{\alpha}}^{\boldsymbol{S}}, w_i^{S+1}, \boldsymbol{X}_{J'})\right)$.
  -Sample $\tilde{\boldsymbol{\alpha}}^{S+1} \sim p\left(\tilde{\boldsymbol{\alpha}} \mid X_{ji}^S(t), \sigma_F^{2^{S+1}}, w_i^{S+1}, \boldsymbol{X}_{J'}\right)$.
  -Sample $X_{ji}^{S+1}(t) = \begin{cases} X_{ji}^S(t), & \text{if } X_{ji}(t) \textit{ is observed}, \\ p\left(X_{ji}(t) \mid Y_i^{*^{S+1}}(t), \tilde{\boldsymbol{\beta}}^{S+1}, u_i^{S+1}, \tilde{\boldsymbol{\alpha}}^{S+1}, w_i^{S+1}, \sigma_A^{2^{S+1}}, \sigma_F^{2^{S+1}}, \boldsymbol{X}_{J'}\right), & \text{if } X_{ji}(t) \textit{ is missing.} \end{cases}$
```

* truncated normal distribution.

6.6 Creating Synthetic Data for Simulation

The aim of creating simulated data is to evaluate the performance of the proposed GCRE-MAR method in dealing with non-ignorable missingness in the response and ignorable missingness in the predictor. The simulation setup employed in this chapter aligns with the simulated data detailed in Section 5.5, with the exception of the form of the response missingness process model. In this case, we assume that the incomplete predictor is independent of the missing response model for simplicity, as mentioned before. Therefore, the missing values on the response variable $Y_i(t)$ were generated based on the following model:

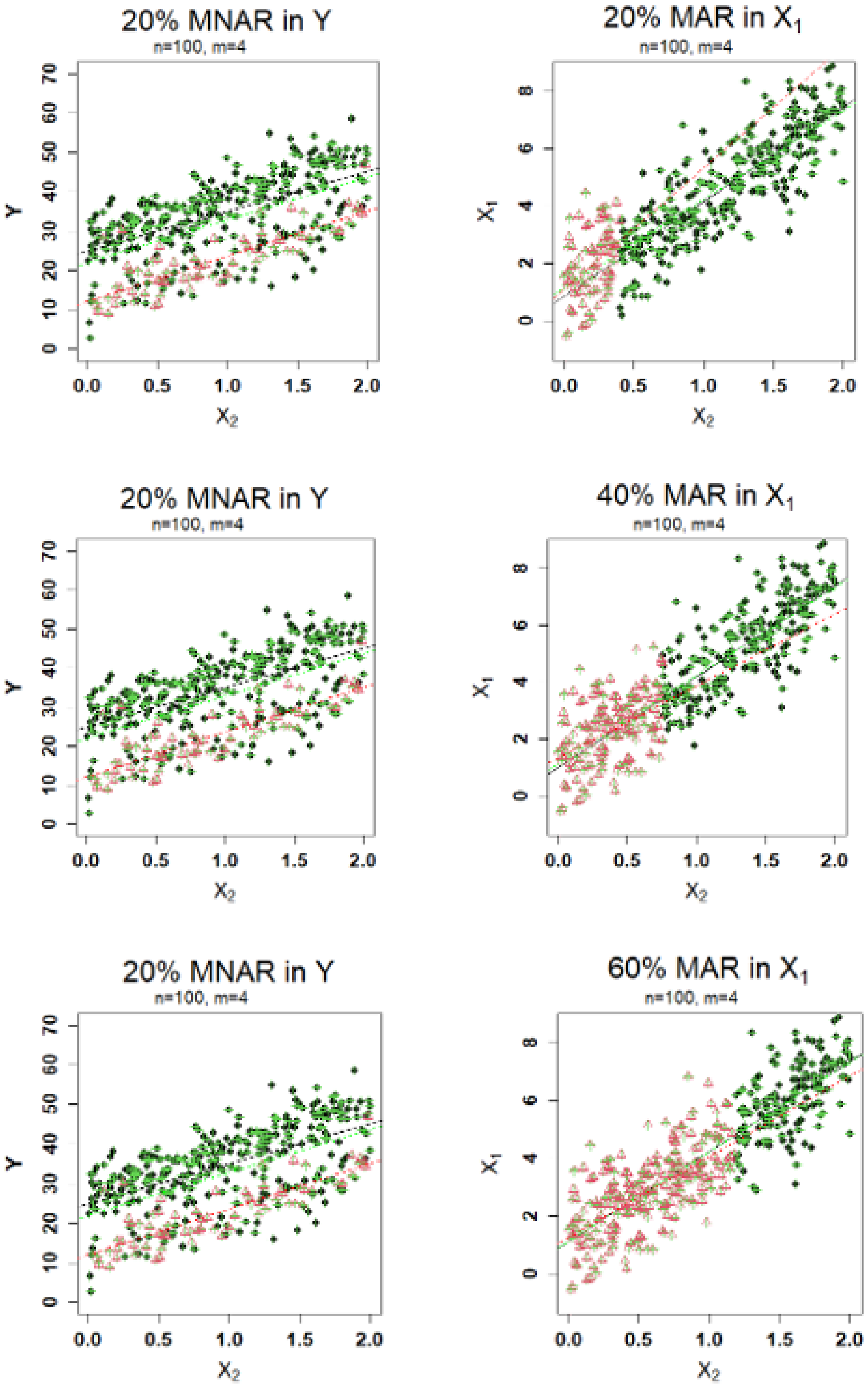
$$U_i^*(t) = \theta_0 + \theta_1 X_{2i}(t) + \theta_2 X_{3i}(t) + v_i \tilde{Z}_i(t) + \varepsilon_i(t) \tag{6.6.1}$$

Table 6.6.1 displays the values of $\boldsymbol{\theta}$ for each missingness proportion and the number of repeated measures to achieve the desired missing data proportion.

	$\mathbf{p_y = 20\%}$			$\mathbf{p_y = 40\%}$			$\mathbf{p_y = 60\%}$		
Parameter	m=2	m=4	m=8	m=2	m=4	m=8	m=2	m=4	m=8
θ_0	-2	-0.6	-0.6	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5
θ_1	3	0.7	0.7	0.4	0.4	0.4	0.4	0.4	0.4
θ_2	7	4	4	2	2	2	1.2	1.2	1.2

Table 6.6.1: The table presents the values of the coefficients $\boldsymbol{\theta}$ for the missing response indicator regression model, for various proportions of missingness and a number of repeated measures. This information is used to generate missing response values for the GCRE-MAR method, where $\mathbf{p_y}$ represents the percentage of missing values in the model response, and m indicates the number of repeated measures.

The missingness of the generated data is shown across different percentages of incomplete predictors and responses in Figure 6.6.1. The observed and missing values of the simulated datasets are illustrated for four repeated measures. Missing values in the response occur when the response variable values themselves are lower, confirming that MNAR is the missing mechanism. Additionally, the missing values in the incomplete predictor X_1 are associated with small values in the continuous complete predictor X_2, indicating that MAR is the missing mechanism. Comparable patterns are found in different values of the repeated measures $m = 2$ & 8.



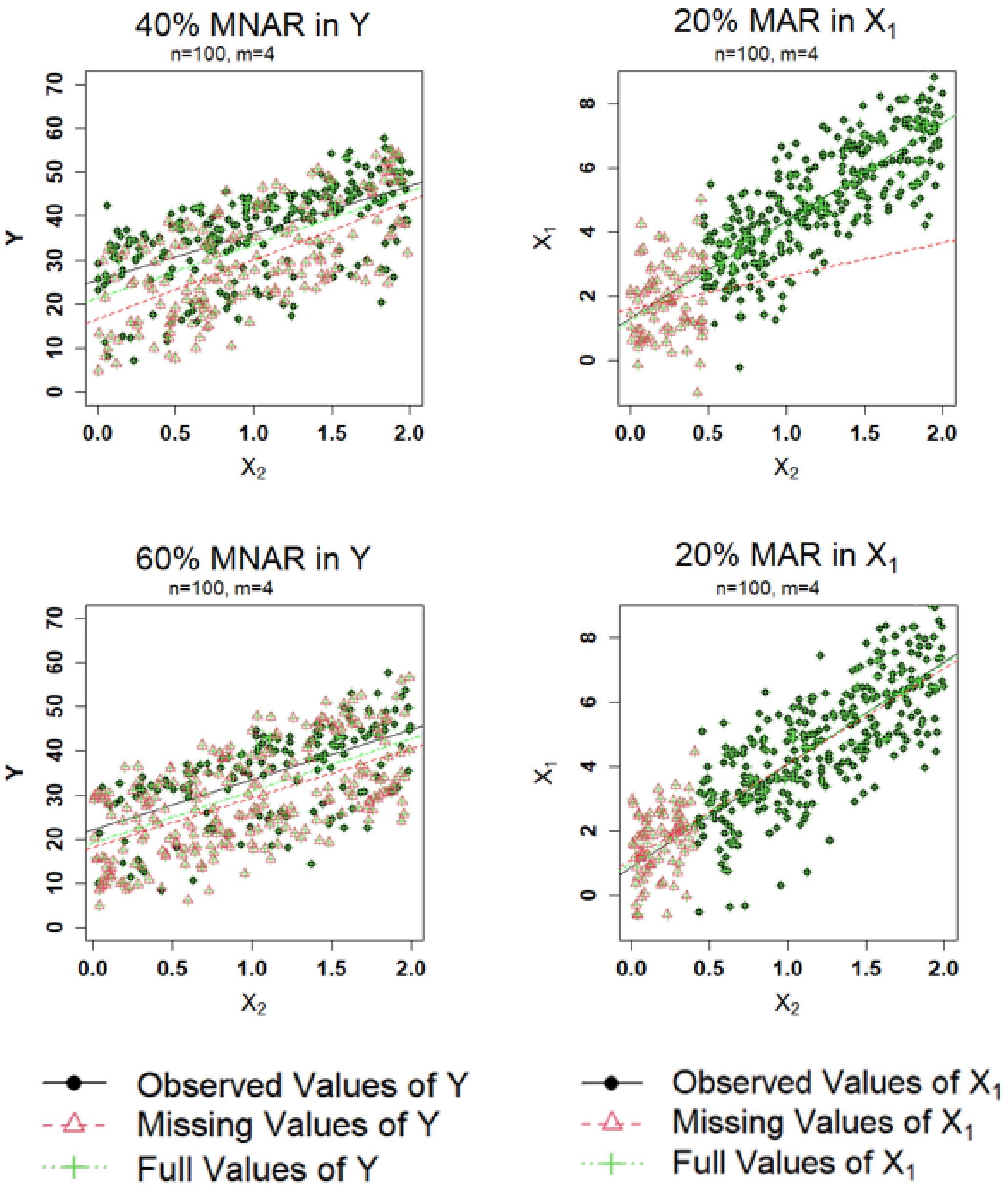


Figure 6.6.1: Scatter plot showing the association between the response (y-axis) and the complete predictor (x-axis) on the left-hand side plots and between the incomplete predictor (y-axis) and the complete predictor (x-axis) on the right-hand side plots. The black dots represent observed values after removing missing data (i.e. the available data), while the red triangles represent missing values. The green crosses show the full data, including both observed and missing values. The plot also displays three regression lines, representing different observed and missing values and illustrating the varying trends for MNAR in the response and MAR in the incomplete predictor. The missingness in the response is dependent on the small values of the response itself, while the incomplete predictor's missingness is dependent on the small values of the complete predictor. Both represent the desired missingness mechanism.

We will test the performance of the GCRE-MAR method using the simulated data, which allows us to compare the proposed method with other baseline methods, which are the full data and the available data. The full data uses the fully observed data before eliminating the missing values, representing the best-case scenario. The available data method involves using only the observed values after eliminating any missing values in the response or predictor variables. Additionally, we will compare the GCRE-MAR method with the Two-Step method and the imputation method that uses the MICE Algorithm to impute the model response and predictor; these are discussed in Chapter 5. We will also compare it with the CRE method with fully observed predictors as discussed in Chapter 4. Moreover, we will apply the GCRE-MAR method with fully observed predictors to compare the results of the generalised version of the CRE method with the original CRE method and to assess the performance of the GCRE-MAR method when there are no missing values in the predictor variable.

All of the methods are implemented in `R` (R Core Team et al., 2013) and Gibbs sampler is carried out for the proposed method. The `brm` function (Bürkner, 2017) is used to fit the full and available data methods using the HMC method as defined in Section 2.4.5, while the MICE method pooled results are fitted using the `brm_multiple` function. The MCMC simulations will run for 50,000 iterations, with half of that designated as a burn-in phase and a ten-thinning rate applied. To ensure convergence, three chains are initialized with different starting values to calculate the Gelman-Rubin (explained in Section 2.4.6) convergence statistic for individual parameters. We consider convergence achieved if all values were below 1.1. We also visually inspect the trace plots. Based on these criteria, we deemed that all discussed runs are converged.

We will evaluate the GCRE-MAR method's performance by comparing the accuracy of its parameter estimates. We will use the Root Mean Square Error (RMSE), Relative Bias (RB), and Coverage Rate (CR) as measures of accuracy for individual parameters. However, different parameter magnitudes could distort the results and hence to ensure equal weighting across all model of interest parameters, we will use Weighted Root Mean Square Errors (WRMSE) for overall method accuracy, in which we assigned weights based on the data-generating parameters. We defined these criteria in Section 2.7, and we will evaluate them across 100 replications of the simulated data.

6.7 Results

This section presents the results of the proposed GCRE-MAR method and the comparative methods, applied to simulated data and the BIOSTAT-CHF dataset.

6.7.1 Simulation Data Results

The simulated data values and average imputed values of each iteration of the Gibbs sampler using the GCRE-MAR method of the response and predictor closely match across different sample sizes and the proportion of missingness, indicating alignment. Moreover, the simulated data values fall within the ±2 Standard Deviation (SD) range of the average imputed values. Figure 6.7.1 demonstrates one scenario out of the 100 replications where the proposed method imputed values match the simulated data values. However, similar conclusions were drawn for the remaining 99 replications and different combinations of missingness proportion. These results demonstrate the effectiveness and reliability of the proposed GCRE-MAR method in capturing the data-generating values that were not observed.

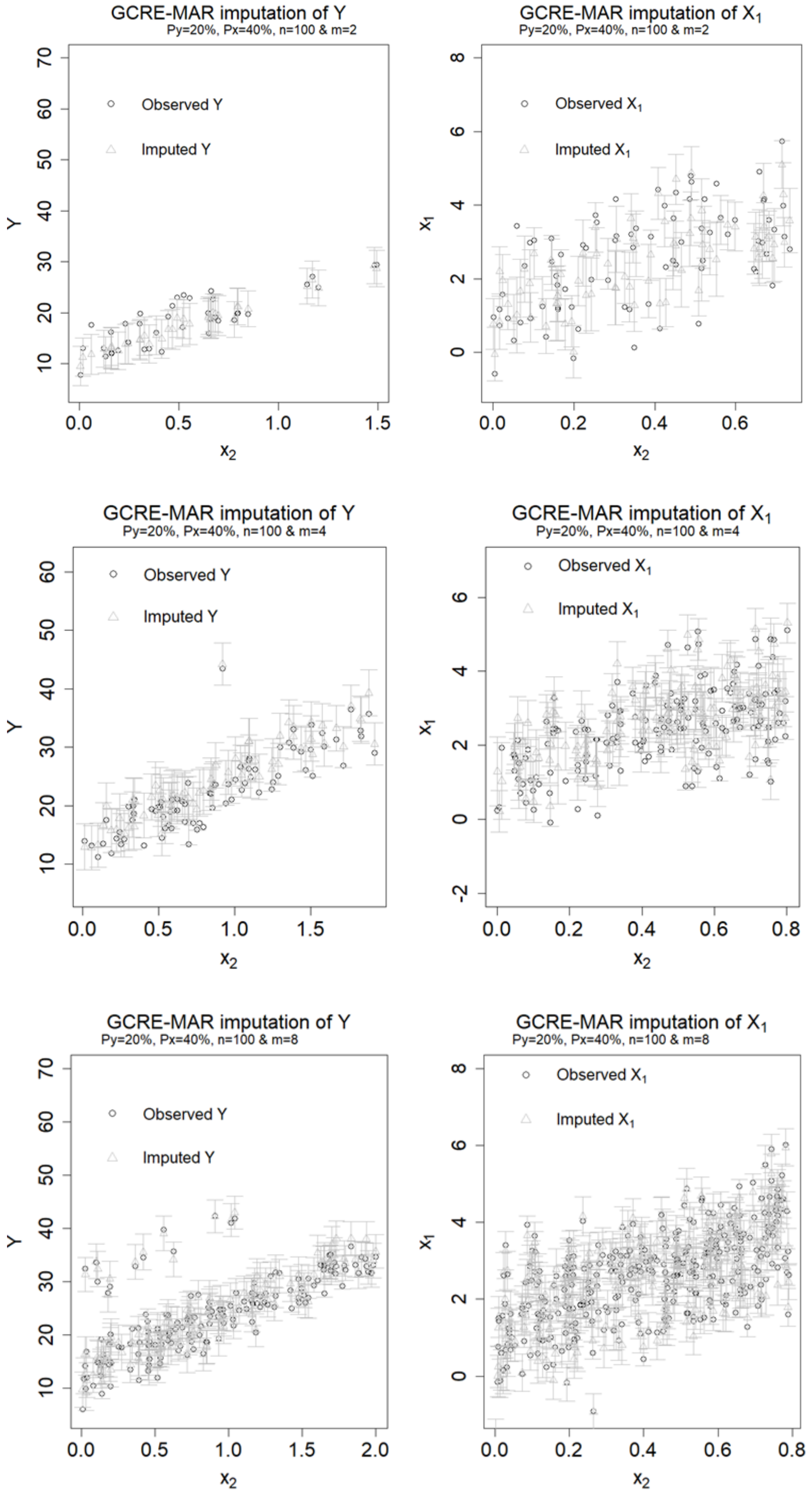


Figure 6.7.1: Scatter plots represent the model response Y on the y-axis and the fully observed predictor X_2 on the x-axis (on the left-hand side). The model incomplete predictor X_1 on the y-axis and the fully observed predictor X_2 on the x-axis (on the right-hand side). These plots are used to assess the simulated data values (depicted as black circles) against the average of imputed values and the ± 2 SD (represented by grey triangles and vertical lines) using the GCRE-MAR method for various repeated measures, with 20% missingness in the model response and 40% missingness in the model predictor. The simulated data values are mostly enclosed within the imputed values region.

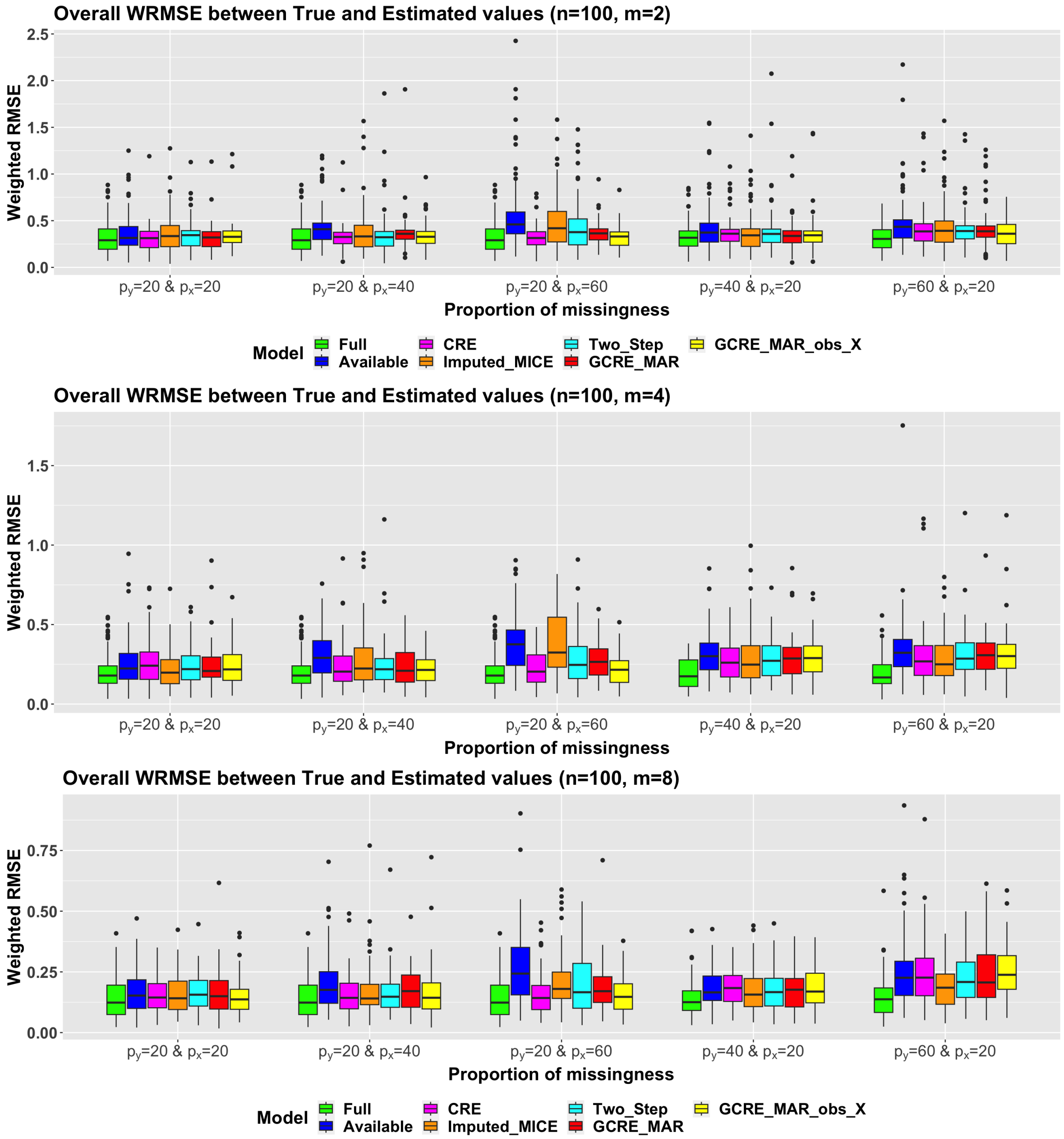


Figure 6.7.2: Boxplots represent the overall WRMSE for each method across different proportions of missing data and repeated measures. The y-axis represents the WRMSE values, while the x-axis represents one of the proportions of missing data in the model response and incomplete predictor. Each boxplot corresponds to one of the applied methods, and each plot represents different repeated measures. The results show that all methods have similar WRMSE, with a slight increase in the available data method.

The overall WRMSE across all model parameters in Figure 6.7.2, illustrates the method's performance across various sample sizes and proportions of missing data. When there is 20% missingness in both the model response and incomplete predictor, the GCRE-MAR method WRMSE aligns with the available data method and has less WRMSE uncertainty with $m = 2$ & 4. Meanwhile, the available data, Two-Step, and MICE Algorithm methods have larger WRMSE values and uncertainty when the incomplete predictor has a higher proportion of missingness (60%). Overall, the available data method has slightly larger WRMSE values than the proposed GCRE-MAR method.

Considering the apparent overall inequalities of WRMSE across methods, let's explore the specific differences across each model parameter. The following plots explore these differences in detail. The RMSE analysis provides insights into the accuracy of our estimates relative to the data-generating parameters, indicating the degree of deviation between the estimated and data-generating parameters.

To illustrate our findings, Figure 6.7.3 shows the RMSE for β_0 and β_1, while β_2 and β_3 are available in Section C in the Appendix. The RMSE of the response model coefficients are higher in the available data method, particularly with 60% missingness in the incomplete predictor. The CRE method has smaller model coefficients' RMSE than the GCRE-MAR method, except for β_3. This is because the CRE method is working with more observed data than the GCRE-MAR method. Recall that the CRE method cannot handle scenarios where there is missingness in the predictors, so we have run the method with missingness in the response only. This presents an easier scenario for the CRE method and it is able to perform slightly better. Nevertheless, it is a useful benchmark for the GCRE-MAR method. When there is 60% missingness in the incomplete predictor, the Two-Step and MICE Algorithm methods have a higher RMSE than the GCRE-MAR method. The GCRE-MAR method, when applied to scenarios with no missingness in the predictors, shows a lower RMSE in the model coefficients, except similar results for β_3. This is because the method considers more observed data without considering any missingness in the model predictor.

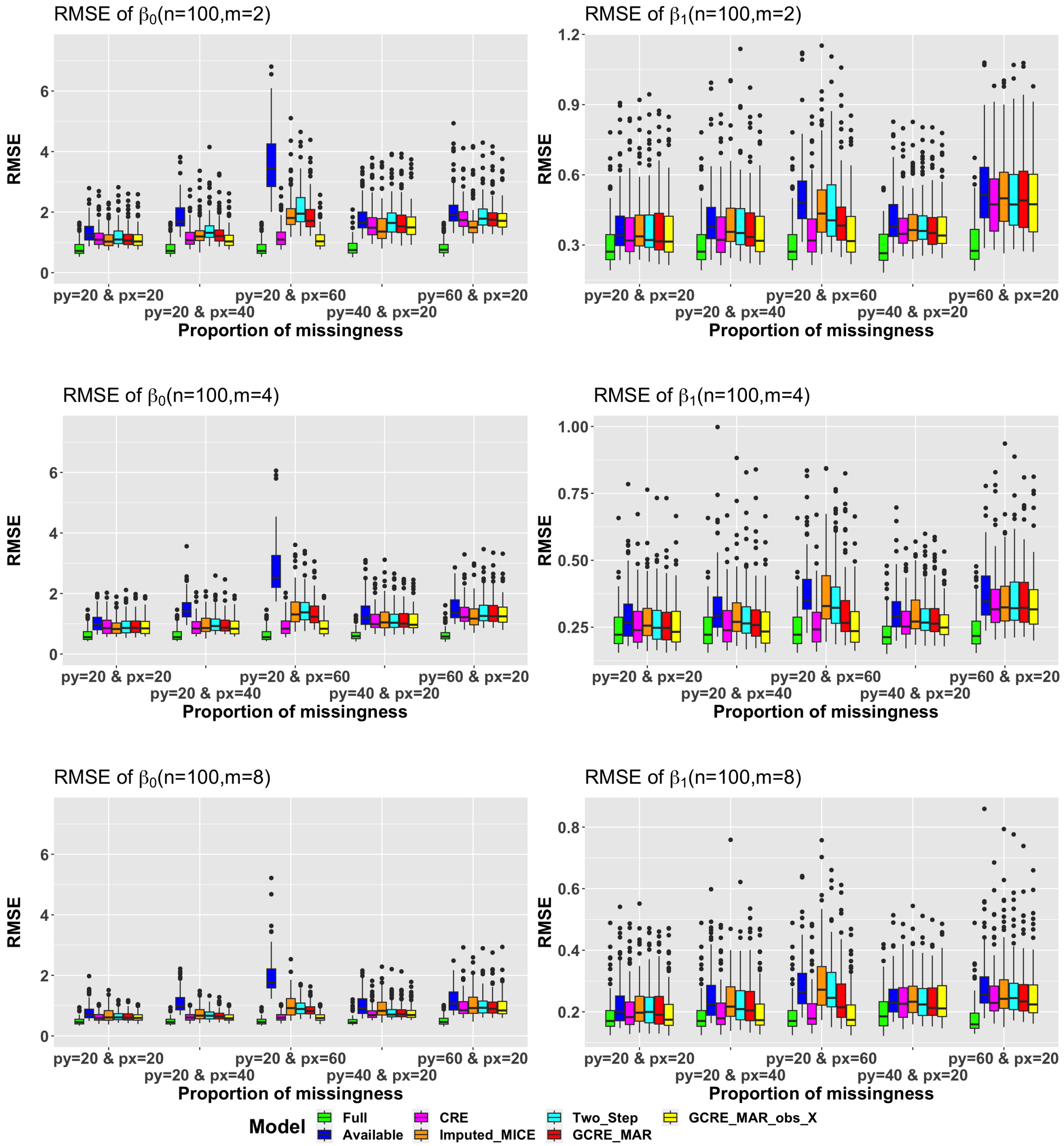


Figure 6.7.3: Boxplots illustrate the RMSE of β_0 in the left-hand side plots, and β_1 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RMSE values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. It's obvious that the available data method has quite larger RMSE values than other methods.

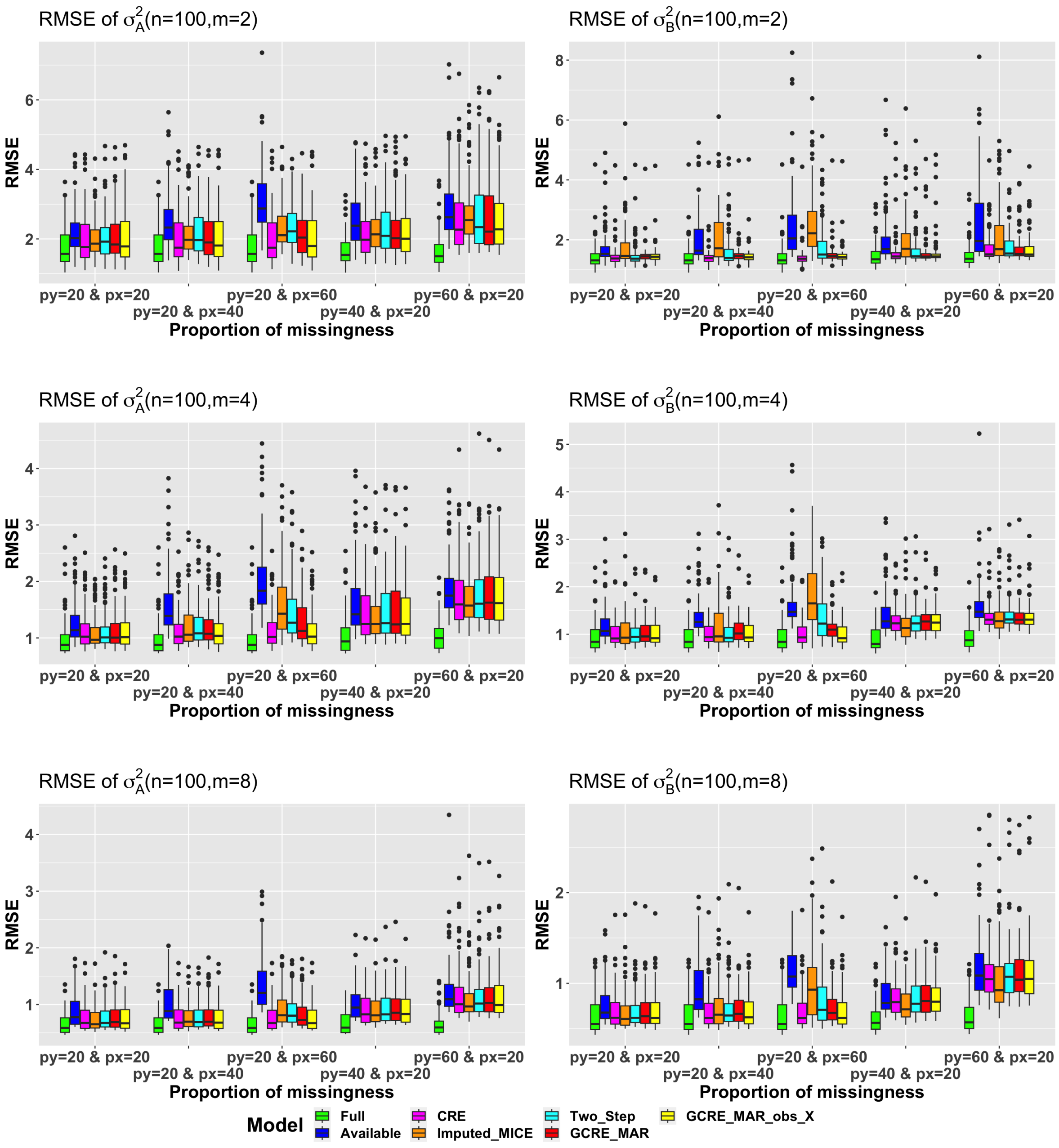


Figure 6.7.4: Boxplots illustrate the RMSE of σ_A^2 in the left-hand side plots, and σ_B^2 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RMSE values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. It is evident that the available data method produces larger RMSE values compared to other methods and is even larger in σ_B^2.

The RMSE values of the variance components in the response model indicate that the variance components have a large RMSE in the available data method compared with other applied methods. This increases further when 40% and

60% are missing in the incomplete predictor. The GCRE-MAR method with no missingness in the predictors, and the CRE method have a similar RMSE of the variance components as the GCRE-MAR method. However, the Two-Step and MICE imputed methods have a large RMSE of the variance components in 60% missingness in the incomplete predictor. Additionally, the available data and the MICE imputed methods has a larger RMSE of the between-individual variance with $m = 2$. These findings are presented in Figure 6.7.4.

Figure 6.7.5 is used to showcase the findings for β_0 and β_1 parameters. Note that Section C in the Appendix contains RB of β_2 and β_3. The RB assesses the model estimates for tendencies of overestimation or underestimation, highlighting nuances in the accuracy of estimating the data-generating parameters across all applied methods. The response model's slope parameters β_1 and β_2 have similar RB of the available data method and the GCRE-MAR method. However, there is a larger RB uncertainty of the available data method with 60% missingness in the incomplete predictor, and it overestimates β_3 in cases where $m = 4\,\&\,8$. The response model intercept is underestimated using the available data method when there are large proportions of missingness in the response variable (40% and 60%) with $m = 4\,\&\,8$, and as the proportion of missingness in the incomplete predictor increases, the RB uncertainty increases using the available data method. The CRE, Two-Step, MICE Algorithm, observed predictor GCRE-MAR, and GCRE-MAR have similar RB, except the MICE Algorithm tends to underestimate β_0 and overestimate β_3 with $m = 4\,\&\,8$.

The RB values of the variance components in the response model in Figure 6.7.6 indicate that the σ_A^2 is being overestimated with $m = 2$ while the σ_B^2 is being underestimated when using the GCRE-MAR method and this bias decreases with the increase of repeated measures. This estimation is comparable to the observed predictor GCRE-MAR and CRE methods. Although the available data provides an unbiased estimation of the variance component, it has a larger RB uncertainty. The MICE Algorithm method tends to overestimate the σ_B^2 and underestimate the σ_A^2 with larger RB uncertainty, except when $m = 8$. On the other hand, the Two-Step method tends to overestimate the σ_B^2 and underestimate the σ_A^2 when $m = 4$ and $m = 8$. The MICE Algorithm and the Two-Step methods tend to underestimate σ_A^2 and overestimate σ_B^2 with a large proportion of missingness in the incomplete predictor (40% and 60%); however, this bias decreases as repeated measures increase.

Our assessment of parameter estimation accuracy of the response model parameters has revealed valuable insights into the performance of the GCRE-MAR method compared to other methods used. The GCRE-MAR method has been shown to have a smaller RMSE than the available data method, which is typically used when dealing with longitudinal data. This indicates that the GCRE-MAR method has an overall superior precision in estimating parameters. It's worth noting that the GCRE-MAR method and available data method demonstrate unbiased estimation according to the relative bias analysis. The smaller RMSE of the GCRE-MAR method suggests that, on average, its estimates are closer to the data-generating parameters compared to the available data method. This indicates that the GCRE-MAR method is more effective in minimizing the spread of estimation errors. Additionally, the unbiased estimation revealed by the relative bias for the GCRE-MAR method underscores its consistency in providing estimates that are centred around the data-generating values.

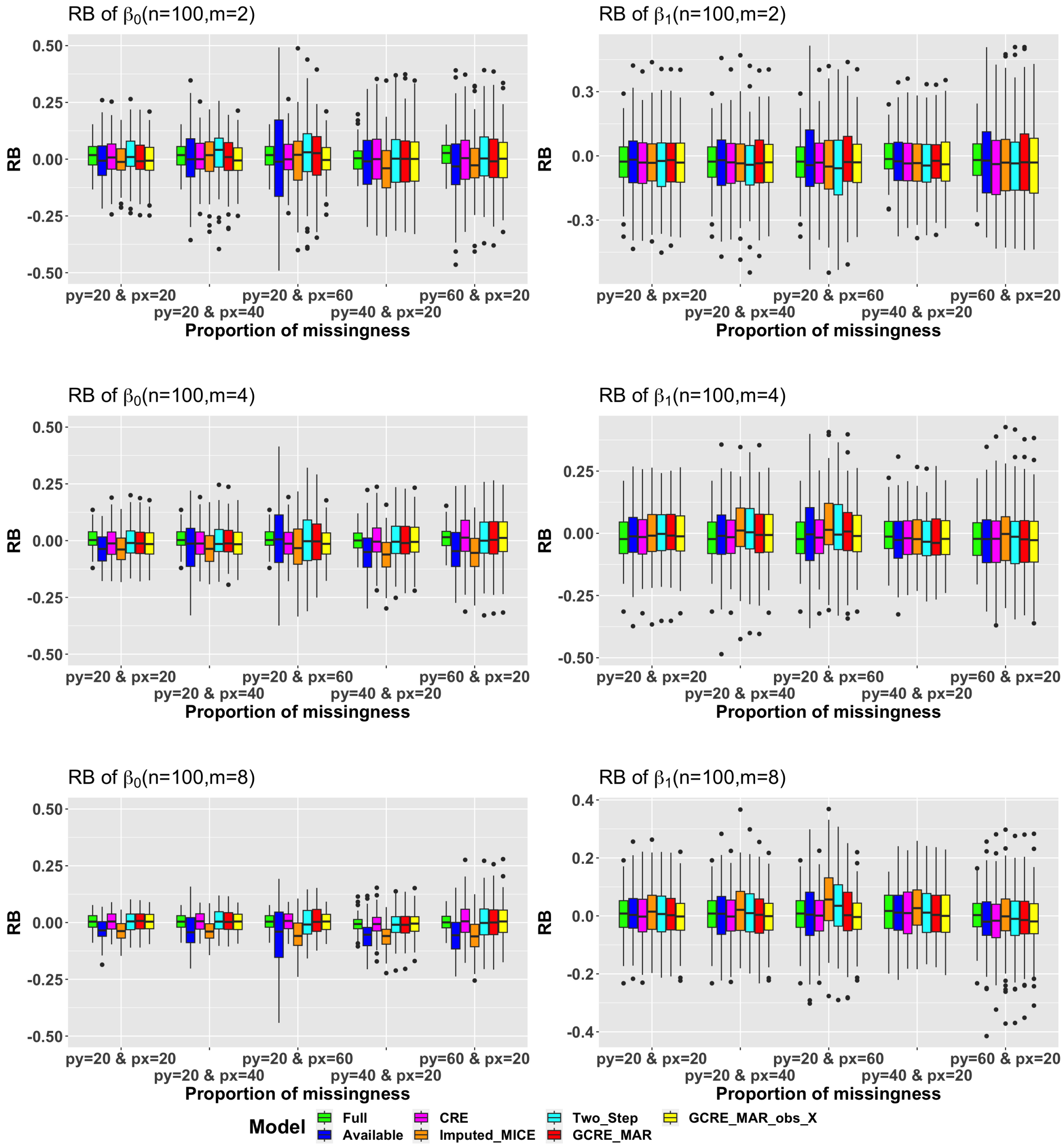


Figure 6.7.5: Boxplots illustrate the RB of β_0 in the left-hand side plots, and β_1 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RB values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. It's obvious that the available data method RB uncertainty increases when there is 60% missingness in the incomplete predictor.

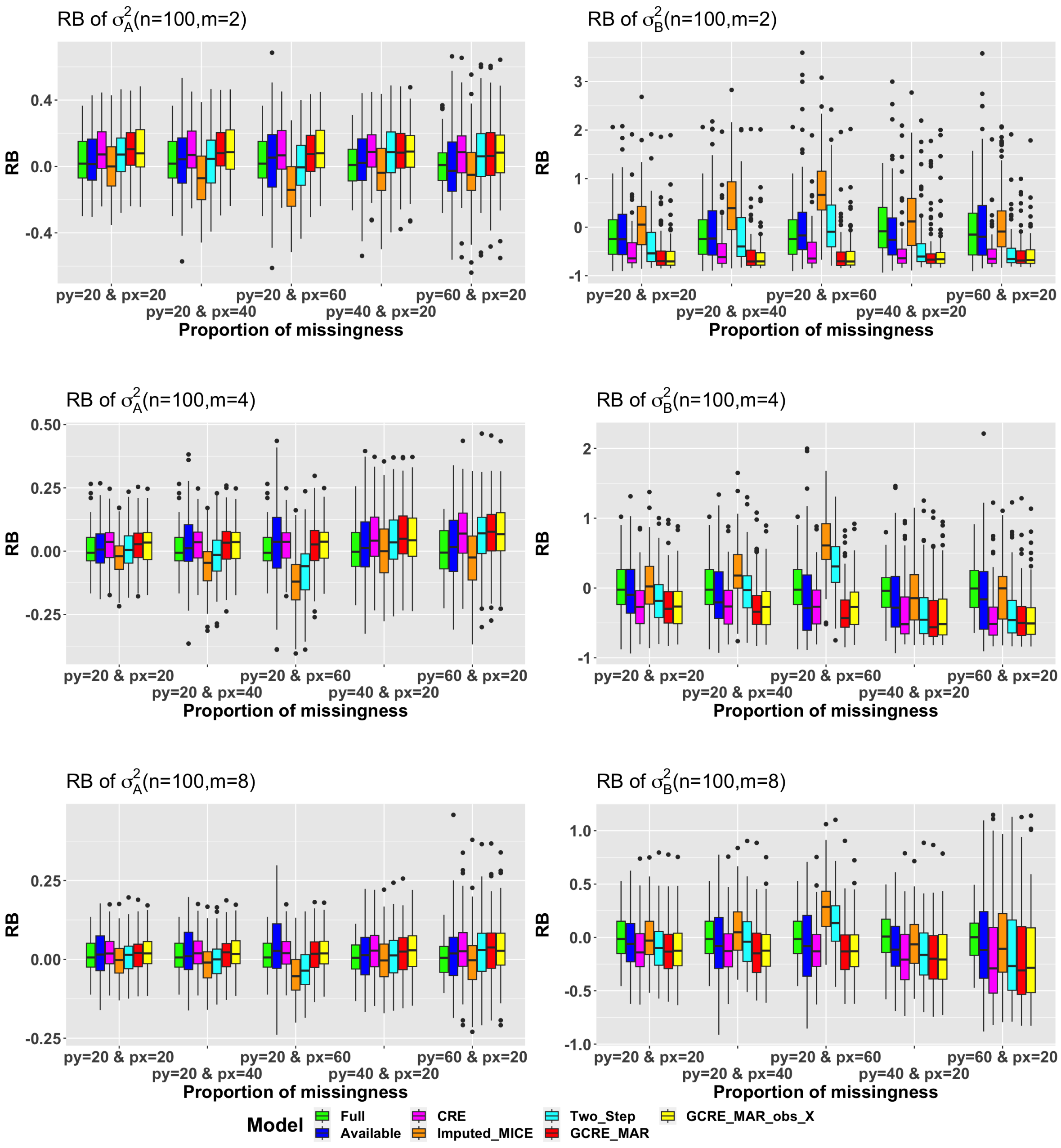


Figure 6.7.6: Boxplots illustrate the RB of σ_A^2 in the left-hand side plots, and σ_B^2 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RB values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. The available data method has unbiased results with large RB uncertainty compared with other methods.

To evaluate how well each method covers the data-generating parameters of the response model parameters, we report the CR. The CR fluctuates between 0.9 and 0.99 for most fixed effects coefficients, with a few exceptions, which are β_3 with $m = 8$, where all methods see a decrease in CR when there is 60% missingness in the response variable. The GCRE-MAR method shows a CR of β_1 less than 0.9 in the case of 40% missingness in the incomplete predictor and with $m = 8$. The MICE Algorithm method shows some decreases in CR, specifically in β_0 and $m = 8$, except when there is 40% missingness in the incomplete predictor, and it shows a CR decrease in β_0 with $m = 4$ when there is 40% missingness in the response variable. The plots for the CR of the response model parameters can be found in Section C in the Appendix.

The CR of the variance component ranges between 0.9 to 0.99, except for some cases where certain methods show a low CR in specific conditions. For instance, the MICE Algorithm method has a low CR in σ_A^2 when there is 40% or 60% missingness in the incomplete predictors, and when there is 40% missingness in the response variable, with $m = 4$. Similarly, the Two-Step method displays a low CR in σ_A^2 with 60% missingness in the incomplete predictors and $m = 4$.

The GCRE-MAR method, the GCRE-MAR method with no considered missingness in the model predictors, and the CRE method shows lower CR for σ_B^2 in all missingness proportions except for when there is 60% missingness in the model response with $m = 2$, where in this case, the full data method has a lower CR. When there is 60% missingness in the model response with $m = 8$, the GCRE-MAR with no considered missingness in the model predictors exhibits a lower CR for σ_B^2 except when there is 60% missingness in the response variable with $m = 4$ and lower CR with $m = 8$ when there is 60% missingness in the response variable. The MICE Algorithm method showed a substantial decrease in the CR for σ_B^2 when there is 60% missingness in the incomplete predictor, which indicates that the MICE Algorithm method struggled to cover the data-generating values of variance parameters when a substantial proportion of data was missing in the incomplete predictor.

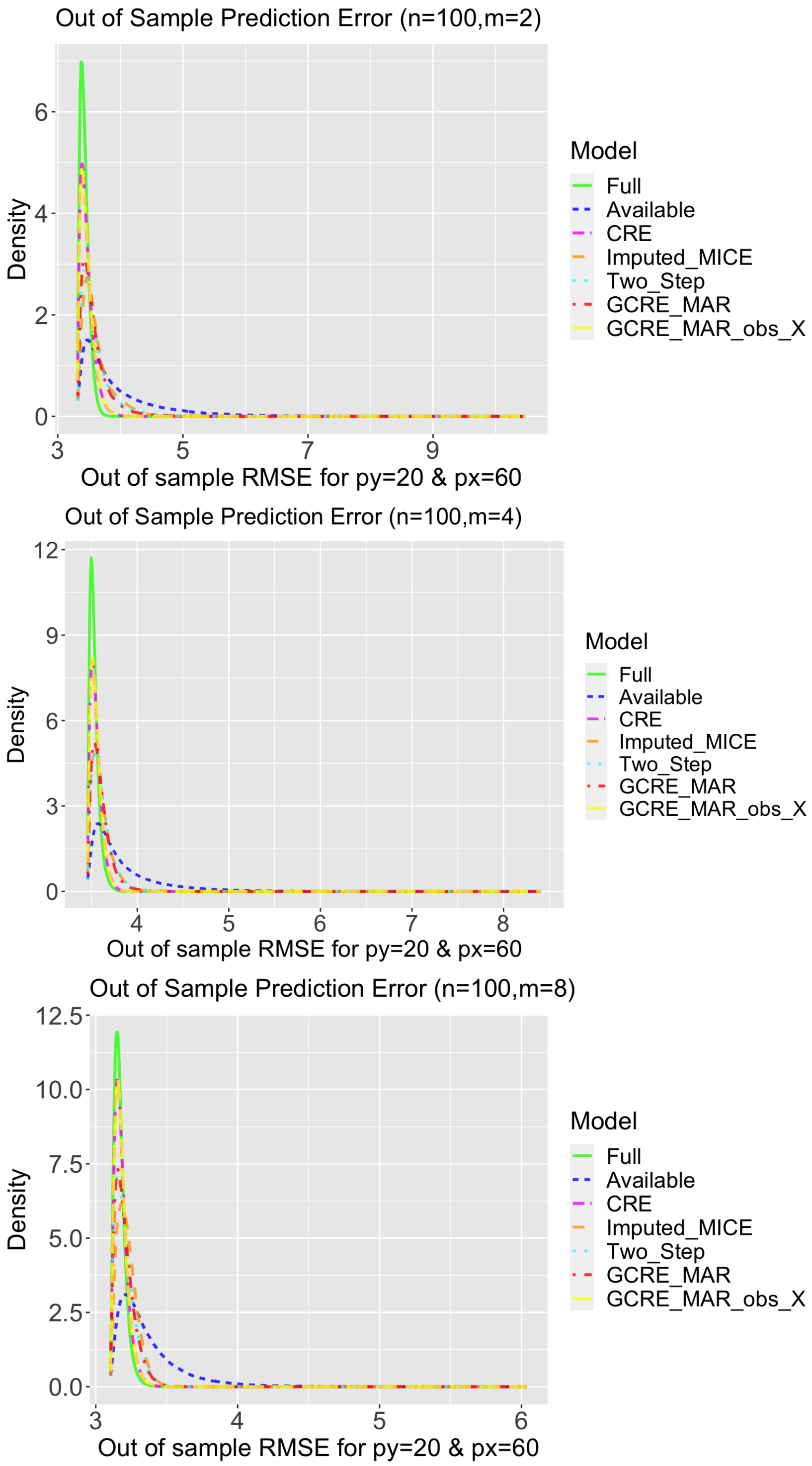


Figure 6.7.7: The density plots of the out-of-sample RMSE for different methods across various repeated measures with 20% missingness in the response and 60% in the incomplete predictor. Each density curve corresponds to one of the methods used, and each plot corresponds to a different value of repeated measures. The available data method appears to have less density and slightly shifted to the right.

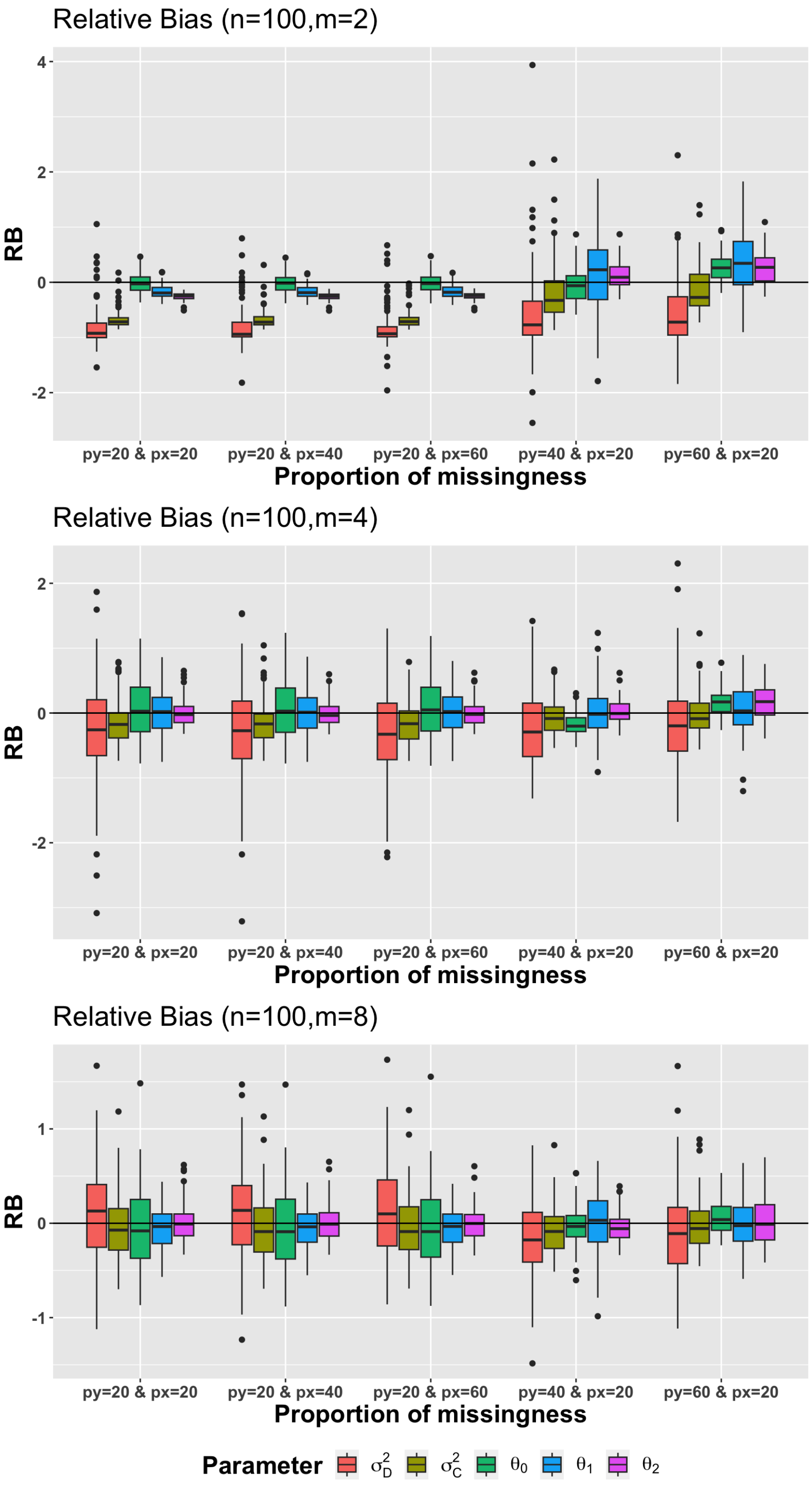


Figure 6.7.8: Boxplots illustrate the RB of the missingness response model parameters using the GCRE-MAR method across various proportions of missingness and repeated measures. The y-axis represents the RB values, while the x-axis shows the proportion of missingness, and each boxplot represents a specific missingness response model parameter. Each plot corresponds to a different repeated measure value. As the repeated measure increases, the model parameter becomes more unbiased.

In the context of out-of-sample prediction performance, the GCRE-MAR method performs better than the available data method, while the MICE Algorithm method has the second-worst out-of-sample performance. Figure 6.7.7 displays the out-of-sample performance for 20% missingness in the model response and 60% missingness in the incomplete predictor for different repeated measures and across different methods. The results are consistent with the remaining proportion of missingness, the plots can be found in Section C in the Appendix.

The RB plots of the missingness response model parameters using the GCRE-MAR method are expressed in Figure 6.7.8, and the RMSE plots of the missingness response model parameters can be found in Section C in the Appendix. The coefficient parameters of the missingness response model using the GCRE-MAR method have a small RMSE, except for θ_2 and the variance components, which have a large RMSE. However, as the proportion of missingness in the response parameter increases, the RMSE of the variance components decreases. Regarding RB, the parameters show unbiasedness except for the variance components in a small number of repeated measures ($m = 2$).

The RB plots of the incomplete predictor model parameters using the GCRE-MAR method are expressed in Figure 6.7.9, and the RMSE of the incomplete predictor model parameters can be found in Section C in the Appendix. The incomplete predictor model parameters using the GCRE-MAR method have a small RMSE overall and a larger RMSE of α_0 when there is 60% missingness in the incomplete predictor. The RMSE uncertainty of α_0, α_2, and σ_E^2 are larger compared with the remaining parameters. Regarding RB, the parameters show unbiasedness except for σ_F^2 with $m = 2$ and 20% of missingness in the response variable. Furthermore, α_2 has a considerably large RB uncertainty overall. The difference between α_2 having larger RB uncertainty and larger RMSE values compared with α_1 is that the variable associated with α_2 is used to generate the missingness in the incomplete predictor.

The plots illustrating these findings can be found in Section C in the Appendix. The CR of the covariance parameter between the random effects of the response model and the missingness response model (σ_D^2) using the GCRE-MAR method is lower when the number of repeated measures is either 2 or 4. This pattern holds for all combinations of missing data proportions, except when the model response has 60% missingness. The same pattern is observed for the missingness response model's individual variance (σ_c^2) when there are two repeated measures. Additionally, σ_c^2 has a lower CR when there is 20% missingness in both variables. When there are two repeated measures, the incomplete predictor model's residual variance (σ_F^2) has a lower CR when there is 20% missingness in the response variable, regardless of the proportion of missingness in the incomplete predictor. The incomplete predictor model's intercept (α_0) has a lower CR when there is 40% missingness in the incomplete predictor and $m = 4$. The remaining variables in both missing variables models (the missingness response model and the incomplete predictor model) using the GCRE-MAR have reasonable CR.

The GCRE-MAR method is effective in capturing missingness response model parameters but may not capture the covariance variable between the random effects when there are fewer repeated measures. This is possible only with a large proportion of missingness in the response. With more repeated measures and missingness in the response, the data-generating covariance parameters value can be captured. The proposed method appears to face difficulties when dealing with a smaller proportion of missing data and a smaller number of repeated measurements. This suggests that the method may require a larger proportion of missing data and more repeated measurements to provide accurate estimates for the missingness response model and incomplete predictor model, which can better represent the data-generating values.

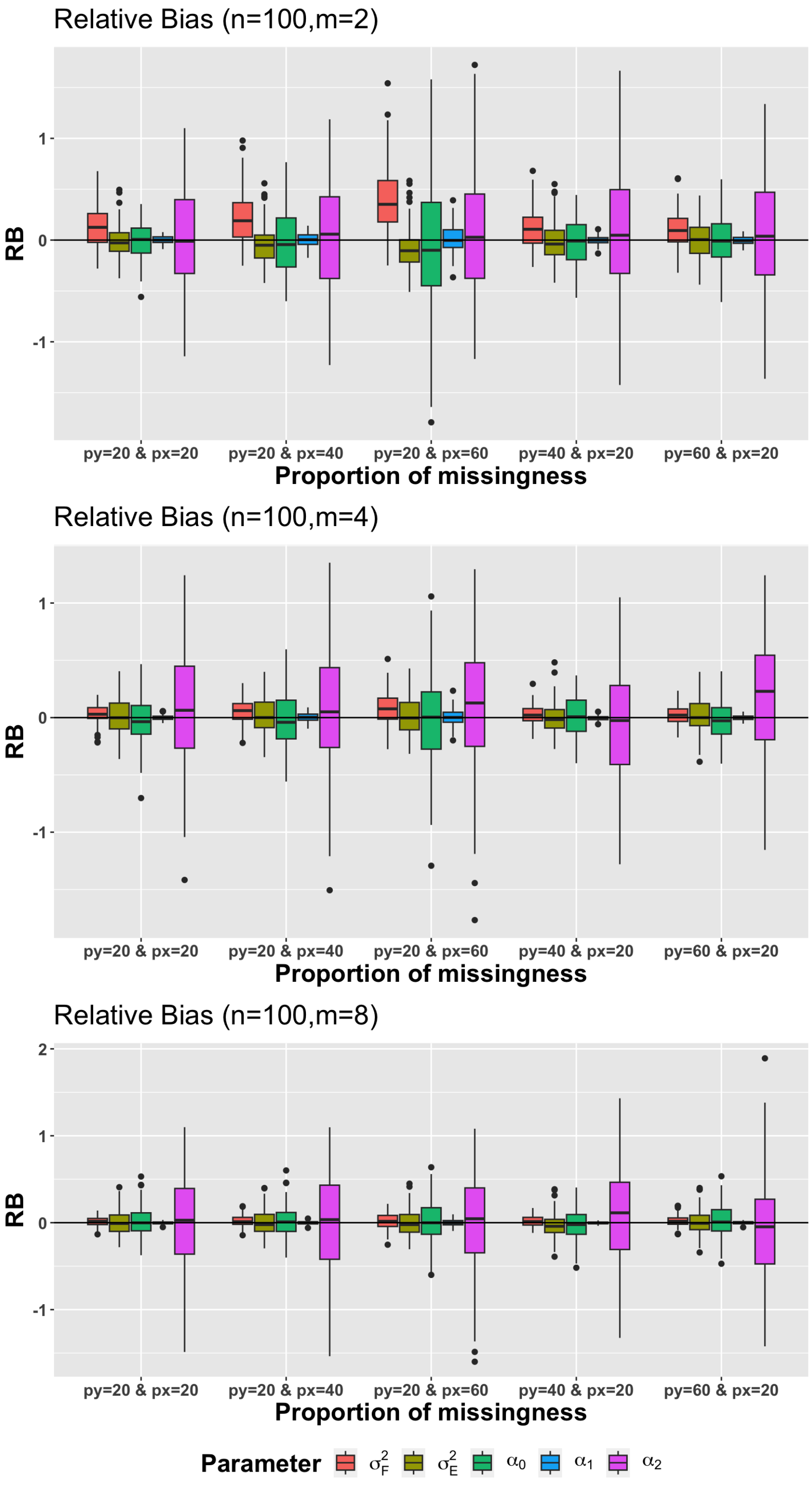


Figure 6.7.9: Boxplots illustrate the RB of the incomplete predictor model parameters, using the GCRE-MAR method across various proportions of missingness and repeated measures. The y-axis represents the RB values, while the x-axis shows the proportion of missingness, and each boxplot represents a specific missingness response model parameter. Each plot corresponds to a different repeated measure value. As the repeated measure increases, the model parameter becomes more unbiased, with larger RB uncertainty in α_2.

6.7.2 Real Data Results

In this section, we will be using the GCRE-MAR method to tackle the challenge of dealing with missing values in both the response and predictor variables of the BIOSTAT-CHF dataset, which is a real-world dataset. The purpose of this analysis is to test the effectiveness of the proposed method on actual data. Moreover, we will also use the BHM using the available data as the baseline method, which is the default approach when it comes to dealing with repeated measures (mixed models). The model employed in this section was introduced in Equation 4.2.2 in Section 4.2. As the true parameter values are unknown, we will utilize Kolmogorov-Smirnov (KS) test statistics defined in Section 2.7 to compare the GCRE-MAR and available data methods.

Figure 6.7.10 displays a noticeable difference in the model's intercept density curves in the two methods, and the disparity in the density curves of the other model's parameters suggests a meaningful distinction between the performance of the two methods. The Kolmogorov-Smirnov test confirms this observation, indicating that only centred eGFR and the Pacemaker variables have no significant difference between the two methods. The random effects' covariance is $\sigma_D^2 = -0.241$, suggesting weak/ moderate evidence of MNAR (Missing Not at Random) of the model response. The imputed values of the model response and predictor using the GCRE-MAR method are presented in Figure 6.7.11, which shows similarities between the observed and imputed data density curves, suggesting that the imputation process successfully preserved the overall data distribution.

Next, we will separate the data into a training set and a test set to further evaluate the proposed method's effectiveness on unseen data. We will make two assumptions for the test data to explore diverse scenarios and assess the model's performance under various conditions.

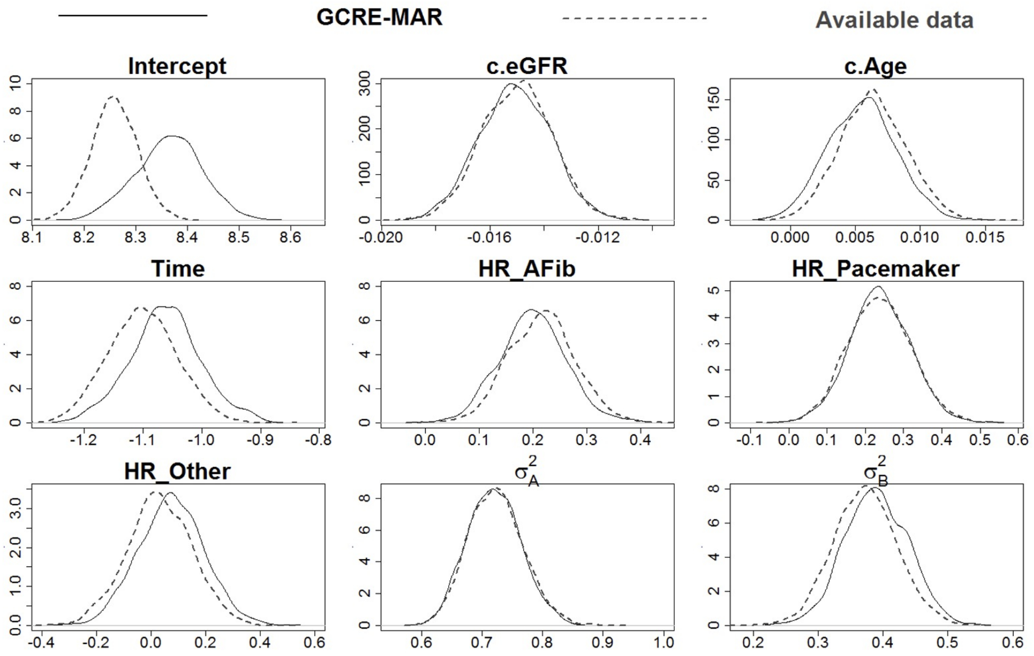


Figure 6.7.10: Posterior distributions of the BIOSTAT-CHF model parameters using the GCRE-MAR method (black solid curve) and the available data method (grey dashed curve). Each plot represents one of the BIOSTAT-CHF model parameters. Density curves overlay, except the model intercept and perhaps to some extent, the coefficient for the Time variables.

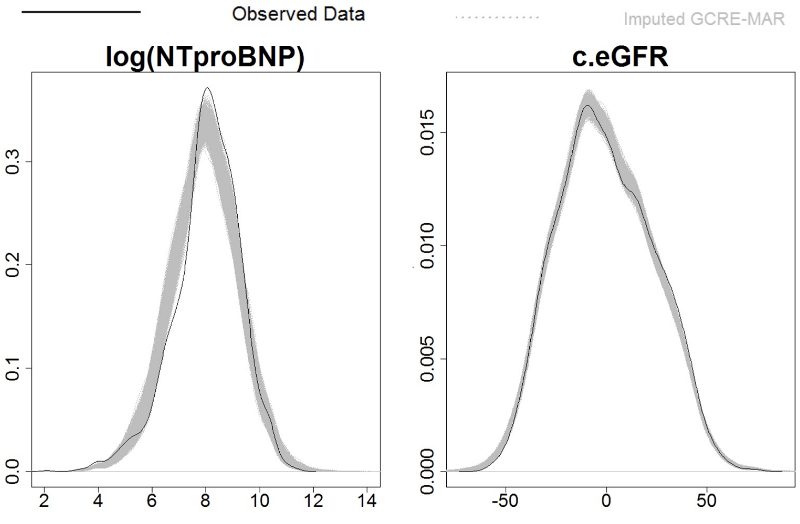


Figure 6.7.11: Density plots show the observed values in the black solid curve and the GCRE-MAR imputed response (left-hand) and predictor (right-hand) values in the grey dashed curves, where each curve is a different draw of the latent variable from the posterior using Gibbs sampler. The density of the imputed values using the GCRE-MAR method at each Gibbs sampling iteration is similar to the density of the observed values.

Case 1: Split the data into test and training data, where the test data has fully observed values.

In order to evaluate the performance of the proposed method and compare it with the baseline method, which can't impute missing values, we need to split the BIOSTAT-CHF data into test and training sets. To avoid advantaging the GCRE-MAR method, we will treat the test data as fully observed and the training data with missing values in the response and predictor. However, only 395 participants in the dataset have complete data, which accounts for only 15% of all 2516 participants. Therefore, we will allocate 85% of the data to the training set and 15% to the test set.

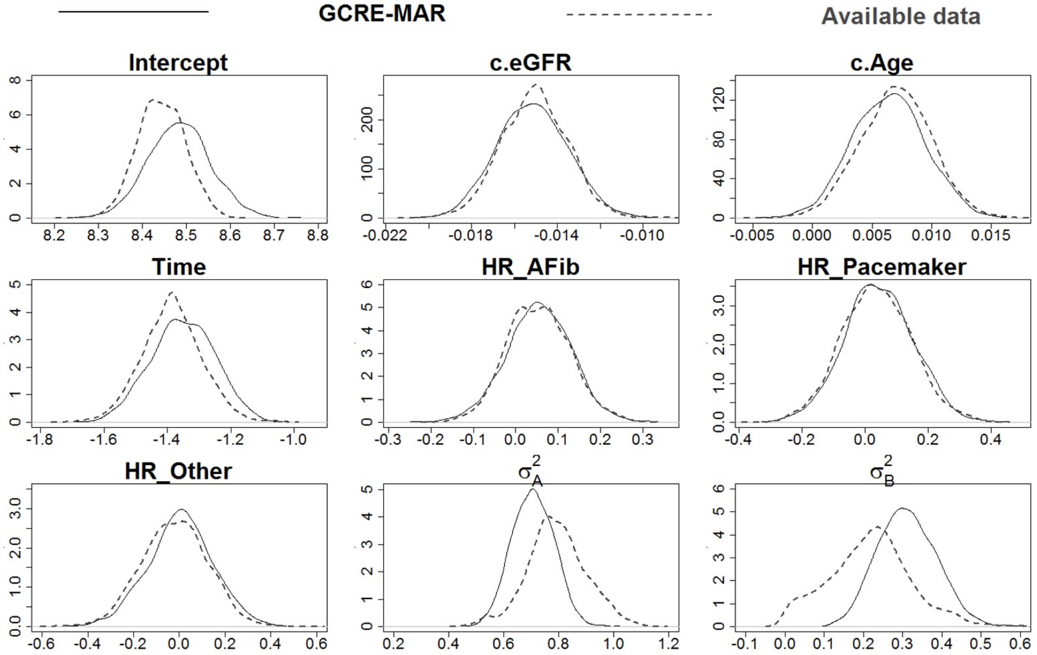


Figure 6.7.12: Posterior distributions of the BIOSTAT-CHF model parameters using the GCRE-MAR method is in a black solid curve, and the available data method is in a grey dashed curve. Each plot represents one of the BIOSTAT-CHF model parameters in Case 1. Density curves overlay except for the model intercept and the variance parameters.

The training data were processed using GCRE-MAR and available data methods. The resulting posterior distributions are shown in Figure 6.7.12, which shows that the model's intercept and variances exhibit distinct density curves between the two methods. Furthermore, the Kolmogorov-Smirnov test was computed to assess the differences with a significance level of 0.05; the null hypothesis was rejected for all variables except for Pacemaker and Atrial fibrillation.

The covariance between the random effects is $\sigma_D^2 = -0.056$, which suggests a very weak indication of MNAR (Missing Not at Random) of the response in the training data. The out-of-sample performance using RMSE between both methods appears similar, as indicated by a Kolmogorov-Smirnov test with a p-value of 0.08, which implies that both methods perform comparably. However, it's noticeable that the RMSE density plot of the available data method exhibits a slight shift towards smaller values. Nonetheless, this shift is negligible, as presented in Figure 6.7.13.

The GCRE-MAR method's imputed response overlaps well with the test data, but they exhibit noticeably different density patterns compared to the training data. This suggests that the imputation is similar to the complete case, as shown in Figure 6.7.14. On the other hand, we notice that the imputed data for the incomplete predictor variable using the GCRE-MAR method aligns well with the training data distribution but differs from the test data distribution. This indicates that the imputed incomplete predictor data aligns well with the training set, demonstrating a satisfactory imputation process for the observed data. Overall, the plot highlights the contrasting behaviour between the response and incomplete predictor variables regarding imputation performance. The imputed response data show a greater disparity between training and test, while the imputed incomplete predictor data resemble the training set more closely.

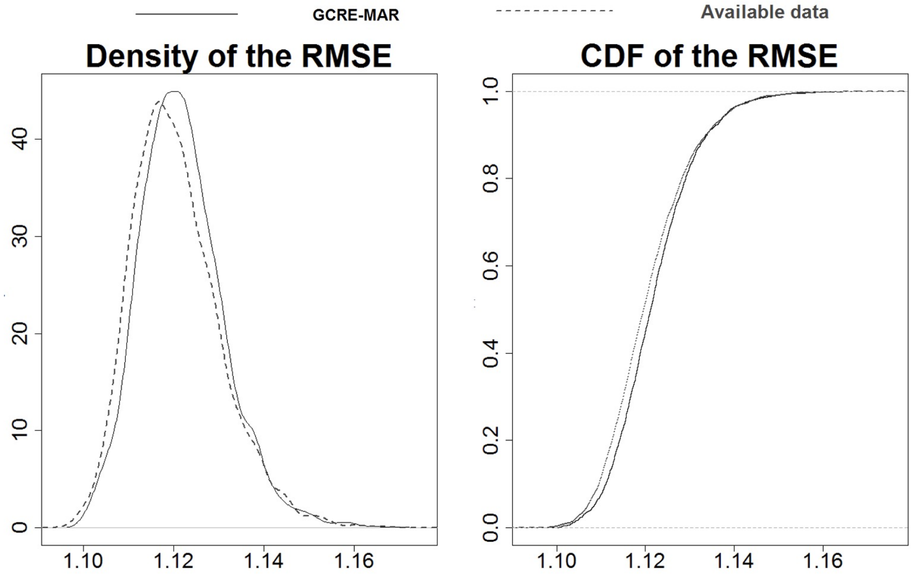


Figure 6.7.13: The RMSE density (on the left-hand side) and CDF (on the right-hand side) of the out-of-sample prediction in BIOSTAT-CHF training data in Case 1, where the black solid curve represents the GCRE-MAR method, and the grey dashed curve represents the available data method. The RMSE of the available data method is shifted toward a lower RMSE values.

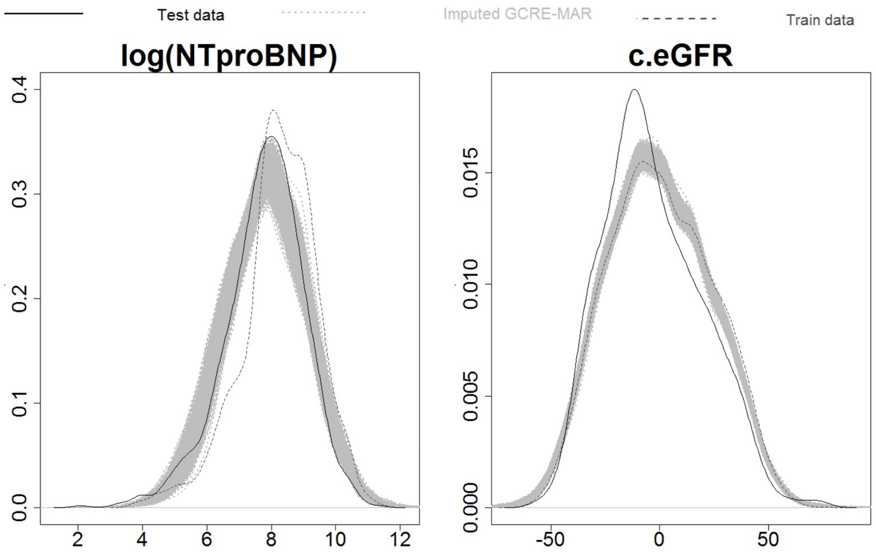


Figure 6.7.14: Density plots show the observed values in the black solid curve and the GCRE- MAR imputed response (left-hand) and predictor (right-hand) values in the grey dashed curves, where each curve is a different draw of the latent variable from the posterior using Gibbs sampler in Case 1. The imputed response overlaps with the test data, and the imputed incomplete predictor overlaps with the training data.

Case 2: Split the data into test and training data, where the test data has missing values in the predictor.

In previous Case 1, we advantaged the available data approach in terms of evaluation, which is an unfair comparison with the GCRE-MAR method. In Case 2, we will discuss handling missing values in predictor variables while evaluating a model. It is essential to consider this aspect as it enables us to test the model's performance under realistic conditions. In this scenario, the training data contains missing values in both the response and predictor variables, while the test data only has missing values in the predictor variable. The percentage of data split for both the training and test data is the same as in Case 1.

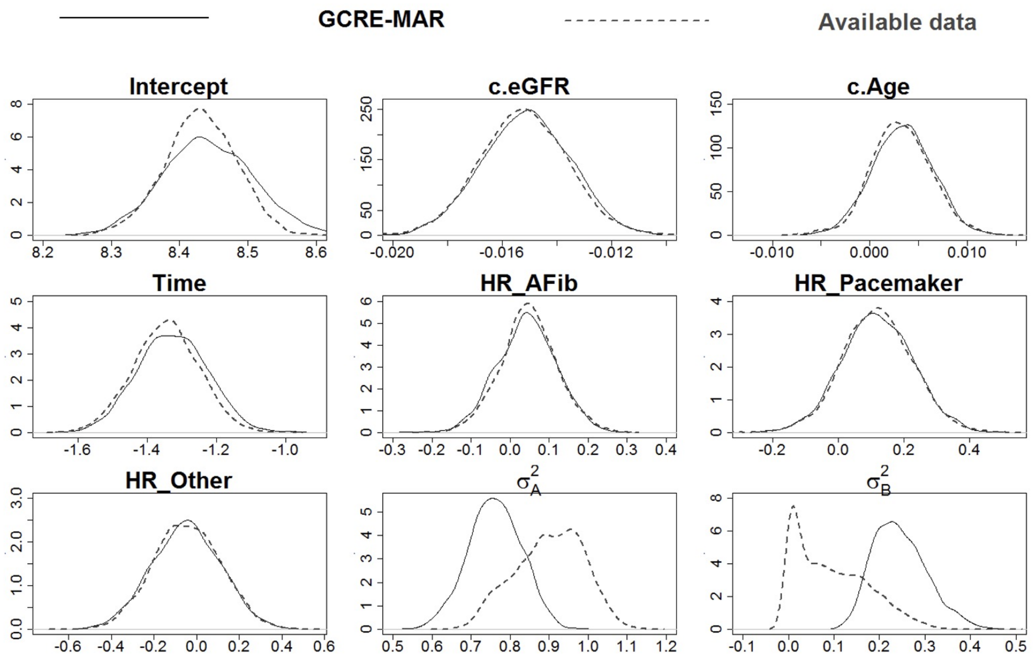


Figure 6.7.15: Posterior distributions of the BIOSTAT-CHF model parameters using the GCRE-MAR method is in a black solid curve, and the available data method is in a grey dashed curve. Each plot represents one of the BIOSTAT-CHF model parameters in Case 2. Density curves overlay except for the variance parameters.

The resulting posterior distributions are shown in Figure 6.7.15, where the posterior distribution of model parameters for the GCRE-MAR and available data methods overlaps, indicating similarity. However, it is worth noting that the variance components of both methods exhibit distinct posterior distributions; this indicates noticeable differences between the two methods in terms of variance estimation. The Kolmogorov-Smirnov test reveals that Pacemaker, Others, eGFR and age have p-values greater than 0.05, suggesting no significant difference between the methods in these parameters.

The random effects' covariance is $\sigma_D^2 = -0.01$, suggesting weak evidence of MNAR (Missing Not at Random) in the model response. The GCRE-MAR method was applied to the test data to impute the missing values of the predictor variable "eGFR", and only fully observed values of the predictor variable "eGFR" are used in the available data method to assess the method's out-of-sample performance.

The density plot of the RMSE presented in Figure 6.7.16 shows that the available data method slightly shifts towards larger values compared with the GCRE-MAR method. The Kolmogorov-Smirnov test indicates significant differences between the GCRE-MAR and available data methods with a significance level of 0.05. Comparing this RMSE distribution to the RMSE distribution in Case 1 (Figure 6.7.13), where there was no missingness in the test data, we observe that both distributions are similar and centred around similar values. However, the GCRE-MAR method outperformed the available data method regarding RMSE in Case 2, which suggests that the performance of the GCRE-MAR method in terms of RMSE improved in the presence of missing values in the test data.

Moreover, the conclusions drawn from comparing the training data distribution, test data distribution, and GCRE-MAR imputed values distribution in Case 1 remain the same in Case 2 (as seen in Figure 30 in the Appendix). The consistency of the findings suggests that the imputation process using the GCRE-MAR method enables reliable assessments of the training, test, and imputed values.

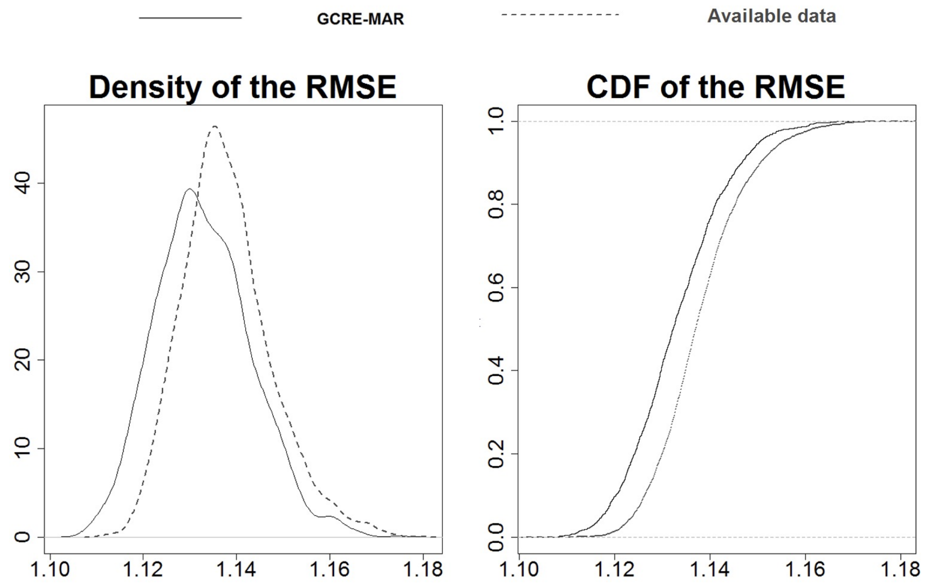


Figure 6.7.16: The RMSE density (on the left-hand side) and CDF (on the right-hand side) of the out-of-sample prediction in BIOSTAT-CHF training data in Case 2, where the black solid curve represents the GCRE-MAR method, and the grey dashed curve represents the available data method. The RMSE of the GCRE-MAR method is shifted toward a lower RMSE values.

6.8 Discussion

In this chapter, we introduced a statistical methodology that is a generalisation of the existing method used to handle non-ignorable missingness in the model response, known as the CRE method discussed in Chapter 4, which was implemented using the Correlated Random Effects and the Gibbs sampling introduced by Bhuyan (2019). Our proposed method, called GCRE-MAR, can handle the missingness in the model predictor in addition to the missingness in the model response. Missingness is common in both the response and predictors when dealing with longitudinal studies; where repeated measures of both the response and predictors are collected over time.

The proposed GCRE-MAR method assumes that missingness in the model predictor is ignorable, while the model response missingness is non-ignorable. This assumption is similar to the proposed method introduced in Chapter 5, but the GCRE-MAR method has advantages in that it can handle missingness simultaneously and jointly using the Gibbs sampler. In contrast, the Two-Step method in Chapter 5 has two different steps and is computationally expensive compared to our approach, the GCRE-MAR. The computational time for the Two-Step method is about three times longer than the GCRE-MAR method, based on a specific setup using ten imputed datasets in the Two-Step method. As the number of imputed datasets increases, the difference in computational time for the Two-Step method becomes even more pronounced.

The simulation study analysis in Section 6.7.1 is limited to a setting involving one incomplete predictor and two fully observed predictors. The results demonstrate that the proposed GCRE-MAR method, when used with a fully observed predictor, generally performs equally well as the CRE method. This indicates that the GCRE-MAR method is a generalisation of the CRE method, as both can handle missingness in the model response while the model predictors are fully observed. Additionally, the GCRE-MAR method was able to accurately impute missing values of the response and predictor variables for different sample sizes and proportions of missingness, similar to the simulated data values. This demonstrates a high level of accuracy and reliability in handling missing data, making it a suitable and effective option for practical use.

Based on our analysis, we have found that the proposed GCRE-MAR method performed better than the available data method in terms of the overall performance of the parameter estimates based on WRMSE. The model coefficients' RMSE of the CRE method was found to be lower than that of the GCRE-MAR method due to the extra error introduced by GCRE-MAR to address missingness in the model predictor. In contrast, the CRE method only deals with the missingness in the model response, whereas the GCRE-MAR method can flexibly handle both scenarios. Which are missingness in the response and missingness in both the response and predictors. The GCRE-MAR method offers an additional ability in this regard.

The GCRE-MAR method was found to have a lower RMSE for the response model parameters in comparison to the available data method. For the response model coefficients, the GCRE-MAR method showed a comparable bias to the compared methods. On the other hand, the available data method produced unbiased results but with considerable uncertainty, especially when there was a high proportion of missing data (60%) in the model incomplete predictor. When there were many repeated measures (m=8), the available data method resulted in more biased results of β_0 and β_3 compared to the GCRE-MAR method.

Although the GCRE-MAR method has a lower RMSE as compared to the available data method, indicating higher overall accuracy in estimating the response model parameter values, it exhibits bias for the variance parameters as repeated measures decrease. This means that while the GCRE-MAR method may perform better in estimating the response model coefficients, it may not accurately capture the variability in the model with a low number of repeated measures. The difference in RB values between the GCRE-MAR method and the available data method, with respect to variance parameters, suggests that the GCRE-MAR method's estimation tends to deviate more from the data-generating variance than the available data method but with tighter estimation bounds. Furthermore, the GCRE-MAR method was able to capture the data-generating parameters except for the σ_B^2 with $m = 2$ and with 60% missingness in the response with $m = 8$.

Dealing with a large proportion of missing data in the incomplete predictor can be challenging for the Two-Step and MICE methods which is consistent with the findings presented in Chapter 5. This results in high RMSE when estimating the response model parameters, overestimation of σ_B^2, underestimation of σ_A^2, and an inability to capture the estimation of the data generating parameters' values. This could be the nature of the MICE Algorithm, which can lower the imputation quality when there are more missing values. Since the algorithm relies heavily on observed values, it can limit the information available for imputation.

The missingness response model estimates using the GCRE-MAR method showed unbiasedness, except for variance parameters with $m = 2$. The GCRE-MAR method captured the data-generating covariance parameters between the random effects of the response models for a higher proportion of missingness in the response and with large repeated measures. It was generally effective in capturing the data-generating parameters of missingness response model parameters. However, when there are fewer repeated measures (e.g. 2 or 4), the covariance between the response model and the missingness response model may not be captured. This is only observed when there is a large proportion of missingness in the response (60%). The covariance variable indicates how likely the missingness in the response is to be MNAR. With a larger number of repeated measures and a larger proportion of missingness in the response, the method was able to capture the data-generating covariance parameters. The proposed method may require a larger proportion of missing data and more repeated measurements to provide accurate estimates of the covariance parameter.

On the other hand, the incomplete predictor model parameters show unbiased results overall, with a large uncertainty of α_2. Except for biased estimates and failed to capture the data-generating parameter value of σ_F^2 with $m = 2$ and 20% missingness in the response variable. It suggests that the GCRE-MAR method struggles to capture the between individual variances with a small number of repeated measures and a low proportion of missingness in the model response variable. It is worth mentioning that out-of-sample prediction was similar to the full data model. So, the effect of not estimating these parameters (and the parameters in the previous paragraphs) is small.

Based on the application of the available data method and the GCRE-MAR method on a real-world (BIOSTAT-CHF) dataset, it was found that the GCRE-MAR method performed better than the available data method in terms of out-of-sample performance when evaluated on test data that had missingness in the incomplete predictor (eGFR). This raises concerns about the reliability of the available data method as a tool for decision-makers and physicians to make informed decisions and improve patient care outcomes using data with missingness.

The GCRE-MAR method has an advantage over the available data, which can impute unseen data. The imputation samples of the response (NT-proBNP) are distributed similarly to the test data, which only contains complete case data. Furthermore, the imputed sample values of the incomplete predictor (eGFR) are distributed similarly to the training data, which contains data that are not complete cases. This indicates that the missingness in the response is ignorable. Furthermore, the covariance between the response model and the missingness response model confirms that the missingness in the response (NT-proBNP) is less likely to be MNAR.

On the other hand, it is possible that eGFR (or other predictors we may want to work with in the future) might have non-ignorable missingness. Therefore, this method needs to be generalised further to estimate and handle the model parameters when the model's incomplete predictor has non-ignorable missingness. This will be introduced in Chapter 7. We will perform a sensitivity analysis to ensure that the GCRE-MAR method produces reliable results for decision-makers and ensure the method's consistency across various situations. This will be presented in Chapter 8.

Chapter 7

A Further Extension to The CRE Method with Non-Ignorable Predictors (GCRE-MNAR)

7.1 Introduction

So far, our focus has been on modelling non-ignorable missing response using the existing CRE method discussed in Chapter 4 and assuming the missingness of the model predictor is ignorable, which we introduced two approaches to deal with that in Chapter 5 and Chapter 6. This chapter aims to enhance our approach by introducing a predictor missingness model into our joint model by allowing predictor missingness to depend on the incomplete predictor itself. Also, the incomplete predictor can depend on the other predictors and is correlated with the predictor missingness model via random effects, following the Correlated Random Effects method.

Previously, Correlated Random Effects are assumed for the response model and the response missingness procedure. However, in the proposed approach, we also assume a Correlated Random Effects for the predictor model and the predictor missingness procedure. This allows us to simultaneously impute the missing predictors and response, estimate the probability of the missingness of model response and predictor to be MNAR, and analyse the main model, missingness model of the response and predictor parameters. We explore this by extending the GCRE-MAR joint model introduced in Chapter 6 by incorporating more realistic predictor missingness assumptions. We assume MNAR (Missing Not At Random) for the model response and predictor as a more complex scenario because the reason behind the missingness is not always available. The proposed method, Generalised Correlated Random Effects with Missing Not At Random predictor (GCRE-MNAR), has the ability to indicate the probability of MNAR occurring in both the model response and predictor. MNAR is the most common scenario and is often seen in longitudinal studies involving repeated measures (Ibrahim and Molenberghs, 2009).

There are numerous studies discussed methods to handle non-ignorable missingness in the model predictors; for example, Huang et al. (2005) proposes Bayesian methods to estimate parameters in generalised linear models with nonignorable missingness in the predictors and addresses issues of posterior impropriety by introducing proper priors. Roy and Lin (2005) proposes a selection model for estimation in generalised linear mixed models for longitudinal data with informative dropouts. The model allows for missing time-varying predictors alongside the response variable and employs the EM algorithm for inference. Stubbendick and Ibrahim (2006) proposes two models for estimating parameters in nonignorable missing responses and predictors for discrete longitudinal data. These are generalised linear mixed models with maximum likelihood estimation and a multivariate probit model using the EM method.

Moreover, Fang et al. (2018) proposed an estimator for predictors in generalised linear models with nonignorable missing data using imputation-based adjustments and a semiparametric approach to estimate parameters. Wu (2008) proposed an approximation technique for non-ignorable missing model predictors for a non-linear mixed effect model. Li and Grace (2013) proposed a method that can handle missingness in the model response and predictor for longitudinal data using a pairwise likelihood method; however, they assume non-simultaneous missingness within observations. In real-world longitudinal datasets, not recording both the response and predictor variables is common when a participant doesn't show up is common. The proposed method aims to simultaneously handle missingness in the response and predictor, where one or both variables' observations are missing.

This chapter will cover the proposed model in Section 7.2, the joint distribution's structure in Section 7.3, the prior distribution setup in Section 7.4, and a description of the sampler procedure in Section 7.5. We will set up simulation data in Section 7.6, then use it to test the proposed method in Subsection 7.7.1. Additionally, we will apply the proposed method to the data provided data, BIOSTAT-CHF data, in Subsection 7.7.2. Finally, we will discuss the pros and cons of this approach over the baseline approaches in Section 7.8.

7.2 Proposed Model

The proposed method GCRE-MNAR will incorporate changes to the incomplete predictor model while preserving the response and missing response indicator model structure as in the GCRE-MAR method, which is described in Section 6.2. The modifications made to the incomplete predictor model and introduction of the missing predictor indicator model are intended to improve its ability to handle non-ignorable missingness in the model predictor, bringing it into closer alignment with the types of missing data that could be encountered in longitudinal data. The GCRE-MNAR uses the same equations in Chapter 6, Equations 6.2.1 through 6.2.4, which is the response and missing response indicator equation, assuming Correlated Random Effects. Nevertheless, for handling time-varying predictors in this way, as a linear mixed model, the partially observed predictor model $X_{ji}(t)$ is presented as follows:

$$X_{ji}(t) = \delta + \sum_{j'=1}^{J'} \alpha_{j'} X_{j'i}(t) + w_i \tilde{Z}_i(t) + r_i(t), \tag{7.2.1}$$

where $X_{ji}(t)$ is the j^{th} partially observed predictor for subject i at time t. $\alpha_{j'}$ denotes the regression coefficients of the j'^{th} fully observed fixed effects, and δ is the fixed intercept representing the mean of the overall population. The random intercepts w_i are assumed to be independent and identically distributed (i.i.d) following a Normal distribution $N(0,\sigma_E^2)$ and the residuals $r_i(t)$ are assumed to be independent and identically distributed (i.i.d) following a Normal distribution $N(0,\sigma_F^2)$. This is known as the *separate specification* since a unique regression model is needed for each incomplete predictor, this is discussed in Section 6.2. We will assume one partially observed predictor and two fully observed predictors.

Since we assume non-ignorable missingness for the partially observed predictor $X_{ji}(t)$, we consider the corresponding missingness model as:

$$R^*_{ji}(t) = \chi + \sum_{j'=1}^{J'} \psi_{j'i} X_{j'i}(t) + q_i \tilde{Z}_i(t) + \xi_i(t), \tag{7.2.2}$$

where $R^*_{ji}(t)$ is the latent incomplete predictor missingness model for predictor j of subject i at time t, $\psi_{j'}$ denotes the regression coefficients of the j'^{th} fully observed fixed effects, χ is the fixed intercept representing the mean of the overall population. Subject-specific random effects q_i capture the longitudinal dependence and are assumed to be independent and identically distributed (i.i.d) following a Normal distribution $N(0,\sigma_G^2)$. The residuals $\xi_i(t)$ are assumed to be i.i.d following a Normal distribution $N(0,1)$.

In this approach, the missing indicator is a binary variable associated with an underlying latent variable that follows a Standard Normal distribution. Thresholds are specified to determine whether the original variable's value is missing or observed. Therefore, the missing predictor indicator $R_{ji}(t)$ is conditional on the propensity $R^*_{ji}(t)$ through a probit regression as follows:

$$p(R_{ji}(t) = 1 | X_{ji}(t)) = \Phi\left(\chi + \psi_{j'}(t) X_{j'i}(t) + q_i \tilde{Z}_i(t)\right), \tag{7.2.3}$$

where:

$$R_{ji}(t) = \begin{cases} 1, & \text{if } R^*_{ji}(t) > 0, \\ 0, & \text{if } R^*_{ji}(t) \leq 0. \end{cases} \tag{7.2.4}$$

The right-hand side of Equation 7.2.3 is a regressive predictor model that defines the association between the missingness in the partially observed predictor $X_{ji}(t)$ and other observed predictors $\boldsymbol{X}_{J'}$. If the corresponding partially observed predictor value is observed, then $R_{ij}(t) = 1$ and $R_{ji}(t) = 0$ if it is missing. The latent incomplete predictor variable can be written as:

$$X_{ji}(t) = \begin{cases} X^*_{ji}(t), & \text{if } R_{ji}(t) = 1, \\ missing, & \text{if } R_{ji}(t) = 0. \end{cases} \tag{7.2.5}$$

The regression model given in Equation 7.2.1 can be re-written as follows:

$$X^*_{ij}(t) = \delta + \sum_{j'=1}^{J'} \alpha_{j'} X_{j'i}(t) + w_i \tilde{Z}_i(t) + r_i(t), \tag{7.2.6}$$

where $X^*_{ij}(t)$ is the latent incomplete predictor variable. Following the Correlated Random Effects approach, we will incorporate a correlation between the missing predictor model and the predictor model following Bhuyan (2019). We consider w_i and q_i are correlated random vectors following a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma_{x_j} = \begin{pmatrix} \sigma^2_E & \sigma^2_H \\ \sigma^2_H & \sigma^2_G \end{pmatrix}$. Nevertheless, the equations follow the same basic form of the response model as the GCRE-MAR method. This approach falls into the category of selection models defined by Bhuyan (2019).

7.3 Joint Distribution

To distinguish between the observed and partially observed predictor, we use $X_j(t)$ for predictors that contain missing values (partially observed predictor) and $X_{j'}(t)$ for the fully observed predictors (predictors that don't contain any missing values). We will assume that only one predictor is partially observed. Let $\boldsymbol{X}_j = (X_{11}(t), \ldots, X_{nm}(t))$, $\boldsymbol{X}_j^* = \left(X_{11}^*(t), \ldots, X_{nm}^*(t)\right)$, $\boldsymbol{R}_j = (R_{11}(t), \ldots, R_{nm}(t))$ and $\boldsymbol{R}_j^* = \left(R_{11}^*(t), \ldots, R_{nm}^*(t)\right)$. $\boldsymbol{Y}, \boldsymbol{Y}^*, \boldsymbol{U}$ and $\boldsymbol{U}^*$ are as mentioned in Section 6.3. The joint posterior is expressed as follows:

$$p\left(\Theta_{Y,U}, \Theta_{x_{mis,R}}, \boldsymbol{X}_j^*, \boldsymbol{R}_j^*, \boldsymbol{Y}^*, \boldsymbol{U}^* \mid \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{X}_j, \boldsymbol{R}_j, \boldsymbol{X}_{j'}(t)\right) \propto p\left(\Theta_{Y,U}\right) \times p\left(\Theta_{x_{mis,R}}\right) \times$$
$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{t=1}^{m} f\left(Y_i^*(t), U_i^*(t) \mid u_i, v_i, X_{ji}^*(t), \boldsymbol{X}_{j'}(t)\right) \times f\left(X_{ji}^*(t), R_j^*(t) \mid w_i, q_i, \boldsymbol{X}_{j'}(t)\right) \times$$
$$\{I(U_i^*(t) > 0) I(U_i(t) = 1) + I(U_i^*(t) \leq 0) I(U_i(t) = 0)\} \times g(u_i, v_i) \times$$
$$\{I(R_i^*(t) > 0) I(R_i(t) = 1) + I(R_i^*(t) \leq 0) I(R_i(t) = 0)\} \times g(w_i, q_i)$$
$$du_i dv_i dw_i q_i dX_{ji}(t), \tag{7.3.1}$$

where $\{\Theta_{Y,U} = \mu, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau, \boldsymbol{\theta}, \sigma_A^2, \Sigma\}$ and $\{\Theta_{x_{mis,R}} = \delta, \boldsymbol{\alpha}, \sigma_F^2, \chi, \boldsymbol{\psi}, \Sigma_{x_j}\}$. The joint priors are defined as $p(\Theta)$. The joint distribution of random effects $g(.)$ follows a multivariate normal distribution, $f(.)$ is the joint distribution and $I(A)$ is an indicator variable which takes value 1 if A occurs and zero otherwise.

7.4 Prior Distribution

Consider the following priors for $\Theta_{x_{mis,R}}$:

$$p\left(\tilde{\boldsymbol{\alpha}}, \sigma_F^2\right) \propto \frac{1}{\sigma_F^2}; \;\; p(\tilde{\boldsymbol{\psi}}) \propto N(0, b); \;\; p\left(\Sigma_{x_j}\right) \propto IW(\nu, \Lambda); \tag{7.4.1}$$

and the priors of $\Theta_{Y,U}$ defined as:

$$p\left(\tilde{\boldsymbol{\beta}}, \sigma_A^2\right) \propto \frac{1}{\sigma_A^2}; \;\; p\left(\tilde{\boldsymbol{\theta}}\right) \propto N(0, b); \;\; p(\Sigma) \propto IW(\nu, \Lambda), \tag{7.4.2}$$

where $\tilde{\boldsymbol{\alpha}} = [\delta, \boldsymbol{\alpha}]$ is a vector of the overall intercept and regression coefficients of fully observed predictors in the incomplete predictor model, $\tilde{\boldsymbol{\psi}} = [\chi, \boldsymbol{\psi}]$ is a vector of the overall intercept and regression coefficients of fully observed predictors in the incomplete predictor missingness model, $\tilde{\boldsymbol{\beta}} = [\mu, \boldsymbol{\beta}, \boldsymbol{\lambda}]$ is a vector of the overall intercept and regression coefficients of fully and partially observed predictors in the response model and $\tilde{\boldsymbol{\theta}} = [\tau, \boldsymbol{\theta}]$ is a vector of the overall intercept and regression coefficients of fully observed predictors in the response missingness model. To ensure convergence, we use weakly informative priors for the matrix variable $p(\Sigma)$ that we discovered and resolved in Chapter 4 and $p\left(\Sigma_{x_j}\right)$, as well as for the missingness model coefficients $p(\theta)$ and $p(\psi)$. Prior information would encourage convergence in the missingness model coefficients (Du et al., 2022), so we set a relatively large prior variance value of $b = 10$ for this purpose. As discussed in Section 4.4.5, the hyperparameters for the Inverse Wishart distribution are set to $\Lambda = (\nu - p - 1)I$ and the degrees of freedom to $\nu = 4$. We use non-informative priors for all other priors. In order to facilitate the Gibbs sampler, we chose conjugate priors, which allow us to derive the full conditional distributions of each variable in a closed form.

7.5 Gibbs Sampler

The first steps are 1-7 steps mentioned in Section 4.4.4, followed by generating estimates for the incomplete predictor model and the incomplete predictor missingness model. The step-by-step Gibbs sampler procedure when the response and predictor are MNAR is given below:

8. Estimate the predictor model's random intercept, w.
9. Estimate the missing predictor values, $X^*_{ji}(t)$.
10. Estimate the predictor model's residual variance, σ^2_F.
11. Estimate the predictor model's regression coefficient, $\tilde{\boldsymbol{\alpha}}$.
12. Estimate the predictor missingness model's random intercept, q.
13. Estimate the predictor missingness model's regression coefficient, $\tilde{\boldsymbol{\psi}}$.
14. Estimate the covariance matrix that represents the correlation between the random intercept in the incomplete predictor and missing predictor model, Σ_{x_j}.

Repeat steps 1-14 until the MCMC chains converge and produce enough posterior samples. The Gibbs sampler produces a posterior distribution for each variable, which is used to conduct Bayesian inference. The structure of the Gibbs sampler is given in Algorithm 8 and Algorithm 9 overleaf.

Algorithm 8 GCRE-MNAR Method Gibbs Sampling Algorithm-Part 1

```
Choose initial $\{\Theta^0_{Y,U} = \tilde{\boldsymbol{\beta}}^0, \tilde{\boldsymbol{\theta}}^{\mathbf{0}}, \sigma_A^{2^0}, \Sigma^0\}$, $\{\Theta^0_{mis,R} = \tilde{\boldsymbol{\alpha}}^{\mathbf{0}}, \tilde{\boldsymbol{\psi}}^{\mathbf{0}}, \sigma_F^{2^0}, \Sigma_x^0\}$
and $\{u^0, v^0, Y^{*0}, U^{*0}, R^{*0}, X_j^{*0}, q^0, w^0\}$.

for $1, \ldots, S$ iterations do

    -Sample $u_i^{S+1} \sim p\left(u_i \mid Y_i^{*^S}(t), \tilde{\boldsymbol{\beta}}^S, \sigma_A^{2^S}, \Sigma^S, v_i^S, \boldsymbol{X}_{J'}\right)$.

    -Sample $Y_i^{*^{S+1}}(t) = \begin{cases} Y_i^{*^S}(t), & \text{if } Y_i(t) \textit{ is observed} \\ p\left(Y_i^*(t) \mid \tilde{\boldsymbol{\beta}}^S, X_{ji}^S(t), u_i^{S+1}, \sigma_A^{2^S}, \boldsymbol{X}_{J'}\right), & \text{if } Y_i(t) \textit{ is missing.} \end{cases}$

    -Sample $\sigma_A^{2^{S+1}} \sim p\left(\sigma_A^2 \mid \tilde{\boldsymbol{\beta}}^S, X_{ji}^S(t), Y_i^{*^{S+1}}(t), u_i^{S+1}, \boldsymbol{X}_{J'}\right)$.

    -Sample $\tilde{\boldsymbol{\beta}}^{S+1} \sim p\left(\tilde{\boldsymbol{\beta}} \mid Y_i^{*^{S+1}}(t), u_i^{S+1}, X_{ji}^S(t), \sigma_A^{2^{S+1}}, \boldsymbol{X}_{J'}\right)$.

    -Sample $v_i^{S+1} \sim p\left(v_i \mid U_i^{*^S}(t), \tilde{\boldsymbol{\theta}}^S, \Sigma^S, u_i^{S+1}, \boldsymbol{X}_{J'}\right)$.

    -Sample $U_i^{*^S}(t) = \begin{cases} p\left(U_i^*(t) \mid \tilde{\boldsymbol{\theta}}^S, v_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ left truncated}^* \text{ at } 0, & \text{if } Y_i(t) \textit{ is observed}, \\ p\left(U_i^*(t) \mid \tilde{\boldsymbol{\theta}}^S, v_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ right truncated}^* \text{ at } 0, & \text{if } Y_i(t) \textit{ is missing.} \end{cases}$

    -Sample $\tilde{\boldsymbol{\theta}}^{S+1} \sim p\left(\tilde{\boldsymbol{\theta}} \mid U_i^{*^{S+1}}(t), v_i^{S+1}, \boldsymbol{X}_{J'}\right)$.

    -Sample $\Sigma^{S+1} \sim p\left(\Sigma \mid u_i^{S+1}, v_i^{S+1}\right)$.
```

* truncated normal distribution.

Algorithm 9 GCRE-MNAR Method Gibbs Sampling Algorithm-Part 2

```
-Sample $w_i^{S+1} \sim p(w_i \mid X_{ji}^{*S}(t), \sigma_F^{2^S}, \Sigma_x^S, q_i^S, \boldsymbol{X}_{J'})$.
-Sample $X_{ji}^{*S+1}(t) = \begin{cases} X_{ji}^{*S}(t), & \text{if }, X_{ji}(t) \textit{ is observed}, \\ p(X_{ij}^{*S}(t) \mid Y_i^{*^{S+1}}(t), \tilde{\boldsymbol{\beta}}^{S+1}, \tilde{\boldsymbol{\alpha}}^S, \sigma_A^{2^{S+1}}, \sigma_F^{2^S}, u_i^{S+1}, w_i^{S+1}, \boldsymbol{X}_{J'}) & \text{if }, X_{ij}(t) \textit{ is missing}. \end{cases}$
-Sample $\sigma_F^{2^{S+1}} \sim p(\sigma_F^2 \mid X_{ji}^{*S+1}, \tilde{\boldsymbol{\alpha}}^S, w_i^{S+1}, \boldsymbol{X}_{J'})$.
-Sample $\tilde{\boldsymbol{\alpha}}^{S+1} \sim p(\tilde{\boldsymbol{\alpha}} \mid X_{ij}^{*S+1}, \sigma_F^{2^{S+1}}, w_i^{S+1}, \boldsymbol{X}_{J'})$.
-Sample $q_i^{S+1} \sim p(q_i \mid R^{*^S}, \tilde{\boldsymbol{\psi}}^S, \Sigma_x^S, w_i^{S+1}, \boldsymbol{X}_{J'})$.
-Sample $R_i^{*^{S+1}}(t) = \begin{cases} p\left(R_i^*(t) \mid \tilde{\boldsymbol{\psi}}^S, q_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ left truncated}^* \text{ at } 0, & \text{if } X_{ji}(t) \textit{ is observed}, \\ p\left(R_i^*(t) \mid \tilde{\boldsymbol{\psi}}^S, q_i^{S+1}, \boldsymbol{X}_{J'}\right) \text{ right truncated}^* \text{ at } 0, & \text{if } X_{ji}(t) \textit{ is missing}. \end{cases}$
-Sample $\tilde{\boldsymbol{\psi}}^{S+1} \sim p(\tilde{\boldsymbol{\psi}} \mid R^{*^{S+1}}, q_i^{S+1}, \boldsymbol{X}_{J'})$.
-Sample $\Sigma_x^{S+1} \sim p(\Sigma_x \mid w_i^{S+1}, q_i^{S+1})$.
```

* truncated normal distribution.

7.6 Creating Synthetic Data for Simulation

In this section, our primary objective is to generate simulated data deliberately designed to exhibit MNAR in the model response and predictor in a longitudinal context. This will make it suitable for testing the proposed method's performance in handling non-ignorable missingness. The simulation setup employed in this chapter aligns with the simulated data detailed in Section 3.4. The missing values for the response Y were generated based on:

$$U_i^*(t) = \theta_0 + \theta_1 X_{2i}(t) + \theta_2 X_{3i} + v_i \tilde{Z}_i(t) + \varepsilon_i(t), \tag{7.6.1}$$

where θ_0 is the overall intercept, θ_1& θ_2 are regression coefficients associated with the fully observed fixed effects. The missing data indicator U^* for each observation is sampled from the binomial distribution with a success rate equal to the observation's missingness probability from the probit model, where the value is one if the corresponding Y is observed and zero if missing. The values of $\boldsymbol{\theta}$ were derived to produce the desired missing data proportion; Table 7.6.1 shows the values of the missing response indicator model for each missingness percentage and number of repeated measures.

	$\mathbf{p_y = 20\%}$			$\mathbf{p_y = 40\%}$			$\mathbf{p_y = 60\%}$		
Parameter	m=2	m=4	m=8	m=2	m=4	m=8	m=2	m=4	m=8
θ_0	-2	-0.6	-0.6	-2.5	-2.5	-2.5	-1.5	-1.5	-1.5
θ_1	3	0.7	0.7	1.4	1.4	1.4	0.4	0.4	0.4
θ_2	7	4	4	2.2	2.2	2.2	1.2	1.2	1.2

Table 7.6.1: The table presents the values of the coefficients $\boldsymbol{\theta}$ for the missing response indicator regression model, for various proportions of missingness and a number of repeated measures. This information is used to generate missing response values for the GCRE-MNAR method, where $\mathbf{p_y}$ represents the percentage of missing values in the model response, and m indicates the number of repeated measures.

Moreover, v_i are assumed to be $N(0,2)$, the residuals ε_i follow $N(0,1)$ and the covariance matrix associated with the random effects is $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.

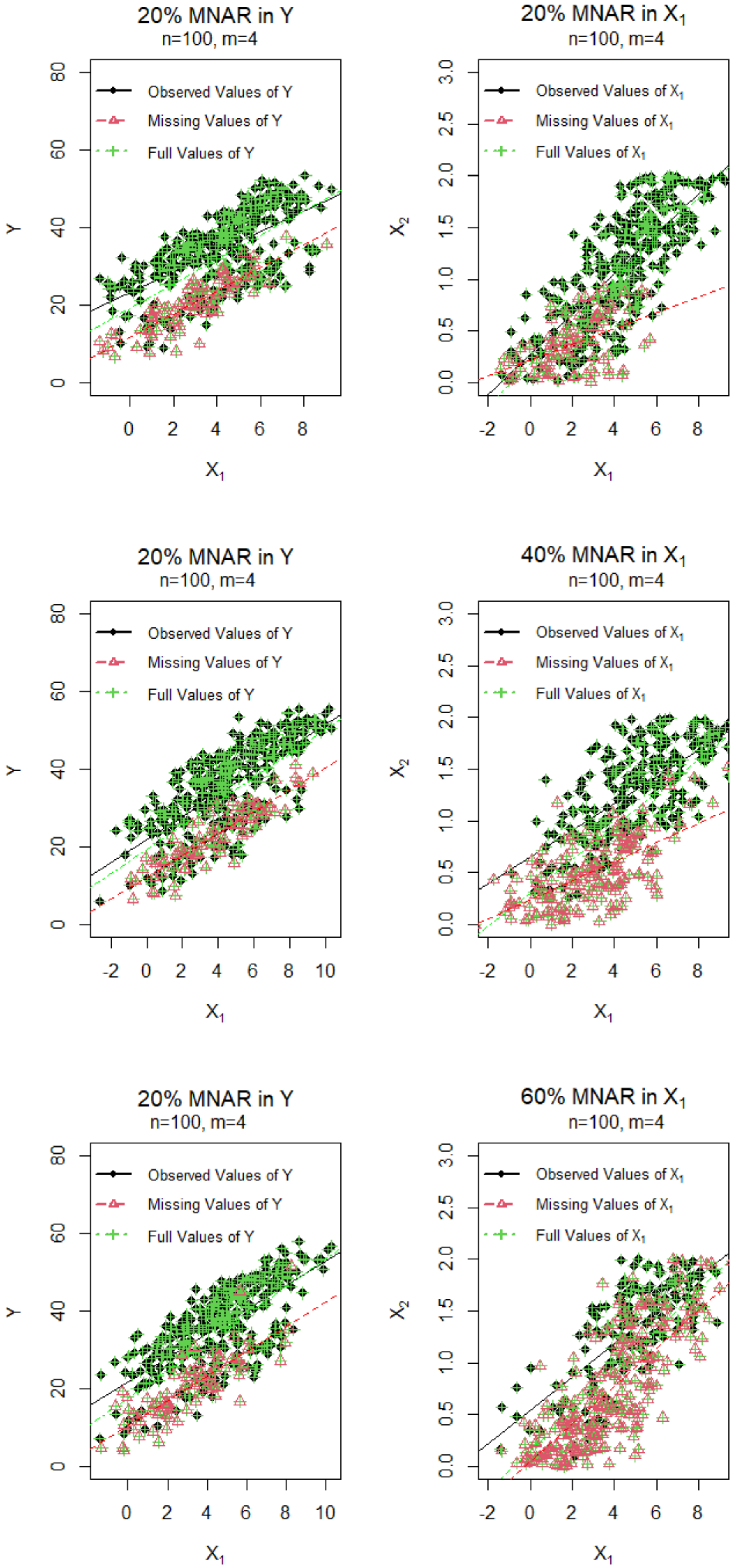
The probit regression equation is used to connect missingness probabilities of the response Y to values of Y through the random component based on the latent missingness indicator regression model U^* for non-ignorable missingness. Furthermore, a different probit regression equation is used to connect missingness probabilities of the incomplete predictor X_1 to the values of X_1 through the random component based on the latent missingness indicator regression equation R_1^*, assuming the incomplete predictor X_1 is missing not at random. These are modelled using the latent variable form for probit regression. The incomplete predictor X_1 missing values were generated based on the missingness indicator $R^*_{i1}(t)$ which is defined as:

$$R^*_{i1}(t) = \psi_0 + \psi_1 X_{2i}(t) + \psi_2 X_{3i} + q_i \tilde{Z}_i(t) + \xi_i(t), \qquad (7.6.2)$$

where ψ_0 is the overall intercept, $\psi_1 \& \psi_2$ are regression coefficients associated with the fully observed fixed effects and the values of $\boldsymbol{\psi}$ will differ based on the desired proportion of missingness and repeated measures as presented in Table 7.6.2. Each observation's missing data indicator $R^*_{i1}(t)$, is sampled from the binomial distribution with a success rate equal to the observation's missingness probability from the probit model, which has a value of one if corresponding $X_{i1}(t)$ is observed and zero otherwise, with a success rate equal to the observation's missingness probability specified using the probit model. The individual's random intercept is assumed $q_i \sim N(0,2)$, and the residuals are assumed $\xi_i \sim N(0,1)$. Since we assume a Correlated Random Effects between the incomplete predictor regression model and the incomplete predictor missing indicator model, the covariance matrix associated with the random effects in the incomplete predictor X_1 can be expressed as $\Sigma_{x_1} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.

	$\mathbf{p_{x_1} = 20\%}$			$\mathbf{p_{x_1} = 40\%}$			$\mathbf{p_{x_1} = 60\%}$		
Parameter	m=2	m=4	m=8	m=2	m=4	m=8	m=2	m=4	m=8
ψ_0	-6	-6	-6	-4	-4	-4	-9	-3	-3
ψ_1	8	8	8	5	5	5	8	2	2
ψ_2	5	5	5	0.1	0.1	0.1	0.4	1	1

Table 7.6.2: The table presents the values of the coefficients $\boldsymbol{\psi}$ for the missing predictor indicator regression model, for various proportions of missingness and a number of repeated measures. This information is used to generate missing predictor values for the GCRE-MNAR method, where $\mathbf{p_{x_1}}$ represents the percentage of missing values in the model predictor, and m indicates the number of repeated measures.


20% MNAR in Y
n=100, m=4
Observed Values of Y
Missing Values of Y
Full Values of Y
Y
X_1
0
20
40
60
80
0
2
4
6
8
20% MNAR in X_1
n=100, m=4
Observed Values of X_1
Missing Values of X_1
Full Values of X_1
X_2
X_1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-2
0
2
4
6
8
20% MNAR in Y
n=100, m=4
Observed Values of Y
Missing Values of Y
Full Values of Y
Y
X_1
-2
0
2
4
6
8
10
40% MNAR in X_1
n=100, m=4
Observed Values of X_1
Missing Values of X_1
Full Values of X_1
X_2
X_1
20% MNAR in Y
n=100, m=4
Observed Values of Y
Missing Values of Y
Full Values of Y
Y
X_1
0
2
4
6
8
10
60% MNAR in X_1
n=100, m=4
Observed Values of X_1
Missing Values of X_1
Full Values of X_1
X_2
X_1

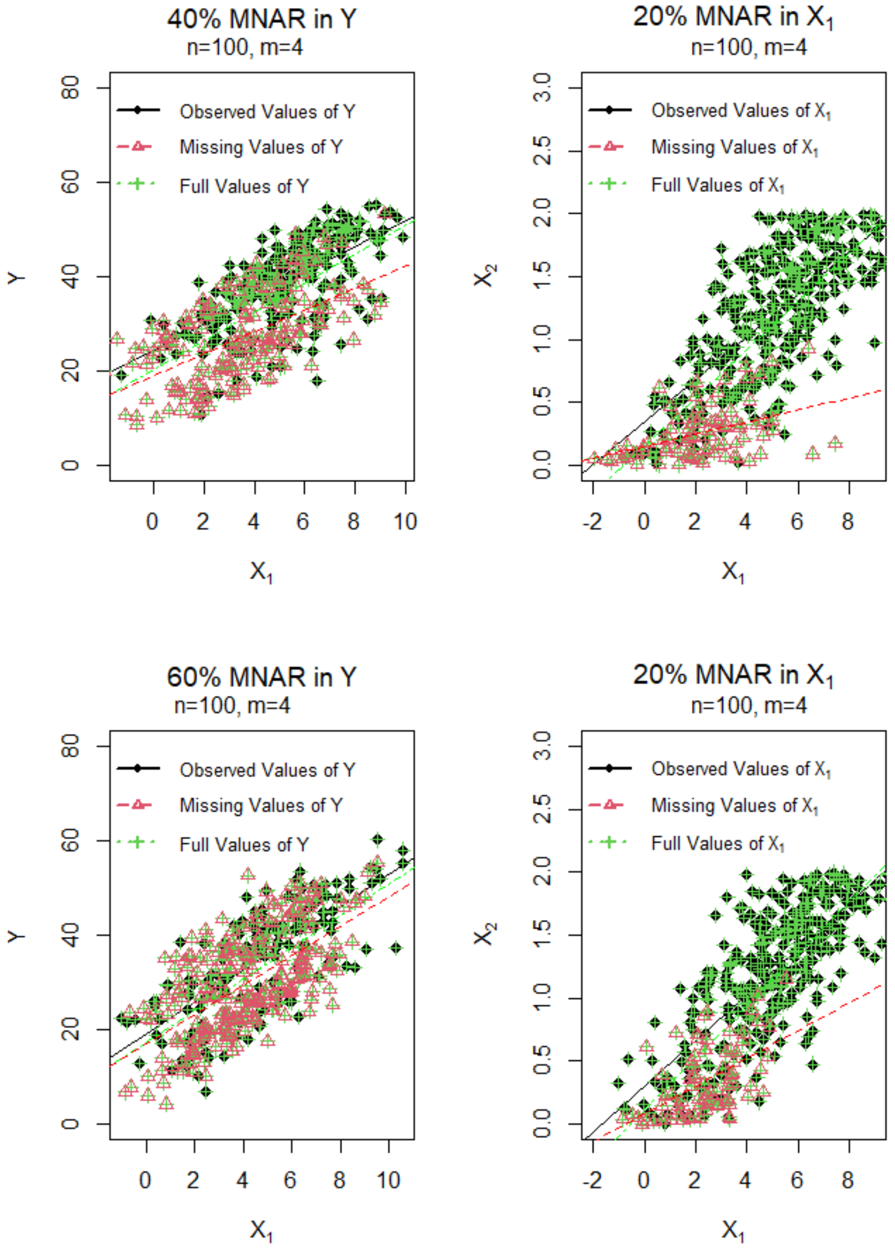


Figure 7.6.1: A scatter plot describing the association between the response (y-axis) and the incomplete predictor (x-axis) on the left-hand side plots and between the complete predictor (y-axis) and the incomplete predictor (x-axis) on the right-hand side plots. The black dots represent observed values after removing missing data, while the red triangles represent missing values. The green crosses show the full data, including both observed and missing values. The plot also displays three regression lines, representing different values and illustrating the varying trends for MNAR in the response and the incomplete predictor. The missingness in the response and incomplete predictor are dependent on the small values of the variable itself, which represent the desired missingness mechanism.

To provide insights into generated data missingness across different proportions of missingness in the model response and the incomplete predictor, Figure 7.6.1 visually represents observed and missing values across the simulated datasets at varying percentages and four repeated measures. The missing values happen with lower values of the variables, which indicates that the missing mechanism is MNAR. Additionally, similar patterns are observed across different repeated measures which are $m = 2$ & 8.

We will use the simulated data to evaluate the GCRE-MNAR method performance so that we are able to assess the proposed approach with the baseline methods. These are the full data method, which uses the fully observed data before eliminating the missing values and capturing an ideal scenario and the available data method, which uses the observed values by discarding any missing data. Additionally, the proposed method will be compared to the CRE method with fully observed predictors which was discussed in Chapter 4 and the GCRE-MNAR method with fully observed predictors in order to assess the effectiveness of the GCRE-MNAR method in the absence of missing values in the incomplete predictor variable and to compare the results of the generalised version of the CRE with the original CRE method. This evaluation will provide valuable insights into these methods' effectiveness and suitability under conditions where both the model's response and predictor missingness are non-ignorable.

R (R Core Team et al., 2013) is used to fit all of the methods, and it performs Markov Chain Monte Carlo (MCMC) is carried out for the proposed method. The brm function in the brms package (Bürkner, 2017) is used to fit the full and available data methods using the HMC method as defined in Section 2.4.5. The MCMC simulations will run in three chains, each with different starting values and a total of 50,000 iterations, with the first half of iterations being regarded as burn-in and a thinning rate of 10 is applied. Convergence is assumed to have been reached if all values were below 1.1, based on the Gelman-Rubin convergence statistic for each individual parameter as mentioned in Section 2.4.6 and through visual trace plots across iterations. All upcoming results are considered to have converged.

We will evaluate the GCRE-MNAR method's performance by comparing the accuracy of its parameter estimates. We will use the Root Mean Square Error (RMSE), Relative Bias (RB), and Coverage Rate (CR) as measures of accuracy for individual parameters. However, different parameter magnitudes could distort the results and hence to ensure equal weighting across all model of interest parameters, we will use Weighted Root Mean Square Errors (WRMSE) for overall method accuracy, in which we assigned weights based on the data-generating parameters. We defined these criteria in Section 2.7, and we will evaluate them across 100 replications of the simulated data.

7.7 Results

This section presents the results of the proposed GCRE-MNAR method and the comparative methods, applied to simulated data and the BIOSTAT-CHF dataset.

7.7.1 Simulation Data Results

Across different sample sizes and missingness proportions, the simulated data values and averaged imputed values from the Gibbs sampler iterations using the GCRE-MNAR method of the response and incomplete predictor approximately match, showing alignment. Furthermore, the averaged imputed values' ± 2 Standard Deviation (SD) range contains simulated data values. Figure 7.7.1 represents one case scenario when there is 20% missingness in the analysis model response and 40% missingness in the incomplete predictor in the analysis model with different numbers of repeated measures. This shows that the imputed values from the proposed GCRE-MNAR method match the simulated data values. Comparable observations were also drawn for other scenarios of different combinations of proportion of missingness and repeated measures. These findings show that the GCRE-MNAR approach is effective in capturing missing values that resemble the data-generating values, indicating its reliability and efficiency.

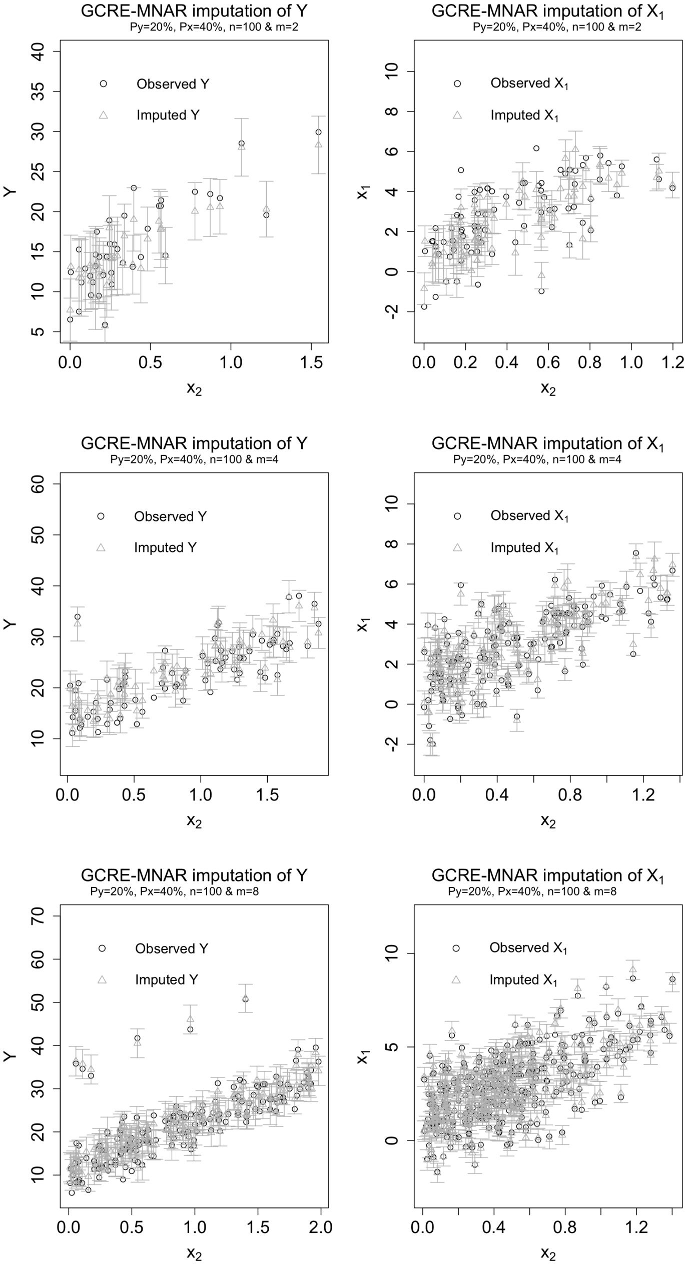


Figure 7.7.1: Scatter plots represent the model response Y on the y-axis and the fully observed predictor X_2 on the x-axis (on the left-hand side). The model incomplete predictor X_1 on the y-axis and the fully observed predictor X_2 on the x-axis (on the right-hand side). These plots are used to assess the simulated data values (depicted as black circles) against the average of imputed values and the ± 2 SD (represented by grey triangles and vertical lines) using the GCRE-MNAR method for various repeated measures, with 20% missingness in the model response and 40% missingness in the model predictor. The simulated data values are mostly enclosed within the imputed values region.

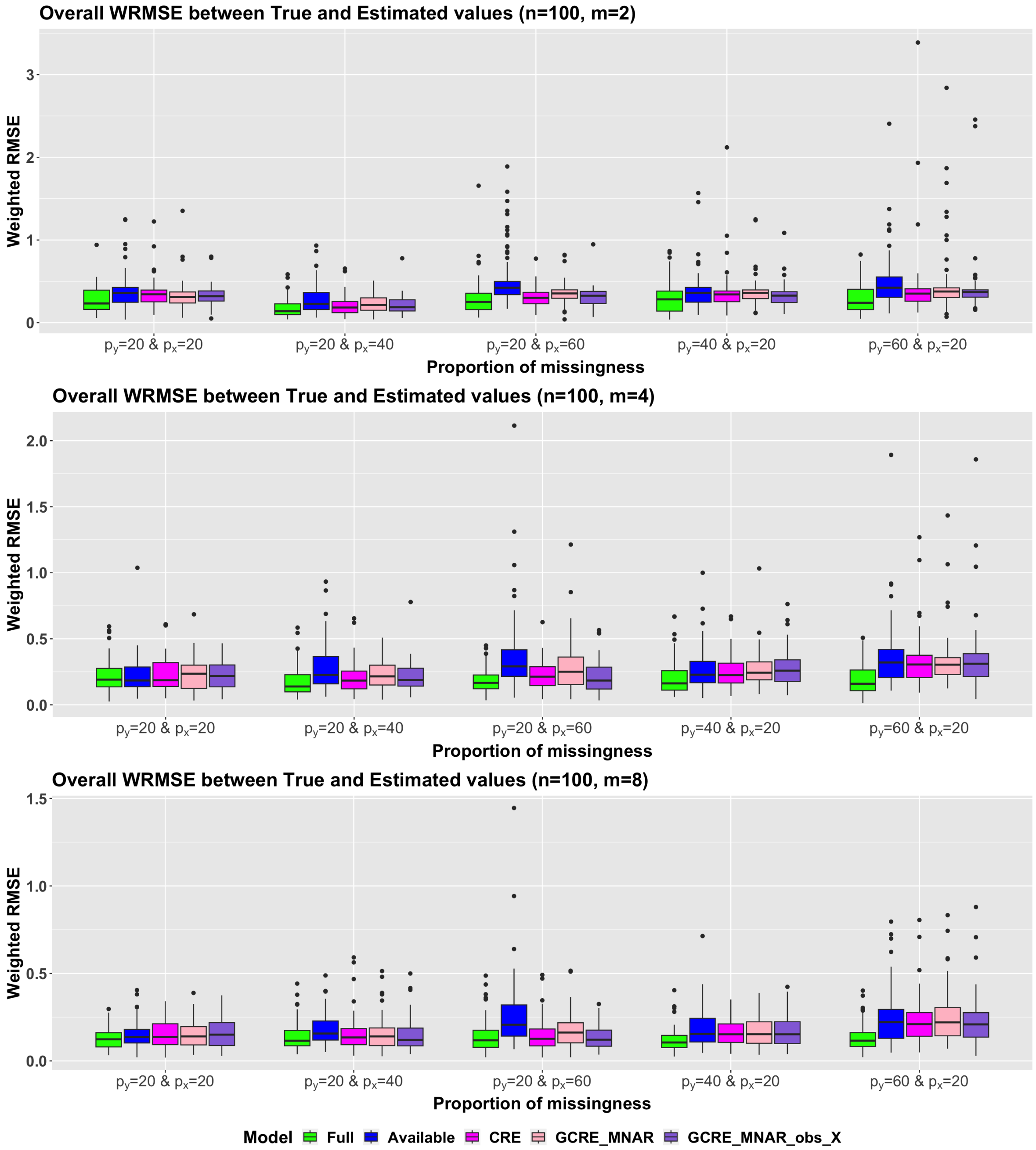


Figure 7.7.2: Boxplots represent the overall WRMSE for each method and different proportions of missing data and repeated measures. The y-axis represents the WRMSE values, while the x-axis represents one of the proportions of missing data in the model response and incomplete predictor. Each boxplot corresponds to one of the applied methods, and each plot represents different repeated measures. The results show that all methods have similar WRMSE, with a slight increase in the available data method.

Figure 7.7.2 displays the performance of the applied methods across different sample sizes and proportions of missing data. The boxplots show the overall WRMSE across all model parameters. The proposed GCRE-MNAR method generally outperforms the available data. The available data shows higher WRMSE uncertainty when the incomplete predictor has large proportions of missingness (40% and 60%). However, the GCRE-MNAR method produces comparable results with the available data when there is 20% missingness in the analysis model response and incomplete predictor, regardless of the number of repeated measures. On the other hand, the GCRE-MNAR, GCRE-MNAR with fully observed predictors, and CRE methods have very similar WRMSE performance. However, with 60% missingness in the incomplete predictor, the GCRE-MNAR method slightly shifts towards a larger WRMSE. GCRE-MNAR encounters an additional challenge than the CRE method, as the CRE method has fully observed predictor data that would not typically be available in practice.

Figure 7.7.3 shows the RMSE distribution for β_0 and β_1, while β_2 and β_3 plots are included in Section D the Appendix. We will explore the differences for each analysis model parameters. The RMSE analysis, which shows the degree of variation between the estimated and data-generating parameters, is used to evaluate the accuracy of the methods. It is generally observed that the available data method produces higher RMSE for the analysis model fixed effects coefficients as compared to other applied methods. GCRE-MNAR, GCRE-MNAR with observed predictors, and CRE methods have very similar RMSE values, except when there is 60% missingness in the analysis model incomplete predictor, where the GCRE-MNAR produces larger RMSE values. It's worth noting that for the slope β_1, the GCRE-MNAR produced lower RMSE values and uncertainty compared with GCRE-MNAR with observed predictor and CRE methods when there is 20% missingness in the analysis model response and incomplete predictor with two repeated measures.

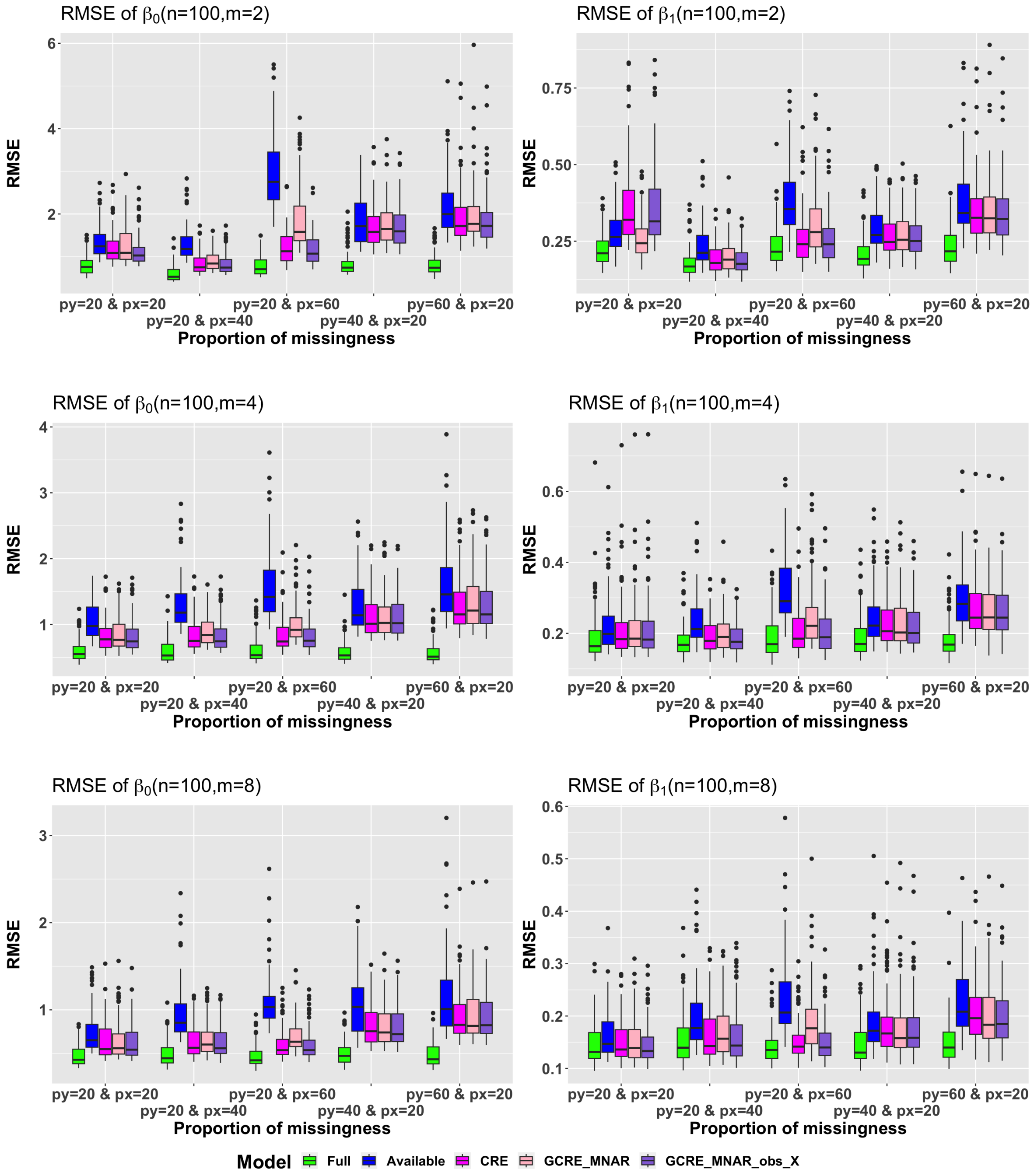


Figure 7.7.3: Boxplots illustrate the RMSE of β_0 in the left-hand side plots, and β_1 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RMSE values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. The available data method produces larger RMSE values compared to other methods.

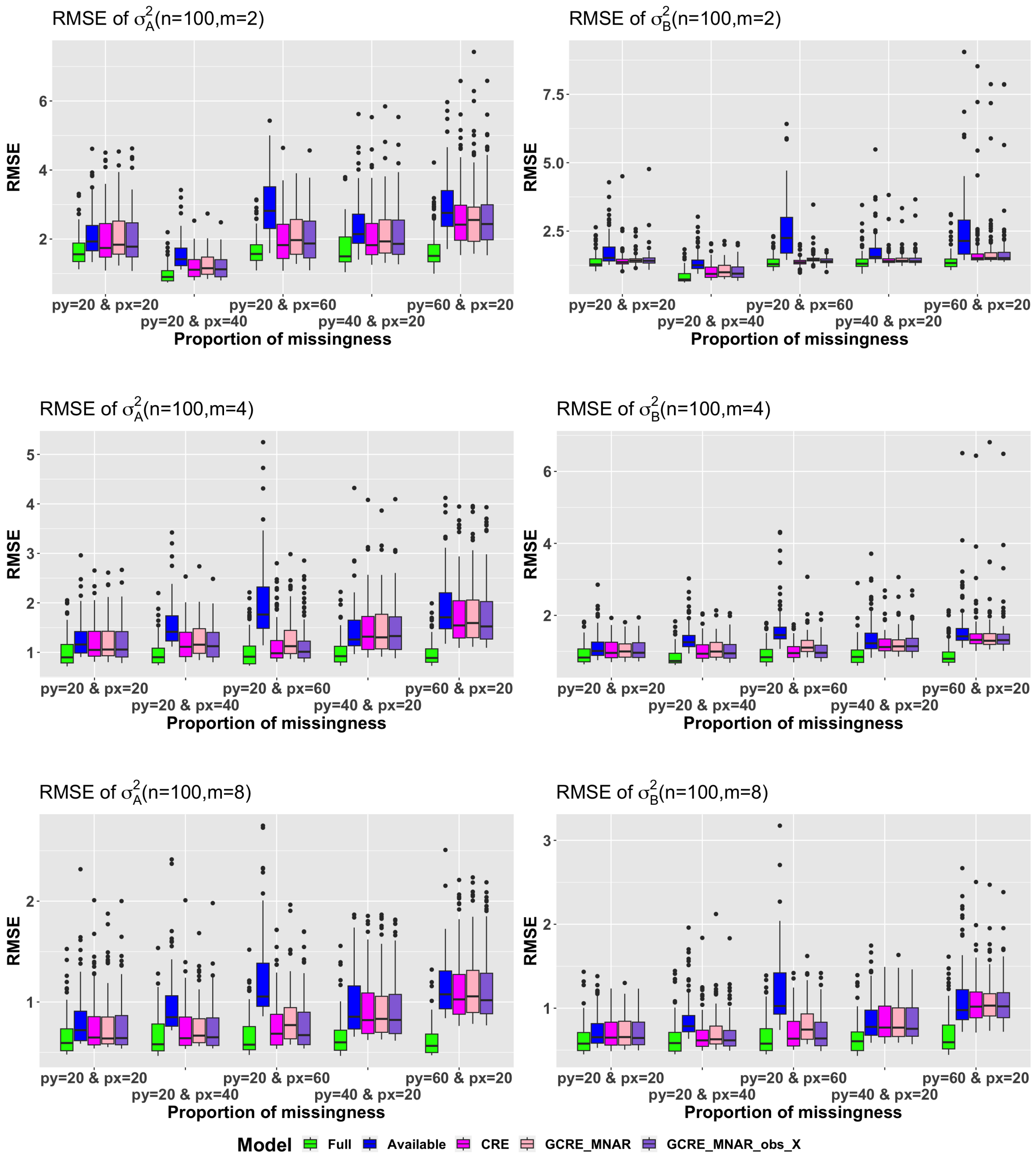


Figure 7.7.4: Boxplots illustrate the RMSE of σ_A^2 in the left-hand side plots, and σ_B^2 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RMSE values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. The available data method produces larger RMSE values and even larger for σ_B^2 with $m = 2$ compared with other methods

Regarding the variance parameters in the analysis model, the available data method generally produces higher RMSE values overall, especially when there is 60% missing data in the incomplete predictors, and larger RMSE values and uncertainty for σ_B^2 with two repeated measures. The GCRE-MNAR method tends to have even larger RMSE values than the GCRE-MNAR with fully observed predictor and CRE methods when there is 60% missingness in the incomplete predictor but is similar otherwise. It is worth noting that Figure 7.7.4 shows that with 60% missingness in the response, the applied methods generally result in larger RMSE extreme values, except for the full data method. The comparable RMSE between GCRE-MNAR with fully observed predictor and CRE methods indicates that the CRE method is a special case of the GCRE-MNAR when no missing values exist in the analysis model predictor.

The distribution of the RB for β_0 and β_1 in Figure 7.7.5, and β_2 and β_3 in Section D in the Appendix. The RB analyses the model's parameters estimates in order to spot trends in over- or underestimation of the parameters that generate the data across all applied methods. Overall, the applied methods produce unbiased estimates of the analysis model fixed effects parameters. However, the available data method tends to have larger RB uncertainty when there is 60% missingness in the incomplete predictor compared to other applied methods. It also underestimates the intercept β_0 and overestimates slope β_3. The over and underestimations of these parameters increase as the repeated measures increase.

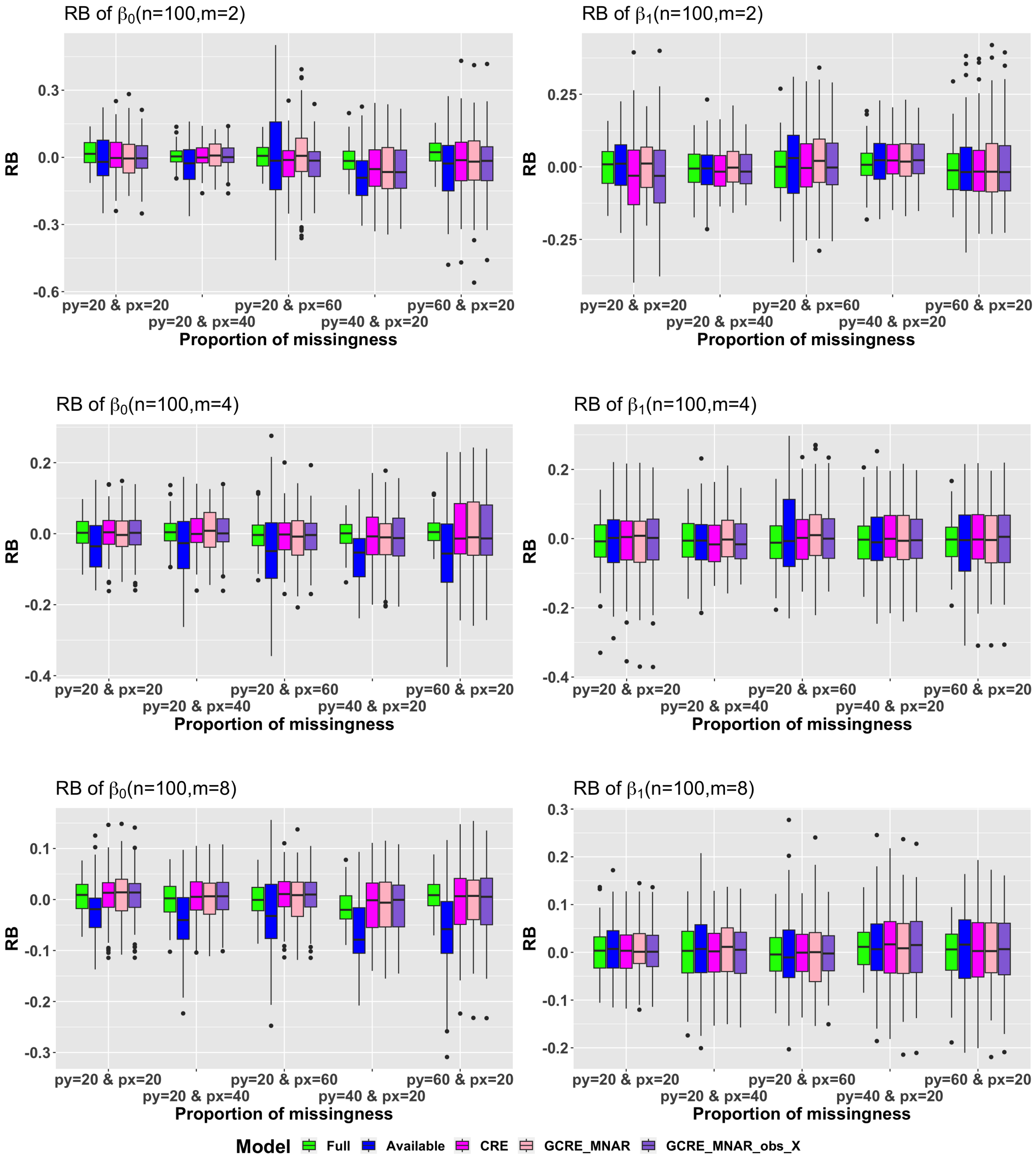


Figure 7.7.5: Boxplots illustrate the RB of β_0 in the left-hand side plots, and β_1 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RB values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. The available data method underestimates the intercept β_0 as repeated measures increase, though the other methods produce unbiased estimates.

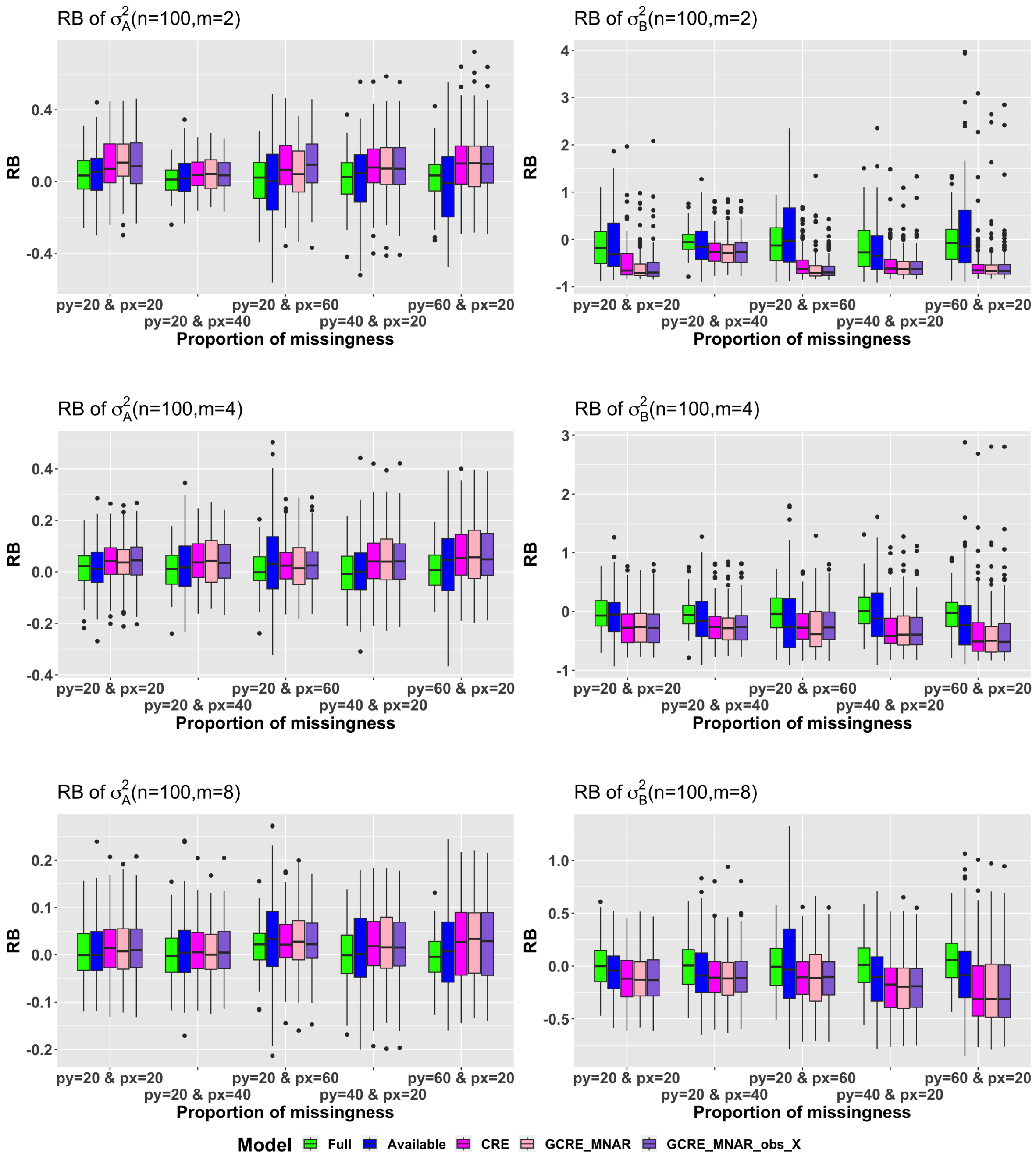


Figure 7.7.6: Boxplots illustrate the RB of σ_A^2 in the left-hand side plots, and σ_B^2 in the right-hand side plots for each method applied to various proportions of missingness and different repeated measures. The y-axis shows the RB values, and the x-axis represents one of the proportions of missingness in the model response and incomplete predictor. Each boxplot represents one of the applied methods. The GCRE-MNAR method produces unbiased estimates of σ_A^2 and underestimates σ_B^2 as the repeated measure decreases.

Based on the RB values of the response model's variance components shown in Figure 7.7.6, the applied methods yield unbiased estimates for the residual variance of the analysis model, σ_A^2. However, the available data method has a large RB uncertainty with 60% in the incomplete predictor and a high proportion of missingness in the response (40% and 60%) with $m = 2$ & 4. The GCRE-MNAR, GCRE-MNAR with fully observed predictor, and CRE methods have similar performance in underestimating the between-individual variance of the analysis model, σ_B^2. However, this biasedness decreases as the number of repeated measures increases. The available data method produces unbiased estimates of σ_B^2 but with larger RB uncertainty compared to other methods.

Our results compare the accuracy of the GCRE-MNAR method and the available data method for estimating analysis model parameters, which showed that the GCRE-MNAR method had a smaller RMSE, indicating more precise parameter estimation than the available data method. The available data method had biased estimates for some analysis model coefficients, while the GCRE-MNAR method was generally unbiased. However, the GCRE-MNAR method produced biased estimates for between-individual variance with fewer repeated measures. In summary, the GCRE-MNAR method reduces the spread of estimation errors and generally provides consistent, unbiased estimates.

To assess how accurately each method captures the data-generating parameters of the analysis model, we used the CR, the corresponding plots can be found in Section D in the Appendix. In most cases, the CR ranges from 0.9 to 0.99. However, there are a few exceptions where the available data method was unable to capture the data-generating parameters of β_0 and β_3. Specifically, for β_0, the available data method failed to capture the actual parameter values when there is 60% missingness in the response with $m = 4$ & 8 and when there is 40% missingness in the response with $m = 8$. For β_3, the available data method was unable to capture the actual parameter values when there is 40% missingness in the response with $m = 8$ and when there is 40% missingness in the incomplete predictor with $m = 2$ & 4.

The proposed GCRE-MNAR method was not able to capture the actual parameter values for β_0 and β_1 with $m = 2$ when there is 60% missingness in the incomplete predictor, but the CR is only slightly below 0.9. Also, for σ_B^2 with $m = 2$ in all combinations of proportions except when there is 60% missingness in the response and 40% missingness in the incomplete predictor. With $m = 4$, the GCRE-MNAR, GCRE-MNAR with fully observed predictor and CRE methods were unable to capture the true parameter values of σ_B^2 in all combinations of proportions except when there is 40% missingness in the incomplete predictor. With $m = 8$, both the GCRE-MNAR, GCRE-MNAR with fully observed predictor and CRE methods were unable to capture the actual parameter values of σ_B^2 when there is 40% missingness in the response.

Regarding the out-of-sample prediction performance, the GCRE-MNAR method outperforms the available data method in various combinations of missingness in the model response, incomplete predictor, and repeated measures. However, when there are 20% missing values in both the model response and incomplete predictor with two repeated measures, the proposed method performs similarly to the available data method. Figure 7.7.7 presents the out-of-sample performance for 20% missingness in the model response and 60% missingness in the incomplete predictor across different methods and repeated measures. The plots displaying the remaining proportion of missingness can be found in Section D of the Appendix.

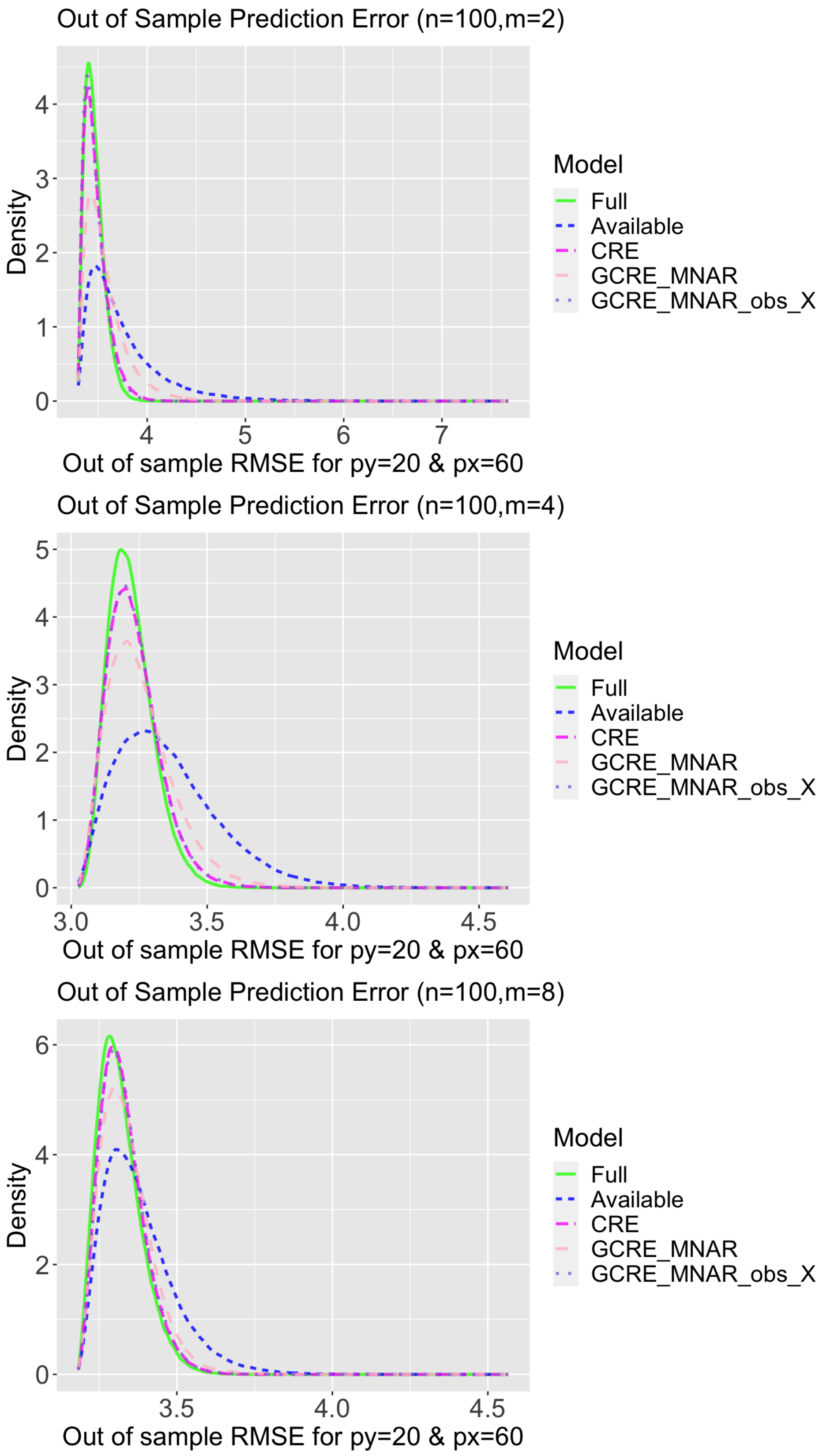


Figure 7.7.7: The density plots of the out-of-sample RMSE for different methods across various repeated measures with 20% missingness in the response and 60% in the incomplete predictor. Each density curve corresponds to one of the methods used, and each plot corresponds to a different value of repeated measures. The available data method appears to have less density and slightly shifted to to higher RMSE values.

RB boxplots of the missingness response model parameters across generated simulated datasets, various proportions of missingness and repeated measures are in Figure 7.7.8, and the plots for RMSE and CR can be found in Section D in the Appendix. The RMSE of the coefficient parameters for the missingness response model using the GCRE-MNAR method decreases as the number of repeated measures increases. Out of all the missingness response model parameters, θ_1 has lower RMSE values and uncertainty. As the number of repeated measures increases, the missingness response model parameters become less biased, and the σ_D^2, which is the between response models covariance parameter, has larger RB uncertainty.

When there are only two repeated measures, the GCRE-MNAR method is unable to capture the data-generating parameters for the missingness response model fixed effect coefficients, especially when there is 40% missingness in the incomplete predictor, and the variance parameters when there is 20% missingness in the model response and incomplete predictor and 60% missingness in the incomplete predictor.

The intercept θ_0 shows low CR with $m = 4$ & 8 when there is 40% missingness in the response. σ_D^2 has a slightly lower CR than 0.9 when there is 40% missingness in the incomplete predictor with four repeated measures. With 60% missingness in response σ_D^2 has a low CR when there are four and eight repeated measures. Otherwise, the CR ranges between 0.9 and 0.99.

Figure 7.7.9 shows the RB plot associated with the incomplete predictor model parameters, and the RMSE and CR plots can be found in section D in the Appendix. According to the incomplete predictor model parameters, the GCRE-MNAR model produced low RMSE values for the residual variance σ_F^2 and the slope α_1 parameter of the incomplete predictor model. These values had lower RMSE uncertainty compared to other model parameters. Relative bias was also found to be unbiased; however, RB uncertainty for α_2 was wider.

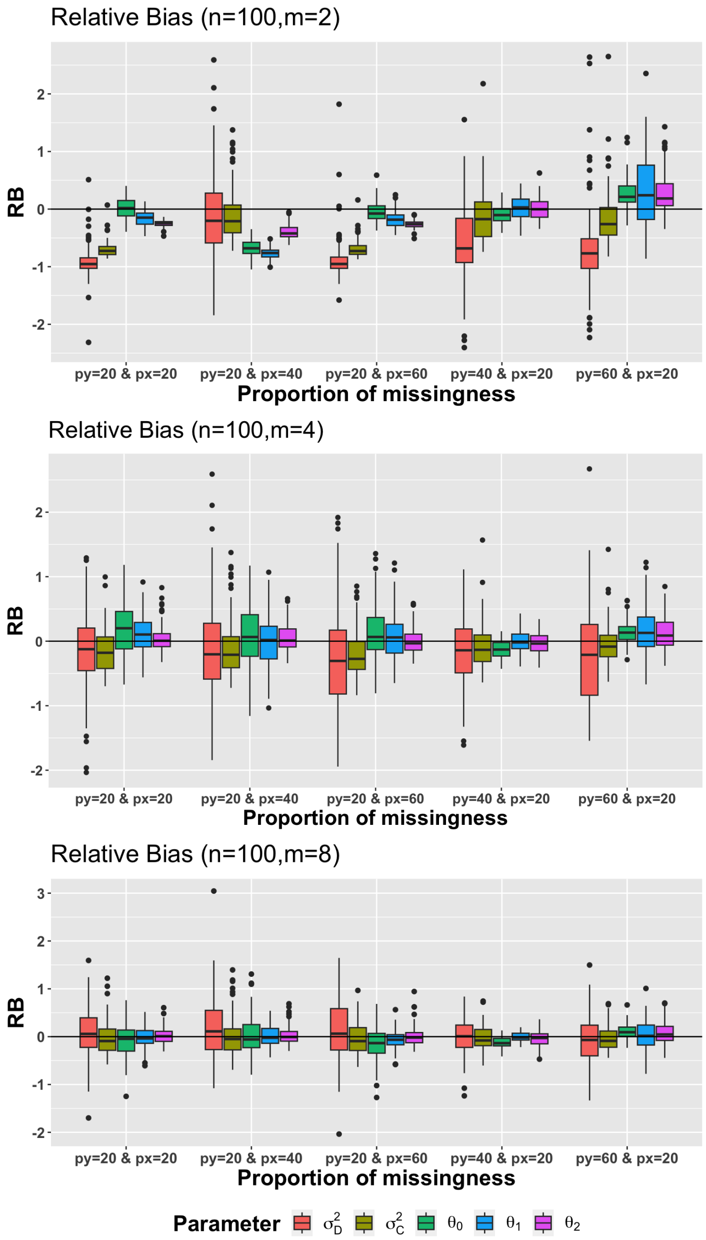


Figure 7.7.8: Boxplots illustrate the RB of the missingness response model parameters using the GCRE-MNAR method across various proportions of missingness and repeated measures. The y-axis represents the RB values, while the x-axis shows the proportion of missingness, and each boxplot represents a specific missingness response model parameter. Each plot corresponds to a different repeated measure value. As the repeated measures increase, the missingness response model parameters become unbiased.

For the incomplete predictor model, the GCRE-MNAR method was able to capture the data-generating parameter values where CR varied between 0.9 and 0.99, except for when there were 20% missing values in the model response and incomplete predictor with $m = 8$, σ_E^2, α_1, and α_2, which had low

CR. When there were four repeated measures, α_2 had low CR when there were 60% missing values in the incomplete predictor, and σ_E^2 had low CR when there were 60% values in the response. The residual variance σ_F^2 had low CR with $m = 2$ when there were 20% missing values in the model response and incomplete predictor, as well as when there were 60% missing values in the incomplete predictor.

As the repeated measures increase, the RMSE and the RB (in Figure 7.7.10) associated with the incomplete predictor missingness model parameters decrease. Interestingly, when the incomplete predictor has a high proportion of missing values (40% and 60%), the RMSE and RB are lower than when the incomplete predictor has only 20% missing values, regardless of the proportion of missingness in the response. The GCRE-MNAR method can capture the data-generating parameters of incomplete missingness model parameters when there is a higher proportion of missingness in the incomplete predictor (40% and 60%) with $m = 4$ & 8. Additionally, the covariance parameter σ_H^2 is captured regardless of the proportion of missingness with $m = 8$. However, with two repeated measures, the GCRE-MNAR method struggles to capture the data-generating parameters for the incomplete missingness model parameters.

Using the GCRE-MNAR method, the missingness response model parameters, the incomplete predictor model parameters and the missingness incomplete predictor model parameters can be estimated effectively with larger sample sizes (larger number of repeated measures). As the sample size increases, the accuracy and precision of the estimates improve, as indicated by the low RB and RMSE values and high CR values. These results suggest that the method is capable of producing reliable estimates that closely match the actual values, especially when working with more repeated measurements taken over time.

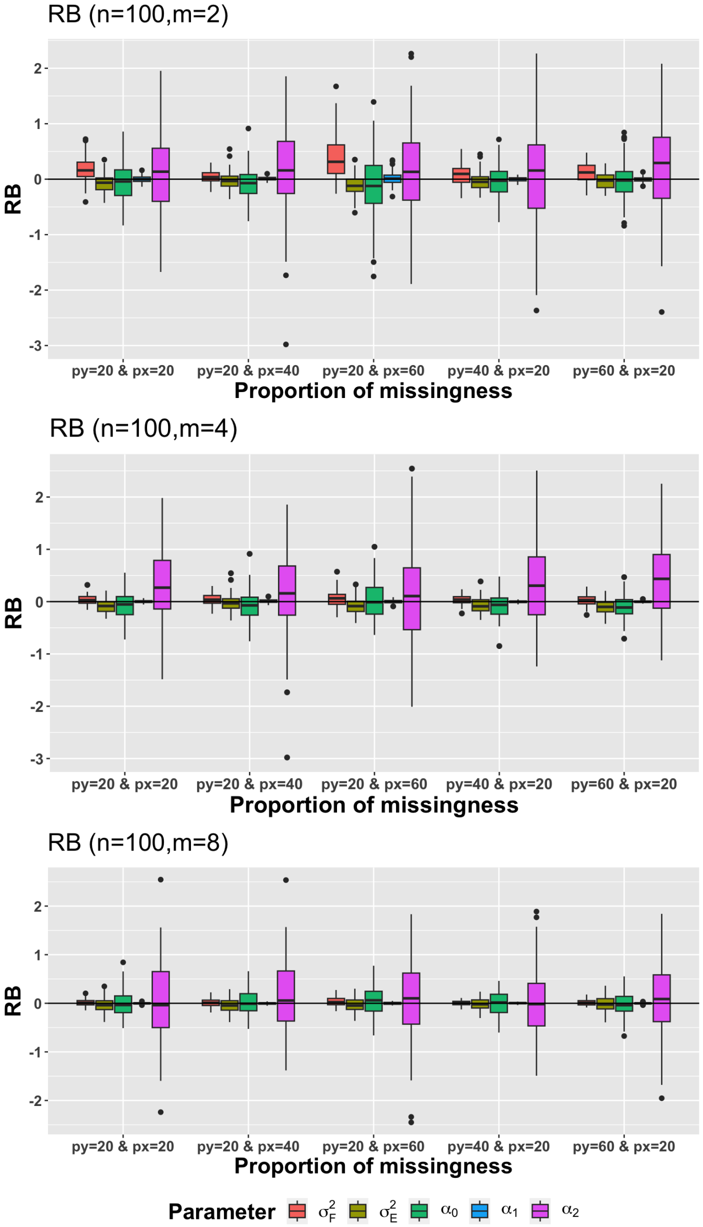


Figure 7.7.9: Boxplots illustrate the RB of the incomplete predictor model parameters, using the GCRE-MNAR method across various proportions of missingness and repeated measures. The y-axis represents the RB values, while the x-axis shows the proportion of missingness, and each boxplot represents a specific missingness response model parameter. Each plot corresponds to a different repeated measure value. The GCRE-MNAR method produces unbiased estimates of the incomplete predictor model parameters, with larger RB uncertainty for α_2.

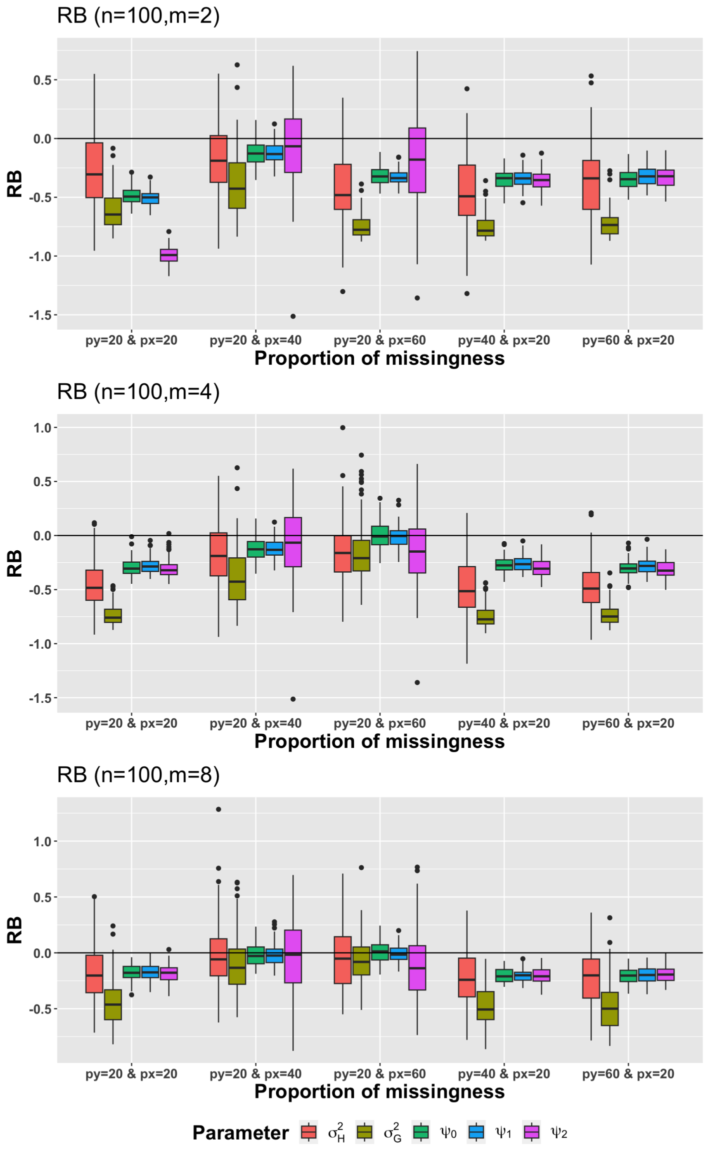


Figure 7.7.10: Boxplots illustrate the RB of the incomplete predictor missingness model parameters, using the GCRE-MNAR method across various proportions of missingness and repeated measures. The y-axis represents the RB values, while the x-axis shows the proportion of missingness, and each boxplot represents a specific missingness response model parameter. Each plot corresponds to a different repeated measure value. With large missingness in the incomplete predictor, the incomplete predictor missingness model parameters tend to be more unbiased compared with a smaller proportion of missingness.

7.7.2 Real Data Results

We will apply the GCRE-MNAR method to the real-world BIOSTAT-CHF dataset. We also employed the available data approach, a default strategy for handling repeated measures. To implement the GCRE-MNAR and available data approaches on the BIOSTAT-CHF dataset, we used the model described in Equation 4.2.2 in Section 4.2.

Figure 7.7.11 shows the parameter's posterior distribution for the GCRE-MNAR and available data approaches. There is a clear distinction between the two approaches' intercept density curves. This finding is supported by the Kolmogorov-Smirnov test, which shows that only eGFR and Pacemaker have p-values higher than the 0.05 significance level. This suggests that the null hypothesis (there is no difference between the two methods) is not rejected. The missingness response model and the model of interest have a random effects covariance of $\sigma_D^2 = -0.24$, which indicates weak to moderate evidence of MNAR (Missing Not at Random) of the model's response. However, the random effects covariance between the incomplete predictor model and the missingness incomplete predictor process is $\sigma_H^2 = -0.84$, which indicates strong evidence of MNAR (Missing not at random) in the model's incomplete predictor, "eGFR".

We explore the performance of the GCRE-MNAR approach for data imputation in Figure 7.7.12, which displays the density for the observed and GCRE-MNAR imputed response and predictor variables for each draw of the latent variable from the posterior using Gibbs sampler. The similarities between the observed and imputed data from the GCRE-MNAR approach density curves imply that the general data distribution was successfully retained during the imputation procedure.

Next, we will separate the data into a training set and a test set to evaluate further the proposed GCRE-MNAR method's effectiveness with unseen data.

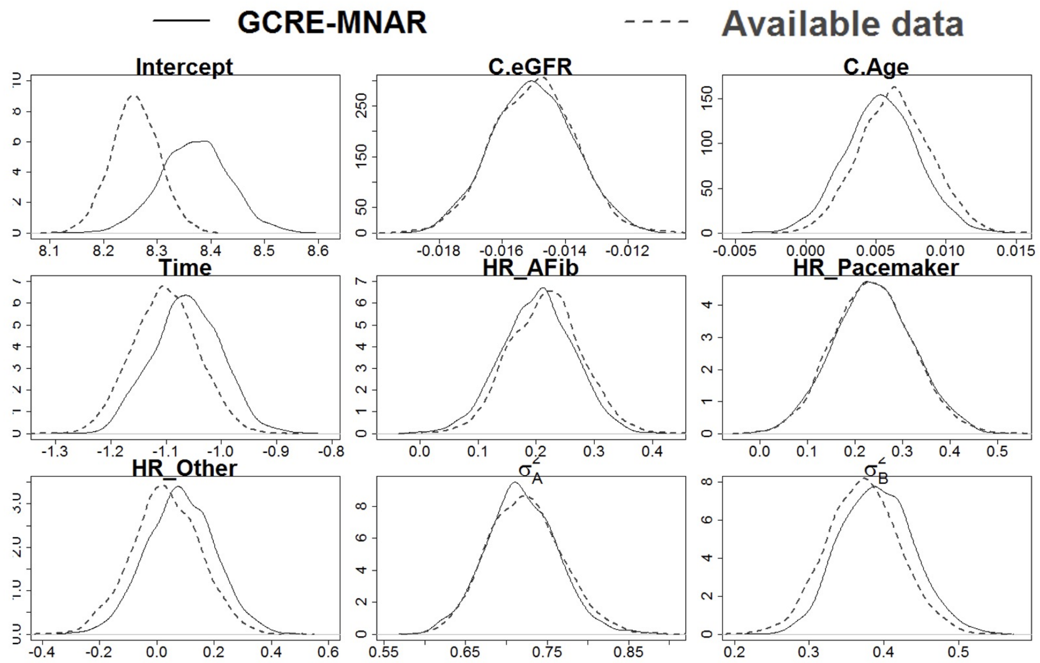


Figure 7.7.11: Posterior distribution of the GCRE-MNAR method in a black solid curve and the available data method in a grey dashed curve, each plot represents one of the BIOSTAT-CHF model parameters. Density curves overlay except for the model intercept and perhaps to some extent, the coefficient for the Time variables and HR Other.

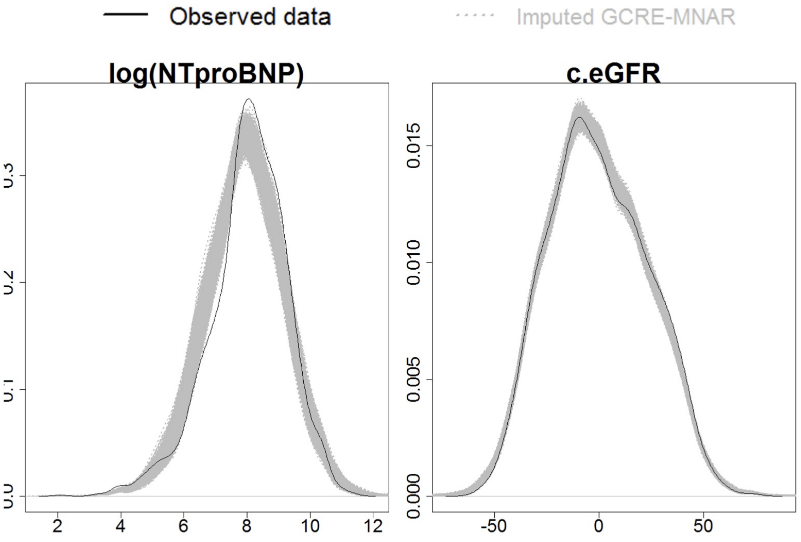


Figure 7.7.12: Density plots show the observed values in the black solid curve and the GCRE-MNAR imputed response (left-hand) and predictor (right-hand) values in the grey dashed curves, where each curve is a different draw of the latent variable from the posterior using Gibbs sampler. The density of the imputed values using the GCRE-MNAR method at each Gibbs sampling iteration is similar to the density of the observed values.

Case 1: Split the data into test and training data, where the test data has fully observed values.

This section aims to assess the performance of the proposed method for handling missing values in comparison to the available data method. We will split the BIOSTAT-CHF data into test and training sets to assess the approaches' ability to generalise to data it has not been built upon. The test data will be considered fully observed, while the training data will have missing values for both the predictor and response variables. Out of the total 2516 participants in the study, only 395 have complete data, which accounts for 15% of the total population. Therefore, we will allocate 85% of the data to the training set and 15% to the test set.

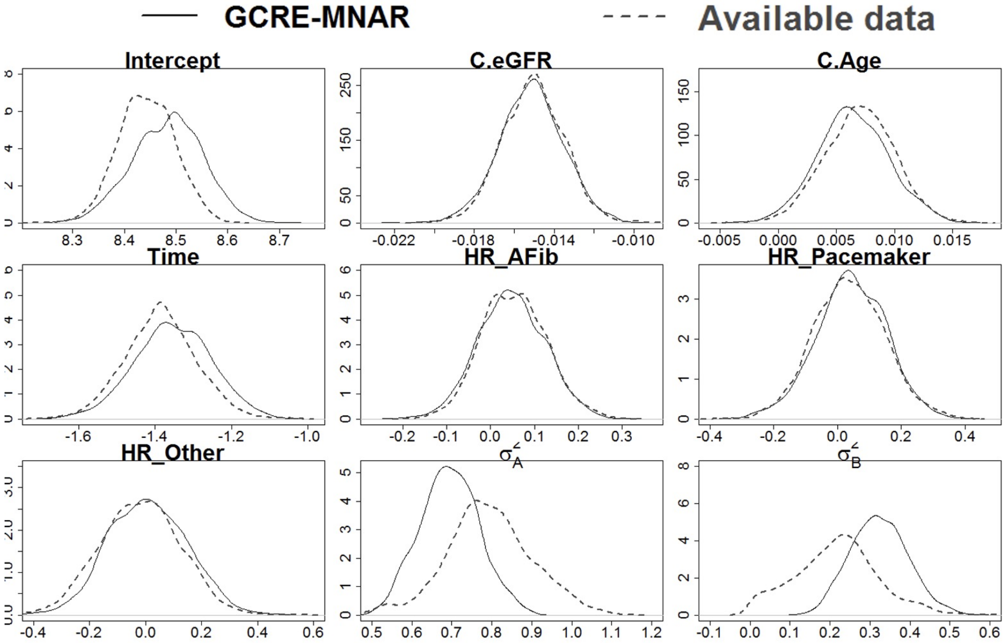


Figure 7.7.13: Posterior distribution of the GCRE-MNAR method in black solid curve and the available data method in grey dashed curve, each plot represents one of the BIOSTAT-CHF training data model parameters in case 1. Density curves overlay except for the model intercept and the variance parameters.

In this analysis of case 1, we compare the GCRE-MNAR and available data approaches' posterior distributions of the training data to gain insights into their performance. The resulting posterior distributions are shown in Figure

7.7.13, which shows that the model's intercept and variances exhibit distinct density curves between the two methods. Furthermore, the Kolmogorov-Smirnov test assessed the differences between the two approaches. For a significance level of 0.05, the null hypothesis was rejected for all variables except for eGFR. The covariance between the random effects of the response is $\sigma_D^2 = -0.052$, which suggests a weak indication of MNAR (Missing Not at Random) of the response "NT-proBNP" in the training data. On the other hand, the covariance between the random effects of the incomplete predictor "eGFR" is $\sigma_H^2 = -0.57$, which suggests a moderate/ strong indication of MNAR (Missing Not at Random) of the incomplete predictor "eGFR" in the training data.

Figure 7.7.14 displays a visual comparison between the RMSE of the GCRE-MNAR and the available data approaches. The out-of-sample performance between both methods is statistically significantly different, as indicated by a Kolmogorov-Smirnov test with a p-value of less than 0.05. The density plot of the available data RMSE exhibits a slight shift towards smaller values, suggesting a higher predictive performance.

The density for the observed and GCRE-MNAR imputed response, and predictor variables are shown in Figure 31 in the Appendix. The GCRE-MNAR method's imputed response overlaps well with the test data, indicating a reasonable imputation similar to the complete case. On the other hand, the training data shows considerably different density patterns. We observe a similarity between the training data distribution and the imputed data for the incomplete predictor variable using the GCRE-MNAR method while displaying differences from the test data distribution; this demonstrates an effective imputation procedure for the observed data, suggesting that the imputation of the incomplete predictor data matches well with the training set. The imputed response data shows a larger deviation from the training data, whereas the training data is more similar to the imputed incomplete predictor data. This is similar to the findings in the GCRE-MAR results in Chapter 6.

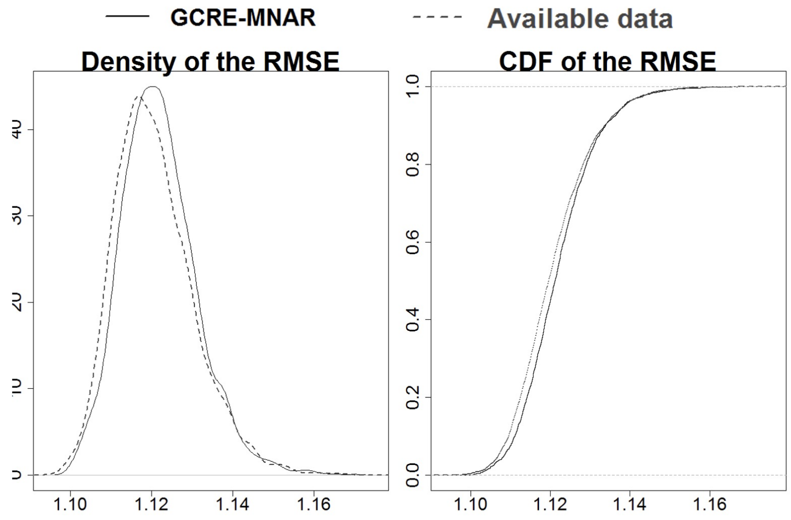


Figure 7.7.14: The RMSE density (on the left-hand side) and CDF (on the right-hand side) of the out-of-sample prediction in BIOSTAT-CHF training data in case 1, where the black solid curve represents the GCRE-MNAR method and the grey dashed curve represents the available data method. The RMSE of the available data method is shifted toward lower RMSE values.

Case 2: Split the data into test and training data, where the test data has missing values in the predictor.

In this case, the training data contains missing values in both the response and predictor variables, while the test data only has missing values in the predictor variable. The split percentages for the training and test data are the same as in Case 1.

To understand the performance of the GCRE-MNAR and available data approaches, we compare their posterior distributions of the training data in this analysis of case 2. Figure 7.7.15 displays the posterior distributions that originated from these methods. The similarity is observed by looking at the overlap in the posterior distribution of model parameters between the GCRE-MNAR and the available data approaches. Heart Rhythms: Atrial Fibrillation and Others and eGFR have p-values greater than 0.05 according to the Kolmogorov-Smirnov test results, indicating no significant difference between the two methods in these parameters. It is interesting to note that the variance components of the two approaches show quite a difference in regard to the posterior distributions. The random effects' covariance is $\sigma_D^2 = -0.01$, suggesting weak evidence of MNAR (Missing Not at Random) in the model's response and moderate/strong evidence of MNAR (Missing Not at Random) in the model's incomplete predictor $\sigma_H^2 = -0.65$.

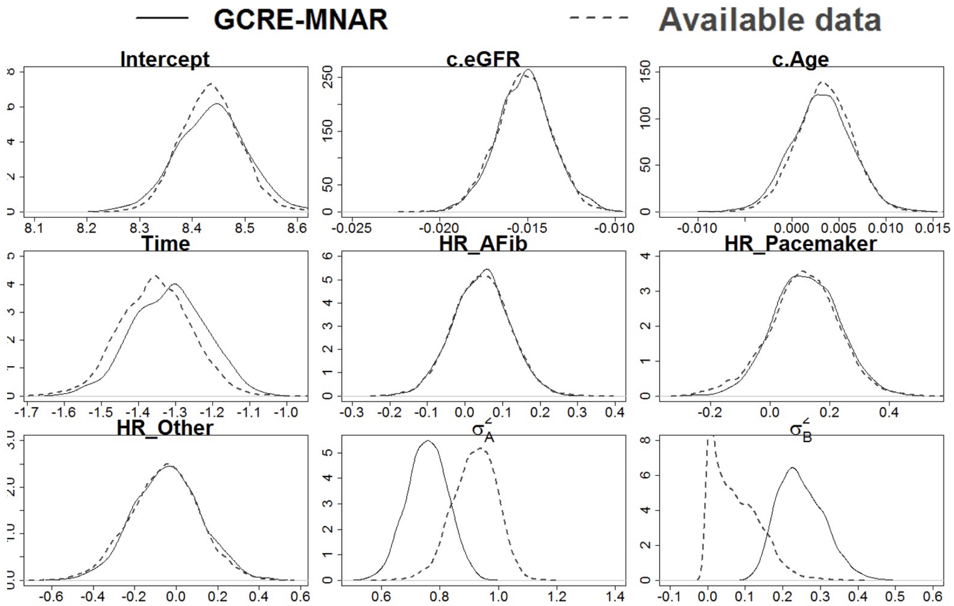


Figure 7.7.15: Posterior distribution of the GCRE-MNAR method in black solid curve and the available data method in grey dashed curve, each plot represents one of the BIOSTAT-CHF training data model parameters in case 2. Density curves overlay except for the variance parameters.

To assess the method's out-of-sample performance, the GCRE-MNAR method was applied to the test data to impute the missing values of the incomplete predictor, eGFR, and only fully observed values of the eGFR are used in the available data method to calculate the out-of-sample-performance. The RMSE of the available data method exhibits a slight shift towards larger values compared with the GCRE-MNAR method, as expressed in Figure 7.7.16. The Kolmogorov-Smirnov test indicates that both methods don't perform comparably. Comparing RMSE distribution in Case 2 and Case 1, we observe that both distributions are similar and centred around very similar values, but in Case 2, the GCRE-MNAR method outperformed the available data method in terms of RMSE. This suggests that the performance of the GCRE-MNAR method in terms of RMSE is preferable since it can handle missingness and outperforms the available data method in out-of-sample prediction. Furthermore, the conclusions of comparing the training data distribution, test data distribution, and GCRE-MNAR imputed values in case 1 remain consistent (as displayed in Figure 32 in the Appendix). The consistency of the results indicates that the imputation procedure employing the GCRE-MNAR approach provides reliable analyses.

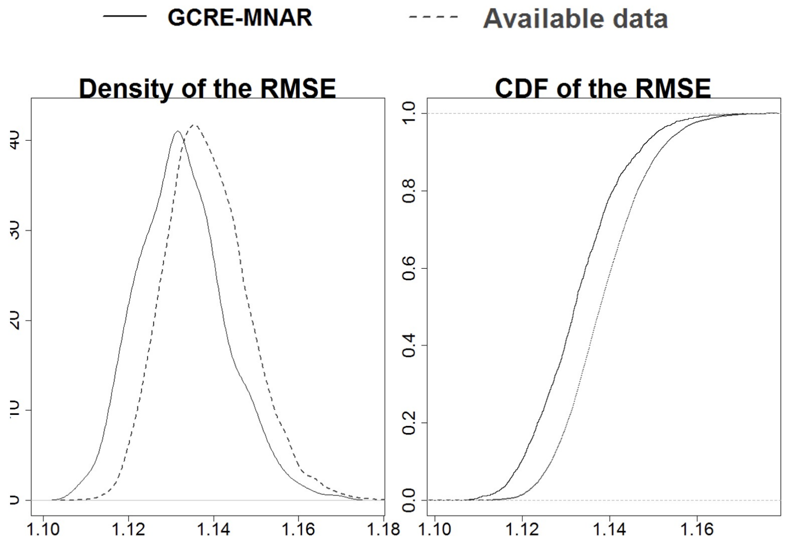


Figure 7.7.16: The RMSE density (on the left-hand side) and CDF (on the right-hand side) of the out-of-sample prediction in BIOSTAT-CHF training data in case 2, where the black solid curve represents the GCRE-MNAR method and the grey dashed curve represents the available data method. The RMSE of the GCRE-MNAR method is shifted toward lower RMSE values.

7.8 Discussion

In this chapter, we further generalise the GCRE-MAR method introduced in Chapter 6, which is a generalisation of the CRE introduced by Bhuyan (2019) that was discussed in Chapter 4. The aim is to allow for missing values in the analysis model response and predictors while assuming non-ignorable missingness in both. Additionally, the proposed GCRE-MNAR method estimates two covariance parameters to determine the possibility of the missingness in the response and the predictor being MNAR by using Correlated Random Effects, which is a statistical technique used to model the relationship between the missing model and the observed model. This information is useful for data analytics since the reason behind the cause of missing data cannot be determined beforehand and can help improve the accuracy of results. We evaluated the GCRE-MNAR method using simulated data with different combinations of missingness in the response and predictor and across different repeated measures.

We have assessed the accuracy of parameter estimation in our analysis of model parameters using simulated data, which is limited to a setting involving one incomplete predictor and two fully observed predictors. This has provided valuable insights into the performance of the GCRE-MNAR method compared to the available data method. The available data method is the default method for addressing linear mixed models in longitudinal data, which only uses the observed data. Our findings show that the GCRE-MNAR method has a smaller RMSE of the analysis model's parameters than the available data method, indicating that the GCRE-MNAR method is more effective in reducing the spread of estimation errors.

Moreover, the GCRE-MNAR has similar performance to the GCRE-MNAR with the fully observed predictor and the CRE methods except when there is a large proportion of missingness in the incomplete predictor (60%). In such cases, the GCRE-MNAR method tends to have larger RMSE values because the GCRE-MNAR can handle missingness in the model predictor. In contrast, the GCRE-MNAR with a fully observed predictor and CRE methods consider a fully observed predictor, and thus, the GCRE-MNAR introduces more error in the method.

The applied methods to the simulated data produced unbiased estimates of the analysis model parameters. However, the available data underestimates the analysis model intercept β_0 and overestimates β_3 as the number of repeated measures increases. It also has a large RB uncertainty when there is 60% missingness in the incomplete predictor. The proposed GCRE-MNAR method and its special cases (the GCRE-MNAR with a fully observed predictor and the CRE methods) tend to produce biased σ_B^2. However, this bias is reduced with the increase in repeated measures, indicating that the GCRE-MNAR struggles with low sample size and requires a larger sample size to produce accurate estimates. The unbiased estimation revealed by the relative bias for the GCRE-MNAR method underscores its ability to provide estimates that are centred around the actual values.

The results indicate that the GCRE-MNAR method consistently performs better than the available data method in predicting out-of-sample accuracy. It suggests that the GCRE-MNAR method could potentially enhance the reliability of conclusions drawn from medical longitudinal studies and handle missingness in the analysis model response and incomplete predictor. Providing more accurate forecasts of results outside the training data and imputing missing values in the analysis model response and incomplete predictor could be particularly helpful in medical research, where decisions often affect patient care and treatments. The ability to make reliable predictions is essential in such research.

The GCRE-MNAR method is a reliable and precise method for estimating missingness response, incomplete predictor, and missingness incomplete predictor model parameters with larger sample sizes. As the number of repeated measures increases, the precision and reliability of the estimates improve, as indicated by low RB, low RMSE values and high CR values. Furthermore, the GCRE-MNAR method can produce reliable estimates of the incomplete predictor missingness model parameters' data-generating parameters when there is a large proportion of missingness in the incomplete predictor (40% and 60%). This may be because a larger proportion of missingness provides the method with more information about the missingness pattern to estimate the parameter values of the missingness incomplete predictor model.

The GCRE-MNAR method was applied to the BIOSTAT-CHF dataset, and it was found that the estimation of analysis model parameters differed between the GCRE-MNAR method and the available data method. Based on the covariance parameter, missingness in the response is less likely to be MNAR, whereas missingness in the incomplete e.GFR is more likely to be MNAR. The GCRE-MNAR imputed values of the response and incomplete e.GFR predictors were found to be similar to the observed ones.

Two scenarios were applied to split the dataset to evaluate the performance of the available data method and the proposed GCRE-MNAR method in the BIOSTAT-CHF dataset. In the first scenario, the methods were trained using the dataset with missing values in the response and predictor, and these methods were then tested on a fully observed test dataset. In the second scenario, the test dataset contained missing values in the eGFR. The out-of-sample performance was different in both scenarios. In the first scenario, the available data method outperformed the GCRE-MNAR method, but in the second case, the GCRE-MNAR outperformed the available data method. This implies that the GCRE-MNAR method offers advantages in handling missing data by predicting unseen data and providing imputed values for the response and predictor. Additionally, in both scenarios, the covariance parameter values concluded that the missingness in the response is MAR and eGFR is MNAR.

Regarding imputation performance, it’s worth noting that the behaviour of response and predictor variables differs between training and test datasets. Specifically, the training set is more similar to the imputed predictor data than the imputed response. This disparity may be linked to the underlying cause of the missing data. Furthermore, the GCRE-MNAR has the added advantage of providing us with the missingness mechanism of the incomplete predictor along with the response variable.

Overall, the GCRE-MNAR method is a promising approach for handling missing data in statistical models, particularly in cases where the missingness is non-ignorable. The method can provide valuable insights into the probability of missingness being MNAR and can improve the accuracy of the result. Since the assumption of MNAR relies on unobserved data, it is essential to apply sensitivity analysis to assess the robustness of the results and conclusions to assumptions made about the missing data mechanism. Analysts can use it to address potential bias, discover the uncertainty associated with their conclusions, and evaluate the impact of assumptions about the missing data process on their findings. This helps improve the clarity of the research, which is especially when dealing with human lives. Additionally, the GCRE-MNAR method could improve various fields beyond medical research. For example, longitudinal studies in education research and social science surveys, where missing data is a common challenge. In Chapter 8, we will address how to perform sensitivity analysis in this context.

Chapter 8

Sensitivity Analyses for the Proposed Methods

8.1 Introduction

In real-world data, it is impossible to determine the real model and missingness mechanism. Therefore, it is important to assess how the proposed methods perform when their assumptions do not hold. We can evaluate whether the model fits the observed data, but we cannot assess how well it fits the unobserved data based on the observed data alone. Conducting sensitivity analysis is crucial in this regard. Staudt et al. (2022) recommended using sensitivity analysis in randomized controlled trials to evaluate the effects of various missing data assumptions on study results. The reliability of the conclusions depends on the consistency of conclusions across different models (Mason, 2010).

For valid inference, appropriate distributional assumptions and a model for the missing data mechanism are required. Since these assumptions impact the results, conducting sensitivity analysis is essential (Ibrahim and Molenberghs, 2009) to demonstrate how assumptions that differ from those in the primary analysis influence the results (Morris et al., 2014). Changes in the model's distributions and the missing indicator model ideally should minimally affect the response model regression coefficients estimate. This is because there are various ways to model these distributions, and each model has a distinct concept. The missingness model investigates the relationship between missing data and other variables. The partially observed variable distributions help to handle missing data by imputing it. The response model is used to analyse the relationship between variables of interest and predicting outcomes. The possible models can typically be determined by additional information about the missing data or through expert elicitation (Stubbendick and Ibrahim, 2003). It is not possible to know the parametric forms of the partially observed variable model and the assumed missing data mechanism using the data at hand (Ibrahim and Molenberghs, 2009). Due to various options for the distribution assumptions and the missing data mechanism, it requires evaluating different models (Ibrahim and Molenberghs, 2009).

Extensive literature suggests various approaches to reviewing the models. For example, Birmingham et al. (2003) discussed estimating the parameters of interest across a range of plausible sensitivity parameter values, where the relationship between the response and the response missingness process is called the sensitivity parameter (Minini and Chavance, 2004). According to Ibrahim et al. (2005), there are two ways to approach the missing data selection model. The first is to let the data decide the selection model empirically by starting with the main effects and progressively adding terms while assessing each model's fit using the likelihood ratio or AIC. However, data from alternative nonignorable models does not provide much information. On the other hand, sensitivity analyses may be more appropriate, particularly in cases where the data is unable to distinguish between alternative nonignorable missing-data methods. According to Molenberghs and Kenward (2007), one strategy is to fit several plausible MNAR models or perform a primary analysis with different adjustments, such as the inclusion of predictors.

Specifying the selection model predictors is not straightforward. Du et al. (2022) studied the consequences of including too few or too many predictors in the missingness model misspecification. It is recommended to include all variables in the response model into the missingness model as it can be challenging to choose which factors to consider (Du et al., 2022), and to avoid making the missingness too complex to prevent the model from being non-identifiable (Ibrahim et al., 2005). Bhuyan (2019) and Lin et al. (2010) evaluated the sensitivity of the CRE approach using a misspecified missingness response model that assumes the association between the missingness response process and the response variable is specified through a fixed effect. The results showed that the response model estimators were biased, which is not surprising.

This chapter discusses various sensitivities of the proposed Two-Step, GCRE-MAR, and GCRE-MNAR approaches to investigate the reliability of conclusions. To provide practical recommendations, we will conduct two types of sensitivity analysis to obtain a complete picture of the robustness of the inference. Firstly, in Section 8.2, we will test the sensitivity of the proposed approaches to different types of misspecified missing data mechanisms (e.g. MAR and MNAR). Secondly, in Section 8.3, we will examine each proposed approach using misspecified missing data process models (i.e. missingness model structure that is misspecified). Finally, Section 8.4 discusses the main results.

8.2 Missing mechanism Sensitivity Analysis

We will create scenarios to test different sensitivities of the proposed approaches. These scenarios will have missingness mechanisms that depart from the approach's assumptions for the response and predictor. We will use the simulation set-up described in Section 3.4. We will generate the missing MNAR values using each method's process structure. The MNAR values for the incomplete predictor are described in Section 7.6, and the MNAR values for the response are described in Section 6.6. To generate MAR values, we will employ the `deleteMARcensoring()` function in the `missMethod` package (Rockel, 2020) in `R` by assuming that the missingness is related to the values of the fully observed continuous predictor $X_{i2}(t)$. Specifically, the values of the model response or incomplete predictor will be missing whenever the corresponding $X_{i2}(t)$ value is within the p^{th} quantile, where p is the proportion of missingness. Furthermore, test data will be generated using fully observed data to evaluate out-of-sample prediction performance. To ensure a comprehensive analysis, we will generate 100 datasets with different numbers of repeated measures and the proportion of missingness, as carried out in previous analyses (Chapters 5-7). The methods' performance will be compared using Root Mean Square Error (RMSE), Relative Bias (RB), and Coverage Rate, as explained in Section 2.7.

8.2.1 Two-Step and GCRE-MAR Methods

The Two-Step and GCRE-MAR methods deal with two types of missing data: MNAR in the model response and MAR in the incomplete predictor. We will apply these methods to simulated data that includes MNAR in both the model response and incomplete predictor $(MNAR\_Y \ \& \ MNAR\_X)$, which misspecified the missingness mechanism for the model's incomplete predictor. To misspecify the missingness mechanism for the model response, we will generate data that have MAR in both the model response and the incomplete predictor $(MAR\_Y \ \& \ MAR\_X)$. Additionally, to test the misspecification of the missingness mechanism for the model response and incomplete predictor, we will simulate data with MAR missingness in the model response and MNAR missingness in the model incomplete predictor $(MAR\_Y \ \& \ MNAR\_X)$. We will compare the misspecification results with baseline methods, which are the full and available data methods. The plots representing the following findings are in Section E in the Appendix.

Misspecified the Missingness Mechanism for the Model's Incomplete Predictor by Assuming *MNAR_Y* & *MNAR_X*

The misspecification of the missingness mechanism in the model's incomplete predictor yields similar results between the Two-Step and the GCRE-MAR methods, with a few exceptions that we will mention throughout. The RMSE between the analysis model's estimated parameter values and the data generating values is large overall, with an even larger RMSE when using the available data with 60% missingness in the model's incomplete predictor. Moreover, the Two-Step and GCRE-MAR methods resulted in lower RMSE for σ_B^2 compared to the available data. However, the Two-Step method has a larger RMSE and uncertainty than the available data method when there is 60% missingness in the model's incomplete predictor for the random intercept variance σ_B^2.

The parameter estimates are unbiased overall, except for the model's intercept β_0 with 60% missingness in the incomplete predictor using the Two-Step method. The available data method resulted in a larger RB uncertainty of β_0 and a more biased estimate with a larger proportion of missingness in the response variable. The random intercept variance σ_B^2 becomes unbiased as repeated measures increase. The Two-Step method showed that the variance components are biased with 60% missing values in the model's incomplete predictor.

In this case, the out-of-sample prediction performs well in predicting new data, with larger uncertainty when there is 60% missing data in the response variable. Additionally, the CR is mostly reasonable, except for the poor CR of σ_B^2 when using the GCRE-MAR method with a small number of repeated measures and as the number of repeated measures increases, the Two-Step method with 60% missing values in the model's incomplete predictor.

Misspecified the Missingness Mechanism for the Model's Response Variable by Assuming *MAR_Y & MAR_X*

In cases involving both *MAR_Y & MAR_X*, misspecifying the missingness in the model's response yields similar results between the Two-Step and the GCRE-MAR methods. The overall RMSE is high, especially with a higher proportion of missingness in the model's response and 60% missingness in the incomplete predictor separately for β_0 and β_2. The available data resulted in high RMSE with 60% missingness in the incomplete predictor. However, when the proportion of missingness in the model's response and the incomplete predictor is low (20%), there are low RMSE and negligible differences between the RMSE values of the applied methods.

The parameter estimates are mostly unbiased, with the larger RB uncertainty for β_0 and β_2, with 60% missingness in the model's response. As the number of repeated measures increases, the variance components become unbiased. On the other hand, the Two-Step method resulted in a biased estimate of σ_A^2, with a high proportion (40% and 60%) of missingness in the model response and 60% missingness incomplete predictor. However, it becomes more biased as the repeated measures decrease. The between-individual variance σ_B^2 estimates are mostly biased.

The out-of-sample prediction performs well with unseen data but has larger uncertainty when there is a high proportion of missing data, 60%, in the response variable. The available data performs poorly, particularly when 60% of missing data is in the incomplete predictor. The CR for σ_B^2 was poor but reasonable for other variables.

Misspecified the Missingness Mechanism for the Model's Response and Incomplete Predictor by Assuming *MAR_Y & MNAR_X*

The results of misspecification of the missingness mechanisms of the model's response and incomplete predictor for the Two-Step and GCRE-MAR methods that deal with *MAR_Y & MNAR_X* have similar outcomes with a few exceptions, which will be mentioned. The RMSE resulted in high values overall and even higher when there is a high proportion of missingness (60%) in the response or incomplete predictor variables separately. The RMSE values of β_1, β_3 and σ_B^2 increase using the Two-Step method with 60% missingness in the incomplete predictor. When there is a low proportion of missingness in the model response and incomplete predictor (20%), the RMSE values are similar to the best-case scenario (the full data).

The proposed methods resulted in unbiased estimates overall with larger uncertainty when there are 60% missing values in the model response for β_0 and β_2. The Two-Step method produced biased β_0 and a larger uncertainty of β_1 with 60% missing values in the incomplete predictor. The variance components become unbiased as the number of repeated measures increases. However, the Two-Step method resulted in biased σ_B^2 with 60% missingness in the incomplete predictor and 40% and 60% missingness in the response variable.

The proposed method performs well in predicting unseen data but with larger uncertainty when the proportion of missingness in the response variable is high. The CR of σ_B^2 is poor and reasonable for other parameters.

8.2.2 GCRE-MNAR Method

The GCRE-MNAR method deals with non-ignorable missing (MNAR) data in the model response and incomplete predictors. We will apply this method to different simulated data: MAR in the model's incomplete predictor and MNAR in the model's response $(MNAR\_Y \,\&\, MAR\_X)$, which misspecifies the missingness assumption for the model's incomplete predictor. To misspecify the missingness assumption for the model's response, we will generate data with MAR missingness in the model response and MNAR missingness in the model incomplete predictor $(MAR\_Y \,\&\, MNAR\_X)$. Furthermore, we will simulate data with MAR missingness in both the model response and incomplete predictor (MAR_Y & MAR_X) to misspecify the missingness assumption of the model response and incomplete predictor. We will compare the proposed method's misspecification results with those of baseline methods, i.e. the full and available data methods. Section E in the Appendix contains plots that illustrate the following results.

Misspecified the Missingness Mechanism for the Model's Incomplete Predictor by Assuming *MNAR_Y & MAR_X*

When using the GCRE-MNAR method with a misspecified missingness mechanism for the incomplete predictor with $MNAR\_Y$ & MAR_X, it's not surprising that the RMSE values for the parameter estimates are high. They are even higher with 60% missing values in the incomplete predictor using the available data method. The estimates are unbiased overall, but there is a larger uncertainty of β_3 when there is a larger proportion of missingness in the response variable (40% and 60%). The available data produces biased estimates for β_0 and β_3. As the number of repeated measures increases, the between-individual variance, σ_B^2, becomes unbiased but with larger uncertainty. The out-of-sample prediction performance is not affected by the misspecification of $MNAR\_Y$ & MAR_X case. However, the available data has a lower density of out-of-sample RMSE with a longer tail towards larger values when there is a 60% missing value in the incomplete predictor. Regarding CR, most parameters have a reasonable rate, except for σ_B^2.

Misspecified the Missingness Mechanism for the Model's Response Variable by Assuming *MAR_Y & MNAR_X*

The GCRE-MNAR method produced large RMSE overall and even higher values with a large proportion of missingness (60%) in the response. However, with 20% missingness in the response and incomplete predictor, the RMSE values were similar to the full data method. Generally, the GCRE-MNAR method produced unbiased estimates with large uncertainty for β_0

and β_2 when 60% of the response values were missing. As the number of repeated measures increased, the variance parameters became unbiased. The out-of-sample prediction performed well under this misspecification assumption, with larger uncertainty when 60% of the response variable was missing. The CR for σ_B^2 was poor compared to other analysis model parameter estimates.

Misspecified the Missingness Mechanism for the Model's Response and Incomplete Predictor by Assuming *MAR_Y & MAR_X*

In the context of specifying the missingness mechanism for the model's response and incomplete predictor, the GCRE-MNAR method resulted in high overall RMSE, which increased even further with a large proportion of missingness, specifically 60% in the model response. The available data method produced higher RMSE than the GCRE-MNAR method with 60% missing values in the incomplete predictor. However, when there is 20% missingness in the response and incomplete predictor, the RMSE values are not large. In terms of RB, the analysis model parameter estimates are unbiased in most cases, except the variance parameters, which become unbiased as the number of repeated measures increases. However, σ_B^2 is still biased with a large number of repeated measures. Generally, out-of-sample prediction performs well with unseen data, with a higher level of uncertainty when there is 60% missingness in the response. Additionally, the available data method performs more poorly than the GCRE-MNAR method with 60% missing values in the incomplete predictor. The values of the CR are generally reasonable, except for σ_B^2.

8.3 Missingness Model Sensitivity Analysis

In this section, we will conduct additional simulations to evaluate the impact of applying a misspecified missingness process model for the proposed methods. As the proposed methods are generalisations of the existing CRE method, we will follow a similar misspecification model set-up proposed by Bhuyan (2019) and Lin et al. (2010) for Correlated Random Effects. The missingness model involves a fixed effect (η) that determines how missingness depends on the response. Consequently, the analysis and missingness process models are no longer correlated via a random effect but rather through the fixed effect η. We set the values of the missingness process model according to each model's simulation data for missingness and centred the incomplete variable (either the response or the incomplete predictor) in the missingness process model around its mean value (ζ) to keep the proportion of missingness fixed, which resulted in 35% missingness. We will present the misspecification model for each proposed method below, along with the results. We will generate 100 datasets for each scenario with different repeated measures using the same simulated data set-up described in Section 3.4.

8.3.1 Two-Step Method

As there is a missingness process model for the response in the Two-Step method, we will misspecify this and generate missingness in the response based on the following model:

$$U_i^*(t) = \theta_0 + \theta_1 X_{1i}(t) + \theta_2 X_{2i}(t) + \theta_3 X_{3i}(t) + \eta\,(Y_i(t) - \zeta) + \varepsilon_i(t) \quad (8.3.1)$$

where ζ is the mean value of the response variable, $\eta = 0.5$ and values of fixed effects coefficients are: $\boldsymbol{\theta} = \{-0.8, -0.4, 3, 4\}$.

8.3.2 GCRE-MAR Method

As the GCRE-MAR method involves a missingness process model for the response, we will generate missingness in the response based on the following misspecified model:

$$U_i^*(t) = \theta_0 + \theta_1 X_{2i}(t) + \theta_2 X_{3i}(t) + \eta\,(Y_i(t) - \zeta) + \varepsilon_i(t) \quad (8.3.2)$$

where ζ is the mean value of the response variable, $\eta = 0.5$ and values of fixed effects coefficients are: $\boldsymbol{\theta} = \{-0.6, 0.7, 4\}$.

Figure 8.3.1 displays the relative bias of the parameter estimates in the analysis model using the GCRE-MAR method as a representative example. The RB using the Two-Step method, RMSE, out-of-sample prediction, and coverage rate plots are in Section E of the Appendix.

The Two-Step, GCRE-MAR and available data methods have large RMSE overall and biased parameter estimates for β_0 and β_3 with a low coverage rate, which decreases as the repeated measures increase. However, as the repeated measures increase, the estimates of the between-individual variance parameter become more unbiased. In terms of out-of-sample prediction, the Two-Step method and the GCRE-MAR method perform similarly, showing low overall densities with noticeable tails extending towards larger values. Although the estimates of β_0 and β_3 appear to exhibit some bias, the magnitude of the bias is small, resulting in a lower density of out-of-sample prediction performance due to using a misspecified response missingness model to test their sensitivity.

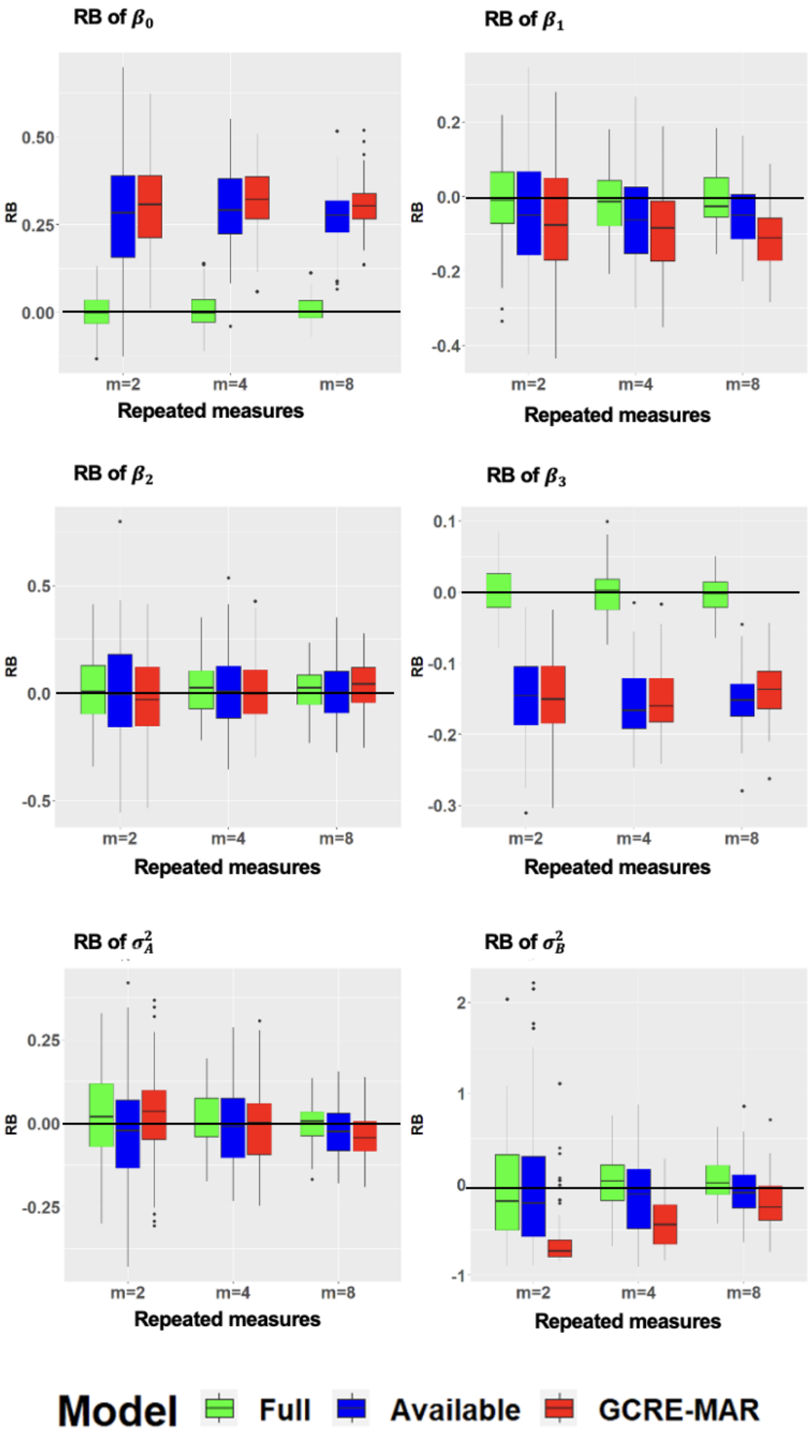


Figure 8.3.1: Boxplots illustrate the RB of parameter estimates in an analysis model when using a misspecified missingness response model in the GCRE-MAR method. The y-axis displays the RB values, while the x-axis represents the number of repeated measures. Each boxplot corresponds to a specific method used. It is evident that the estimates of β_0 and β_3 are biased using the GCRE-MAR and available data methods.

8.3.3 GCRE-MNAR Method

The GCRE-MNAR method includes separate models for response and incomplete predictor missingness processes, which can be misspecified individually and together.

Misspecified Response Missingness Process Model

The response's missing values are generated based on the misspecified missingness process model as follows:

$$U_i^*(t) = \theta_0 + \theta_1 X_{2i}(t) + \theta_2 X_{3i}(t) + \eta\,(Y_i(t) - \zeta) + \varepsilon_i(t) \qquad (8.3.3)$$

where ζ is the mean value of the response variable, $\eta = 0.5$ and values of fixed effects coefficients are: $\boldsymbol{\theta} = \{-0.6, 0.7, 4\}$. The missingness model for the incomplete predictor is specified correctly, as described in Section 7.6.

Misspecified Incomplete Predictor Missingness Process Model

The incomplete predictor's missing values are generated based on the misspecified missingness process model as follows:

$$R_{i1}^*(t) = \psi_0 + \psi_1 X_{2i}(t) + \psi_2 X_{3i} + \eta\,(X_{i1}(t) - \zeta_x) + \xi_i(t) \qquad (8.3.4)$$

where $\eta = 2.5$, ζ_x is the mean value of the incomplete predictor $(X_{i1}(t))$ and values of fixed effects coefficients are: $\boldsymbol{\psi} = \{-6, 8, 5\}$. The missingness model for the response is specified correctly, as described in Section 7.6.

Misspecified Response Missingness Process Model and Incomplete Predictor Missingness Process Model

Finally, we will test the proposed GCRE-MNAR method when both the missingness process model for the response and incomplete predictor are misspecified using Equation 8.3.3 and 8.3.4.

The results of misspecification in either the missingness response model or both the missingness response model and the incomplete predictor missingness model, are similar to the case when the missingness response model in the Two-Step and the GCRE-MAR method was misspecified. Except when the incomplete predictor missingness process model is misspecified. In this case, the analysis model's parameter estimates for β_0 and β_3 (as shown in Figure 8.3.2) are not biased, and the out-of-sample performance is about as good as the full data. Moreover, the coverage rate of the random intercept variance σ_B^2 is low when there is a low number of repeated measures and become unbiased as the number of repeated measures increases. The remaining plots for each case can be found in Section E in the Appendix.

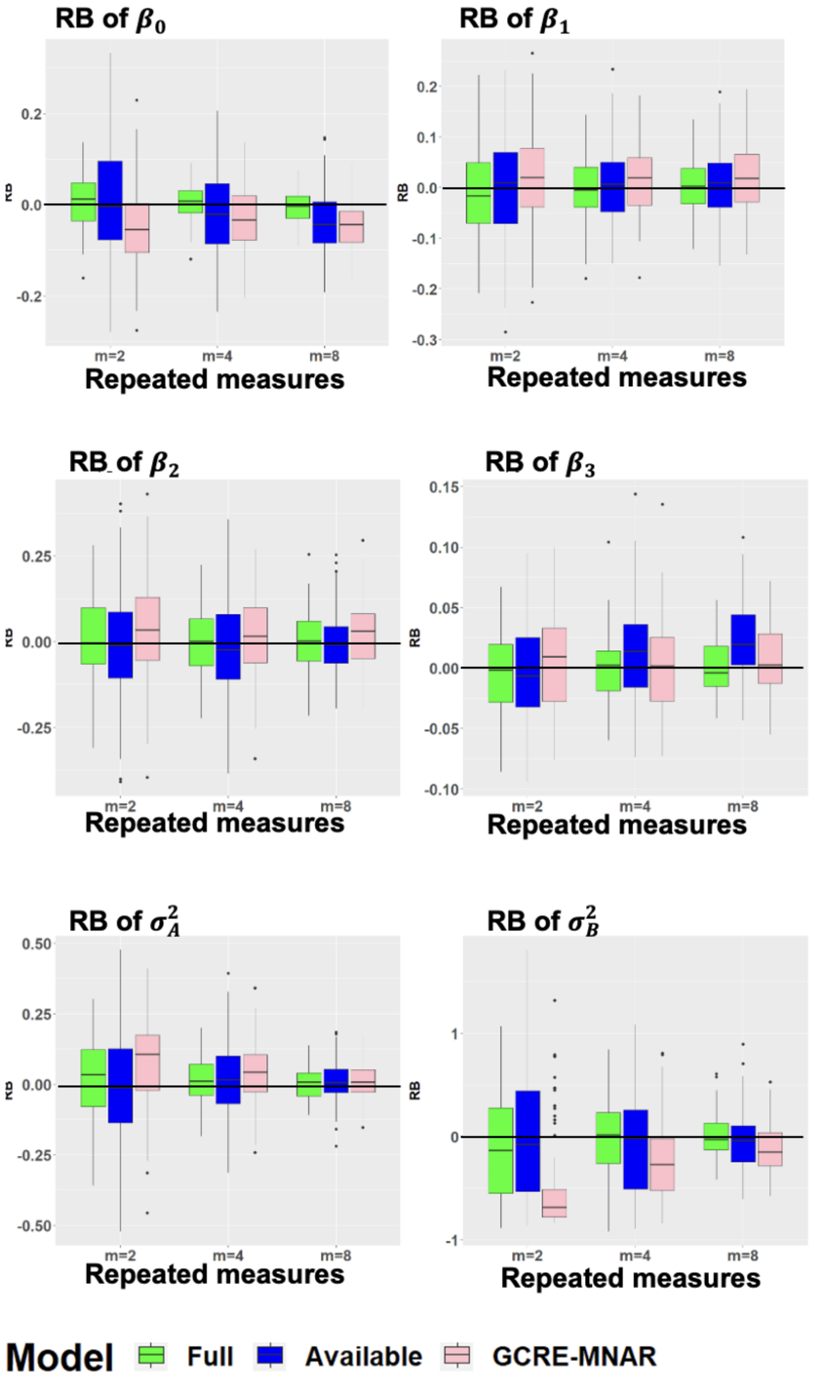


Figure 8.3.2: Boxplots illustrate the RB of parameter estimates in an analysis model when using a misspecified missingness incomplete predictor model in the GCRE-MNAR method. The y-axis displays the RB values, while the x-axis represents the number of repeated measures. Each boxplot corresponds to a specific method used. It is evident that the variance parameters estimators are biased as the repeated measures decrease using the GCRE-MNAR method.

8.4 Discussion

In this chapter, we examined the proposed methods for misspecified missing data mechanisms and missingness process models. Since these assumptions of the proposed models cannot be verified in practice. The aim of this exploration is to determine whether using these methods under different assumptions of missing data will affect the analysis model parameter estimates and consequently change the conclusion of the results.

When the missing mechanism is misspecified for each proposed method, it results in large RMSE values and uncertainty for out-of-sample prediction performance when there is a high proportion of missingness (60%) in the response variable and with a small number of repeated measures. This can lead to a risk of overfitting the outcome and an inability to generalise the results to unseen data. The available data mostly performs poorly in terms of out-of-sample prediction when there is a high proportion of missingness in the incomplete predictor (60%). If the missing mechanism of the missingness in the response variable is misspecified, either alone or with the incomplete predictor missingness, it results in a large RMSE when there is a high proportion of missingness (60%) in the response. This indicates that the model's estimates deviate considerably from the data-generating parameter values.

On the other hand, when there is a low proportion of missingness (20%) in the response and incomplete predictor, the RMSE values are similar to the full data, which is the best-case scenario. This indicates that the proposed methods under the misspecified missingness assumption do not perform well with a high proportion of missingness and do not struggle with a low proportion of missingness. The model intercept parameter is mostly affected by misspecified missingness assumption, resulting in biased estimates, especially with a high proportion of missingness. The variance parameter estimates become unbiased as the number of repeated measures increases, which is consistent with the findings of the proposed methods with correct specification of missingness assumptions. However, the coverage rate of the between-individual variance is consistently low, indicating that the proposed methods are not capturing this data-generating parameter value as frequently as they should.

The misspecified missingness process model affects some parameter estimates, resulting in biased estimates. The number of repeated measures affects the between-individual variance estimates, where performance increases as the number of repeated measures increases. This result is consistent with the results of the proposed methods under the correct specification of the missingness process model. The out-of-sample prediction with the misspecified missingness process model results in larger uncertainty toward larger values. However, this is not the case when the incomplete predictor missingness model is misspecified using the GCRE-MNAR method. In this scenario, the model misspecification does not affect the parameter estimates and the out-of-sample prediction.

The use of misspecified missingness assumptions and missingness process models did not considerably impact the overall conclusions. The impact decreases as the number of repeated measures increases, indicating that misspecification has less impact on the results with a larger sample size. However, the proposed methods struggle to capture the actual between-individual variance, which may affect the robustness of the conclusions given that variability among individuals is an essential part of longitudinal data analysis.

It's important to note that the out-of-sample performance tends to be skewed towards larger values, especially with a higher proportion of missingness in the response variable and a low number of repeated measures. This skewness should be considered as it might have an impact on how reliable and effective the model is seen to be in real-world situations.

Chapter 9

Conclusions and Extensions

9.1 Summary

Longitudinal data is a type of study in which data is collected repeatedly over time for each individual. It is used to study long-term effects and has applications in various fields of study. Mixed models are commonly used to account for individual variability and produce more precise estimates. Longitudinal studies require an extended recruitment period and collection of repeated measures, which makes them subject to missing data, potentially caused by unknown reasons.

This thesis begins with a general introduction to the research interest, provides an overview of the statistical literature used to analyse longitudinal data, missing data, statistical techniques used in the thesis and explains the data used in the analysis. It then compares two approaches in the statistical field for analysing longitudinal data: the Linear Mixed Effects model from the Frequentist approach and the Bayesian Hierarchical model from the Bayesian approach. This comparison is to understand how analogous approaches perform when applied to simulated and real-world data used in this thesis.

Incomplete data is common for various reasons in longitudinal studies. A recent method has been proposed by Bhuyan (2019), called the Correlated Random Effects method, which assumes nonignorable missingness in the response and fully observed predictor variables. We introduced a solution to overcome non-convergence in the CRE method and then proposed three methods to accommodate missingness in the explanatory variables and model response (where the current CRE method can only handle missingness in the model response). The three methods are the Two-Step method, the GCRE-MAR method, and the GCRE-MNAR method.

The proposed methods consider different combinations of missingness mechanisms in both the model predictor and the response. These methods aim to impute missing data and estimate parameters of interest in cases where the response has non-ignorable missingness and predictors have ignorable or non-ignorable missingness, by adapting the Correlated Random Effects method. This is achieved using a Bayesian estimation procedure to simultaneously estimate the analysis model parameters and the missingness models, providing a useful alternative to solving the intractable log-likelihood function using approximation methods.

The proposed methods were evaluated using simulated data with different factors: the number of repeated measures and the proportion of missingness. Additionally, the proposed methods were compared with baseline methods; the full data (containing no missing values), the available data (missing values remain in the dataset and no imputation is applied), and the CRE method. The proposed methods were also applied to real-world data to create predictive models for heart failure patients (BIOSTAT-CHF dataset). As the reason behind the missing values in real-world data is unknown, sensitivity analysis was applied to misspecify the missingness mechanism and missingness model for the proposed methods to examine how this might affect the overall performance of each method.

9.2 The Strengths and Limitations

Each subsection will briefly explain the different methods used in this thesis, along with the overall results and the main strengths and limitations.

9.2.1 Frequentist and Bayesian Comparison

In Chapter 4, we compared the Frequentist approach using Linear Mixed Effects modelling and the Bayesian approach using Hierarchical Bayesian modelling. We used simulated and real-world data, as described in Chapter 3. This comparison study found that the point estimate obtained from the Frequentist approach and the average of the posterior distribution from the Bayesian approach are similar and close to the data-generating parameters. This similarity is likely due to the use of a non-informative prior, which has less influence on the posterior distribution. Thus, the parameter estimates will depend more on the likelihood, which is derived from the observed data. This is similar to Frequentist inference, which also focuses on the likelihood. In addition, the out-of-sample prediction was similar, but the Bayesian approach yielded a smaller average RMSE value when applied to the BIOSTAT-CHF dataset, although the difference was not significant.

The Frequentist approach provides a confidence interval and a point estimate. The confidence interval is challenging to interpret in terms of probabilistic statements about the model parameters. On the other hand, the Bayesian approach offers a point estimate and a probability distribution, providing insight into the chances of a particular value for a parameter. This makes it easier for decision-makers to interpret the uncertainty of parameter estimates. Furthermore, Bayesian methods allow the incorporation of prior knowledge, which can be valuable given the wide range of available information today. However, Bayesian analysis can be computationally intensive for large datasets or complex models.

Using 100 repetitions in the simulated data balances the accuracy and the computational feasibility. However, achieving a lower MCSE (e.g. less than 0.001) will improve the precision of the estimates, which would require a substantially larger number of repetitions (Morris et al., 2019).

9.2.2 CRE Method

The CRE method (Bhuyan, 2019), proposed in the literature, addresses non-ignorable missingness in the response variable for longitudinal data models. It has shown good performance using Legendre polynomials (LP) for semi-parametric modelling of time-varying variables. However, when using a Linear Mixed Effects model, which is an appropriate structure to model the data used in this thesis, the covariance matrix parameters did not converge even with a large number of iterations. A weakly informative prior was introduced in Chapter 4, Section 4.4.5, to address this issue, which resolved this problem.

The advantages of the CRE method include using Bayesian inference, which avoids the computational challenge of intractable numerical integration in the log-likelihood function. Additionally, it incorporates Correlated Random Effects between the response model and the missingness response model, allowing for the estimation of a covariance parameter that can indicate the response's missingness mechanism. As the value of the parameter increases, the response's missingness is more likely to be MNAR, which is a useful technique when applied to real data, as the missingness mechanism is often unknown.

However, a drawback of the CRE method is that it assumes the model predictors are fully observed, whereas, in longitudinal studies, often the predictors also have missingness.

9.2.3 Two-Step Method

The Two-Step method introduced in Chapter 5 is the first proposed method in this thesis for dealing with missing data in both the model predictors and the model response. It is designed to handle situations where the missingness in the model response is non-ignorable, while the missingness in the incomplete predictors is ignorable. The Two-Step method involves two steps. First, it imputes the missing values for the incomplete predictors using the MICE algorithm, which assumes that MAR is the missing mechanism for the incomplete predictors. Then, the CRE method is applied to estimate the analysis model parameters, considering that the missing mechanism for the response is MNAR.

The Two-Step method was evaluated using simulated data and compared to how the CRE method performed. The Two-Step method considers missingness in the model's predictors, which is an advantage over the CRE method. However, the CRE method gains an unfair advantage in this comparison because it does not handle missingness in the predictors. Therefore, the CRE method was run with no missing data in the predictors. This advantages the CRE method, because it has data that would not otherwise be available in practice. Although there is an apparent similarity in performance between the Two-Step and CRE methods, the Two-Step method is additionally handling missingness in the predictors without a loss of performance in the majority of scenarios.

The RMSE of the parameter estimates indicated that the Two-Step method resulted in larger RMSE values than the CRE method when there were 60% missing values in the incomplete predictor. This finding makes sense because the CRE method does not have missing data in the model predictors. In other words, the Two-Step method produces results similar to the CRE method when the CRE method has additional data that is not available in practice, and the Two-Step method has up to about 60% missing data for the incomplete predictor variable. In contrast, the available data method's RMSE for parameter estimates is larger than the Two-Step and CRE methods.

Overall, the Two-Step method produced unbiased parameter estimates, except for the response model variance parameters, in the case of 60% of the missing values in the incomplete predictor. This proportion of missing data with multiple imputations can introduce additional variability, which may affect the estimation of variance parameters. However, this bias decreased as the repeated measures increased. Additionally, the Two-Step method's estimates for the response missingness process model's parameters become unbiased as the repeated measures increased. Misspecifying the missingness mechanism did not considerably affect the out-of-sample performance. However, it did impact the variance parameter estimates by resulting in biased estimates when there was a higher proportion of missing data in the incomplete predictor. This result is unsurprising, as it aligns with findings when the missingness mechanism is correctly specified.

The Two-Step method involves applying the CRE method to multiple imputed datasets, which makes it more computationally costly and time-consuming than the CRE method. The computational time and storage space required depends on the number of MICE-imputed datasets, the dataset sample size, and the availability of parallel computing. These factors may make the process prohibitive. On the other hand, the Two-Step method could benefit from adding auxiliary variables in the incomplete predictor imputation model during the MICE step. These variables could help predict the missingness in the incomplete predictor. This is because adding these variables will have a minimal effect on the analysis of the two-step method. However, these variables will affect the analysis of the GCRE-MAR and GCRE-MNAR methods due to joint modelling.

The Two-Step method resulted in biased parameter estimates and low out-of-sample performance when the missingness response model was misspecified. Additionally, the Two-Step method considers specific missingness mechanisms. The model's response is considered to be MNAR, and the incomplete predictors are considered to be MAR, which may not accurately reflect real-world data in all cases.

9.2.4 GCRE-MAR Method

The GCRE-MAR is a generalisation of the CRE method, proposed in Chapter 6 to address missing data in both the model's response and predictors and overcome the Two-Step method's computational burden. This method takes into account non-ignorable missingness in the model's response and ignorable missingness in the model's predictors. The proposed GCRE-MAR method can impute missing data and estimate the analysis model parameters simultaneously. Compared to the CRE method, the GCRE-MAR method incorporates an incomplete predictor model using Gibbs sampling as an additional step.

The analysis of the GCRE-MAR method using simulated data showed that, in general, the GCRE-MAR applied to fully observed predictors performs similarly to the CRE method, both of which did not consider missingness in the incomplete predictors. This suggests that the CRE method is a special case of the GCRE-MAR method when the analysis model predictors have no missing data. The GCRE-MAR method outperforms the available data method for the response model's parameter estimates and out-of-sample performance. This is because the proposed method can impute missing data that is similar to the actual values that were generated, which helps prevent the loss of valuable information that occurs with the available data method.

For the response model variance parameters, the GCRE-MAR method outperforms the Two-Step method when there was 60% missingness in the model's incomplete predictor. The GCRE-MAR method can produce unbiased parameter estimates for the response model variance parameters when there is 60% missing data in the model's incomplete predictor. The GCRE-MAR underestimates the between-individual variance; however, as the repeated measures increase, it results in an unbiased estimate. Additionally, the GCRE-MAR produced unbiased estimates of the response missingness process model as repeated measures increased. This suggests that the proposed method is able to capture the actual parameter values as the sample size increases (more repeated measures).

The missing data mechanism misspecification did not greatly affect the out-of-sample performance of the proposed method. However, it produced biased variance parameter estimates with fewer number of repeated measures, similar to the results obtained with the correctly specified missing data mechanism. Generally, the GCRE-MAR method performs well when the sample size is larger, especially with larger numbers of repeated measures. It accounts for specific missingness mechanisms in the model's response and incomplete predictors, where the model's response is MNAR and the incomplete predictors are MAR. This may not accurately reflect real data in all cases. The GCRE-MAR method produced biased parameter estimates and showed poor performance in out-of-sample prediction when the missingness response model was misspecified. This suggests that the proposed method is sensitive to the choice of missingness response model structure.

9.2.5 GCRE-MNAR Method

The GCRE-MNAR method introduced in Chapter 7 is an extension of the CRE and the GCRE-MAR methods. It is designed to handle missing data in both the model response and predictors. This method considers non-ignorable missingness in both the model response and incomplete predictor variables. It uses the Gibbs sampler and Correlated Random Effects to model the relationship between the incomplete variables (the model response and incomplete predictors) and their corresponding missingness process model.

The GCRE-MNAR method was evaluated using simulated data. The results showed that the performance of the response model's parameter estimates for the CRE method, the GCRE-MNAR method, and the GCRE-MNAR method applied to fully observed predictors are similar, except when there were 60% missing values in the incomplete predictor. In that case, the GCRE-MNAR method tended to have a larger RMSE, although it was still smaller than the available data method. This comparison suggests that the GCRE-MNAR performs as well as the CRE method. However, this comparison is unfair because the CRE method cannot handle missing data in the predictors. Therefore, the CRE method was run with no missing data in the predictors, which allows it to access the data that would otherwise be missing. On the other hand, the GCRE-MNAR method has the advantage of handling missing data in both the model response and incomplete predictors.

Comparing the proposed method applied to fully observed predictors with the CRE method indicates that the GCRE-MNAR method is a generalisation of the CRE method when there is no missing data in the model predictors. Furthermore, the proposed GCRE-MNAR method performs as well as the CRE method when the CRE method has extra, unavailable data in practice, while the GCRE-MNAR method can handle up to about 60% missingness for the explanatory variables. Generally, the GCRE-MNAR method produced unbiased model parameter estimates and outperformed the available data method, especially for the response model's intercept and the time-invariant predictor parameters estimate.

However, the GCRE-MNAR method underestimated the between-individual variance with a small number of repeated measures. As the number of repeated measures increased, the GCRE-MNAR method produced unbiased parameter estimates for the response missingness process model, the model for the incomplete predictor, and the incomplete predictor missingness process model. This indicates that the proposed method might require more data points (more repeated measures) to accurately capture the data-generating parameter values of the between-individual variance and the missingness process models. Except for the covariance matrix parameters for the incomplete predictor model, the GCRE-MNAR method produced biased estimates with a low (20%) proportion of missingness in the incomplete predictor. This indicates that the method is able to capture the covariance matrix parameters for the incomplete predictor model when there is a substantial proportion of missingness in the incomplete predictor variable. Its out-of-sample performance outperformed the available data method.

The misspecified missingness mechanism affects the parameter estimates of the between-individual variance, which becomes unbiased as the number of repeated measures increases. This is unsurprising, as it aligns with the GCRE-MNAR method's results with a correctly specified missingness mechanism. On the other hand, the misspecification of the incomplete predictor missingness model did not affect the out-of-sample performance and the parameter estimates. This means the GCRE-MNAR method produces unbiased results and performs well in terms of out-of-sample prediction.

The distinctive feature of the GCRE-MNAR method over the GCRE-MAR method is the covariance parameter between the incomplete predictor model and the incomplete predictor missingness process model. The covariance parameter indicates the probability that the missingness in the incomplete predictor is MNAR. However, for a moderate to large proportion of missingness in the incomplete predictor, this parameter is estimated more accurately. The missingness mechanism is usually unknown. An advantage of using the GCRE-MNAR method is that it can be helpful for data analytics in estimating the model's parameters and determining the probability of the missingness mechanism in the model response and predictors when there is a considerable amount of missing data in the model's incomplete predictor.

9.3 BIOSTAT-CHF Dataset Results

The BIOSTAT-CHF dataset was used throughout the thesis as a real-world data application for the proposed methods. We first used it to compare the Frequentist and Bayesian approaches using only complete case data (patients with fully observed response and predictors variables). Next, we applied the CRE method from the literature after filtering the data, so that the missing values only appeared in the response variable (NT-proBNP) and with fully observed predictors. Then, for each proposed method (Two-Step, GCRE-MAR, and GCRE-MNAR methods), we carried out inference when there were missing values in the response variable and a predictor (eGFR) to understand how the proposed methods handle real-life scenarios.

The results from the different applications mentioned were consistent. We discovered a negative relationship between eGFR and log(NT-proBNP). This means that for every one unit increase in eGFR from its average value (centred eGFR), the log(NT-proBNP) will decrease by approximately 0.01 units on average while holding all other variables constant. The centred age shows a positive correlation with log(NT-proBNP). This indicates that for every one year increase in age from its average value, the log(NT-proBNP) will increase by approximately 0.01 units on average, holding all other variables constant.

Additionally, patients with Sinus heart rhythm have the lowest rate of log(NT-proBNP) compared to Atrial fibrillation and Pacemaker heart rhythm categories. In contrast, the Other category shows a log(NT-proBNP) rate similar to Sinus rhythm. Furthermore, patients' log(NT-proBNP) decreased by approximately one unit on average during their second visit, while all other variables were constant. The within-individual variance is greater than the between-individual variance, indicating that the differences in observations within each individual are greater than the differences between the individuals' means. Note that the majority (73%) of patients in this study (BIOSTAT-CHF) were male, which represents a limitation in the generalizability of the results. According to Timmis et al. (2022), males have a higher incidence and worse risk factors for cardiovascular disease compared with females.

The advantage of using the Correlated Random Effects employed by the proposed methods (Two-Step, GCRE-MAR, and GCRE-MNAR) is that it can estimate the possibility of missingness in the model response being MNAR. An approximate estimate of -0.2 suggests that the missingness in the NT-proBNP is less likely to be MNAR. Additionally, GCRE-MNAR provides an additional advantage by estimating the possibility of missingness in the incomplete predictor of the model as being MNAR, with an estimate of -0.8 indicating that missingness in the eGFR is more likely to be MNAR. The negative values suggest that patients with higher NT-proBNP and eGFR are more likely to have missing values.

9.4 Extensions

The proposed methods introduced in this thesis open up several opportunities for further exploration. Future studies could benefit from the incorporation of expert knowledge as a prior distribution in a Bayesian setting, which is an advantage of the Bayesian approach over the Frequentist approach. This technique is called prior elicitation. For further reading, see Martin et al. (2012).

One possible future direction could involve exploring nonlinear models, such as using the semi-parametric model used by Bhuyan (2019) for the CRE method or assimilating the Gaussian Process or other flexible functional forms. To reduce the complexity of the nonlinear longitudinal model with a large number of repeated measures, consider only one variable as nonlinear in the model. This simplifies the model by concentrating the nonlinearity on the most relevant variable while keeping the rest of the model linear.

The linear mixed effects model focuses on estimating the mean of the response variable. However, in medical studies, different patients may behave differently in the extreme values of the outcome distribution. Future studies could benefit from considering the Quantile Regression (QR), which provides a more comprehensive understanding of how the analysis model predictors could affect the analysis model response variable at different quantiles. To adapt the quantile regression in the proposed methods, the relationship between the QR check function and the Asymmetric Laplace Distribution can be reformulated into the standard likelihood framework (Yuan and Yin, 2010). By incorporating QR into the proposed methods, the missingness model can reflect the probability of observing the response at different quantiles. Additionally, the correlated random effects will differ across the different quantiles of the response.

A possible future direction could involve investigating a binary or count response variable using Generalised Linear Mixed Models (GLMMs). This approach may introduce complexity due to the need for link functions and incorporating latent variables with binary responses utilising the probit model, which could increase the computational burden. Estimating the variance of the random effect is a challenge of the probit mixed effects regression model for longitudinal binary response data (Wu et al., 2018). Additionally, overdispersion is a common challenge when working with longitudinal count data using the Poisson model (Rizzato et al., 2016). Additionally, future work could explore and incorporate interaction terms in the analysis and missingness models.

Another interesting area for further investigation is the real-world application of these methods, which involves scaling up the proposed model for the BIOSTAT-CHF dataset. This can be achieved by applying variable selection techniques to employ all available variables in the dataset. Additionally, it would be beneficial to examine other real-world data with more repeated measures.

The proposed methods were tested using only one continuous predictor with missing data. Future research could involve testing the proposed methods with a nominal incomplete predictor, multiple and mixed types of incomplete predictors, interaction terms with incomplete predictors and when the response variable is binary and subject to missingness. The Two-Step method can include auxiliary variables that are predictive of missingness in the imputation model in step one (MICE algorithm), which we don't want to condition the analysis on these variables (by using the GCRE-MAR method or the GCRE-MNAR method).

Future research could explore the performance of the proposed methods on a finer grid. This would involve a larger number of different combinations of missing proportions, the number of individuals, and repeated measures. The results of the proposed methods showed unbiased parameter estimates as the number of repeated measures increased. Therefore, it is expected that with a larger number of repeated measures (more than eight) and a larger number of participants (more than 100), the proposed methods will yield even more unbiased results; however, this will increase the computational time.

9.5 Practical Recommendations for Analysts

The thesis discusses three proposed methods to address missing data in both the model response and incomplete predictor in longitudinal data. A practical recommendation for analysts is to begin with the GCRE-MNAR method, which provides two parameter estimates on the possibility of missing not at random data for both the model response and the incomplete predictor. This can help understand the missing data mechanism in both the model response and the incomplete predictor. If the results from the GCRE-MNAR inference indicate that the incomplete predictor is less likely to be missing not at random, then it is recommended for analysts to use the GCRE-MAR method. If analysts want to incorporate predictive auxiliary variables of the incomplete predictor with minimal effect on the analysis and don't prioritise time, they could consider using the Two-Step method. In general, it is recommended for analysts to conduct sensitivity analysis when transitioning from one method to another to test the consistency of results.

9.6 Proposed Methods Implementation

The algorithms used in this thesis were implemented using the `R` programming language. These include the Two-Step method, the GCRE-MAR method, and the GCRE-MNAR method. The code for these methods has been developed by the author and is available upon request.

Bibliography

Agiashvili, G. et al. (2021). Probabilistic predictive elicitation.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, Volume 656. John Wiley & Sons.

Agresti, A. (2018). *An Introduction to Categorical Data Analysis*. John Wiley & Sons.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Alvarez, I., J. Niemi, and M. Simpson (2014). Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*.

Alzahrani, H., B. Macdonald, C. Haig, and J. Cleland (2024). Adapting a correlated random effects method to handle ignorable missingness in the predictors and non-ignorable missingness in the response. *The 6th International Conference on Statistics: Theory and Applications (ICSTA'24)*. (Accepted for Publication).

Ambrosy, A. P., G. C. Fonarow, J. Butler, O. Chioncel, S. J. Greene, M. Vaduganathan, S. Nodari, C. S. Lam, N. Sato, A. N. Shah, et al. (2014). The global health and economic burden of hospitalizations for heart failure: lessons learned from hospitalized heart failure registries. *Journal of the American College of Cardiology 63*(12), 1123–1133.

Austin, P. C., I. R. White, D. S. Lee, and S. van Buuren (2021). Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology 37*(9), 1322–1331.

Azzolina, D., P. Berchialla, D. Gregori, and I. Baldi (2021). Prior elicitation for use in clinical trial design and analysis: A literature review. *International Journal of Environmental Research and Public Health 18*(4), 1833.

Bayes, T. (1763). Lii. An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions of the Royal Society of London* (53), 370–418.

Best, N., M. Cowles, and K. Vines (1995). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output. *MRC Biostatistics Unit, Institute of Public Health, Cambridge University*.

Bhuyan, P. (2019). Estimation of random-effects model for longitudinal data with nonignorable missingness using gibbs sampling. *Computational Statistics 34*(4), 1693–1710.

Birmingham, J., A. Rotnitzky, and G. M. Fitzmaurice (2003). Pattern–mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society Series B: Statistical Methodology 65*(1), 275–297.

Boedeker, P. (2017). Hierarchical linear modeling with maximum likelihood, restricted maximum likelihood, and fully bayesian estimation. *Practical Assessment, Research, and Evaluation 22*(1), 2.

Bonnini, S., L. Corain, M. Marozzi, and L. Salmaso (2014). *Nonparametric hypothesis testing: rank and permutation methods with applications in R*. John Wiley & Sons.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software 80*, 1–28.

Burton, A., D. G. Altman, P. Royston, and R. L. Holder (2006). The design of simulation studies in medical statistics. *Statistics in Medicine 25*(24), 4279–4292.

Cai, J.-H., X.-Y. Song, and Y.-I. Hser (2010). A bayesian analysis of mixture structural equation models with non-ignorable missing responses and covariates. *Statistics in Medicine 29*(18), 1861–1874.

Calman, L., L. Brunton, and A. Molassiotis (2013). Developing longitudinal qualitative designs: lessons learned and recommendations for health services research. *BMC Medical Research Methodology 13*, 1–10.

Carrière, I. and J. Bouyer (2002). Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. *BMC Medical Research Methodology 2*(1), 1–10.

Caruana, E. J., M. Roman, J. Hernández-Sánchez, and P. Solli (2015). Longitudinal studies. *Journal of Thoracic Disease 7*(11), E537.

Caruana, L., M. C. Petrie, A. P. Davie, and J. J. McMurray (2000). Do patients with suspected heart failure and preserved left ventricular systolic function suffer from "diastolic heart failure" or from misdiagnosis? a prospective descriptive study. *BMJ 321*(7255), 215–218.

Chaudhuri, S. and V. Agiwal (2024). Strategies for preventing and addressing missing data in research. *Current Medical Issues 22*(3), 181–183.

Choy, S. L., R. O'Leary, and K. Mengersen (2009). Elicitation by design in ecology: using expert opinion to inform priors for bayesian statistical models. *Ecology 90*(1), 265–277.

Collins, L. M., J. L. Schafer, and C.-M. Kam (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods 6*(4), 330.

Cook, C., G. Cole, P. Asaria, R. Jabbour, and D. P. Francis (2014). The annual global economic burden of heart failure. *International Journal of Cardiology 171*(3), 368–376.

Cro, S., J. H. Roger, and J. R. Carpenter (2024). Handling partially observed trial data after treatment withdrawal: Introducing retrieved dropout reference-base centred multiple imputation. *Pharmaceutical Statistics*.

Daniels, M. J. and J. W. Hogan (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC.

De Leeuw, J., E. Meijer, and H. Goldstein (2008). *Handbook of multilevel analysis*, Volume 401. Springer.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Diggle, P., P. Heagerty, K. Liang, and S. Zeger (2013). Analysis of longitudinal data.

Donald, H. and G. R. D. (2006). *Longitudinal Data Analysis*. Chichester: John Wiley and Sons.

Dong, Y. and C.-Y. J. Peng (2013). Principled missing data methods for researchers. *SpringerPlus 2*, 1–17.

Du, H., C. Enders, B. T. Keller, T. N. Bradbury, and B. R. Karney (2022). A bayesian latent variable selection model for nonignorable missingness. *Multivariate Behavioral Research 57*(2-3), 478–512.

Enders, C. K. (2022). *Applied Missing Data Analysis*. Guilford Publications.

Enders, C. K., H. Du, and B. T. Keller (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods 25*(1), 88.

Enders, C. K. and D. Tofighi (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods 12*(2), 121.

Erler, N. (2019, June). *Bayesian Imputation of Missing Covariates*. Ph. D. thesis, Erasmus University Rotterdam.

Erler, N. S., D. Rizopoulos, J. v. Rosmalen, V. W. Jaddoe, O. H. Franco, and E. M. Lesaffre (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full bayesian approach. *Statistics in Medicine 35*(17), 2955–2974.

Fang, F., J. Zhao, and J. Shao (2018). Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. *Statistica Sinica 28*(4), 1677–1701.

Ferreira, J. P., M. Metra, I. Mordi, J. Gregson, J. M. Ter Maaten, J. Tromp, S. D. Anker, K. Dickstein, H. L. Hillege, L. L. Ng, et al. (2019). Heart failure in the outpatient versus inpatient setting: findings from the biostat-chf study. *European Journal of Heart Failure 21*(1), 112–120.

Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs (2008). *Longitudinal Data Analysis*. CRC Press.

Fitzmaurice, G. M. (2003). Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica 57*(1), 75–99.

Gałecki, A., T. Burzykowski, A. Gałecki, and T. Burzykowski (2013). *Linear Mixed-Effects Model*. Springer.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. CRC Press.

Gao, S. (2004). A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. *Statistics in Medicine 23*(2), 211–219.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), 515–534.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. and J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A. and D. B. Rubin (1992). Using multiple sequences. *Statistical Science 7*(4), 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6), 721–741.

George, B. and I. Aban (2015). Selecting a separable parametric spatiotemporal covariance structure for longitudinal imaging data. *Statistics in Medicine 34*(1), 145–161.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics 4*, 641–649.

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling 10*(1), 80–100.

Graham, J. W. et al. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology 60*(1), 549–576.

Groenewegen, A., F. H. Rutten, A. Mosterd, and A. W. Hoes (2020). Epidemiology of heart failure. *European Journal of Heart Failure 22*(8), 1342–1356.

Hajian, A. (2007). Efficient cosmological parameter estimation with hamiltonian monte carlo technique. *Physical Review D 75*(8), 083525.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Hayati Rezvan, P., K. J. Lee, and J. A. Simpson (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology 15*(1), 1–14.

Hoffman, M. D., A. Gelman, et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res. 15*(1), 1593–1623.

Hsu, C.-H., Y. He, C. Hu, and W. Zhou (2023). A multiple imputation-based sensitivity analysis approach for regression analysis with a missing not at random covariate. *Statistics in Medicine 42*(14), 2275–2292.

Huang, A. and M. P. Wand (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis 8*(2), 439–452.

Huang, L., M.-H. Chen, and J. G. Ibrahim (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics 61*(3), 767–780.

Huque, M. H., J. B. Carlin, J. A. Simpson, and K. J. Lee (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology 18*(1), 1–16.

Ibrahim, J. G., M.-H. Chen, and S. R. Lipsitz (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics 30*(1), 55–78.

Ibrahim, J. G., M.-H. Chen, S. R. Lipsitz, and A. H. Herring (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association 100*(469), 332–346.

Ibrahim, J. G. and G. Molenberghs (2009). Missing data methods in longitudinal studies: a review. *Test 18*(1), 1–43.

Jakobsen, J. C., C. Gluud, J. Wetterslev, and P. Winkel (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC Medical Research Methodology 17*(1), 1–10.

James, A., S. L. Choy, and K. Mengersen (2010). Elicitator: an expert elicitation tool for regression in ecology. *Environmental Modelling & Software 25*(1), 129–145.

James, S. L., D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet 392*(10159), 1789–1858.

Janssen, K. J., A. R. T. Donders, F. E. Harrell Jr, Y. Vergouwe, Q. Chen, D. E. Grobbee, and K. G. Moons (2010). Missing covariate data in medical research: to impute is better than to ignore. *Journal of Clinical Epidemiology 63*(7), 721–727.

Jiang, J. and T. Nguyen (2007). *Linear and Generalized Linear Mixed Models and their Applications*, Volume 1. Springer.

Johnson, V. E. and J. H. Albert (2006). *Ordinal Data Modeling*. Springer Science & Business Media.

Jost, A., B. Rauch, M. Hochadel, R. Winkler, S. Schneider, M. Jacobs, C. Kilkowski, A. Kilkowski, H. Lorenz, K. Muth, et al. (2005). Beta-blocker treatment of chronic systolic heart failure improves prognosis even in patients meeting one or more exclusion criteria of the MERIT-HF study. *European Heart Journal 26*(24), 2689–2697.

Kinnersley, N. and S. Day (2013). Structured approach to the elicitation of expert beliefs for a bayesian-designed clinical trial: a case study. *Pharmaceutical Statistics 12*(2), 104–113.

Knol, M. J., K. J. Janssen, A. R. T. Donders, A. C. Egberts, E. R. Heerdink, D. E. Grobbee, K. G. Moons, and M. I. Geerlings (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology 63*(7), 728–736.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.

Lee, K. J. and J. B. Carlin (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology 171*(5), 624–632.

Lee, K. J. and J. A. Simpson (2014). Introduction to multiple imputation for dealing with missing data. *Respirology 19*(2), 162–167.

Lee, Y. and J. A. Nelder (2004). Conditional and marginal models: another view. *Statistical Science 19*(2), 219–238.

Lesaffre, E. and A. B. Lawson (2012). *Bayesian Biostatistics*. John Wiley & Sons.

Levey, A. S., J. Coresh, E. Balk, A. T. Kausz, A. Levin, M. W. Steffes, R. J. Hogg, R. D. Perrone, J. Lau, and G. Eknoyan (2003). National kidney foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Annals of Internal Medicine 139*(2), 137–147.

Levey, A. S., L. A. Stevens, C. H. Schmid, Y. Zhang, A. F. Castro III, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene, et al. (2009). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine 150*(9), 604–612.

Li, H. and Y. Y. Grace (2013). A pairwise likelihood approach for longitudinal data with missing observations in both response and covariates. *Computational Statistics & Data Analysis 68*, 66–81.

Lin, H., D. Liu, and X.-H. Zhou (2010). A correlated random-effects model for normal longitudinal data with nonignorable missingness. *Statistics in Medicine 29*(2), 236–247.

Link, W. A. and M. J. Eaton (2012). On thinning of chains in mcmc. *Methods in Ecology and Evolution 3*(1), 112–115.

Littell, R. C., J. Pendergast, and R. Natarajan (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine 19*(13), 1793–1819.

Little, R. J. and D. B. Rubin (2019). *Statistical Analysis with Missing Data*, Volume 793. John Wiley & Sons.

Lombardi, C. M., J. P. Ferreira, V. Carubelli, S. D. Anker, J. G. Cleland, K. Dickstein, G. Filippatos, C. C. Lang, L. L. Ng, P. Ponikowski, et al. (2020). Geographical differences in heart failure characteristics and treatment across europe: results from the biostat-chf study. *Clinical Research in Cardiology 109*, 967–977.

Lüdtke, O., A. Robitzsch, and S. G. West (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using bayesian estimation. *Psychological Methods 25*(2), 157.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer Science & Business Media.

Ma, Z. and G. Chen (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society 47*(3), 297–313.

Malmberg, I. and U. Persson (2000). Primary health care costs in connection with heart failure surveyed: increased use of ace inhibitors would be beneficial. *Lakartidningen 97*(20), 2465–2470.

Martin, T. G., M. A. Burgman, F. Fidler, P. M. Kuhnert, S. Low-Choy, M. McBride, and K. Mengersen (2012). Eliciting expert knowledge in conservation science. *Conservation Biology 26*(1), 29–38.

Mason, A., N. Best, I. Plewis, and S. Richardson (2010). Insights into the use of bayesian models for informative missing data. Technical report, Imperial College London.

Mason, A. J. (2010). *Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies*. Ph. D. thesis, Imperial College London.

Metra, M., G. Cotter, M. Gheorghiade, L. Dei Cas, and A. A. Voors (2012). The role of the kidney in heart failure. *European Heart Journal 33*(17), 2135–2142.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Michiels, B., G. Molenberghs, L. Bijnens, T. Vangeneugden, and H. Thijs (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine 21*(8), 1023–1041.

Minini, P. and M. Chavance (2004). Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biostatistics 5*(4), 531–544.

Mokkink, L. B., H. de Vet, S. Diemeer, and I. Eekhout (2023). Sample size recommendations for studies on reliability and measurement error: an online application based on simulation studies. *Health Services and Outcomes Research Methodology 23*(3), 241–265.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2014). *Handbook of Missing Data Methodology*. CRC Press.

Molenberghs, G. and M. Kenward (2007). *Missing Data in Clinical Studies*. John Wiley & Sons.

Moons, K. G., R. A. Donders, T. Stijnen, and F. E. Harrell Jr (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology 59*(10), 1092–1101.

Morris, T. P., B. C. Kahan, and I. R. White (2014). Choosing sensitivity analyses for randomised trials: principles. *BMC Medical Research Methodology 14*, 1–5.

Morris, T. P., I. R. White, and M. J. Crowther (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine 38*(11), 2074–2102.

Muff, S., L. Held, and L. F. Keller (2016). Marginal or conditional regression models for correlated non-normal data? *Methods in Ecology and Evolution 7*(12), 1514–1524.

Murphy, J. I., N. E. Weaver, and A. E. Hendricks (2022). Accessible analysis of longitudinal data with linear mixed effects models. *Disease Models & Mechanisms 15*(5).

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology 47*(1), 90–100.

National Kidney Foundation (2002). National kidney foundation K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratication. *American Journal of Kidney Diseases 39*(1), S66–S1.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics 31*(3), 705–767.

Neal, R. M. et al. (2011). MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*(11), 2.

Nevalainen, J., M. G. Kenward, and S. M. Virtanen (2009). Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Statistics in Medicine 28*(29), 3657–3669.

Nishio, M. and A. Arakawa (2019). Performance of hamiltonian monte carlo and no-u-turn sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution 51*, 1–12.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, Volume 698. John Wiley & Sons.

O'Kelly, M. and B. Ratitch (2014). *Clinical Trials with Missing Data: a Guide for Practitioners*. John Wiley & Sons.

Oremus, M., R. McKelvie, A. Don-Wauchope, P. L. Santaguida, U. Ali, C. Balion, S. Hill, R. Booth, J. A. Brown, A. Bustamam, et al. (2014). A systematic review of bnp and nt-probnp in the management of heart failure: overview and methods. *Heart Failure Reviews 19*, 413–419.

Paccagnella, O. (2006). Centering or not centering in multilevel models? the role of the group mean and the assessment of group effects. *Evaluation Review 30*(1), 66–85.

Patton, K. K., P. T. Ellinor, S. R. Heckbert, R. H. Christenson, C. DeFilippi, J. S. Gottdiener, and R. A. Kronmal (2009). N-terminal pro-b-type natriuretic peptide is a major predictor of the development of atrial fibrillation: the cardiovascular health study. *Circulation 120*(18), 1768–1774.

Perkins, J. and D. Wang (2004). A comparison of bayesian and frequentist statistics as applied in a simple repeated measures example. *Journal of Modern Applied Statistical Methods 3*(1), 24.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology 48*(1), 85–112.

Pinheiro, J. and D. Bates (2006). *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, S. Heisterkamp, B. Van Willigen, and R. Maintainer (2017). Package 'nlme'. *Linear and Nonlinear Mixed Effects Models, version 3*(1), 274.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, R. C. Team, et al. (2007). Linear and nonlinear mixed effects mmodels. *R package version 3*(57), 1–89.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). Coda: convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Ponikowski, P., S. D. Anker, K. F. AlHabib, M. R. Cowie, T. L. Force, S. Hu, T. Jaarsma, H. Krum, V. Rastogi, L. E. Rohde, et al. (2014). Heart failure: preventing disease and death worldwide. *ESC Heart Failure 1*(1), 4–25.

Porter, E. K. and J. Carré (2014). A hamiltonian monte–carlo method for bayesian inference of supermassive black hole binaries. *Classical and Quantum Gravity 31*(14), 145004.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team, R. et al. (2013). R: A Language and Environment for Statistical Computing.

Raphael, C., C. Briscoe, J. Davies, Z. I. Whinnett, C. Manisty, R. Sutton, J. Mayet, and D. P. Francis (2007). Limitations of the new york heart association functional classification system and self-reported walking distances in chronic heart failure. *Heart 93*(4), 476–482.

Raudenbush, S. W. and A. S. Bryk (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, Volume 1. SAGE Publications.

Resche-Rigon, M. and I. R. White (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research 27*(6), 1634–1649.

Rizzato, F. B., R. A. Leandro, C. G. Demétrio, and G. Molenberghs (2016). A bayesian approach to analyse overdispersed longitudinal count data. *Journal of Applied Statistics 43*(11), 2085–2109.

Robert, C. P., N. Chopin, and J. Rousseau (2009). Harold jeffreys's theory of probability revisited. *Statistical Science 24*(2), 141–172.

Rockel, T. (2020). *missMethods: Methods for Missing Data*. R package version 0.2.0.

Rosenthal, J. S. et al. (2011). Optimal proposal distributions and adaptive mcmc. *Handbook of Markov Chain Monte Carlo 4*.

Roy, J. and X. Lin (2005). Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. *Biometrics 61*(3), 837–846.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Survey Nonresponse*. New York: Wiley.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*, Volume 81. John Wiley & Sons.

Rubin, D. B. and N. Schenker (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association 81*(394), 366–374.

Savarese, G., P. M. Becher, L. H. Lund, P. Seferovic, G. M. Rosano, and A. J. Coats (2022). Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovascular Research 118*(17), 3272–3287.

Schafer, J. L. and J. W. Graham (2002). Missing data: our view of the state of the art. *Psychological Methods 7*(2), 147.

Schafer, J. L. and R. M. Yucel (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics 11*(2), 437–457.

Schuurman, N., R. Grasman, and E. Hamaker (2016). A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research 51*(2-3), 185–206.

Seide, S. E., K. Jensen, and M. Kieser (2020). A comparison of bayesian and frequentist methods in random-effects network meta-analysis of binary data. *Research Synthesis Methods 11*(3), 363–378.

Shahid, F. and G. Y. Lip (2016). Atrial fibrillation and heart failure: How should we manage our patients? *Arrhythmia & Electrophysiology Review 5*(3), 162.

Smeets, M., B. Vaes, P. Mamouris, M. Van Den Akker, G. Van Pottelbergh, G. Goderis, S. Janssens, B. Aertgeerts, and S. Henrard (2019). Burden of heart failure in flemish general practices: a registry-based study in the intego database. *BMJ Open 9*(1).

Staudt, A., J. Freyer-Adam, T. Ittermann, C. Meyer, G. Bischof, U. John, and S. Baumann (2022). Sensitivity analyses for data missing at random versus missing not at random using latent growth modelling: a practical guide for randomised controlled trials. *BMC Medical Research Methodology 22*(1), 250.

Stavseth, M. R., T. Clausen, and J. Røislien (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine 7*.

Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ 338*.

Stirrup, O. T. (2016). *Extending mixed effects models for longitudinal data before and after treatment*. Ph. D. thesis, UCL (University College London).

Stubbendick, A. L. and J. G. Ibrahim (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics 59*(4), 1140–1150.

Stubbendick, A. L. and J. G. Ibrahim (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, 1143–1167.

Tang, N.-S. and H. Zhao (2014). Bayesian analysis of nonlinear reproductive dispersion mixed models for longitudinal data with nonignorable missing covariates. *Communications in Statistics-Simulation and Computation 43*(6), 1265–1287.

Thisted, R. A. (1988). *Elements of Statistical Computing*. Chapman and Hall.

Thomas, S. and W. Tu (2021). Learning hamiltonian monte carlo in r. *The American Statistician 75*(4), 403–413.

Timmis, A., P. Vardas, N. Townsend, A. Torbica, H. Katus, D. De Smedt, et al. (2022). European society of cardiology: cardiovascular disease statistics 2021. *European Heart Journal 43*, 716–799.

Tsonaka, R., G. Verbeke, and E. Lesaffre (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics 65*(1), 81–87.

Tsutsui, H., N. M. Albert, A. J. Coats, S. D. Anker, A. Bayes-Genis, J. Butler, O. Chioncel, C. R. Defilippi, M. H. Drazner, G. M. Felker, et al. (2023). Natriuretic peptides: role in the diagnosis and management of heart failure: a scientific statement from the heart failure association of the european society of cardiology, heart failure society of america and japanese heart failure society. *European Journal of Heart Failure 25*(5), 616–631.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press.

Van Buuren, S. et al. (2011). Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis 10*, 173–196.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software 45*(1), 1–67.

Varsi, A. (2021). *Streaming Multi-core Sample-based Bayesian Analysis*. Ph. D. thesis, The University of Liverpool, United Kingdom.

Voors, A. A., S. D. Anker, J. G. Cleland, K. Dickstein, G. Filippatos, P. van der Harst, H. L. Hillege, C. C. Lang, J. M. Ter Maaten, L. Ng, et al. (2016). A Systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure: rationale, design, and baseline characteristics of BIOSTAT-CHF. *European Journal of Heart Failure 18*(6), 716–726.

Voors, A. A., W. Ouwerkerk, F. Zannad, D. J. van Veldhuisen, N. J. Samani, P. Ponikowski, L. L. Ng, M. Metra, J. M. Ter Maaten, C. C. Lang, et al. (2017). Development and validation of multivariable models to predict mortality and hospitalization in patients with heart failure. *European Journal of Heart Failure 19*(5), 627–634.

Weber, M. and C. Hamm (2006). Role of b-type natriuretic peptide (bnp) and nt-probnp in clinical routine. *Heart 92*(6), 843–849.

Werhahn, S. M., C. Becker, M. Mende, H. Haarmann, K. Nolte, U. Laufs, S. Zeynalova, M. Löffler, N. Dagres, D. Husser, et al. (2022). Nt-probnp as a marker for atrial fibrillation and heart failure in four observational outpatient trials. *ESC Heart Failure 9*(1), 100–109.

White, I. R., J. Carpenter, S. Evans, and S. Schroter (2007). Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials 4*(2), 125–139.

White, I. R., P. Royston, and A. M. Wood (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine 30*(4), 377–399.

Wilson Tang, W. (2007). B-type natriuretic peptide: a critical review. *Congestive Heart Failure 13*(1), 48–52.

Wong, A. Y., S. Warren, and G. N. Kawchuk (2010). A new statistical trend in clinical research–bayesian statistics. *Physical Therapy Reviews 15*(5), 372–381.

Wu, J., J. G. Ibrahim, M.-H. Chen, E. D. Schifano, and J. D. Fisher (2018). Bayesian modeling and inference for nonignorably missing longitudinal binary response data with applications to hiv prevention trials. *Statistica Sinica 28*, 1929.

Wu, L. (2008). An approximate method for nonlinear mixed-effects models with nonignorably missing covariates. *Statistics & Probability Letters 78*(4), 384–389.

Yang, M. and S. E. Maxwell (2014). Treatment effects in randomized longitudinal trials with different types of nonignorable dropout. *Psychological Methods 19*(2), 188.

Yang, Z. and C. E. Rodríguez (2013). Searching for efficient markov chain monte carlo proposal kernels. *Proceedings of the National Academy of Sciences 110*(48), 19307–19312.

Yu, F., M.-H. Chen, L. Huang, and G. J. Anderson (2013). Hierarchical bayesian analysis of repeated binary data with missing covariates. In *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, pp. 311–322. Springer.

Yuan, Y. and G. Yin (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics 66*(1), 105–114.

Zapata-Vázquez, R. E., A. O'Hagan, and L. Soares Bastos (2014). Eliciting expert judgements about a set of proportions. *Journal of Applied Statistics 41*(9), 1919–1933.

Zhang, Z. (2021). A note on wishart and inverse wishart priors for covariance matrix. *Journal of Behavioral Data Science 1*(2), 119–126.

Zhou, X. and J. P. Reiter (2010). A note on bayesian inference after multiple imputation. *The American Statistician 64*(2), 159–163.

Zhu, J. and T. E. Raghunathan (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association 110*(511), 1112–1124.