University of Southampton

Faculty of Social and Human Sciences

Multiple imputation for missing data and statistical disclosure control for mixed-mode data using a sequence of generalised linear models

Min Cherng Lee

Doctor of Philosophy

June 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Statistics

Doctor of Philosophy

MULTIPLE IMPUTATION FOR MISSING DATA AND STATISTICAL DISCLOSURE CONTROL FOR MIXED-MODE DATA USING A SEQUENCE OF GENERALISED LINEAR MODELS

by Min Cherng Lee

Multiple imputation is a commonly used approach to deal with missing data and to protect confidentiality of public use data sets. The basic idea is to replace the missing values or sensitive values with multiple imputation, and we then release the multiply imputed data sets to the public. Users can analyze the multiply imputed data sets and obtain valid inferences by using simple combining rules, which take the uncertainty due to the presence of missing values and synthetic values into account. It is crucial that imputations are drawn from the posterior predictive distribution to preserve relationships present in the data and allow valid conclusions to be made from any analysis. In data sets with different types of variables, e.g. some categorical and some continuous variables, multivariate imputation by chained equations (MICE) (Van Buuren (2011)) is a commonly used multiple imputation method. However, imputations from such an approach are not necessarily drawn from a proper posterior predictive distribution. We propose a method, called factored regression model (FRM) to multiply impute missing values in such data sets by modelling the joint distribution of the variables in the data through a sequence of generalised linear models. We use data augmentation methods to connect the categorical and continuous variables and this allows us to draw imputations from a proper posterior distribution. We compare the performance of our method with MICE using simulation studies and on a breast-feeding data. We also extend our modelling strategies to incorporate different informative priors for the FRM to explore robust regression modelling and the sparse relationships between the predictors. We then apply our model to protect confidentiality of the current population survey (CPS) data by generating multiply imputed, partially synthetic data sets. These data sets comprise a mix of original data and the synthetic data where values chosen for synthesis are based on an approach that considers unique and sensitive units in the survey. Valid inference can then be made using the combining rules described by Reiter (2003). An extension to the modelling strategy is also introduced to deal with the presence of spikes at zero in some of the continuous variables in the CPS data.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, MIN CHERNG LEE, declare that the thesis entitled

MULTIPLE IMPUTATION FOR MISSING DATA AND STATISTICAL DISCLOSURE CONTROL FOR MIXED-MODE DATA USING A SEQUENCE OF GENERALISED LINEAR MODELS

and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signed .....

Date .....

Acknowledgements

I would firstly like to thank my supervisor, Dr. Robin Mitra for his guidance, enthusiasm and encouragement throughout. I am also grateful for Professor Jennifer Hill for providing the breast-feeding data used in Chapter 3 and Dr. Jörg Drechsler for providing the current population survey (CPS) data used in Chapter 6.

Finally, a thank you also to all my colleagues at the university that I have shared offices with through the years; it has been an enormous pleasure.

List of Abbreviations

CART	Classification and regression tree
CPS	Current population survey
FCS	Fully conditional specification
FRM	Factored regression model
JM	Joint modelling
LASSO	Least absolute shrinkage and selection operator
LFRM	Lasso factored regression model
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MFRM	Modified factored regression model
MICE	Multiple imputation via chained equations
MVN	Multivariate normal distribution
NMAR	Not missing at random
ODA	Open data alliance
ODI	Open data institute
OLS	Ordinary least squares
RFRM	Robust factored regression model
SDL	Statistical disclosure limitation
SFRM	Skip factored regression model

Chapter 1

Introduction

This thesis develops imputation modelling strategies that can be used to handle the problems of missing data and to protect confidentiality of public use data sets. We propose a method, called factored regression model (FRM) to multiply impute missing values in data sets that contain categorical and continuous variables. We model the joint distribution of the variables in the data through a sequence of generalised linear models, we then connect the categorical and continuous variables through data augmentation methods. This allows us to draw imputations from a proper posterior predictive distribution in order to preserve the statistical properties in the imputed data sets. This model can also be applied to generate partially synthetic data to protect confidentiality of data sets. The key idea is to treat the values being selected for synthesis as missing data, and then use the proposed model to replace the selected values with synthetic values, hence releasing multiply imputed, partially synthetic data sets. In this chapter, we first describe the general problems of missing data and some common strategies to handle them, and we then describe our modelling strategy to impute the missing values in general terms. In the second part of this chapter we will review the area of statistical disclosure control and the use of partially synthetic data to protect confidentiality of public use data sets, and describe the benefits of our modelling approach.

1.1 Missing data problems

Survey data sets often consist of large numbers of variables which have different measurement scales. Typically, such data sets have some continuous, binary, ordinal, nominal, and semi-continuous variables. The presence of missing values in such data sets can cause problems for the data users who don't have the required skills and knowledge to deal with the missing data. Common statistical procedures used by the data users normally remove units with missing values, and the users can then perform analysis on the remaining data. This method is called the complete case analysis and it might lead to biased estimates of parameters if the missing mechanism is not missing completely at random. Ad-hoc methods such as mean imputation, or single imputation from regression models have been

considered. However, studies have shown that while these ad-hoc methods are simple to use, they can lead to biased estimates of parameters (Schafer and Graham (2002)). Both the complete case analysis and ad-hoc methods fail to reflect any uncertainty due to presence of missing data, as an alternative, we use multiple imputation, which was proposed by Rubin (1987) to tackle this problem.

Imputation is the process where data imputers replace the missing values with some simulated values. An important reason why imputation is used is because data imputers and analysts can be distinct entities. The analysts might not have the required knowledge to tackle the problems arising from the presence of missing values in the data sets. Hence, it is the imputers' responsibility to deal with the missing data problems. Multiple imputation is a statistical technique for handling incomplete data and allow valid inferences to be made. It simply extends single imputation by replacing each missing value by m sets of imputed values, which are drawn from a plausible distribution. More specifically, the m sets of the imputations will fill in the missing values m times to create m completed data sets with no missing values. The data analysts can apply the complete-data analysis to each of the completed data set and combine the results using simple combining rules proposed by Rubin (1987). The data analysts can then obtain valid statistical inferences using the combined results which reflect the uncertainty due to presence of missing values.

There are two common strategies used to impute missing values: the joint modelling (JM) and the fully conditional specification (FCS). The joint modelling approach assumes that all variables that are included in the imputation model jointly follow a multivariate distribution. The commonly used multivariate distributions are the multivariate normal distribution and the general location model (Schafer (1997)). The main advantage of this approach is that it leads to imputation procedures whose statistical properties are known, as the joint modelling is based on parametric statistical theory. However, any important relations between the variables in the data set that are outside that theoretical structure of the imputation model might not be adequately captured. Hence the imputations for missing data depend heavily on the imputation model that we define, and as a result, might lead to biased inference if the imputation model is mis-specified. Also, large data sets usually consist of a mixture of categorical, continuous and semi-continuous variables, and the assumption of multivariate normality is often not plausible, as it lacks flexibility needed to represent data with such features.

An alternative approach to impute missing data is called fully conditional specification (FCS) proposed by Van Buuren (2007). FCS is a flexible alternative that specifies the multivariate model by a series of conditional regression models, one for each incomplete variable. The regression models are specified for each variable with missing values, conditional on all of the other variables in the imputation model. Imputations are performed on a variable by variable iterative basis. The main advantage of FCS is that it is flexible to use as it does not restrict the conditional distributions to being normal, so that univariate regression models can be tailored appropriately for data sets with categorical, continuous and semi-continuous variables. Common options for the univariate conditional distributions are normal regression models for continuous variables, logistic regressions for binary variables and ordered logistic regressions for ordinal variables. Other choices of conditional regression models can be found in Van Buuren (2007). FCS is flexible and easy to apply, but its statistical properties are difficult to establish. This method is available as a package, called multivariate imputation by chained equations (MICE) in

the software package R (http://cran.r-project.org/web/packages/mice/index.html). This method is easy to apply but has no guarantee of generating imputations from a proper posterior predictive distribution. A similar idea has also been proposed by Raghunathan et al. (2001). For more details on comparison between both strategies, please see Lee and Carlin (2010) and Li et al. (2012).

One of the main challenges in using the joint modelling approach is defining a joint model for complex data sets that contain categorical and continuous variables with a non-monotone missing pattern. This motivates us to propose a joint modelling approach to multiply impute missing values in data sets containing mixed categorical (binary, ordinal and nominal) and continuous variables with a non-monotone missing pattern. In order to preserve statistical properties in the imputed data sets, imputed values should be drawn from a plausible distribution. The missing values are drawn from a posterior predictive distribution that is the distribution of the missing values conditional on the observed values. However, this distribution is often not determined analytically in data sets with mixed binary, ordinal, nominal and continuous variables with a non-monotone missing pattern. Hence, we use Markov chain Monte Carlo (MCMC) techniques to sample the missing values from their posterior predictive distributions. For a given a data set, we model the joint distribution of the variables by decomposing the joint distribution into a sequence of univariate conditional distributions. Each of these univariate conditional distributions is a generalised linear model. This approach of modelling the joint distribution has been considered by others such as Lipsitz and Ibrahim (1996) and Ibrahim et al. (2005).

By modelling the joint distribution of all the variables in the data through a sequence of generalised linear models, we can draw imputations from a proper posterior distribution. Specifically, we introduce latent variables underlying the binary and ordinal variables in our data set (Albert and Chib (1993)), allowing us to use a data augmentation technique proposed by Tanner and Wong (1987) to impute the missing values by connecting the latent variables to the continuous variables through a multivariate normal distribution. In data sets containing only binary, ordinal and continuous variables, the full conditional distributions of all the unknowns (missing data and parameters) are available in closed form and hence a Gibbs sampler can be used to sample the unknowns from their joint posterior distributions. In the case where we also have nominal variables, certain full conditional distributions are not available in closed form, hence a Metropolis-Hastings step is needed to sample certain parameters from their posterior distributions. We propose an innovative independent Metropolis-Hastings proposal that performs well in updating the necessary parameters from their full conditional distributions. We will discuss this modelling approach further in Chapter 3.

We also extend our modelling strategies to include robust regression modelling as well as exploring the potential sparse relationships between the variables in the data sets. When we decompose the joint distribution into a sequence of univariate linear regression models, the error terms of the univariate linear regression models are assumed to be normally and independently distributed, each with zero mean and common variance. A known limitation with this normality assumption is its non-robustness, which does not allow for heavy tailed error distributions. To deal with the heavy tailed features, we assume that the errors of the linear regression models follow a marginal t-distribution, which are more robust to the outliers in the data. We also assign a marginal t-prior to the regression parameters in our Gibbs sampler, and this will still give us a set of tractable full conditional distributions

which allow us to sample the unknowns (missing data and parameters) from their joint posterior distributions.

Another approach that has been considered in extending our modelling strategies is incorporating the Bayesian Lasso (Park and Casella (2008)) into our model. This is the approach where we are exploring the potential sparse relationships between the variables in the data set. The Bayesian Lasso has features such as shrinking the regression coefficients toward zero, resulting in some regression coefficients identically equal to zero. Such shrinkage features can reduce the variance of the estimates by sacrificing a little bias, and hence may increase the prediction accuracy which could potentially lead to more plausible imputations for our model.

Finally, we have considered a modelling strategy where we can explore the robustness of the linear regression model and the sparse relationship between variables at the same time. We assign a marginal t-prior for the errors of the linear regression models, which are more robust to the outliers in the data, and we also assign Bayesian Lasso prior for the regression parameters. This strategy allows us to include robust modelling as well as exploring potential sparse relationship between variables in the data set under one modelling strategy. Details on these 3 approaches can be found in Chapter 5.

1.2 Statistical disclosure control

For the last two decades, open access to data has become more and more popular among statistical agencies and government organizations around the world. (Streeter et al. (1996) and Reichman et al. (2011)). Open data is the concept where data users can access certain data for free, re-use as well as redistribute the data for any scientific research, or commercial purposes without restrictions from copyright and infringement. The US and UK governments are among the leaders in promoting the open data concept globally, by launching the open-data government initiatives such as www.data.gov and www.data.gov.uk. More recently, the UKs Open Data Institute (ODI) and Taiwans Open Data Alliance (ODA) have signed a letter of intent for future cooperation. This collaboration will promote the open data access holds in both countries for the academic, public, and private sectors. With better access to data, both sides can strengthen cooperation in the area of economies, government policies, climate change, health care issues as well as developing open data technologies.

So why is open data so important in the modern era? One of the main reasons is transparency. We are living in a democratic world, where every citizen has the right to know what their government is doing. In order to do that, we need to have free access to government data because these data contain information on areas such as the economy, innovation, defense, and other services that are provided by the government. Using these data, we can then do some research and decide whether the government is acting out of the best interests of its citizens. Also, business communities can use the opportunity of open data access for the development of new and innovative ideas that can deliver commercial values.

However, when statistical agencies or the public can access the data freely, it does raise the question of confidentiality of the released data. Most of the data sets that have private information on individuals (name or national insurance number) as well as sensitive information (income, race or personal wealth) are normally collected under confidentiality pledges, which restricts the release of such information to the public. One of the solutions to address this problem is to forbid remote access to the original data. Any data users that wish to perform analysis on the data can send request to the statistical agencies for analyses, and will only receive the results of the statistical analyses submitted. Since the data users can't access the original data, the data agencies can protect confidentiality of the data set. However, different data users have different desirable statistical analyses, by responding to each of them might become a burden to the statistical agencies.

Alternatively, data users are only allowed to access the data in designated data centers. This method can limit the chances of disclosure of confidential information from the released data as data users are not allowed to bring the data outside the data centers, but occasionally this is inconvenient for the data users as they might have to travel for a long distance to arrive at the data center. So if restricted access to data is not viable, then the data agencies can consider statistical disclosure control for protecting the confidentiality of public use data sets.

Statistical disclosure control refers to the methodology used to protect the confidentiality of public use data sets. Statistical agencies that release public data have legal obligations to protect the confidentiality of the respondents' identities and sensitive attributes they collect. At the same time, they have to preserve the utility of the data for the benefit of analysis of the data. Agencies have therefore developed different methods to protect confidentiality of public use data sets. One method is to remove uniquely identifying information such as name and social security numbers of the respondents. Such a method can minimize the damage done to the utility of the data; however, this is normally considered insufficient for confidentiality protection. This is due to fact that intruders may still be able to link remaining identifying information in the data to another external publicly available data source. A common approach is then to modify the released public data by applying statistical disclosure limitation (SDL) methods to those variables that will potentially be used to make identifications. However, these SDL methods can complicate analyses for users and can severely reduce the utility of the released data. We will discuss these further in Chapter 2.

An alternative approach to SDL methods is to use multiple imputation for statistical disclosure control. The early idea, which was proposed by Rubin (1993) is to release multiply imputed, fully synthetic data sets which comprise entirely of synthetic data rather than the original data. In this approach, the statistical agencies first propose a model that describes the original data, the agencies then generate synthetic data using this model and hence release multiple versions of these synthetic data sets to the public. Since the released data are comprised entirely of synthetic values, the identification of individuals and their sensitive information is very unlikely and hence this can protect the confidentiality of the released data sets. This idea has been developed further to the concept of the partially synthetic data, which propose that only sensitive units or variables need to be synthesized and replaced by synthetic values. The partially synthetic data approach was proposed by Little (1993b), where statistical agencies release multiply imputed, partially synthetic data sets.

In this approach, the statistical agencies release multiply imputed, partially synthetic data sets which comprise some original data values with sensitive data replaced by synthetic values. The approach first defines a model that describes the complete data, and then selects the values that have high disclosure risk to be replaced by synthetic data. There are similar challenges here to those with multiple imputation for missing data, complex large data sets containing mixed categorical (binary, ordinal and nominal) and continuous variables, as well as patterns of synthesis that can have a non-monotone pattern. Hence, in this thesis, we will apply our modelling strategy used to impute missing data to generate synthetic data. The key idea is to treat the sensitive units or values in a given data set, as missing values, and thus impute them. The synthetic values are drawn from a posterior predictive distribution that is the distribution of the synthetic values conditional on the observed values, so that the statistical properties in the imputed data sets can be preserved. Hence, the users can then obtain valid inference from these synthetic data sets using the combining rules described in Reiter (2003). We adapt the modelling strategy proposed to impute missing values to generate partially synthetic data for the current population survey (CPS) data set and assess the risk and utility of the synthetic data. We also extend our modelling strategy to address the presence of spikes at zero in some of the continuous variables in the CPS data. Details on this approach can be found in Chapter 6.

The structure of the thesis will be as follows: first, Chapter 2 reviews the literature about missing data and statistical disclosure control. Chapter 3 then looks at the overview of Markov chain Monte Carlo (MCMC), a necessary computational technique that we need to draw from the posterior distribution. We also propose our modelling strategy to multiply impute missing data in Chapter 3 and briefly describe a commonly used alternative known as MICE. Chapter 4 will illustrate the performance of the modelling strategy through simulations and on a real data set taken from a breast-feeding study and compares the results to the performance using MICE. We then extend our modelling strategies to use informative priors in Chapter 5. We then apply our model to generate multiply imputed, partially synthetic data to protect confidentiality of the CPS data in Chapter 6. We finish the thesis with conclusions and future work in Chapter 7.

Chapter 2

Literature review

2.1 Introduction to missing data

In this section we review the literature on missing data including some of the ad-hoc strategies for handling them.

2.1.1 Missing patterns

We can consider two important missing data patterns in our data sets, namely monotone and non-monotone missing patterns. Suppose we have a fully observed $n \times p$ data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})'$, $j = 1, \ldots, p$ is the j-th variable in $\mathbf{X}$. With missing data, we denote an $n \times p$ missing data indicator matrix $\mathbf{M} = (m_{1,j}, \ldots, m_{n,j})'$, where $m_{i,j} = 1$ indicates $x_{i,j}$ is missing and $m_{i,j} = 0$ indicates $x_{i,j}$ is observed, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We can then denote the observed and missing portions of $\mathbf{X}$ by $\mathbf{X}_{obs} = \{x_{i,j} : m_{i,j} = 0\}$ and $\mathbf{X}_{mis} = \{x_{i,j} : m_{i,j} = 1\}$ respectively. The matrix $\mathbf{M}$ then defines the pattern of missing data. A monotone missing pattern is where we can reorder the data set in such a way that if $m_{i,j} = 1$ for $j = 2, \ldots, p$, then $m_{i,k} = 1$ for $k = j+1, \ldots, p$. If this is not possible then the missing pattern is said to be non-monotone. Figure 2.1 shows a monotone pattern (left) and non-monotone pattern (right) where ticked boxes represent observed data while missing data are denoted by boxes with question marks.

x_1	x_2	x_3	x_4	x_5
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✓	✓	✓	✓	?
✓	✓	✓	?	?
✓	✓	?	?	?
✓	?	?	?	?
✓	?	?	?	?

x_1	x_2	x_3	x_4	x_5
?	✓	✓	✓	✓
✓	✓	?	✓	✓
✓	✓	?	✓	?
?	✓	✓	?	?
?	✓	?	?	✓
✓	?	✓	?	?
✓	?	?	?	✓

Figure 2.1: Illustration of missing patterns

2.1.2 Missing mechanisms

2.1.2.1 Missing at random (MAR)

We now formalise the idea of missing mechanisms. Despite the name, missing at random (MAR) does not imply that the missing values are comprised of a simple random sample from the data set. Given a data set $\mathbf{X}$, we partition $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$ where the observed part of $\mathbf{X}$ is $\mathbf{X}_{obs}$, i.e. $\mathbf{X}_{obs} = \{x_{i,j} : m_{i,j} = 0\}$ and the missing part is $\mathbf{X}_{mis}$, i.e. $\mathbf{X}_{mis} = \{x_{i,j} : m_{i,j} = 1\}$. MAR is then defined as a mechanism where the probability that an observation is missing may depend on $\mathbf{X}_{obs}$, but not on $\mathbf{X}_{mis}$. That is,

$$f(\mathbf{M}|\mathbf{X}, \boldsymbol{\xi}) = f(\mathbf{M}|\mathbf{X}_{obs}, \boldsymbol{\xi}) \quad \text{for all } \mathbf{X} \text{ and } \boldsymbol{\xi},$$

where $f(\mathbf{M}|\mathbf{X}, \boldsymbol{\xi})$ is a probability model for the indicator function $\mathbf{M}$ with unknown parameter $\boldsymbol{\xi}$. A practical example of missing at random in a survey is where the variable "income" has some missing values. The variable "income" is assumed to be missing at random if the probability that an observation is missing depends only on other fully observed variables such as race, education and age in the survey.

2.1.2.2 Missing completely at random (MCAR)

Missing completely at random (MCAR) is a special case of MAR. This mechanism is such that the missing values are a simple random sample from the data set. More formally, it means that the missingness does not depend on the observations of the data set, rather than that the missingness itself is random. That is,

$$f(\mathbf{M}|\mathbf{X}, \boldsymbol{\xi}) = f(\mathbf{M}|\boldsymbol{\xi}) \quad \text{for all } \mathbf{X} \text{ and } \boldsymbol{\xi}.$$

In the example above, the missing values in the variable "income" would be assumed to be missing completely at random if the missing values are just a simple random sample

from the survey.

2.1.2.3 Not missing at random (NMAR)

The mechanism is called not missing at random (NMAR) if the distribution of $\mathbf{M}$ depends, in part, on the missing values in the data set $\mathbf{X}$. Then the distribution of $\mathbf{M}$ is given by

$$f(\mathbf{M}|\mathbf{X}, \boldsymbol{\xi}) = f(\mathbf{M}|\mathbf{X}_{mis}, \mathbf{X}_{obs}, \boldsymbol{\xi}).$$

For example, suppose that survey respondents with higher income (more than \$100,000) are less likely to respond to the income question. Then, the variable "income" is assumed to be not missing at random. If the missing mechanism is not missing at random, it must be explicitly modelled, or else it will lead to biased inferences. See Little and Rubin (2002) Chapter 2 and Rubin (1976) for more details on missing mechanisms.

2.1.3 Ad-hoc methods for dealing with missing data

There are different methods to deal with missing data, which include ad-hoc methods such as complete case analysis and single imputation methods. These approaches, together with multiple imputation will be discussed in the following sections.

2.1.3.1 Complete case analysis

This is the most common method where we simply discard the rows of a data set that contain missing values and analyze only the completely observed values. The advantages of this method are simplicity and that standard statistical analyses can be applied without modification. When the missing values comprise only a small fraction of all cases, then the complete case analysis may be a reasonable solution to deal with the missing data problem. In a data set where more than one variable has missing values, the number of incomplete cases are often a substantial portion of the entire dataset. Deleting them will cause a large amount of information being discarded, and complete case analysis can also lead to biased estimates of parameters if the missing mechanism is not MCAR.

2.1.3.2 Mean imputation

Mean imputation is another method for handling missing data. We impute the missing values for each variable with the mean of the observed values of the variable in the data set. It therefore assumes that the mean of the variable is the best estimate for any observation that has missing information on that variable. In contrast to the complete case analysis, the mean imputation method does not alter the sample mean of the variable and does not discard observed information in the data set. The main disadvantage of this method is

that we would underestimate the variability in the data as the same value is being imputed for each of the missing values.

2.1.3.3 Conditional mean imputation

This method imputes the missing values with imputed values generated from the fitted regression model where the regression model is formed based on the observed values in the data set. One of the advantages of this method over mean imputation is that the imputed values are in some way conditional on other variables in the data set. Similar to mean imputation, conditional mean imputation has the advantage of not discarding cases with missing data and maintaining the sample size of the data. The main disadvantage of conditional mean imputation is that it can be difficult to apply fitted regression model to data sets where more than one variable has missing values. Also, the problem in underestimating variability remains.

2.1.3.4 Single imputation from the posterior predictive distributions

Based on a Bayesian approach, we draw the missing values $\mathbf{X}_{mis}$ from their posterior predictive distributions $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$. Such distributions are based on a model for the complete data $\mathbf{X}$ and we will discuss this approach further in Chapter 3. However, this method tends to underestimate the uncertainty due to the presence of missing values (Schafer and Graham (2002), Little and Rubin (2002)), leading to biased confidence intervals.

In general, mean imputation, conditional mean imputation, and single imputation from the posterior predictive distributions address the missing data problem by imputing missing values once; thus they are referred to as single imputation methods. The single imputation approach unfortunately does not reflect the uncertainty due to presence of missing values. As an alternative we use multiple imputation, which will be discussed in the next section, to tackle this problem.

2.2 Multiple imputation for missing data

Suppose an analyst wishes to use the data to infer about some population quantity Q, this might be for example the mean of j-th variable or it could be the coefficient from a regression model for one of the variables on some other subset of variables. To do this they obtain a point estimate, q, for Q, and an estimate of its variance u. With the presence of missing data, analysts may no longer be able to obtain these estimates. Alternative approaches that are more statistically principled have been proposed to tackle this problem and multiple imputation is one of them. In multiple imputation, we generate imputed values from their posterior predictive distributions based on a model for the complete data. We do this m times, to generate m multiply imputed data sets. Figure 2.2 illustrates this procedure in a case where we have an observed data set with variables $\mathbf{x}_1$ and $\mathbf{x}_2$ where $\mathbf{x}_1$

is fully observed (all the boxes are ticked) and $\mathbf{x}_2$ has some missing data (denoted by empty boxes). To apply multiple imputation, we generate 3 sets of independent missing values from their posterior predictive distribution, i.e. $p(\mathbf{x}_2|\mathbf{x}_1)$. These independent missing values are denoted by 3 different colours: orange, pink and green. Here we have imputed the missing values 3 times, to generate 3 imputed data sets.

Figure 2.2: Illustration of multiple imputation

In multiple imputation, the missing values are imputed from their posterior predictive distribution $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$, this is done m times to generate m completed data sets $\mathbf{X}_{com}^{(1)}, \ldots, \mathbf{X}_{com}^{(m)}$. Analysts can then treat each completed data set as a fully observed data set to obtain m sets of point and variances estimates (q_k, u_k) from each $\mathbf{X}_{com}^{(k)}$, $k = 1, \ldots, m$. Appropriate inference for Q can then be made following simple combining rules (Rubin (1987)). Specifically the analyst computes the following quantities,

$$\bar{q}_m = \frac{1}{m}\sum_{k=1}^{m} q_k, \qquad \overline{u}_m = \frac{1}{m}\sum_{k=1}^{m} u_k,$$
$$B_m = \frac{1}{m-1}\sum_{k=1}^{m}(q_k - \bar{q}_m)^2, \qquad T_m = \overline{u}_m + \left(1 + \frac{1}{m}\right) B_m, \tag{2.1}$$

where $\bar{q}_m$ is the mean of the m imputed data point estimates and $\overline{u}_m$ represents the average within imputation variance. The estimate B_m is known as the between imputation variance and is the variance between the m imputed data point estimates. This variance arises as we are now analysing m imputed data sets instead of one and B_m reflects the additional uncertainty due to the presence of the imputed values. The estimate T_m represents the total variance, which combines the within and between imputation variance. This thus takes account of the extra variability in our estimate for Q. Analysts can then use $\bar{q}_m$ as a point estimate for Q, and a $100(1-\alpha)\%$ interval estimate for Q is given by

$$\left[\bar{q}_m \pm t_{\nu_m}(1-\alpha/2)\sqrt{T_m}\right].$$

where the degrees of freedom of the t-distribution in the above expression are given by

$$\nu_m = (m-1)\left(1 + \frac{m}{m+1}\frac{\overline{u}_m}{B_m}\right)^2. \tag{2.2}$$

Modified degrees of freedom have been proposed by Rubin and Barnard (1999) when n is small, while Steele et al. (2010) explore alternative strategies to making inferences when m is small. How do we choose the number of imputations m? The common advice is to use a low number of imputations, such as $m = 3, 5$ for moderate amounts of missing information. Several authors such as Bodner (2008), Graham et al. (2007) and Carpenter and Kenward (2007) have discussed the issues on selecting the number of imputations that are needed in multiple imputation.

2.3 Statistical disclosure control

Statistical disclosure control refers to the methodology used to protect the confidentiality of public use data sets. When releasing data to the public, statistical agencies, survey organizations, and researchers are required to protect the confidentiality of survey respondents' identities and attribute values. One method is simply to remove uniquely identifying information such as name and social security number of the respondents. By applying such a method, damage done to the utility of the data can be minimized; however, this is normally considered insufficient for protecting the confidentiality of the data set. This is because intruders may still be able to link the remaining identifying information to another external publicly available data source. A common approach is then to modify the released public data by applying statistical disclosure limitation (SDL) methods to those variables that will potentially be used to make identifications. Common strategies include recoding variables, such as releasing ages in five year intervals (Willenborg and de Waal (2001)); reporting exact values only above certain thresholds, for example top coding all incomes above \$100 000 as "\$100 000 or more"(Willenborg and de Waal (2001)); swapping data values for selected records (Fienberg and McIntyre (2004)) and adding noise to numerical data values (Fuller (1993)). We now review the various common SDL methods used for statistical disclosure control.

2.3.1 Statistical disclosure limitation (SDL) methods

In this section we will discuss the various common SDL methods used to protect the confidentiality of the respondents in public use data sets.

2.3.1.1 Topcoding/Bottomcoding

Statistical agencies can apply the method of topcoding in order to prevent disclosure of extreme values in a variable. The method of topcoding (Willenborg and de Waal (2001)) is to replace the value of the variable $\mathbf{x}$ by the value C if $\mathbf{x} \geq C$, where C is some chosen threshold value. For example, top coding all incomes above \$100 000 as "\$100 000 or more". The same can be done for variables whose values are below a certain threshold C, i.e. $\mathbf{x} \leq C$. Hence, instead of the real individual income, the intruder will know only that

the value is greater or smaller than the threshold value C. However, such a method will skew the data distribution and analysts might obtain biased estimates of parameters.

2.3.1.2 Recoding

Statistical agencies might release variables as categories, such as recoding age into 5-year intervals. This method can help reduce the intruders' probability of making identifications. However, the data will provide less detailed information when continuous data are replaced with intervals. Please see Willenborg and de Waal (2001) for more details on this method.

2.3.1.3 Data swapping

Statistical agencies can swap data values of sensitive variables with other records' values in order to reduce the risk of disclosure. This method can preserve univariate distributions. However, one drawback of this method is that it may not maintain multivariate relationships between the variables in the data set. See Fienberg and McIntyre (2004) for more details on the data swapping method.

2.3.1.4 Adding random noise

Statistical agencies can protect confidentiality by modifying the values of the variables in a stochastic way. This involves adding random noise, ϵ, with zero mean and predefined variance to all values of the potentially identifying variables when the variables are continuous. However, this method does not maintain the covariance structure of the original data. Fuller (1993) provides more details on this method.

A drawback in general with all these SDL methods are that they can complicate analyses for users and can severely reduce the utility of the released data by distorting the relationships between variables in the data set. An alternative approach, multiple imputation for fully synthetic data and partially synthetic data, will be discussed in the next section.

2.4 Multiple imputation for fully synthetic data

The fully synthetic data approach was first proposed by Rubin (1993), by releasing multiply imputed fully synthetic data sets which comprise entirely of synthetic data rather than the observed data. This approach protects confidentiality of the data sets, since the identification of unique individuals and their sensitive information is arguably not possible when the released data are synthetic values. This concept is based on the idea of multiple

imputation (Rubin (1987)), where the synthetic data are repeatedly drawn from a posterior predictive distribution and this allows data users to obtain valid inferences using simple combining rules developed by Reiter et al. (2003).

Suppose we have a fully observed $n \times p$ data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})'$, for $j = 1, \ldots, p$ is the j-th variable in $\mathbf{X}$. Let $\mathbf{X}_{syn}$ be the synthetic values generated from the posterior predictive distribution $p(\mathbf{X}_{syn}|\mathbf{X})$. We then make d independent imputations to create d different synthetic data sets, where each of the synthetic data sets, $\mathbf{D}_{syn}^{(k)}, k = 1, \ldots, d$, is comprised only of synthetic data, $(\mathbf{X}_{syn}^{(k)})$. These synthetic data sets are then released to the public. Figure 2.3 illustrates this procedure in a case where we have an observed data set with variables $\mathbf{x}_1$ and $\mathbf{x}_2$. We first generate 2 sets of synthetic values, each for $\mathbf{x}_1$ and $\mathbf{x}_2$ from their posterior predictive distributions respectively, i.e. $p(\mathbf{x}_1|\mathbf{x}_2)$ and $p(\mathbf{x}_2|\mathbf{x}_1)$. These independent synthetic values are denoted by 4 different colours: yellow, dark blue, red and light blue. Here we have imputed the data set 2 times, to generate 2 multiply imputed, fully synthetic data sets.

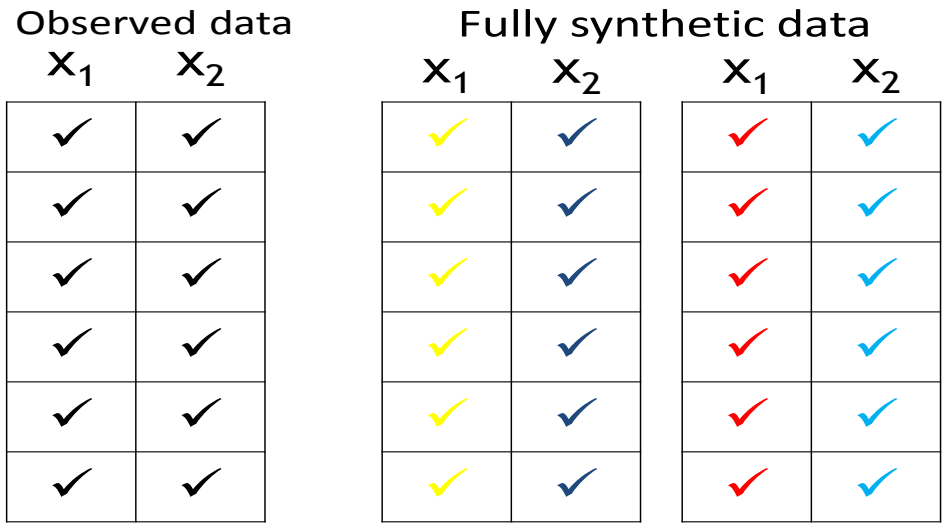


Figure 2.3: Illustration of releasing multiply imputed, fully synthetic data

We now assume that an analyst wishes to use the data to infer about some population quantity Q, this might be for example the mean of j-th variable or it could be the coefficient from a regression model for one of the variables on some other subset of variables. To do this they obtain a point estimate, q, for Q, and an estimate of its variance u. For $k = 1, \ldots, d$, let q_k and u_k be the values of q and u respectively in the synthetic data set $\mathbf{D}_{syn}^{(k)}$. Specifically, the analyst computes the following quantities,

$$\bar{q}_d = \frac{1}{d}\sum_{k=1}^{d} q_k, \qquad \overline{u}_d = \frac{1}{d}\sum_{k=1}^{d} u_k$$

$$B_d = \frac{1}{d-1}\sum_{k=1}^{d}(q_k - \bar{q}_d)^2,$$

where $\bar{q}_d$ is the mean of the m imputed data point estimates and $\overline{u}_d$ represents the average within imputation variance. The estimate B_d is known as the between imputation variance

and is the variance between the d imputed data point estimates. The total variance is given by

$$T_d = (1 + \frac{1}{d})B_d - \overline{u}_d.$$

The term $(1/d)B_d$ represents the adjustments for using only a finite number of synthetic data sets. We notice that the total variance T_d can be negative but by choosing the number of imputations d to be large enough, T_d can remain positive. See Reiter et al. (2003) for further details on the justification of the combining rules.

Confidence intervals can be constructed using a t-distribution with ν_d degrees of freedom (Reiter (2005b)) given by

$$\nu_d = (d-1)\left(1 + \frac{d\overline{u}_d}{B_d}\right)^2,$$

where this is different from the degrees of freedom expression for the multiple imputation of missing data (see Equation 2.2). The fully synthetic data approach has positive features such as protecting confidentiality of the data sets and allowing valid inferences to be made using the combining rules. However, this approach is not commonly used by statistical agencies, instead a variant version has been considered: releasing multiply imputed partially synthetic data sets which comprise original values and synthetic values. This approach, called partially synthetic data was first proposed by Little (1993b) and we will review this approach in the next section. We will focus on this approach to generate synthetic data in this thesis.

2.5 Multiple imputation for partially synthetic data

Suppose we have a fully observed $n \times p$ data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})'$, for $j = 1, \ldots, p$ is the j-th variable in $\mathbf{X}$. We denote an $n \times p$ data indicator matrix $\mathbf{Z} = (z_{1,j}, \ldots, z_{n,j})'$, where $z_{i,j} = 1$ indicates $x_{i,j}$ is selected to be replaced by synthetic value and $z_{i,j} = 0$ indicates $x_{i,j}$ is unchanged, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We can then denote the original and synthetic values of $\mathbf{X}$ by $\mathbf{X}_{ori} = \{x_{i,j} : z_{i,j} = 0\}$ and $\mathbf{X}_{syn} = \{x_{i,j} : z_{i,j} = 1\}$ respectively. Let $\mathbf{X}_{syn}$ be the synthetic values generated from the posterior predictive distribution $p(\mathbf{X}_{syn}|\mathbf{X})$. We then make d independent imputations to create d different synthetic data sets, where each of the synthetic data sets, $\mathbf{D}_{syn}^{(k)}, k = 1, \ldots, d$, is comprised of $(\mathbf{X}_{ori}, \mathbf{X}_{syn}^{(k)})$. These synthetic data sets are then released to the public.

Figure 2.4 illustrates the procedure of releasing multiply imputed, partially synthetic data. Suppose we have a fully observed data set with variables $\mathbf{x}_1$ and $\mathbf{x}_2$. The sensitive values that need to be replaced by synthetic values in $\mathbf{x}_2$ are denoted by S^*. To apply multiple imputation, we generate 2 sets of independent synthetic values from their posterior predictive distribution, i.e. $p(\mathbf{x}_2|\mathbf{x}_1)$. These independent synthetic values are denoted by 2 different colours: red and blue. Here we have imputed the data set 2 times, to generate

2 multiply imputed, partially synthetic data sets.

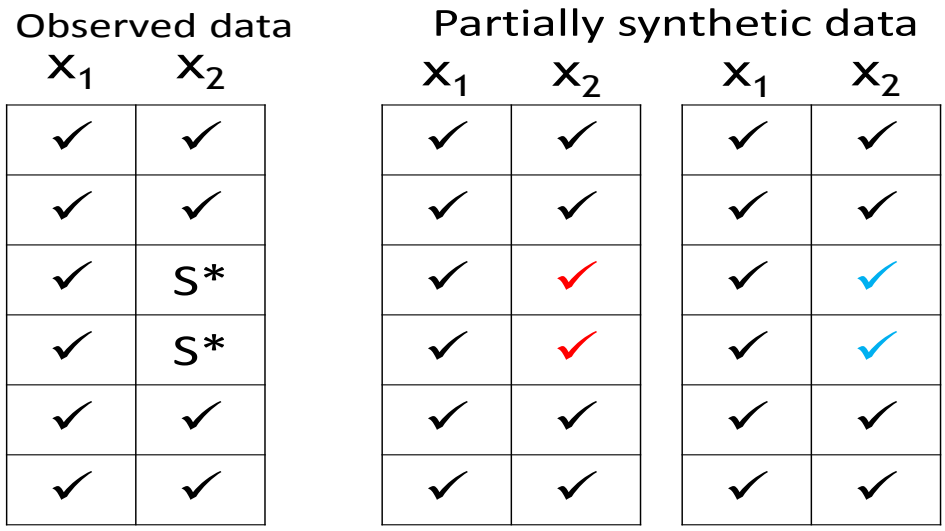


Figure 2.4: Illustration of releasing multiply imputed, partially synthetic data

We now assume that an analyst wishes to use the data to infer about some population quantity Q, this might be for example the mean of j-th variable or it could be the coefficient from a regression model for one of the variables on some other subset of variables. To do this they obtain a point estimate, q, for Q, and an estimate of its variance u. For $k = 1, \ldots, d$, let q_k and u_k be the values of q and u respectively in the synthetic data set $\mathbf{D}_{syn}^{(k)}$. Specifically, the analyst computes the following quantities,

$$\bar{q}_d = \frac{1}{d}\sum_{k=1}^{d} q_k, \qquad \overline{u}_d = \frac{1}{d}\sum_{k=1}^{d} u_k,$$
$$B_d = \frac{1}{d-1}\sum_{k=1}^{d}(q_k - \bar{q}_d)^2, \qquad T_d = \overline{u}_d + \frac{B_d}{d}. \tag{2.3}$$

Analysts can use $\bar{q}_d$ as a point estimate for Q, and T_d as an estimate of the variance for Q. B_d is known as the between imputation variance and is the variance between the d imputed data sets. The total variance T_d differs from the variance estimator for multiple imputation of missing data, $T_m = \bar{u}_m + (1+1/m)B_m$ (Equation 2.1). There is an additional B_m in Equation 2.1 due to the presence of the missing data. See Reiter (2003) for further details on the justification of the combining rules.

Confidence intervals can be constructed using a t-distribution with ν_d degrees of freedom (Reiter (2003)) given by

$$\nu_d = (d-1)\left(1 + \frac{d\overline{u}_d}{B_d}\right)^2,$$

where this is different from the degrees of freedom expression for the multiple imputation

of missing data (see Equation 2.2).

2.6 Simultaneous use of multiple imputation for missing data and statistical disclosure control

The multiple imputation framework can be extended to handle missing data and disclosure control simultaneously via a two stage imputation procedure. First, we use multiple imputation to impute the missing values in the data set, generating m multiply-imputed data sets, we then replace the selected sensitive values in each imputed data set with d synthetic values using multiple imputation; thus creating $m \times d$ multiply imputed data sets.

Figure 2.5 illustrates a simple example of handling missing data and disclosure control simultaneously via a two stage imputation. Suppose we have an observed data set with variables $\mathbf{x}_1$ and $\mathbf{x}_2$ where $\mathbf{x}_1$ is fully observed (all the boxes are ticked) and $\mathbf{x}_2$ has some missing data (denoted by question marks). We first impute the missing values with 3 sets of independent missing values drawn from their posterior predictive distribution, i.e. $p(\mathbf{x}_2|\mathbf{x}_1)$. These independent missing values are denoted by 3 different colours: red, blue and green. Now we have imputed the missing values 3 times, to generate 3 imputed data sets. Next, for each of the imputed data sets, we generate 2 sets of independent synthetic values for $\mathbf{x}_2$ from the same posterior predictive distribution, i.e. $p(\mathbf{x}_2|\mathbf{x}_1)$. These independent synthetic values are denoted by 6 different colours. Here we have generated 2 multiply imputed, partially synthetic data sets for each of the completed data sets, and we have 6 synthetic data sets in total.

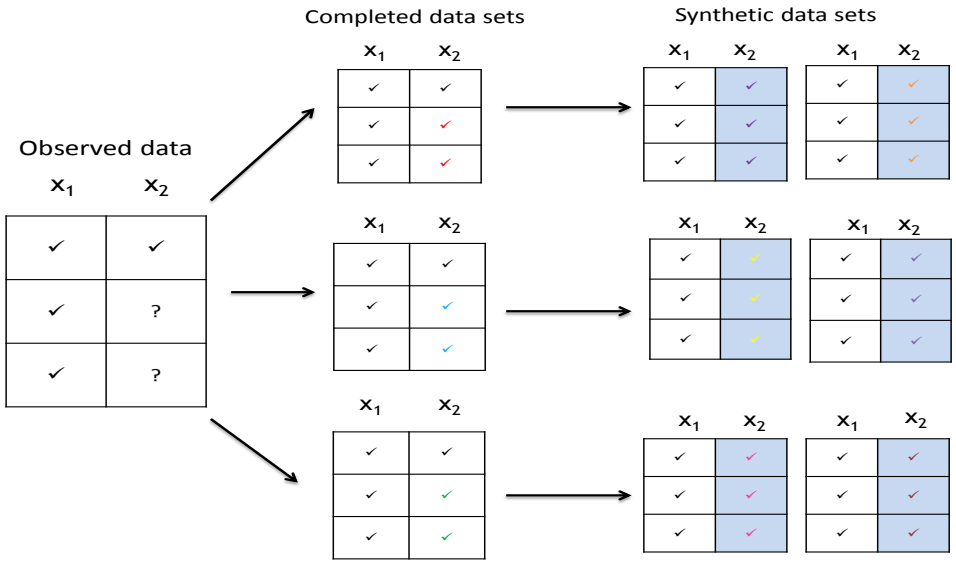


Figure 2.5: Illustration of simultaneous use of multiple imputation for missing data and disclosure control

We now assume that an analyst wishes to use the data to infer about some population

quantity Q, this might be for example the mean of j-th variable or it could be the coefficient from a regression model for one of the variables on some other subset of variables. To do this they obtain a point estimate, q, for Q, and an estimate of its variance u. For $l = 1, \ldots, m,\ k = 1, \ldots, d$, let $q_k^{(l)}$ and $u_k^{(l)}$ be the values of q and u respectively in the synthetic data set $d_k^{(l)}$. Specifically, the analyst computes the following quantities,

$$\bar{q}_M = \frac{1}{md} \sum_{l=1}^{m} \sum_{k=1}^{d} q_k^{(l)} = \frac{1}{m} \sum_{l=1}^{m} \bar{q}^{(l)},$$

$$\overline{u}_M = \frac{1}{md} \sum_{l=1}^{m} \sum_{k=1}^{d} u_k^{(l)}$$

$$\bar{b}_M = \frac{1}{m(d-1)} \sum_{l=1}^{m} \sum_{k=1}^{d} (q_k^{(l)} - \bar{q}^{(l)})^2 = \frac{1}{m} \sum_{l=1}^{m} \bar{b}^{(l)},$$

$$B_M = \frac{1}{m-1} \sum_{l=1}^{m} (\bar{q}^{(l)} - \bar{q}_M)^2,$$

where M is the total number of synthetic data sets ($m \times d$). The average of the point estimates in each group of the synthetic data sets is denoted by $\bar{q}^{(l)}$ and $\bar{q}_M$ is the average of these point estimates average across l. The average of the estimated variances of u across all the synthetic data sets is denoted by $\overline{u}_M$. The $\bar{b}_M$ is the average of the m between variances estimates, $\bar{b}^{(l)},\ l = 1, \ldots, m$ where $\bar{b}^{(l)}$ represents the sample variance of the point estimates for each group of the synthetic data sets. The B_M is the sample variance of the $\bar{q}^{(l)}$ point estimates. The total variance is given by

$$T_M = (1 + \frac{1}{m}) B_M - \frac{\bar{b}_M}{d} + \overline{u}_M.$$

Confidence intervals can be constructed using a t-distribution with ν_d degrees of freedom (Reiter (2003)) given by

$$\nu_M = \left(\frac{((1 + 1/m) B_M)^2}{(m-1) T_M^2} + \frac{(\bar{b}_M/d)^2}{m(d-1) T_M^2} \right)^{-1},$$

where this is different from the degrees of freedom expression for the multiple imputation of missing data (see Equation 2.2). For more details on this approach, please see Reiter (2004). We will now review our modelling strategy in the next chapter.

Chapter 3

Modelling methods for multiple imputation

A key challenge in the multiple imputation approach is to be able to draw imputations from a plausible distribution. To apply multiple imputation, we need to draw the missing values from their posterior predictive distributions. Suppose we have a data set $\mathbf{X}$ which can be decomposed as $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$ (see Chapter 2 for definitions). $\mathbf{X}_{obs}$ represents the observed values of the data set $\mathbf{X}$ while $\mathbf{X}_{mis}$ represents the missing values in $\mathbf{X}$. Determining an analytical expression for $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$ is not typically possible in most practical situations. It may however, be possible to model $p(\mathbf{X}|\Theta)$ where Θ represents a set of (unknown) parameters in the model. Using this model and a suitable prior distribution, $p(\Theta)$, draws from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$ can be accomplished through Markov chain Monte Carlo techniques that takes samples iteratively from $p(\mathbf{X}_{mis}|\Theta^{(t)}, \mathbf{X}_{obs})$ (to generate a completed data set $\mathbf{X}_{com}^{(t)}$) and $p(\Theta^{(t)}|\mathbf{X}_{mis}^{(t)}, \mathbf{X}_{obs})$ at each iteration t. Once samples $(\mathbf{X}_{mis}^{(t)}, \Theta^{(t)})$ have converged to their stationary distribution, these can be assumed to come from the joint distribution $p(\mathbf{X}_{mis}, \Theta|\mathbf{X}_{obs})$, and the imputed data sets, $\mathbf{X}_{com}^{(t)}$, can be assumed to have been created using draws from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$. Imputing missing values in this way is often called data augmentation (Tanner and Wong (1987)). Before we describe our modelling strategy to impute missing values through data augmentation in data sets containing variables with different measurement scales, we first review the general theories behind Markov chain and MCMC techniques.

3.1 Overview of Markov chain Monte Carlo

As mentioned before, we want to find the posterior predictive distribution of the missing data, which is denoted by $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$. This equation can be rewritten as:

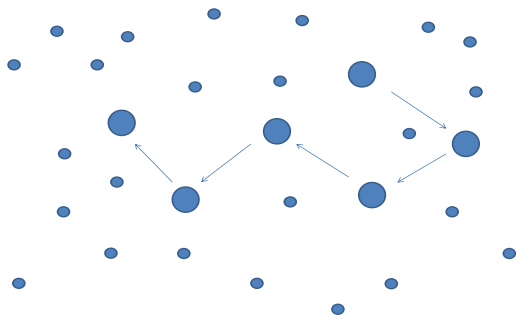
$$\begin{aligned} p(\mathbf{X}_{mis}|\mathbf{X}_{obs}) &= \int p(\mathbf{X}_{mis}, \boldsymbol{\theta}|\mathbf{X}_{obs})d\boldsymbol{\theta} \\ &= \int p(\mathbf{X}_{mis}|\boldsymbol{\theta}, \mathbf{X}_{obs})p(\boldsymbol{\theta}|\mathbf{X}_{obs})d\boldsymbol{\theta}. \end{aligned} \tag{3.1}$$

However, this integral might be multi-dimensional and difficult to compute. In the case where the missing pattern is monotone, we can sample $\mathbf{X}_{mis}$ from integral (3.1). First, we draw $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{X}_{obs})$, then conditional on this $\boldsymbol{\theta}$, we can draw $\mathbf{X}_{mis}$ from $p(\mathbf{X}_{mis}|\boldsymbol{\theta}, \mathbf{X}_{obs})$. By focusing solely on the imputed data sets generated, we are implicitly integrating out the other unknowns from their posterior distributions and are thus obtaining imputations from their posterior predictive distributions. See Gelman et al. (2004) and Khan (2005) for more details. When the missing pattern is non-monotone, Markov chain Monte Carlo (MCMC) technique can be used to sample $\mathbf{X}_{mis}$ from the integral. MCMC is a sampling based algorithm that can be used to sample draws from the distribution of interest (or target distribution). In our case, our target distribution is $p(\mathbf{X}_{mis}, \boldsymbol{\theta}|\mathbf{X}_{obs})$. MCMC will work by specifying an irreducible and aperiodic Markov chain with a unique stationary distribution $\boldsymbol{\pi}$ equal to our target distribution, the joint posterior distribution of the missing data and unknown parameters $p(\mathbf{X}_{mis}, \boldsymbol{\theta}|\mathbf{X}_{obs})$. We now review the general theories behind Markov chain and MCMC techniques.

3.1.1 Markov chain

3.1.1.1 Definition

A Markov chain can be regarded as particle X moving randomly in a state space.



Let $X_t = i$ represent the particle X being in state i at time t. A sequence of discrete

random variables $\{X_t, t \geq 0\}$ is called a Markov chain if it satisfies

$$P\{X_{s+t} = k | X_0 = l, \ldots, X_s = j\} = P\{X_{s+t} = k | X_s = j\},$$

$\forall k, l, \ldots, j$ and $\forall t \geq 0$. This is known as the Markov property where the next state of the particle X depends only on the current state but not any past states.

3.1.2 Basic properties of Markov chain

3.1.2.1 Reducibility

Let p_{ij} be the transition probability from state i to state j and let $p_{ij}^{(n)}$ denote the n-step transition probability. We define $p_{ij}^{(0)} = 1$ if $i = j$, and $p_{ij}^{(0)} = 0$ if $i \neq j$. We say that state i communicates with state j if there exists some $n \geq 0$ such that $p_{ij}^{(n)} > 0$. States i and j are said to intercommunicate if both states communicate with each other. Further, a Markov chain is defined as irreducible if all its states intercommunicate.

3.1.2.2 Periodicity

The probability that a Markov chain ever returns to state j is defined as

$$f_{jj} = \sum_{n=1}^{\infty} f_{jj}^{(n)}.$$

Let $f_{ij}^{(n)}$ be the probability that the first transition from state i to state j occurs at the n step. State j is recurrent if $f_{jj} = 1$ and is transient if $f_{jj} < 1$. The largest common divisor d for all n is called the period of state j. State j is aperiodic if $d = 1$ and state j is periodic if $d > 1$.

3.1.2.3 Recurrence and ergodicity

The mean recurrence time of a recurrent state j is given by

$$\mu_j = \sum_{n=1}^{\infty} n f_{jj}^{(n)}.$$

State j is null recurrent if $\mu_j = \infty$ and is positive recurrent if $\mu_j < \infty$. A Markov chain is called ergodic if it is irreducible and all its states are positive recurrent and aperiodic.

3.1.2.4 Stationary distribution

A distribution $\boldsymbol{\pi}$ is defined as a stationary distribution if $\boldsymbol{\pi}\mathbf{P}_{ij} = \boldsymbol{\pi}$, where $\mathbf{P}_{ij}$ is the transition probability matrix whose (i, j)-th element is p_{ij}. For any finite Markov chain, at least one stationary distribution exists and this distribution is unique if the Markov chain is irreducible.

3.1.2.5 Reversibility

A Markov chain is reversible if

$$\boldsymbol{\pi}_i \mathbf{P}_{ij} = \boldsymbol{\pi}_j \mathbf{P}_{ji} \tag{3.2}$$

for all i, j, where $\mathbf{P}_{ij}$ is the transition probability matrix. This equation is also known as detailed balance. If the stationary distribution $\boldsymbol{\pi}$ satisfies the detailed balance equation, this stationary distribution $\boldsymbol{\pi}$ is the unique stationary distribution. With a Markov chain that converges to its stationary distribution that is unique, we will have a valid MCMC sampling method. A general way to construct a Markov chain with a given stationary distribution $\boldsymbol{\pi}$ can be found in Metropolis et al. (1953).

In practice, we may need to run MCMC for quite a while to achieve convergence to our target distribution. The length of our MCMC is greatly affected by the initial values chosen to initiate the chain. Typically, the values in early iterations from our MCMC are not valid because the Markov chain has not yet stabilized to its stationary distribution. Therefore, the early iterations are discarded in order to diminish the effect of the initial values. These discarded iterations are referred to as the burn-in period.

3.1.3 Metropolis-Hastings algorithm

Now suppose we want to draw samples from our target distribution $p(\widetilde{\boldsymbol{\theta}})$ where $p(\widetilde{\boldsymbol{\theta}}) \equiv p(\mathbf{X}_{mis}, \boldsymbol{\theta}|\mathbf{X}_{obs}) \equiv \frac{f(\boldsymbol{\omega})}{K}$, where the normalizing constant K is hard to determine, and direct sampling from this distribution is also difficult. We can use a Metropolis-Hastings algorithm to sample a sequence of random samples from this distribution $p(\widetilde{\boldsymbol{\theta}})$. We choose a proposal density q, whose form is explicitly available and relatively easy to sample from. The Metropolis-Hastings algorithm then works as follows. We start with any given initial values $\boldsymbol{\omega}^{(0)}$, then at iteration t:

- Generate $\boldsymbol{\omega}^*$ from $q(\boldsymbol{\omega}^*|\boldsymbol{\omega}^{(t)})$.
- Compute the acceptance rate $\alpha(\boldsymbol{\omega}^*|\boldsymbol{\omega}^{(t)})$ given by

$$\min\left(\frac{f(\boldsymbol{\omega}^*)q(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^*)}{f(\boldsymbol{\omega}^{(t)})q(\boldsymbol{\omega}^*|\boldsymbol{\omega}^{(t)})}, 1\right).$$

- Simulate u from a uniform distribution on $[0, 1]$. If $u \leq \alpha$, then accept $\boldsymbol{\omega}^*$ and set $\boldsymbol{\omega}^{(t+1)} = \boldsymbol{\omega}^*$, else reject $\boldsymbol{\omega}^*$ and set $\boldsymbol{\omega}^{(t+1)} = \boldsymbol{\omega}^{(t)}$.

We notice that the algorithm depends on

$$\frac{f(\boldsymbol{\omega}^*)}{f(\boldsymbol{\omega}^{(t)})} \quad \text{and} \quad \frac{q(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^*)}{q(\boldsymbol{\omega}^*|\boldsymbol{\omega}^{(t)})}.$$

Thus, the normalizing constant K is no longer an issue here. A special case where the algorithm depends only on $\frac{f(\boldsymbol{\omega}^*)}{f(\boldsymbol{\omega}^{(t)})}$ is when the proposal density is symmetric, so that $q(\boldsymbol{\omega}^*|\boldsymbol{\omega}^{(t)}) = q(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^*)$. This is the original Metropolis algorithm (Metropolis et al. (1953)). Other special cases include the algorithm suggested by Hastings (1970) which suggests an independent sampler in which $q(\boldsymbol{\omega}^*|\boldsymbol{\omega}^{(t)}) = q(\boldsymbol{\omega}^*)$, so that the proposed move is independent of the current state. For other options for implementing a Metropolis-Hastings algorithm, see Chib and Greenberg (1995).

We need to show that the Metropolis-Hastings algorithm will generate a Markov chain whose stationary distribution is equivalent to our target denstity $p(\widetilde{\boldsymbol{\theta}})$. To demonstrate this, it is sufficient to show that the Metropolis-Hastings transition kernel satisfies the detailed balance equation (Equation 3.2) with $p(\widetilde{\boldsymbol{\theta}})$. First, we sample from $q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)$ and we accept the move with probability $\alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*)$, so that the transition probability kernel is given by,

$$\mathbf{P}_{\omega,\omega^*} = q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)\alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*) = q(\boldsymbol{\omega}|\boldsymbol{\omega}^*) \cdot \min\left(\frac{p(\boldsymbol{\omega})q(\boldsymbol{\omega}^*|\boldsymbol{\omega})}{p(\boldsymbol{\omega}^*)q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)}, 1\right)$$

The detailed balance equation holds if the Metropolis-Hastings kernel satisfies

$$\begin{aligned} \mathbf{P}_{\omega,\omega^*}p(\boldsymbol{\omega}) &= \mathbf{P}_{\omega^*,\omega}p(\boldsymbol{\omega}^*) \quad \text{or} \\ q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)\alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega}) &= q(\boldsymbol{\omega}^*|\boldsymbol{\omega})\alpha(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*). \end{aligned}$$

for all $\boldsymbol{\omega}, \boldsymbol{\omega}^*$. Hence we have three cases to prove here:

1. For $q(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*) > q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega})$, we have

$$\alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*) = 1 \quad \text{and} \quad \alpha(\boldsymbol{\omega}^*|\boldsymbol{\omega}) = \frac{p(\boldsymbol{\omega})q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)}{p(\boldsymbol{\omega}^*)q(\boldsymbol{\omega}^*|\boldsymbol{\omega})}.$$

Hence

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{\omega}^*,\boldsymbol{\omega}} p(\boldsymbol{\omega}^*) &= q(\boldsymbol{\omega}^*|\boldsymbol{\omega})\alpha(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*) \\
&= q(\boldsymbol{\omega}^*|\boldsymbol{\omega})\frac{q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega})}{q(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*)}p(\boldsymbol{\omega}^*) \\
&= q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega}) \\
&= q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)\alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega}) \\
&= \mathbf{P}_{\boldsymbol{\omega},\boldsymbol{\omega}^*} p(\boldsymbol{\omega}).
\end{aligned}$$

Therefore, the detailed balance equation holds.

2. For $q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega}) > q(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*)$, we have

$$\alpha(\boldsymbol{\omega}^*|\boldsymbol{\omega}) = 1 \quad \text{and} \quad \alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*) = \frac{p(\boldsymbol{\omega}^*)q(\boldsymbol{\omega}^*|\boldsymbol{\omega})}{p(\boldsymbol{\omega})q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)}.$$

Hence

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{\omega},\boldsymbol{\omega}^*} p(\boldsymbol{\omega}) &= q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)\alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega}) \\
&= q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)\frac{q(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*)}{q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega})}p(\boldsymbol{\omega}) \\
&= q(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*) \\
&= q(\boldsymbol{\omega}^*|\boldsymbol{\omega})\alpha(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*) \\
&= \mathbf{P}_{\boldsymbol{\omega}^*,\boldsymbol{\omega}} p(\boldsymbol{\omega}^*).
\end{aligned}$$

The detailed balance equation also holds in this case.

3. For $q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega}) = q(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*)$ we have $\alpha(\boldsymbol{\omega}|\boldsymbol{\omega}^*) = \alpha(\boldsymbol{\omega}^*|\boldsymbol{\omega}) = 1$ implying that

$$\mathbf{P}_{\boldsymbol{\omega},\boldsymbol{\omega}^*} p(\boldsymbol{\omega}) = q(\boldsymbol{\omega}|\boldsymbol{\omega}^*)p(\boldsymbol{\omega}) \quad \text{and} \quad \mathbf{P}_{\boldsymbol{\omega}^*,\boldsymbol{\omega}} p(\boldsymbol{\omega}^*) = q(\boldsymbol{\omega}^*|\boldsymbol{\omega})p(\boldsymbol{\omega}^*).$$

This shows that the detailed balance equation holds. Therefore the stationary distribution from this kernel corresponds to draws from our target distribution.

3.1.4 Gibbs sampler

If the full conditional distributions from all the unknowns are available in closed form, we can use these full conditionals as our proposal distributions in the Metropolis-Hastings algorithm. The proposed values are always accepted. This is known as a Gibbs sampler (Gelfand and Smith (1990)). For example, suppose we want to draw values from a joint distribution $p(\widetilde{\boldsymbol{\theta}}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ that is hard to compute. Then a Gibbs sampler can be used whereby we draw values from a series of conditional distributions. The Gibbs

sampler in this example works as follows:
At iteration t, we draw

$$\begin{aligned}
\boldsymbol{\theta}_1^{(t+1)} &\sim p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}) \\
\boldsymbol{\theta}_2^{(t+1)} &\sim p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}) \\
\boldsymbol{\theta}_3^{(t+1)} &\sim p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_4^{(t)}, \ldots \boldsymbol{\theta}_k^{(t)}) \\
&\vdots \\
\boldsymbol{\theta}_k^{(t+1)} &\sim p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(t+1)}).
\end{aligned}$$

Such conditional distribution, $p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \ldots, \boldsymbol{\theta}_k)$ for $j = 1, \ldots, k$ are often easier to sample from than the joint distribution. The sequence of iterates $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)})$ will converge to a draw from the joint distribution $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. The Gibbs sampler is a valid MCMC method to construct the stationary distribution as the target distribution (Tierney (1994)). The ideas of the Gibbs sampler and Metropolis-Hastings will be revisited in the next section where we describe our modelling strategies.

3.2 Model for continuous, binary and ordinal variables

Before we use multiple imputation to impute the missing values, we need to derive the posterior predictive distribution of the missing data. In this section we will describe the modelling approach used to determine the posterior predictive distributions in data sets with continuous, binary and ordinal variables.

We now present the full conditional distributions in the Metropolis within Gibbs sampler required to generate imputed datasets. We model the joint distribution $p(\mathbf{X}|\Theta)$ using a sequence of conditional regression models:

$$\begin{aligned}
p(\mathbf{X}|\Theta) &= p(\mathbf{x}_1|\boldsymbol{\theta}^{(1)}) \prod_{k=2}^{p} p(\mathbf{x}_k|\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \boldsymbol{\theta}^{(k)}) \\
&= \prod_{i=1}^{n} \left\{ p(x_{i,1}|\boldsymbol{\theta}^{(1)}) \prod_{k=2}^{p} p(x_{i,k}|x_{i,1}, \ldots, x_{i,k-1}, \boldsymbol{\theta}^{(k)}) \right\}, \qquad (3.3)
\end{aligned}$$

where $\boldsymbol{\theta}^{(j)}$ represents the vector of parameters in the regression model for $\mathbf{x}_j$ and $\Theta = \{\boldsymbol{\theta}^{(j)},\ j = 1, \ldots, p\}$. Such a decomposition has been used by Ibrahim et al. (2005) and Ibrahim et al. (1999) in imputing missing values; however, these approaches do not draw imputations from their exact posterior predictive distributions in the case where there are non-monotone patterns of missing data. We propose a modelling strategy that allows imputations to be drawn from their posterior predictive distributions through Markov chain Monte Carlo. As we are using a proper factorisation of the joint distribution to impute missing values, we hereby refer to this modelling strategy for imputation as the

factored regression model (FRM). We assume here that all variables have some missing values, if there were some variables that were fully observed the modelling framework would remain the same, and we would condition on these fully observed variables in every regression model.

We first consider the case when $\mathbf{x}_k$ is either continuous, binary, or ordinal. If $\mathbf{x}_k$ is continuous, we use a normal linear regression model with $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ as covariates in the model (for now assuming $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ are continuous), so that

$$p(x_{i,k}|x_{i,1}, \ldots, x_{i,k-1}, \boldsymbol{\theta}^{(k)}) = \mathrm{N}(\beta_0^{(k)} + \beta_1^{(k)} x_{i,1} + \cdots + \beta_{k-1}^{(k)} x_{i,k-1}, \phi_k^{-1}),$$

where $\boldsymbol{\beta}^{(k)} = (\beta_0^{(k)}, \beta_1^{(k)}, \ldots, \beta_{k-1}^{(k)})$ are the regression coefficients of the model, ϕ_k is the residual precision, and $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \phi_k)$.

Secondly, if $\mathbf{x}_k$ is binary, i.e. $x_{i,k} \in \{0, 1\}\,,\ i = 1, \ldots, n$, we introduce a latent variable, denoted by $\mathbf{x}_k^*$ where $\mathbf{x}_k^* = (x_{1,k}^*, \ldots, x_{n,k}^*)'$ and model the conditional distribution of $\mathbf{x}_k$ through the data augmentation approach suggested by Albert and Chib (1993). We thus model $p(x_{i,k}|x_{i,1}, \ldots, x_{i,k-1}, \boldsymbol{\theta}^{(k)})$ (again for now assuming $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ are continuous) by

$$\begin{aligned} &x_{i,k} = I(x_{i,k}^* > 0), \text{ where} \\ &p(x_{i,k}^*|x_{i,1}, \ldots, x_{i,k-1}, \boldsymbol{\theta}^{(k)}) = \mathrm{N}(\beta_0^{(k)} + \beta_1^{(k)} x_{i,1} + \cdots + \beta_{k-1}^{(k)} x_{i,k-1}, 1), \end{aligned}$$

where $I(\cdot)$ is the indicator function, and $\boldsymbol{\theta}^{(k)} = \boldsymbol{\beta}^{(k)} = (\beta_0^{(k)}, \beta_1^{(k)}, \ldots, \beta_{k-1}^{(k)})$.

Thirdly if $\mathbf{x}_k$ is an ordinal variable, i.e. $x_{i,k} \in \{1, \ldots, J_k\}\,,\ i = 1, \ldots, n$ with J_k be the number of levels that the observations in $\mathbf{x}_k$ can take, we can extend the data augmentation representation above to model the conditional distribution of $\mathbf{x}_k$. The distribution $p(x_{i,k}|x_{i,1}, \ldots, x_{i,k-1}, \boldsymbol{\theta}^{(k)})$ (again for now assuming $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ are continuous) is then given by

$$\begin{aligned} &x_{i,k} = j^k \ I(\gamma_{j^k-1}^{(k)} < x_{i,k}^* < \gamma_{j^k}^{(k)}), \text{ where} \\ &p(x_{i,k}^*|x_{i,1}, \ldots, x_{i,k-1}, \boldsymbol{\theta}^{(k)}) = \mathrm{N}(\beta_0^{(k)} + \beta_1^{(k)} x_{i,1} + \cdots + \beta_{k-1}^{(k)} x_{i,k-1}, 1), \end{aligned}$$

with $j^k \in \{1, \ldots, J_k\}$, $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)})$, where $\boldsymbol{\beta}^{(k)} = (\beta_0^{(k)}, \ldots, \beta_{k-1}^{(k)})$ as before, and $\boldsymbol{\gamma}^{(k)} = \left\{\gamma_{j^k}^{(k)} : j^k \in \{1, \ldots, J_k\}\right\}$ are threshold parameters, with $\gamma_0^{(k)} = -\infty$, $\gamma_1^{(k)} = 0$ and $\gamma_{J_k}^{(k)} = \infty$.

Now, if $x_{i,q},\ q \in \{1, \ldots, k-1\}$ is not continuous we replace $x_{i,q}$ with $x_{i,q}^*$ in the model for $p(x_{i,k}|x_{i,1}, \ldots, x_{i,k-1}, \boldsymbol{\theta}^{(k)})$. Thus, in our modelling strategy, we place a multivariate normal model on the joint distribution of the continuous variables and latent variables, which were introduced through the data augmentation representation proposed by Albert and Chib (1993). Benefits of latent variable in modelling data sets with these types of variables have been noted by De LEON and Carrièrgre (2007).

To complete a Bayesian specification we place prior distributions on all parameters. We first place independent Jeffreys priors on all regression coefficients and residual preci-

sions for the regression models of the continuous variables, i.e.

$$p(\beta_q^{(k)}) \propto 1 \quad \text{for} \quad q = 0, \ldots, k-1, \quad k = 1, \ldots, p,$$

and

$$p(\phi_k) \propto \frac{1}{\phi_k}, \quad k = 1, \ldots, p.$$

We then place independent Jeffreys priors on regression coefficients for the regression models of the binary and ordinal variables, i.e.

$$p(\beta_q^{(k)}) \propto 1 \quad \text{for} \quad q = 0, \ldots, k-1, \quad k = 1, \ldots, p,$$

and we also place an improper uniform prior on $\boldsymbol{\gamma}^{(k)}$ i.e.

$$p(\boldsymbol{\gamma}^{(k)}) \propto I(\boldsymbol{\gamma}^{(k)} \in \Omega^{(k)}),$$

where $\Omega^{(k)} = \left\{ \gamma_{j^k}^{(k)} : \gamma_0^{(k)} = -\infty < \gamma_1^{(k)} = 0 < \gamma_2^{(k)} < \ldots < \gamma_{J_{k-1}}^{(k)} < \gamma_{J_k}^{(k)} = \infty \right\}$.

With this modelling approach and choice of prior distributions, the full conditional distributions of all unknowns: missing/latent values and model parameters, are available in closed form. Draws from the joint distribution of all unknowns can then be taken using a Gibbs sampler; in particular, the missing values in the data set are imputed from their full conditional distributions on a variable by variable basis. Given the models and prior specifications we can present the full conditional distribution required to impute missing values in a Metropolis within Gibbs sampler. Following Schafer (1997) we present this a data augmentation scheme and so first present the "I-steps" followed by the "P-steps". In the "I-steps", conditional on parameter values, first impute missing continuous values $x_{i,q}^*$ or latent values $x_{i,q}^*$ with missing $x_{i,q}$, from a normal distribution with mean $\widetilde{\mu}_{i,k}$ and variance $\widetilde{\phi}_k^{-1}$ where

$$\widetilde{\mu}_{i,k} = \widetilde{\phi}_k^{-1} \left\{ \frac{\mu_{i,k}}{\phi_k^{-1}} + \sum_{s=k+1}^{p} \frac{\beta_k^{(s)}}{\phi_s^{-1}} \left[x_{i,s}^* - (\mu_{i,s} - \beta_k^{(s)} x_{i,k}^*) \right] \right\},$$

and

$$\widetilde{\phi}_k^{-1} = \left(\frac{1}{\phi_k^{-1}} + \sum_{s=k+1}^{p} \frac{(\beta_k^{(s)})^2}{\phi_s^{-1}} \right)^{-1},$$

where $\mu_{i,k} = \beta_0^{(k)} + \sum_{j=1}^{k-1} \beta_j^{(k)} x_{i,j}^*$ and $\mu_{i,s} = \beta_0^{(s)} + \sum_{j=1}^{s-1} \beta_j^{(s)} x_{i,j}^*$.

Once we have imputed a value for $x_{i,k}^*$, we define a function $g(x_{i,k}^*)$ to map each $x_{i,k}^*$

back to its original measurement scale, where

$$g(x^*_{i,k}) = \begin{cases} I(x^*_{i,k} > 0) & \text{if } x^*_{i,k} \text{ is binary,} \\ j^k \; I(\gamma^{(k)}_{j^k-1} < x^*_{i,k} < \gamma^{(k)}_{j^k}) & \text{if } x^*_{i,k} \text{ is ordinal, } j^k \in \{1, \ldots, J_k\} \\ x^*_{i,k} & \text{if } x^*_{i,k} \text{ is continuous.} \end{cases}$$

Denote an imputed value for $x_{i,q}$ by $x^{(t)}_{i,q}$ at iteration t, and an imputed data set at iteration t by $\mathbf{X}^{(t)}_{com}$. In the "P-steps", conditional on an imputed data set, first sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \phi_q),\ q \in \{1, \ldots, p\}$ when $\mathbf{x}^*_q$ is continuous from the joint posterior distribution of
$\boldsymbol{\beta}^{(q)} = (\beta^{(q)}_0, \beta^{(q),2}_1, \ldots, \beta^{(q),L_1}_1, \ldots, \beta^{(q),2}_k, \ldots, \beta^{(q),L_k}_k, \beta^{(q)}_{k+1}, \ldots, \beta^{(q)}_{q-1})$ and ϕ_q given by

$$p(\boldsymbol{\beta}^{(q)}, \phi_q | \mathbf{X}^{(t)}_{com}) = p(\boldsymbol{\beta}^{(q)} | \phi_q, \mathbf{X}^{(t)}_{com}) p(\phi_q | \mathbf{X}^{(t)}_{com}),$$

where

$$p(\phi_q | \mathbf{X}^{(t)}_{com}) = \text{Gamma}\left(\frac{n-q}{2}, \frac{RSS}{2}\right),$$

and

$$p(\boldsymbol{\beta}^{(q)} | \phi_q, \widetilde{\mathbf{X}}_q, \mathbf{x}^*_q) = \text{N}(\hat{\boldsymbol{\beta}}^{(q)}, (\widetilde{\mathbf{X}}'_q \widetilde{\mathbf{X}}_q \phi_q)^{-1}),$$

with

$$\hat{\boldsymbol{\beta}}^{(q)} = (\widetilde{\mathbf{X}}'_q \widetilde{\mathbf{X}}_q)^{-1} \widetilde{\mathbf{X}}'_q \mathbf{x}^*_q,$$

and

$$RSS = (\mathbf{x}^*_q - \widetilde{\mathbf{X}}_q \hat{\boldsymbol{\beta}}^{(q)})'(\mathbf{x}^*_q - \widetilde{\mathbf{X}}_q \hat{\boldsymbol{\beta}}^{(q)}),$$

and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where
$\widetilde{\mathbf{x}}_{i,q} = (1, I(x^*_{i,1} = 2), \ldots, I(x^*_{i,1} = L_1), \ldots, I(x^*_{i,k} = 2), \ldots, I(x^*_{i,k} = L_k), x^*_{i,k+1}, \ldots, x^*_{i,q-1})$
for $i = 1, \ldots, n$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}),\ q \in \{k+1, \ldots, p\}$ when $\mathbf{x}_q$ is binary from

$$p(\boldsymbol{\beta}^{(q)} | \widetilde{\mathbf{X}}_q, \mathbf{x}^*_q) = \text{N}(\hat{\boldsymbol{\beta}}^{(q)}, (\widetilde{\mathbf{X}}'_q \widetilde{\mathbf{X}}_q)^{-1}),$$

where $\hat{\boldsymbol{\beta}}^{(q)}$ and $\widetilde{\mathbf{X}}_q$ are as above.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\gamma}^{(q)}),\ q \in \{k+1, \ldots, p\}$ when $\mathbf{x}_q$ is ordinal. We sample $\boldsymbol{\beta}^{(q)}$ from

$$p(\boldsymbol{\beta}^{(q)} | \widetilde{\mathbf{X}}_q, \mathbf{x}^*_q) = \text{N}(\hat{\boldsymbol{\beta}}^{(q)}, (\widetilde{\mathbf{X}}'_q \widetilde{\mathbf{X}}_q)^{-1}),$$

where $\hat{\boldsymbol{\beta}}^{(q)}$ and $\widetilde{\mathbf{X}}_q$ are as above. We then update the threshold values. The full conditional distribution of the threshold values $\gamma_{j^q}^{(q)}$ for $j^q \in \{1, \ldots, J_q\}$ is uniformly distributed on the interval

$$\left[\max\left\{\max\left\{x_{i,q}^* : x_{i,q} = j^q\right\}, \gamma_{j^q-1}^{(q)}\right\}, \min\left\{\min\left\{x_{i,q}^* : x_{i,q} = j^q + 1\right\}, \gamma_{j^q+1}^{(q)}\right\}\right],$$

where $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_{J_q}^{(q)} = \infty$, for $j^q \in \{1, \ldots, J_q\}$.

In this way we create imputed data sets within the Gibbs sampler that samples from the joint posterior distribution of all unknowns. By only focusing on the imputed data sets, we are implicitly integrating over the other unknowns in the joint distribution and thus creating imputed data sets from draws of the posterior predictive distribution.

3.2.1 Incorporating nominal variables

Of course in practice, data sets may also contain nominal variables, i.e. variables that have no implicit ordering to their levels. If the variable only has two levels, then the variable can be treated as a binary variable and dealt with using the framework described before; we assume here that a nominal variable implies it has more than two levels.

Suppose in our data set, $\mathbf{x}_k,\ k \in \{1, \ldots, p\}$ is a nominal variable. Then we can use the same decomposition as proposed in Equation 3.3, with a a multinomial logistic regression model for $\mathbf{x}_k$ including covariates $\mathbf{x}_1^*, \ldots, \mathbf{x}_{k-1}^*$ in the model. Specifically, for any nominal observation $x_{i,k} \in \{1, \ldots, L_k\}$, where L_k is the number of levels that the observations in $\mathbf{x}_k$ can take, we model the distribution of $x_{i,k}$ as

$$\begin{aligned} p(x_{i,k} = j^k | x_{i,1}^*, \ldots, x_{i,k-1}^*, \boldsymbol{\theta}_{j^k}^{(k)}) &= \pi_{i,j^k}^{(k)} \\ &= \frac{\exp(\beta_{0,j^k}^{(k)} + \beta_{1,j^k}^{(k)} x_{i,1}^* + \cdots + \beta_{k-1,j^k}^{(k)} x_{i,k-1}^*)}{\sum_{s=1}^{L_k} \exp(\beta_{0,s}^{(k)} + \beta_{1,s}^{(k)} x_{i,1}^* + \cdots + \beta_{k-1,s}^{(k)} x_{i,k-1}^*)}, \end{aligned} \tag{3.4}$$

where $\boldsymbol{\theta}_{j^k}^{(k)} = (\beta_{0,j^k}^{(k)}, \ldots, \beta_{k-1,j^k}^{(k)})$ is the vector of parameters in the regression model for $x_{i,k} = j^k$, $j^k = 1, \ldots, L_k$. We set $\boldsymbol{\theta}_1^{(k)}$ equal to $\mathbf{0}$ for identifiability and the probability that $x_{i,k} = 1$ is thus given by

$$\begin{aligned} p(x_{i,k} = 1 | x_{i,1}^*, \ldots, x_{i,k-1}^*, \boldsymbol{\theta}_1^{(k)}) &= \pi_{i,1}^{(k)} \\ &= \frac{1}{1 + \sum_{s=2}^{L_k} \exp(\beta_{0,s}^{(k)} + \beta_{1,s}^{(k)} x_{i,1}^* + \cdots + \beta_{k-1,s}^{(k)} x_{i,k-1}^*)}. \end{aligned}$$

We assumed in Equation 3.4 that $x_{i,1}^*, \ldots, x_{i,k-1}^*$ were not nominal; if we conditioned on a nominal variable $x_{i,q}^* = j^q$ (for notational convenience we assume if $x_{i,q}$ is nominal, taking values $j^q \in \{1, \ldots, L_q\}$ then $x_{i,q} = x_{i,q}^*$) for $1 \leq q \leq k-1$, then we would replace

$\beta^{(k)}_{q,j^k} x^*_{i,q}$ with $\sum_{j^q=2}^{L_q} \beta^{(k),j^q}_{q,j^k} I(x^*_{i,q} = j^q)$ in Equation 3.4. Note that when $k = 1$ there are no covariates apart from the intercept in the regression model, and so we can write

$$p(x^*_{i,1} = j^1 | \boldsymbol{\theta}^{(1)}_{j^1}) = \pi^{(1)}_{j^1},$$

where $\pi^{(1)}_{j^1} = \frac{\exp(\beta^{(1)}_{0,j^1})}{1+\sum_{s=2}^{L_1} \exp(\beta^{(1)}_{0,s})}$, $j^1 = 1, \ldots, L_1$, and so the distribution of $x^*_{i,1}$ can be written using the regular multinomial model.

To complete the Bayesian specification we place a diffuse multivariate normal prior distribution on $\boldsymbol{\theta}^{(k)} (k > 1)$, i.e.

$$p(\boldsymbol{\theta}^{(k)}) = \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{3.5}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with large entries. For $k = 1$ we place a Dirichlet prior on $\boldsymbol{\pi}^{(1)} = (\pi^{(1)}_1, \ldots, \pi^{(1)}_{L_1})$, i.e.

$$p(\boldsymbol{\pi}^{(1)}) \propto \text{Dirichlet}(\alpha_1, \ldots, \alpha_{L_1}),$$

we set all $\alpha_{j^1} = 0.5$, $j^1 = 1, \ldots, L_1$ (corresponding to the Jeffreys prior), another common choice could be to set all $\alpha_{j^1} = 0$.

With this extension to our modelling framework, parameters in other conditional regression models (where the response variable is not nominal) still retain their original full conditional distributions and can be sampled accordingly. However, a missing continuous covariate value or latent value $x^*_{i,q}$, $q \in \{1, \ldots, k-1\}$, will not have a full conditional distribution available in closed form. This is due to their presence in Equation 3.4. In order to sample values of $x^*_{i,q}$ from their full conditional distribution, a Metropolis sampler would need to be specified. To avoid this we decompose the joint distribution so that if $\mathbf{x}_q$ is nominal then $\mathbf{x}_1, \ldots, \mathbf{x}_{q-1}$ are also nominal. This means that in any multinomial regression model, all the covariates will also be nominal. We assume there are k nominal variables in our dataset, and in this way a missing $x^*_{i,q}$, $q \in \{1, \ldots, k\}$ can be imputed from its full conditional distribution in closed form, which is given using Bayes rule by

$$\begin{aligned} &p(x^*_{i,q} = j^q | x^*_{i,1} = j^1, \ldots, x^*_{i,q-1} = j^{q-1}, x^*_{i,q+1} = j^{q+1}, \ldots, x^*_{i,k} = j^k, x^*_{i,k+1}, \ldots, x^*_{i,p}, \Theta), \\ &= \tilde{\pi}^{(q)}_{i,j^q} \end{aligned}$$

where $j^q \in \{1, \ldots, L_q\}$, and

$$\tilde{\pi}^{(q)}_{i,j^q} = \frac{\displaystyle\prod_{b=q}^{k} \pi^{(b)}_{i,j^b} \prod_{b=k+1}^{p} \exp\left(\frac{-\phi_b}{2} (\tilde{x}^*_{i,b} - \beta^{(b),j^q}_q I(x^*_{i,q} = j^q) - \sum_{t=k+1}^{(b-1)} \beta^{(b)}_t x^*_{i,t})^2 \right)}{\displaystyle\sum_{u=1}^{L_q} \left\{ \pi^{(q)}_{i,u} \prod_{b=q+1}^{k} \pi^{(b)}_{i,j^b} \prod_{b=k+1}^{p} \exp\left(\frac{-\phi_b}{2} (\tilde{x}^*_{i,b} - \beta^{(b),j^q}_q I(x^*_{i,q} = u) - \sum_{t=k+1}^{(b-1)} \beta^{(b)}_t x^*_{i,t})^2 \right) \right\}},$$

where

$$\tilde{x}^*_{i,b} = x^*_{i,b} - \beta_0^{(b)} - \sum_{s=1}^{q-1}\sum_{j^s=2}^{L_s} \beta_s^{(b),j^s} I(x^*_{i,s} = j^s) - \sum_{s=q+1}^{k}\sum_{j^s=2}^{L_s} \beta_s^{(b),j^s} I(x^*_{i,s} = j^s)$$

Missing continuous covariate values or latent variable values can now be imputed from their full conditional distributions in closed form as before. This is because they do not appear as covariates in Equation 3.4 anymore.

We now sample the parameters, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}_1^{(q)}, \ldots, \boldsymbol{\theta}_{L_q}^{(q)})$ with $\boldsymbol{\theta}_{j^q}^{(q)} = (\beta_{0,j^q}^{(q)}, \beta_{1,j^q}^{(q),2}, \ldots, \beta_{1,j^q}^{(q),L_2}, \ldots, \beta_{q-1,j^q}^{(q),2}, \ldots, \beta_{q-1,j^q}^{(q),L_{q-1}})$, for $j^q = 1, \ldots, L_q$, where we set $\boldsymbol{\theta}_1^{(q)} = \mathbf{0}$ for identifiability. Rather than sample $\boldsymbol{\theta}^{(1)}$ we instead sample $\boldsymbol{\pi}^{(1)}$ from its full conditional distribution; this is because we have used the conjugate Dirichlet prior for $\boldsymbol{\pi}^{(1)}$ and so the posterior distribution for $\boldsymbol{\pi}^{(1)}$ is available in closed form, i.e.

$$\boldsymbol{\alpha} = \left(\alpha_1 + \sum_{i=1}^{n} I(x_{i,1} = 1),\ \alpha_2 + \sum_{i=1}^{n} I(x_{i,1} = 2), \ldots, \alpha_{L_1} + \sum_{i=1}^{n} I(x_{i,1} = L_1)\right)'.$$

where $(\alpha_1, \ldots, \alpha_{L_1})$ are the parameters from the Dirichlet prior distribution $\boldsymbol{\pi}^{(1)}$. However, for $q = 2, \ldots, k$, the full conditional distributions for $\boldsymbol{\theta}^{(q)}$ are not available in closed form, and so we use a Metropolis-Hastings proposal to sample from these distributions. The full conditional distribution for $\boldsymbol{\theta}^{(q)}$ is proportional to

$$\prod_{i=1}^{n}\prod_{w=1}^{L_q}\left(\frac{\exp\left(\beta_{0,w}^{(q)} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,w}^{(q),j^b} I(x^*_{i,b} = j^b)\right)}{\sum_{u=1}^{L_q}\exp\left(\beta_{0,u}^{(q)} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,u}^{(q),j^b} I(x^*_{i,b} = j^b)\right)}\right)^{I(x_{i,q}=w)} \pi(\boldsymbol{\theta}^{(q)}),$$

where $I(\cdot)$ is the indicator function and $\pi(\boldsymbol{\theta}^{(q)})$ is the diffuse prior given by Equation 3.5.

There is no closed form expression for this full conditional distribution, and so to sample from this distribution we use a Metropolis-Hastings sampler; thus we develop a Metropolis within Gibbs sampler to impute missing values. One approach would be to specify an independent multivariate normal proposal distribution for $\boldsymbol{\theta}^{(q)}$, $\mathrm{N}(\boldsymbol{\mu}^{(cc)}, \boldsymbol{\Sigma}^{(cc)})$, where the $\boldsymbol{\mu}^{(cc)}$ and $\boldsymbol{\Sigma}^{(cc)}$ are determined using the complete case likelihood and large sample normal approximations to the posterior (Gelman et al. (2004)). This would allow the proposal distribution to be fixed over iterations of the Gibbs sampler. Specifically, the

complete case log-likelihood for $\boldsymbol{\theta}^{(q)}$, $l^{cc}(\boldsymbol{\theta}^{(q)};\mathbf{X}^{(cc)})$ is given by

$$l^{cc}(\boldsymbol{\theta}^{(q)};\mathbf{X}^{(cc)}) =$$
$$\sum_{w=1}^{L_q}\sum_{i=1}^{n}\ln\left(\frac{\exp\left(\beta_{0,w}^{(q)}+\sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,w}^{(q),j^b}I(x_{i,b}^{*}=j^b)\right)}{\sum_{u=1}^{L_q}\exp\left(\beta_{0,u}^{(q)}+\sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,u}^{(q),j^b}I(x_{i,b}^{*}=j^b)\right)}\right)^{I(x_{i,q}=w)} I(m_{i,q}=0),$$

where $I(\cdot)$ is the indicator function. From this we can use the value of $\boldsymbol{\theta}^{(q)}$, that maximises $l^{cc}(\boldsymbol{\theta}^{(q)};\mathbf{X}^{(cc)})$ for $\boldsymbol{\mu}^{(cc)}$, and $-E[\frac{\delta^2}{\delta\beta_{j,w}^{(q)}\delta\beta_{\tilde{j},\tilde{w}}^{(q)}}l^{(cc)}(\boldsymbol{\theta}^{(q)};\mathbf{X}^{(cc)})]^{-1}_{\boldsymbol{\theta}^{(q)}=\boldsymbol{\mu}^{(cc)}}$ for $\boldsymbol{\Sigma}^{(cc)}$, i.e. the inverse of the Fisher information matrix for the complete case log-likelihood evaluated at its maximum likelihood estimate. In practice however, we found that this proposal distribution performed poorly in proposing plausible values. We thus consider an alternative proposal distribution based on the imputed data log-likelihood at each iteration t, $l(\boldsymbol{\theta}^{(q)};\mathbf{X}_{com}^{(t)})$. We consider a proposal distribution for $\boldsymbol{\theta}^{(q)}$ from $\mathrm{N}(\boldsymbol{\mu}^{(t)},\boldsymbol{\Sigma}^{(t)})$ which based on the imputed data log-likelihood at each iteration t, $l(\boldsymbol{\theta}^{(q)};\mathbf{X}_{com}^{(t)})$, and can be expressed as

$$l(\boldsymbol{\theta}^{(q)};\mathbf{X}_{com}^{(t)}) =$$
$$\sum_{w=1}^{L_q}\sum_{i=1}^{n}\ln\left(\frac{\exp\left(\beta_{0,w}^{(q)}+\sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,w}^{(q),j^b}f(x_{i,b}^{(t)},x_{i,b},m_{i,b})\right)}{\sum_{u=1}^{L_q}\exp\left(\beta_{0,u}^{(q)}+\sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,u}^{(q),j^b}f(x_{i,b}^{(t)},x_{i,b},m_{i,b})\right)}\right)^{I(x_{i,q}=w)},$$

where $f(x_{i,b}^{(t)},x_{i,b},m_{i,b}) = (I(x_{i,b}^{(t)}=j^b)m_{i,b}+I(x_{i,b}=j^b)(1-m_{i,b}))$, with $m_{i,b}$ is a missing data indicator with $m_{i,b}=1$ indicating $x_{i,b}$ is missing and $m_{i,b}=0$ indicating $x_{i,b}$ is observed. We generate proposals for $\boldsymbol{\theta}^{(q)}$ from a normal distribution $\mathrm{N}(\boldsymbol{\mu}^{(t)},\boldsymbol{\Sigma}^{(t)})$, and $\boldsymbol{\mu}^{(t)}$ is the value that maximises $l(\boldsymbol{\theta}^{(q)};\mathbf{X}_{com}^{(t)})$ and $\boldsymbol{\Sigma}^{(t)}$ is given by $-E[\frac{\delta^2}{\delta\beta_{j,w}^{(q)}\delta\beta_{\tilde{j},\tilde{w}}^{(q)}}l(\boldsymbol{\theta}^{(q)};\mathbf{X}_{com}^{(t)})]^{-1}_{\boldsymbol{\theta}^{(q)}=\boldsymbol{\mu}^{(t)}}$.

In the simulation studies and applications to the breast-feeding study in Chapter 4, values proposed from a $\mathrm{N}(\boldsymbol{\mu}^{(t)},\boldsymbol{\Sigma}^{(t)})$ distribution were accepted approximately 90% of the time, while proposal values drawn from a $\mathrm{N}(\boldsymbol{\mu}^{(cc)},\boldsymbol{\Sigma}^{(cc)})$ distribution were accepted only approximately 5% of the time. Thus the proposal distribution based on the imputed data likelihood greatly improves the efficiency of the Metropolis within Gibbs sampler, even with the additional computational burden of having to recalculate the proposal density at each iteration. It may seem surprising at first that even when the sample sizes are relatively large, a normal proposal based on the complete case likelihood is inefficient. This is because when data are MAR, parameter estimates based on the complete case likelihood are not closely matched to estimates that would have been obtained from the complete data likelihood.

Thus, in complete data sets that contain categorical and continuous variables we have

proposed a modelling strategy that allows a Metropolis within Gibbs sampler to sample all unknowns from their joint posterior distribution, and hence creates imputed data sets by sampling from the posterior predictive distribution $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$. We note that a related imputation modelling strategy has been developed by Goldstein et al. (2009) in the context of multilevel models, that uses latent normal random variables. Our approach differs fundamentally with Goldstein et al. (2009) in the modelling of nominal variables, and thus also in the proposed method for posterior computations here. We now briefly describe an alternative strategy that imputes missing values in these data sets based on chained equations.

3.2.2 Multivariate imputation via chained equations (MICE)

Multivariate imputation via chained equations (MICE) (Van Buuren (2011), Van Buuren et al. (1999)) is a commonly used approach to impute missing values. The method is also known as sequential regression multiple imputation (SRMI) (Raghunathan et al. (2001)), regression switching (Van Buuren et al. (1999)) and partially incompatible MCMC (Rubin (2003)).

Like the approach described above imputations are performed on a variable by variable iterative basis. Suppose $\mathbf{D}$ represents our set of fully observed covariates then in the first iteration ($t = 1$), missing values in $\mathbf{x}_1$ are imputed using the distribution $p_1(\mathbf{x}_1|\mathbf{D})$, denote the imputed variable $\mathbf{x}_1^{(t)}$. Then missing values in $\mathbf{x}_2$ are imputed using $p_2(\mathbf{x}_2|\mathbf{x}_1^{(t)}, \mathbf{D})$ to create an imputed variable $\mathbf{x}_2^{(t)}$, this continues sequentially until an imputed variable $\mathbf{x}_k^{(t)}$ is created from $p_k(\mathbf{x}_k|\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{k-1}^{(t)}, \mathbf{D})$. Flat prior distributions are used on all model parameters. The form of the model $p_k(\cdot)$ depends on the measurement scale of $\mathbf{x}_k$ as with FRM, so for example if $\mathbf{x}_k$ is continuous $p_k(\cdot)$ will take the form of a normal linear regression, while if $\mathbf{x}_k$ is binary then $p_k(\cdot)$ might take the form of a logistic regression. In subsequent iterations ($t > 1$) the method cycles through a sequence of conditional regressions $g_k(\mathbf{x}_k|\mathbf{D}, \mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{k-1}^{(t)}, \mathbf{x}_{k+1}^{(t-1)}, \ldots, \mathbf{x}_p^{(t-1)})$ to impute missing values in $\mathbf{x}_k$; again, flat prior distributions are used for all model parameters and the form of $g_k(\cdot)$ depends on the measurement scale of $\mathbf{x}_k$. Once the method has cycled through a sufficient number of iterations the imputed values are used to create an imputed data set; typically the number of iterations used is fairly modest, often $t = 5$ or 10. The method is applied at m random starting points to create m imputed data sets.

Imputations generated from this method are not guaranteed to be draws from the posterior predictive distribution $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$, this is is because draws from each of $g_k(\mathbf{x}_k|\mathbf{D}, \mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{k-1}^{(t)}, \mathbf{x}_{k+1}^{(t-1)}, \ldots, \mathbf{x}_p^{(t-1)})$ are not derived from any joint posterior distribution as was the case with FRM. Problems with this modelling approach have been noted by Gelman and Speed (1993), and limitations have also been noted by White et al. (2011). See Azur et al. (2011) for more details about the MICE method and a discussion of its benefits and drawbacks. It would be interesting to see how FRM compares with MICE in imputing missing values in data sets containing variables with different measurement scales. In the next chapter we illustrate the performance of both modelling approaches in a simulation study.

Chapter 4

Simulation study and real data application

In this chapter we will be illustrating the performances of our imputation strategy FRM through a simulation study and a real data set. We also compare the results obtained using FRM to those obtained using MICE in both the simulation study and the real data application. We first describe the simulation study in Section 4.1 and Section 4.2 will present the results on the real data.

4.1 Simulation study

In this simulation study, we simulate data sets that contain variables measured on binary, ordinal, continuous and nominal (greater than two levels) scales. Specifically each data set contains one binary variable, two ordinal variables, four continuous variables, and two nominal variables. Variables are simulated in a sequential manner with each variable conditional on a subset of variables already generated. We then introduce missing values into all but one of the variables using the MAR mechanism, so that each incomplete variable has approximately 30% missing values. Specific details of how we simulated the incomplete dataset are given in Appendix A. We replicate this data generating process 1000 times to generate 1000 incomplete data sets.

Using the FRM imputation strategy proposed in the previous section we multiply impute the missing values in each incomplete data set $m = 10$ times. To generate m independent imputed data sets, we run m Metropolis within Gibbs samplers, each with a different starting value, with each sampler resulting in an imputed data set after convergence. We assume analysts may be interested in making inferences about various types of estimands arising from both univariate analyses, e.g. population means of variables or the proportions in the population taking a particular level of a categorical variable, as well as multivariate analyses, e.g. the coefficients from a regression model. See Appendix A for a full list of the estimands considered. Using the m imputed data sets, we apply the

combining rules described in Equation 2.1 to construct point and interval estimates for these estimands. We also construct estimates for the same estimands when using MICE to multiply impute the missing values. To impute missing values using MICE we use the MICE package in R (Van Buuren (2011)).

When implementing FRM or MICE for imputation, we consider two scenarios representing the state of the imputer's knowledge about the data generation mechanism. In scenario 1, we assume the imputer knows the data generating process and so decomposes the joint distribution of the imputation model as described in Equation 3.3. The analysis of the imputed data sets will also respect the ordering of the variables used to impute the missing values, the analysis model is thus congenial to the imputation model (Meng (1994)).

In scenario 2, we assume the imputer has no prior knowledge about how the data was generated; this will be the case in most practical situations. We thus explore a different ordering of the predictors (to that used to generate the data) to decompose the joint distribution and impute missing values. The analyses performed are the same as those in scenario 1 and the analysis models are not congenial to the imputation model in scenario 2. Details of the decomposition used and analysis models considered are given in Appendix A.

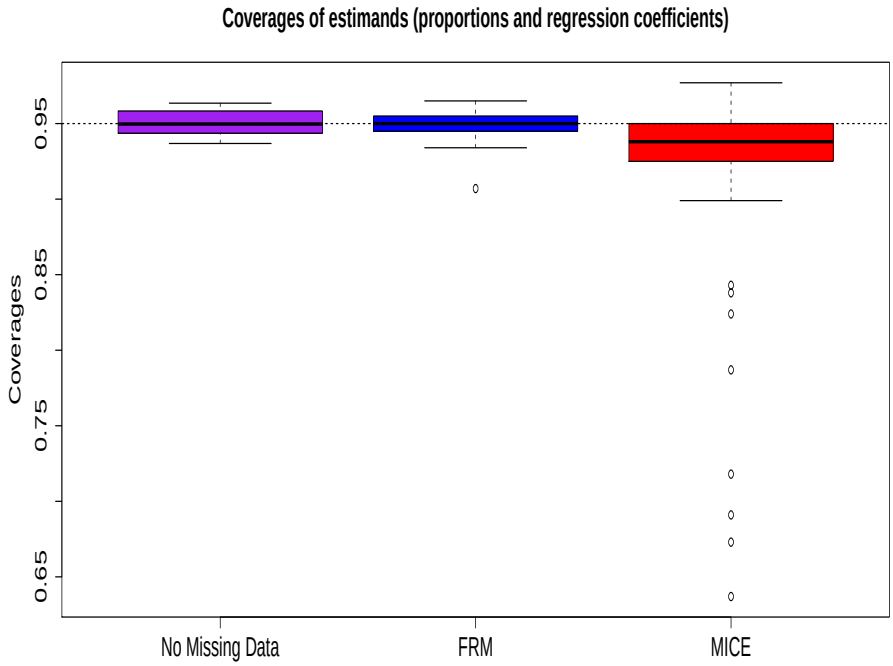


Figure 4.1: The coverages of estimands in scenario 1.

Figure 4.1 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by FRM and MICE over the 1000 datasets. These estimands include those arising from both univariate analyses and regression analyses. The first box plot presents coverages when there are no missing data in the datasets. These coverage are as expected the closest to 0.95. The second box plot

shows the coverages from FRM while the third box plot shows the coverages obtained from MICE. We see that the proposed method obtains coverages much closer to 0.95 than the coverages obtained from using MICE.

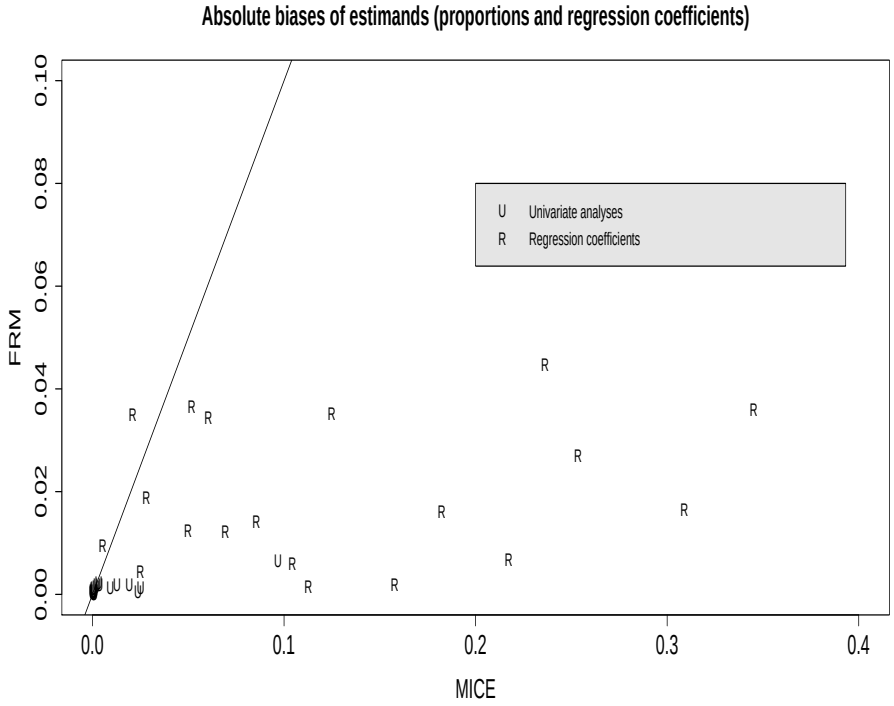


Figure 4.2: Absolute biases of estimands in scenario 1.

To determine whether these low coverages seen from using MICE are due to a result in biases in the estimates, in Figure 4.2 we plot the biases in the estimates obtained from using FRM against the biases obtained from using MICE. We distinguish between estimates arising from univariate analyses and estimates arising from multivariate analyses, i.e. regression coefficient estimates. We see that the majority of the large biases are those arising from multivariate analyses, and are below the $y = x$ line, which indicates that MICE tends to obtain regression coefficient estimates further from the true values than FRM.

Figures 4.3 and 4.4 present similar plots from scenario 2 where we investigate a different decomposition to the joint distribution (from that used to generate the data) to impute missing values. We see that FRM again obtains similar gains over MICE in obtaining coverages much closer to the nominal values, and smaller biases in general.

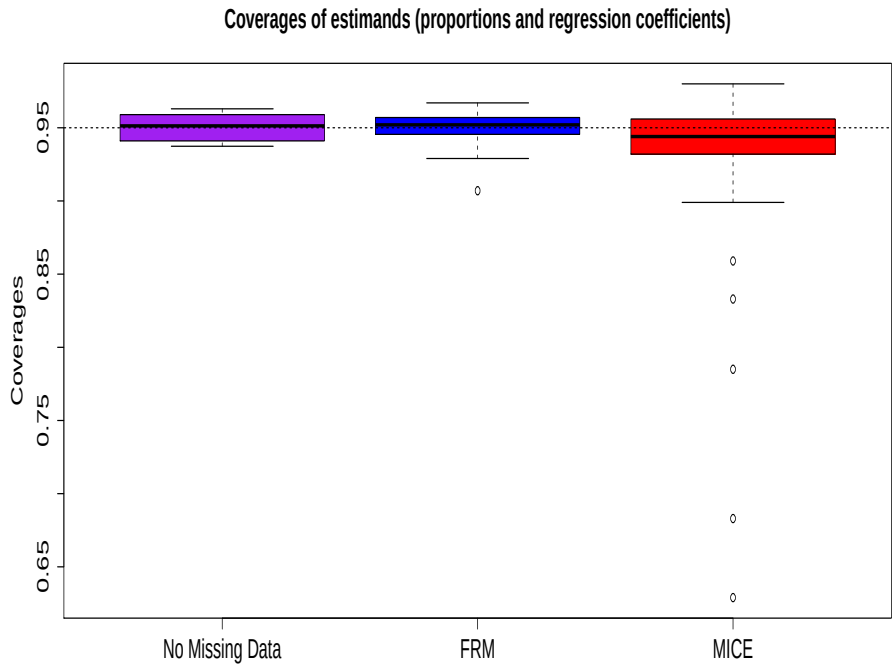


Figure 4.3: The coverages of estimands in scenario 2.

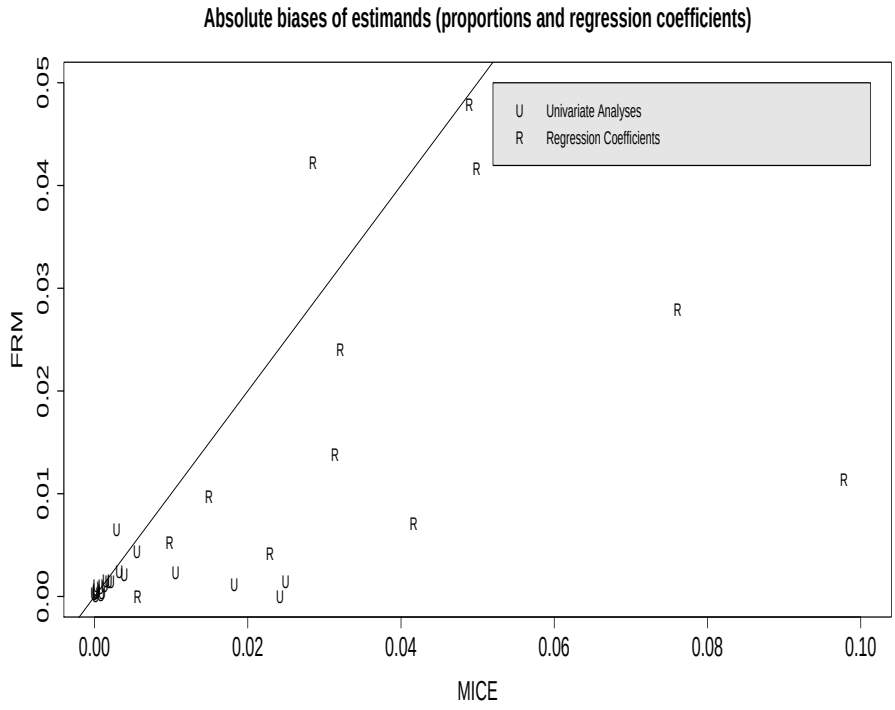


Figure 4.4: Absolute biases of estimands in scenario 2.

4.2 Application to breast-feeding study

We now apply the modelling strategy to impute missing values in a subsample of the 1979 National Longitudinal Survey of Youth (NLSY79). We first describe the data and the analysis of interest, we then apply our modelling strategy to multiply impute the missing data in a simulation based on the complete cases, finally we apply the imputation model on the full data sample.

4.2.1 NLSY dataset

This survey interviewed a sample of 12,686 youths, aged 14-22, on an annual basis from 1979. A separate survey began in 1986 to the children born to female respondents in NLSY79 known as the NLSY79 Children and Young Adults. The data set we consider here is a subsample of the NLSY79 Children and Young Adults which looked at the effect of breast-feeding on children's cognitive development.

The study recorded the Peabody individual assessment math score (PI-ATM) administered at five or six years of age, which we use as the measure of child's cognitive development. Following Mitra and Reiter (2011) we dichomotize the variable that measures duration of breast-feeding so as to split units into two groups; the first group, denoted the control group, comprises those units who were breast-fed for less than 24 weeks, while the second group, denoted the treatment group, comprises those units who were breast-fed for 24 weeks or more. The threshold value, 24 weeks, has been given by the American Academy of Pediatrics (Chantry et al. (2006)) and the World Health Organization as a minimum standard for breast-feeding duration. However, the analysis could be repeated with different threshold values of the breast-feeding duration variable. We assume that an analyst is interested in determining the relationship between PI-ATM and the effect of treatment after adjusting for relevant pre-treatment variables. We adjust on thirteen background pre-treatment variables that are a subset of those used in the analysis by Mitra and Reiter (2011); these were the child's race, whether the spouse or partner was present at birth, child's sex, whether grandparents were present at birth, family income, the number of years between 1979 and when the mother gave birth, mother's intelligence as measured by an armed forces qualification test, mother's highest educational attainment, child's birth weight, number of days that the child spent in hospital, number of days that the mother spent in hospital, number of weeks that the mother worked in the year prior to giving birth, and the number of weeks the child was born premature. The first four variables in this list are categorical. Following Mitra and Reiter (2011) we also categorised the last two variables due to their highly non-normal distributions (see Appendix B for histograms of these variables), we categorised the number of weeks that the mother worked in the year prior to giving birth into four categories, zero weeks, 1-47 weeks, 47-51 weeks and 52 weeks, we categorised the number of weeks the child was born premature into three categories, zero weeks, 1-4 weeks, greater than 5 weeks. This resulted in six categorical variables (comprising binary, ordinal and nominal variables) and seven continuous variables in the data set. Please see Appendix B for more details on the data set.

Analysts can fit a regression model using PI-ATM as the response, and including

treatment plus the other relevant pre-treatment variables as covariates in the model. The effect of treatment could then be estimated by the regression coefficient for treatment. Of course, there is the possibility of unmeasured confounders that we have not adjusted for, which can bias the treatment effect. Hence, we do not seek to make definitive conclusions about the effect of breast-feeding on cognitive development; we are simply using this analysis to illustrate the FRM approach to impute missing values and making subsequent inferences.

We only include the first born children in the analysis to avoid complications arising from family nesting which resulted in 4886 units. Several variables in the study were not fully observed, the response variable contains 48.2% missing values, number of weeks that the mother worked in the year prior to giving birth contains 33.05% missing values, family's income contains 25.69% missing values, both number of days that the child spent in hospital and number of days that the mother spent in hospital contain approximately 10.0% missing values, two variables had no missing values (the number of years between 1979 and when the mother gave birth, and the child's race) and the rest of the variables had missing data rates of less than 10% (see Appendix B for more details). There were 1313 fully observed units in this sample.

To impute the missing values we decompose the joint distribution of the variables in the study using a sequence of regression models as described in Chapter 3. Variables without any missing values are conditioned on in every regression model. The ordering of variables used in the decomposition is given in Appendix B. When a regression model had a continuous response, we considered Box-Cox transformations where necessary to improve the normality assumptions required in the imputation modelling strategy; full details are given in Appendix B. We included the response variable as the last variable in our decomposition. Inclusion of the response variable in imputation models has been recommended by Little (1992), although we note that others have suggested the response should not be included (D'Agostino and Rubin (2000)). We could have repeated the analysis without including the response in the imputations.

4.2.2 Simulation involving the complete cases

Before applying FRM to impute missing values in the full sample, we apply the imputation model to a simulation involving the complete case subsample. We reintroduce missing data patterns in the complete case subsample using the same fractions that appeared in the full sample. We can then run FRM to multiply impute the missing values in this incomplete subsample, perform the analysis using the imputed data, and compare the results to the analysis performed on the subsample prior to introducing missing values. We also compare results to those from using MICE to multiply impute the missing values. As MICE uses a full conditional regression model to impute missing values in each variable, we considered potentially different Box-Cox transformations for variables to improve the model fit here, more details are given in Appendix B.

Table 4.1 presents estimates of the regression coefficients and variances obtained from fitting the regression model for PI-ATM on treatment and other covariates. The first column presents estimates prior to introducing missing values, the second column presents

Table 4.1: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	No missing data	FRM	MICE
Intercept.	86.344 (3.386)	85.430 (6.061)	85.955(4.915)
Treatment effect.	1.609 (0.948)	1.275 (1.586)	2.019 (1.232)
The number of years between 1979 and when the mother gave birth.	-0.025 (0.099)	0.001 (0.182)	-0.054 (0.133)
The child's race-Black	-1.244 (1.117)	-0.933 (2.078)	-0.662(1.437)
The child's race-Other	2.346 (0.971)	4.159 (1.736)	4.114 (1.350)
Spouse present at birth.	-1.595 (1.438)	-1.860 (2.879)	-1.412 (1.782)
Partner present at birth.	-0.159 (1.104)	0.177 (1.818)	0.085(1.554)
Family income.	0.000 (0.000)	0.000 (0.000)	0.000(0.000)
Child's sex.	0.811 (0.676)	0.130 (1.176)	0.125(0.933)
Grandparents were present at birth.	-1.078 (1.151)	-1.175 (1.867)	-0.471 (1.577)
Mother's intelligence.	0.081 (0.018)	0.057 (0.032)	0.052(0.027)
Mother's highest educational.	0.606 (0.212)	0.628 (0.355)	0.579(0.316)
Child's birth weight.	0.023 (0.021)	0.026 (0.032)	0.026(0.027)
Days that the child spent in hospital.	-0.045 (0.052)	-0.047 (0.089)	-0.046(0.066)
Days that the mother spent in hospital.	-0.283 (0.119)	-0.387 (0.214)	-0.417 (0.160)
Weeks that the mother worked-level 2	1.190 (0.984)	2.194 (1.556)	2.349 (1.615)
Weeks that the mother worked-level 3	0.038 (1.261)	1.653 (2.083)	1.575 (2.196)
Weeks that the mother worked-level 4	2.156 (1.102)	3.392 (2.003)	4.873 (1.773)
Child was born premature-level 2	1.265 (0.919)	-0.233 (1.531)	-0.183 (1.147)
Child was born premature-level 3	2.009 (2.248)	-0.226 (3.718)	-0.073(2.969)

results based on the FRM, the final column presents results from MICE. From an analysts point of view, the key estimate is the coefficient on treatment, which gives an estimate of the treatment effect; ideally the estimate from the imputed data should be close to that obtained from the fully observed data. We see that the bias in the treatment effect estimate from using MICE is about 25.5% when compared to the fully observed estimate, while the bias from using FRM is 20.7%. So, in terms of estimating the treatment effect, there is a potential benefit from using FRM over MICE.

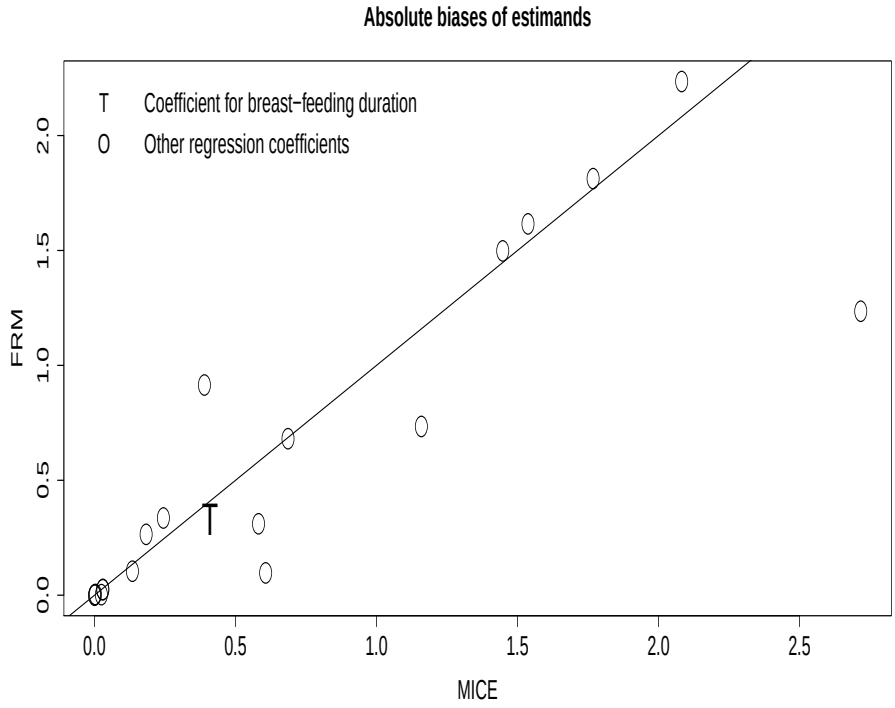


Figure 4.5: Absolute biases of estimands.

In Figure 4.5 we also plot the absolute biases in the regression coefficient estimates from using FRM, again when compared to the fully observed estimates, against similar biases from using MICE. We see that the majority of points (11 out of the 20) lie below the $y = x$ line, indicating that in general using FRM for imputation has the potential to obtain estimates closer to those obtained from the fully observed data than using MICE. We investigated how sensitive results were to three different ordering of variables in the decomposition of the joint distribution, and we found similar results to those obtained above.

4.2.3 Application to the full data sample

We now apply FRM and MICE to the full data sample. We used the same decomposition for the joint distribution as was used in the previous section. For FRM we run $m = 50$ Gibbs samplers at different starting points each for 20000 iterations to generate 50 imputed data sets, we also use MICE to generate 50 imputed data sets. Table 4.2 presents coefficient estimates from the regression model described above using both sets of imputed data. We see that there is no real significant difference in the treatment effect estimates here; they

differ by approximately 0.233 (standard errors from using FRM and MICE are 0.727 and 0.736 respectively).

Table 4.2: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	FRM	MICE
Intercept.	85.222 (2.578)	84.562(2.605)
Treatment effect.	0.903 (0.727)	1.136 (0.736)
The number of years between 1979 and when the mother gave birth.	0.041 (0.071)	0.038(0.072)
The child's race-Black	-0.549 (0.776)	-0.567 (0.811)
The child's race-Other	3.093 (0.726)	3.176(0.765)
Spouse present at birth.	-0.101 (1.227)	0.187(1.092)
Partner present at birth.	0.294 (0.824)	0.061 (0.931)
Family income.	0.000 (0.000)	0.000(0.000)
Child's sex.	0.740 (0.468)	0.787(0.458)
Grandparents were present at birth.	-0.537 (0.806)	0.040 (0.816)
Mother's intelligence.	0.108 (0.014)	0.108 (0.014)
Mother's highest educational.	0.580 (0.162)	0.583(0.161)
Child's birth weight.	0.013 (0.016)	0.016 (0.016)
Days that the child spent in hospital.	-0.055 (0.046)	-0.057 (0.045)
Days that the mother spent in hospital.	-0.104 (0.079)	-0.106 (0.090)
Weeks that the mother worked-level 2	0.960 (0.704)	1.004 (0.793)
Weeks that the mother worked-level 3	1.040 (0.925)	0.828 (1.023)
Weeks that the mother worked-level 4	1.898 (0.871)	1.948 (0.890)
Child was born premature-level 2	0.811 (0.659)	0.749 (0.748)
Child was born premature-level 3	0.958 (1.372)	0.596(1.902)

While we cannot be certain which approach most closely reflects the estimates that would be obtained from the fully observed data, we saw that FRM tended to obtain closer estimates than MICE in the complete case simulation. Of course we cannot also be certain if the treatment effect estimate that would have been obtained from the fully observed data is a reliable estimate of the true treatment effect, the standard problems of unmeasured confounding and model mis-specification still apply. We use this breast-feeding study simply to illustrate the potential gains when using FRM to achieve results closer to the complete data results over MICE.

Chapter 5

Alternative prior specifications for FRM

Until now, we have proposed a modelling strategy to multiply impute missing values in data sets that contain both categorical (ordinal and nominal) and continuous variables. In the FRM strategy, we have used non-informative priors for our univariate linear regression models. We will now extend our modelling strategy to incorporate different informative priors for the FRM.

We noticed that in the real data application, the convenient normality assumption is not quite tenable and heavy tails are called for in order to adequately capture the main features of the data. In FRM, the error terms of the univariate linear regression models are assumed to be normally and independently distributed, each with zero mean and common variance. A known limitation with this normality assumption is its non-robustness, which does not allow for heavy tailed error distributions. Hence we considered the Box-Cox transformations (Box and Cox (1964)) for continuous responses where necessary to improve the normality assumptions required in the imputation models (see Appendix B for details). To extend our modelling approach to deal with the heavy tailed features, we assume that the errors of the linear regression models follow a marginal t-distribution with degrees of freedom ν. This allows for heavier-tailed error distributions which are more robust to the outliers in our data. The assumptions of normally distribution errors can be viewed as a special case within this framework by letting $\nu \to \infty$. This is important as the outlying observations may severely impact any inferences we seek to make from the resulting imputed data sets (Barnett and Lewis (1994)). In the Bayesian approach, we also assign a marginal t-prior to the regression parameters in our Gibbs sampler, and this will still give us a set of tractable full conditional distributions. We then refer to this modelling strategy for imputation as the robust factored regression model (RFRM). We will be comparing the performances of the RFRM and FRM through a simulation study and on the breast-feeding data taken from Chapter 3.

We will also be considering another different approach in extending our model. We noticed that when we perform a regression model on a data set with large numbers of

variables, this will lead to large numbers of predictors present in the regression model, but not all predictors are statistically significant in the model. In the usual regression model, the parameter estimates are obtained by using the ordinary least squares (OLS) method. However, the OLS estimates often have low bias but large variance, and hence could affect the prediction accuracy; also it is difficult to interpret the model with large numbers of predictors. A technique called the Lasso (Tibshirani (1996)) has become a widely used alternative method to ordinary least squares (OLS) in estimating regression parameters (Yuan and Lin (2006), Zou (2006), Zhao and Yu (2006)). The Lasso has features such as shrinking the regression coefficients toward zero, resulting in some regression coefficients identically equal to zero, which leads to a variable selection procedure. Moreover, such shrinkage features can sacrifice a little bias to reduce the variance of the estimates, and hence may increase the prediction accuracy and could potentially lead to more plausible imputations for our model. We will be incorporating the Bayesian Lasso (Park and Casella (2008)) into our model and we refer to this modelling strategy as the Lasso factored regression model (LFRM). Similarly to before, we compare the performances of the LFRM and FRM through a simulation study and on the breast-feeding data set taken from Chapter 3.

Finally, we will be considering combining elements of RFRM and LFRM into one modelling strategy. The aim is to explore the robustness of the linear regression model and the potential sparse relationship between variables. Similar strategy has also been mentioned in Yi and Xu (2008). We refer to this modelling strategy as the modified factored regression model (MFRM). We compare the performances of the MFRM and FRM through a simulation study and on the breast-feeding data taken from Chapter 3. For each of the joint modelling approaches RFRM, LFRM and MFRM, we also compare the results between these 3 approaches with their equivalent fully conditional specifications (MICE) through simulation studies and on the breast-feeding data set.

5.1 Robust factored regression model (RFRM)

In this section, we will be presenting the full conditional distributions in the Metropolis within Gibbs sampler required to generate imputed datasets using RFRM. We first express the joint posterior distribution of all unknowns; these comprise model parameters Θ, and the missing and latent values introduced through data augmentation, denote the set of all unobserved data as $\mathbf{X}_{unobs}$. The joint posterior can then be expressed as,

$$p(\Theta, \mathbf{X}_{unobs}|\mathbf{X}_{obs}) \propto \prod_{i=1}^{n} \left\{ p(x_{i,1}|\boldsymbol{\theta}^{(1)})p(\boldsymbol{\theta}^{(1)}) \prod_{j=2}^{p} p(x_{i,j}|x_{i,1}, \ldots, x_{i,j-1}, \boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)}) \right\}.$$

We assume that there are k nominal variables, we then order the variables so that $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are nominal, each $\mathbf{x}_q$, $q = 1, \ldots, k$, taking a set of possible values $1, \ldots, L_q$. Variables $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_p$ could then be measured on a binary, ordinal or continuous scale. We provide expressions for each conditional regression model $p(x_{i,q}|x_{i,1}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)})$.

First if q = 1 then,

$$p(x_{i,1}|\boldsymbol{\theta}^{(1)}) = \text{Multinomial}(\pi_1^{(1)}, \ldots, \pi_{L_1}^{(1)}),$$

where $\pi_{j^1}^{(1)} = p(x_{i,1} = j^1|\boldsymbol{\theta}^{(1)}) = \frac{\exp(\beta_{0,j^1}^{(1)})}{\sum_{s=1}^{L_1}\exp(\beta_{0,s}^{(1)})}, \quad j^1 = 1, \ldots, L_1$. We place a Dirichlet prior on $\boldsymbol{\pi}^{(1)}$, i.e.

$$p(\boldsymbol{\pi}^{(1)}) \propto \text{Dirichlet}(\alpha_1, \ldots, \alpha_{L_1}),$$

we set all $\alpha_{j^1} = 0.5$ for $j^1 = 1, \ldots, L_1$ (corresponding to the Jeffreys prior), another common choice could be to set all $\alpha_{j^1} = 0$. For $q = 2, \ldots, k$,

$$p(x_{i,q}|x_{i,1}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \text{Multinomial}(\pi_{i,1}^{(q)}, \ldots, \pi_{i,L_q}^{(q)}),$$

where

$$\pi_{i,j^q}^{(q)} = \frac{\exp\left(\beta_{0,j^q}^{(q)} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,j^q}^{(q),j^b} I(x_{i,b} = j^b)\right)}{\sum_{u=1}^{L_q}\exp\left(\beta_{0,u}^{(q)} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b}\beta_{b,u}^{(q),j^b} I(x_{i,b} = j^b)\right)},$$

where $j^q = 1, \ldots, L_q$. Let $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}_1^{(q)}, \ldots, \boldsymbol{\theta}_{L_q}^{(q)})$
with $\boldsymbol{\theta}_{j^q}^{(q)} = (\beta_{0,j^q}^{(q)}, \beta_{1,j^q}^{(q),2}, \ldots, \beta_{1,j^q}^{(q),L_2}, \ldots, \beta_{q-1,j^q}^{(q),2}, \ldots, \beta_{q-1,j^q}^{(q),L_{q-1}})$ for $j^q = 1, \ldots, L_q$ and we set $\boldsymbol{\theta}_1^{(q)} = \mathbf{0}$ for identifiability. We specify the prior for each $\boldsymbol{\theta}_{j^q}^{(q)}, j^q = 2, \ldots, L_q$ by

$$p(\boldsymbol{\theta}^{(q)}) = \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}).$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with large entries.

For $\mathbf{x}_q,\ q \in \{k+1, \ldots, p\}$ is continuous (for now assuming $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{q-1}$ are continuous),

$$p(x_{i,q}|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) =$$
$$N(\beta_0^{(q)} + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b}\beta_b^{(q),j^b} I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1}\beta_b^{(q)} x_{i,b}, (\tau_q\phi_{i,q})^{-1}).$$

We specify the prior distribution for $\boldsymbol{\theta}^{(q)}$ by,

$$p(\boldsymbol{\theta}^{(q)}) = p(\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \Phi_q) = p(\boldsymbol{\beta}^{(q)}|\tau_q, \Psi_q, \Phi_q)p(\tau_q|\Psi_q, \Phi_q)p(\Psi_q|\Phi_q)p(\Phi_q),$$

where $\Psi_q = (\psi_{0,q}, \ldots, \psi_{j,q})'$ for $j = 1, \ldots, q-1$ and $\Phi_q = (\phi_{i,q}, \ldots, \phi_{n,q})'$. The prior distribution for $\boldsymbol{\beta}^{(q)}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance

matrix $\mathbf{\Sigma}_q$,

$$p(\boldsymbol{\beta}^{(q)}|\tau_q, \Psi_q) \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma_q}), \tag{5.1}$$

where $\mathbf{\Sigma}_q$ is a diagonal matrix with entries given by $\frac{1}{\tau_q \psi_{j,q}}$ for $j = 0, \ldots, q-1$ and we assign a non-informative prior for τ_q, i.e.

$$p(\tau_q) \propto \frac{1}{\tau_q}. \tag{5.2}$$

The prior distribution for Ψ_q where $\Psi_q = (\psi_{0,q}, \ldots, \psi_{q-1,q})'$ is a gamma distribution with shape and rate parameters equal to $\frac{\nu}{2}$ where

$$p(\psi_{j,q}) \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}),\ j = 0, \ldots, q-1, \tag{5.3}$$

This results in a marginal t-prior for $\boldsymbol{\beta}^{(q)}$ with degrees of freedom ν. The prior distribution for Φ_q, where $\Phi_q = (\phi_{i,q}, \ldots, \phi_{n,q})'$ follows a gamma distribution with shape and rate parameters equal to $\frac{\nu}{2}$ where

$$p(\phi_{i,q}) \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}),\ i = 1, \ldots, n.$$

and this results in the errors of the linear regression model follow a marginal t-distribution with degrees of freedom ν.

For $\mathbf{x}_q,\ q \in \{k+1, \ldots, p\}$ is binary (for now assuming $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{q-1}$ are continuous),

$$x_{i,q} = I(x^*_{i,q} > 0), \quad \text{where}$$
$$p(x^*_{i,q}|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \text{N}(\beta_0^{(q)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \beta_b^{(q),j^b} I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1} \beta_b^{(q)} x_{i,b}, 1).$$

We also specify a prior distribution for $\boldsymbol{\theta}^{(q)}$ by,

$$p(\boldsymbol{\theta}^{(q)}) = p(\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q) = p(\boldsymbol{\beta}^{(q)}|\tau_q, \Psi_q) p(\tau_q|\Psi_q) p(\Psi_q),$$

where the prior distributions for $\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q$ are given by Equation 5.1, Equation 5.2 and Equation 5.3 respectively. This results in a marginal t-prior for $\boldsymbol{\beta}^{(q)}$ with degrees of freedom ν.

For $\mathbf{x}_q,\ q \in \{k+1, \ldots, p\}$ is ordinal taking values in $\{1, \ldots, J_q\}$ (for now assuming

$\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{q-1}$ are continuous),

$$x_{i,q} = j^q \; I(\gamma^{(q)}_{j^q-1} < x^*_{i,q} < \gamma^{(q)}_{j^q}), \quad \text{where}$$
$$p(x^*_{i,q}|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \mathrm{N}(\beta^{(q)}_0 + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b} \beta^{(q),j^b}_b I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1} \beta^{(q)}_b x_{i,b}, 1),$$

where the $\gamma^{(q)}_{j^q}$ are threshold values, with $\gamma^{(q)}_0 = -\infty$, $\gamma^{(q)}_1 = 0$ and $\gamma^{(q)}_{J_q} = \infty$, for $j^q \in \{1, ..., J_q\}$.

We also specify a prior distribution for $\boldsymbol{\theta}^{(q)}$ by,

$$p(\boldsymbol{\theta}^{(q)}) = p(\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)}) = p(\boldsymbol{\beta}^{(q)}|\tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)})p(\tau_q|\Psi_q, \boldsymbol{\gamma}^{(q)})p(\Psi_q|\boldsymbol{\gamma}^{(q)})p(\boldsymbol{\gamma}^{(q)}),$$

where $\boldsymbol{\gamma}^{(q)} = \left\{\gamma^{(q)}_{j^q} : j^q \in \{1, \ldots, J_q\}\right\}$. The prior distributions for $\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q$ are given by Equation 5.1, Equation 5.2 and Equation 5.3 respectively. This is a marginal t-prior for $\boldsymbol{\beta}^{(q)}$ with degrees of freedom ν. We then place an improper uniform prior on $\boldsymbol{\gamma}^{(q)}$ i.e.

$$p(\boldsymbol{\gamma}^{(q)}) \propto I(\boldsymbol{\gamma}^{(q)} \in \Omega^{(q)}),$$

where $\Omega^{(q)} = \left\{\gamma^{(q)}_{j^q} : \gamma^{(q)}_0 = -\infty < \gamma^{(q)}_1 = 0 < \gamma^{(q)}_2 < \ldots < \gamma^{(q)}_{J_{q-1}} < \gamma^{(q)}_{J_q} = \infty\right\}$.

If $x_{i,q},\ q \in \{k+1, \ldots, q-1\}$ is not continuous then we replace $x_{i,q}$ with $x^*_{i,q}$ in the model for $p(x_{i,q}|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)})$. For notational convenience when $x_{i,q}$ is nominal or continuous we assume $x_{i,q} = x^*_{i,q},\ q \in \{1, \ldots, p\}$.

With this model and prior specification we can present the full conditional distribution required to impute missing values in the Metropolis within Gibbs sampler. Following Schafer (1997) we present this as a data augmentation scheme and so first present the "I-steps" followed by the "P-steps".

In the "I-steps", conditional on parameter values, first impute missing nominal values $x^*_{i,q}$ from the following discrete distribution

$$p(x^*_{i,q} = j^q | x^*_{i,1} = j^1, \ldots, x^*_{i,q-1} = j^{q-1}, x^*_{i,q+1} = j^{q+1}, \ldots, x^*_{i,k} = j^k, x^*_{i,k+1}, \ldots, x^*_{i,p}, \Theta)$$
$$= \tilde{\pi}^{(q)}_{i,j^q},$$

where $j^q \in \{1, \ldots, L_q\}$, and

$$\tilde{\pi}^{(q)}_{i,j^q} = \frac{\displaystyle\prod_{b=q}^{k} \pi^{(b)}_{i,j^b} \prod_{b=k+1}^{p} \exp\left(\frac{-\tau_b \phi_{i,b}}{2}(\tilde{x}^*_{i,b} - \beta^{(b),j^q}_q I(x^*_{i,q} = j^q) - \sum_{t=k+1}^{(b-1)} \beta^{(b)}_t x^*_{i,t})^2\right)}{\displaystyle\sum_{u=1}^{L_q}\left\{\pi^{(q)}_{i,u} \prod_{b=q+1}^{k} \pi^{(b)}_{i,j^b}\right\} \prod_{b=k+1}^{p} \exp\left(\frac{-\tau_b \phi_{i,b}}{2}(\tilde{x}^*_{i,b} - \beta^{(b),j^q}_q I(x^*_{i,q} = u) - \sum_{t=k+1}^{(b-1)} \beta^{(b)}_t x^*_{i,t})^2\right)},$$

where

$$\tilde{x}^*_{i,b} = x^*_{i,b} - \beta^{(b)}_0 - \sum_{s=1}^{q-1}\sum_{j^s=2}^{L_s} \beta^{(b),j^s}_s I(x^*_{i,s} = j^s) - \sum_{s=q+1}^{k}\sum_{j^s=2}^{L_s} \beta^{(b),j^s}_s I(x^*_{i,s} = j^s).$$

Next impute missing continuous values $x^*_{i,q}$ or latent values $x^*_{i,q}$ with missing $x_{i,q}$ from,

$$p(x^*_{i,q}|x^*_{i,1},\ldots,x^*_{i,q-1},x^*_{i,q+1},\ldots,x^*_{i,k},x^*_{i,k+1},\ldots,x^*_{i,p},\Theta) = \mathrm{N}(\widetilde{\mu}_{i,q},\widetilde{\phi}^{-1}_{i,q}),$$

where

$$\widetilde{\mu}_{i,q} = \widetilde{\phi}^{-1}_{i,q}\left\{\frac{\mu_{i,q}}{(\tau_q\phi_{i,q})^{-1}} + \sum_{s=q+1}^{p}\frac{\beta^{(s)}_q}{(\tau_s\phi_{i,s})^{-1}}\left[x^*_{i,s} - (\mu_{i,s} - \beta^{(s)}_q x^*_{i,q})\right]\right\},$$

and

$$\widetilde{\phi}^{-1}_{i,q} = \left(\frac{1}{(\tau_q\phi_{i,q})^{-1}} + \sum_{s=q+1}^{p}\frac{(\beta^{(s)}_q)^2}{(\tau_s\phi_{i,s})^{-1}}\right)^{-1},$$

where $\mu_{i,q} = \beta^{(q)}_0 + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b}\beta^{(q),j^b}_b I(x^*_{i,b} = j^b) + \sum_{b=k+1}^{q-1}\beta^{(q)}_b x^*_{i,b}$ and
$\mu_{i,s} = \beta^{(s)}_0 + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b}\beta^{(s),j^b}_b I(x^*_{i,b} = j^b) + \sum_{b=k+1}^{s-1}\beta^{(s)}_b x^*_{i,b}$.

Next impute latent values $x^*_{i,q}$ with observed ordinal variable $x_{i,q}$ from,

$$p(x^*_{i,q}|x^*_{i,1},\ldots,x^*_{i,q-1},x^*_{i,q+1},\ldots,x^*_{i,p},\Theta) = \frac{I(\gamma^{(q)}_{j^q-1} < x^*_{i,q} < \gamma^{(q)}_{j^q})\mathrm{N}(\widetilde{\mu}_{i,q},\widetilde{\phi}^{-1}_{i,q})}{\Phi(\frac{\gamma^{(q)}_{j^q}-\widetilde{\mu}_{i,q}}{\widetilde{\phi}^{-1}_{i,q}}) - \Phi(\frac{\gamma^{(q)}_{j^q-1}-\widetilde{\mu}_{i,q}}{\widetilde{\phi}^{-1}_{i,q}})},$$

which is a normal distribution $\mathrm{N}(\widetilde{\mu}_{i,q},\widetilde{\phi}^{-1}_{i,q})$ truncated to $(\gamma^{(q)}_{j^q-1},\gamma^{(q)}_{j^q})$.

Next impute latent values $x^*_{i,q}$ with observed binary variable $x_{i,q}$ from,

$$p(x^*_{i,q}|x^*_{i,1},\ldots,x^*_{i,q-1},x^*_{i,q+1},\ldots,x^*_{i,p},\Theta) = \frac{I(\gamma^{(q)}_{j^q-1} < x^*_{i,q} < \gamma^{(q)}_{j^q})\mathrm{N}(\widetilde{\mu}_{i,q},\widetilde{\phi}^{-1}_{i,q})}{\Phi(\frac{\gamma^{(q)}_{j^q}-\widetilde{\mu}_{i,q}}{\widetilde{\phi}^{-1}_{i,q}}) - \Phi(\frac{\gamma^{(q)}_{j^q-1}-\widetilde{\mu}_{i,q}}{\widetilde{\phi}^{-1}_{i,q}})},$$

which is a normal distribution $\mathrm{N}(\widetilde{\mu}_{i,q},\widetilde{\phi}^{-1}_{i,q})$ truncated to $(\gamma^{(q)}_{j^q-1},\gamma^{(q)}_{j^q})$ where the $\gamma^{(q)}_{j^q}$ are fixed threshold values, with $\gamma^{(q)}_0 = -\infty$, $\gamma^{(q)}_1 = 0$ and $\gamma^{(q)}_2 = \infty$.

Once we have imputed values for $x^*_{i,q}$ (with missing $x_{i,q}$), we define a function $g(x^*_{i,q})$ to map each $x^*_{i,q}$ back to its original measurement scale, and thus create an imputed data

set. The mapping function is defined as follows,

$$g(x^*_{i,q}) = \begin{cases} I(x^*_{i,q} > 0) & \text{if } x^*_{i,q} \text{ is binary,} \\ j^q \; I(\gamma^{(q)}_{j^q-1} < x^*_{i,q} < \gamma^{(q)}_{j^q}) & \text{if } x^*_{i,q} \text{ is ordinal, } j^q \in \{1,\ldots,J_q\}, \\ x^*_{i,q} & \text{if } x^*_{i,q} \text{ is continuous,} \\ x^*_{i,q} & \text{if } x^*_{i,q} \text{ is nominal, } j^q \in \{1,\ldots,L_q\}, \end{cases}$$

Denote an imputed value for $x_{i,q}$ by $x^{(t)}_{i,q}$ at iteration t, and an imputed data set at iteration t by $\mathbf{X}^{(t)}_{com}$.

In the "P-steps", conditional on an imputed data set, first sample values for $\boldsymbol{\theta}^{(1)}$ from a Dirichlet($\boldsymbol{\alpha}$) distribution with parameters

$$\boldsymbol{\alpha} = \left(\alpha_1 + \sum_{i=1}^{n} I(x_{i,1} = 1),\ \alpha_2 + \sum_{i=1}^{n} I(x_{i,1} = 2), \ldots, \alpha_{L_1} + \sum_{i=1}^{n} I(x_{i,1} = L_1) \right)'.$$

Next propose values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}^{(q)}_1, \ldots, \boldsymbol{\theta}^{(q)}_{L_q})$ with $\boldsymbol{\theta}^{(q)}_{j^q} = (\beta^{(q)}_{0,j^q}, \beta^{(q),2}_{1,j^q}, \ldots, \beta^{(q),L_2}_{1,j^q}, \ldots, \beta^{(q),2}_{q-1,j^q}, \ldots, \beta^{(q),L_{q-1}}_{q-1,j^q})$, for $j^q = 1, \ldots, L_q$ using a Metropolis-Hastings step, where we set $\boldsymbol{\theta}^{(q)}_1 = \mathbf{0}$ for identifiability. We consider a proposal distribution based on the imputed data log-likelihood at each iteration t, $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}^{(t)}_{com})$, which can be expressed as

$$l(\boldsymbol{\theta}^{(q)}; \mathbf{X}^{(t)}_{com}) = \sum_{w=1}^{L_q} \sum_{i=1}^{n} \ln \left(\frac{\exp\left(\beta^{(q)}_{0,w} + \sum_{b=1}^{q-1} \sum_{j^b=2}^{L_b} \beta^{(q),j^b}_{b,w} f(x^{(t)}_{i,b}, x_{i,b}, m_{i,b}) \right)}{\sum_{u=1}^{L_q} \exp\left(\beta^{(q)}_{0,u} + \sum_{b=1}^{q-1} \sum_{j^b=2}^{L_b} \beta^{(q),j^b}_{b,u} f(x^{(t)}_{i,b}, x_{i,b}, m_{i,b}) \right)} \right)^{I(x_{i,q}=w)},$$

where $f(x^{(t)}_{i,b}, x_{i,b}, m_{i,b}) = (I(x^{(t)}_{i,b} = j^b) m_{i,b} + I(x_{i,b} = j^b)(1 - m_{i,b}))$, with $m_{i,b}$ is a missing data indicator with $m_{i,b} = 1$ indicating $x_{i,b}$ is missing and $m_{i,b} = 0$ indicating $x_{i,b}$ is observed. We generate proposals for $\boldsymbol{\theta}^{(q)}$ from a $\mathrm{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ distribution, and $\boldsymbol{\mu}^{(t)}$ is the value that maximises $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}^{(t)}_{com})$ and $\boldsymbol{\Sigma}^{(t)}$ is given by $-\mathrm{E}[\frac{\delta^2}{\delta\beta^{(q)}_{j,w}\delta\beta^{(q)}_{\tilde{j},\tilde{w}}} l(\boldsymbol{\theta}^{(q)}; \mathbf{X}^{(t)}_{com})]^{-1}_{\boldsymbol{\theta}^{(q)}=\boldsymbol{\mu}^{(t)}}$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \Phi_q),\ q \in \{k+1, \ldots, p\}$ when $\mathbf{x}^*_q$ is continuous from the joint posterior distribution of $\boldsymbol{\beta}^{(q)} = (\beta^{(q)}_0, \beta^{(q),2}_1, \ldots, \beta^{(q),L_1}_1, \ldots, \beta^{(q),2}_k, \ldots, \beta^{(q),L_k}_k, \beta^{(q)}_{k+1}, \ldots, \beta^{(q)}_{q-1})'$, $\Phi_q = (\phi_{i,q}, \ldots, \phi_{n,q})'$, $\Psi_q = (\psi_{0,q}, \ldots, \psi_{q-1,q})'$ and τ_q, which is given by

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}, \tau_q, \Phi_q, \Psi_q | \mathbf{X}^{(t)}_{com}) &= p(\boldsymbol{\beta}^{(q)} | \tau_q, \Psi_q, \Phi_q, \mathbf{X}^{(t)}_{com}) p(\tau_q | \Psi_q, \Phi_q, \mathbf{X}^{(t)}_{com}) p(\Psi_q | \Phi_q, \mathbf{X}^{(t)}_{com}) \\ &\quad p(\Phi_q | \mathbf{X}^{(t)}_{com}). \end{aligned}$$

We sample $\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \Phi_q$ from their full conditional distributions in the following way:

1. We first sample $\boldsymbol{\beta}^{(q)}$ from

$$p(\boldsymbol{\beta}^{(q)}|\tau_q, \Psi_q, \Phi_q, \widetilde{\mathbf{X}}_q, \mathbf{x}_q^*) = \text{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}),$$

with

$$\hat{\boldsymbol{\beta}}^{(q)} = \tau_q \mathbf{V}^{(q)} \widetilde{\mathbf{X}}_q' \mathbf{D}_\phi^{(q)} \mathbf{x}_q^*,$$

and

$$\mathbf{V}^{(q)} = \frac{1}{\tau_q}(\widetilde{\mathbf{X}}_q' \mathbf{D}_\phi^{(q)} \widetilde{\mathbf{X}}_q + \mathbf{D}_\psi^{(q)})^{-1},$$

where $\mathbf{x}_q^* = (x_{1,q}^*, \ldots, x_{n,q}^*)'$, and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, with
$\widetilde{\mathbf{x}}_{i,q} = (1, I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$
for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n \times n$ diagonal matrix with entries $\phi_{i,q}$ for $i = 1, \ldots, n$ and $\mathbf{D}_\psi^{(q)}$ is a $q \times q$ diagonal matrix with entries $\psi_{j,q}$ for $j = 0, \ldots, q-1$.

2. Next sample a value for τ_q from

$$p(\tau_q|\boldsymbol{\beta}^{(q)}, \Psi_q, \Phi_q, \widetilde{\mathbf{X}}_q', \mathbf{x}_q^*) \sim$$
$$\text{Gamma}\left(\frac{n+q}{2}, \frac{(\mathbf{x}_q^* - \widetilde{\mathbf{X}}_q \boldsymbol{\beta}^{(q)})' \mathbf{D}_\phi^{(q)} (\mathbf{x}_q^* - \widetilde{\mathbf{X}}_q \boldsymbol{\beta}^{(q)}) + (\boldsymbol{\beta}^{(q)})' \mathbf{D}_\psi^{(q)} \boldsymbol{\beta}^{(q)}}{2}\right),$$

where $\mathbf{x}_q^* = (x_{1,q}^*, \ldots, x_{n,q}^*)'$, and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, with
$\widetilde{\mathbf{x}}_{i,q} = (1, I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$
for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n \times n$ diagonal matrix with entries $\phi_{i,q}$ for $i = 1, \ldots, n$ and $\mathbf{D}_\psi^{(q)}$ is a $q \times q$ diagonal matrix with entries $\psi_{j,q}$ for $j = 0, \ldots, q-1$.

3. We then sample values for $\psi_{j,q}$ from

$$p(\psi_{j,q}|\beta_j^{(q)}, \tau_q) = \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\tau_q (\beta_j^{(q)})^2 + \nu}{2}\right), \quad j = 0, \ldots, q-1.$$

4. We sample values for $\phi_{i,q}$ from

$$p(\phi_{i,q}|\tau_q, \boldsymbol{\beta}^{(q)}, x_{i,q}^*, \widetilde{\mathbf{x}}_{i,q}) \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + \tau_q (x_{i,q}^* - \widetilde{\mathbf{x}}_{i,q} \boldsymbol{\beta}^{(q)})^2}{2}\right), \quad i = 1, \ldots, n,$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(1, I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$, for
$i = 1, \ldots, n$.

We now sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q),\ q \in \{k+1, \ldots, p\}$ when $\mathbf{x}_q$ is binary from the joint posterior distribution of
$\boldsymbol{\beta}^{(q)} = (\beta_0^{(q)}, \beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$, τ_q,
$\Psi_q = (\psi_{0,q}, \ldots, \psi_{q-1,q})'$, which is given by

$$p(\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q | \mathbf{X}_{com}^{(t)}) \;\; = \;\; p(\boldsymbol{\beta}^{(q)} | \tau_q, \Psi_q, \mathbf{X}_{com}^{(t)}) p(\tau_q | \Psi_q, \mathbf{X}_{com}^{(t)}) p(\Psi_q | \mathbf{X}_{com}^{(t)})$$

We sample $\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q$ from their full conditional distributions in the following way:

1. We first sample $\boldsymbol{\beta}^{(q)}$ from

$$p(\boldsymbol{\beta}^{(q)} | \tau_q, \Psi_q, \widetilde{\mathbf{X}}_q', \mathbf{x}_q^*) = \text{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}), \tag{5.4}$$

with

$$\hat{\boldsymbol{\beta}}^{(q)} = \mathbf{V}^{(q)} \widetilde{\mathbf{X}}_q' \mathbf{x}_q^*,$$

and

$$\mathbf{V}^{(q)} = (\widetilde{\mathbf{X}}_q' \widetilde{\mathbf{X}}_q + \tau_q \mathbf{D}_\psi^{(q)})^{-1},$$

where $\mathbf{x}_q^* = (x_{1,q}^*, \ldots, x_{n,q}^*)'$, and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, with
$\widetilde{\mathbf{x}}_{i,q} = (1, I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\psi^{(q)}$ is a $q \times q$ diagonal matrix with entries $\psi_{j,q}$ for $j = 0, \ldots, q-1$.

2. Next sample a value for τ_q from

$$p(\tau_q | \boldsymbol{\beta}^{(q)}, \Psi_q) \sim \text{Gamma}\left(\frac{q}{2}, \frac{(\boldsymbol{\beta}^{(q)})' \mathbf{D}_\psi^{(q)} \boldsymbol{\beta}^{(q)}}{2}\right), \tag{5.5}$$

where matrix $\mathbf{D}_\psi^{(q)}$ is a $q \times q$ diagonal matrix with entries $\psi_{j,q}$ for $j = 0, \ldots, q-1$.

3. We then sample values for $\psi_{j,q}$ from

$$p(\psi_{j,q} | \tau_q, \beta_j^{(q)}) = \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\tau_q (\beta_j^{(q)})^2 + \nu}{2}\right), \quad j = 0, \ldots, q-1. \tag{5.6}$$

We now sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)}),\ q \in \{k+1, \ldots, p\}$ when $\mathbf{x}_q$ is ordinal from the joint posterior distribution of
$\boldsymbol{\beta}^{(q)} = (\beta_0^{(q)}, \beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$, τ_q,

$\Psi_q = (\psi_{0,q}, \ldots, \psi_{q-1,q})'$ and $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1, \ldots, J_q\}\right\}$, which is given by

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)} | \mathbf{X}_{com}^{(t)}) &= p(\boldsymbol{\beta}^{(q)} | \tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) p(\tau_q | \Psi_q, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) p(\Psi_q | \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) \\ & \quad p(\boldsymbol{\gamma}^{(q)} | \mathbf{X}_{com}^{(t)}) \end{aligned}$$

Now we sample $\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)}$ from their full conditional distributions in the following way:

1. We sample $\boldsymbol{\beta}^{(q)}$ from its full conditional distribution which is given by Equation 5.4.

2. Next sample a value for τ_q from its full conditional distribution which is given by Equation 5.5.

3. We then sample values for $\psi_{j,q},\ j = 0, \ldots, q-1$ from its full conditional distribution which is given by Equation 5.6.

4. We then update the threshold values $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1, \ldots, J_q\}\right\}$. The full conditional distribution of the threshold values $\gamma_{j^q}^{(q)}$ for $j^q \in \{1, \ldots, J_q\}$ is uniformly distributed on the interval

$$\left[\max\left\{\max\left\{x_{i,q}^* : x_{i,q} = j^q\right\}, \gamma_{j^q-1}^{(q)}\right\}, \min\left\{\min\left\{x_{i,q}^* : x_{i,q} = j^q + 1\right\}, \gamma_{j^q+1}^{(q)}\right\}\right],$$

where $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_{J_q}^{(q)} = \infty$, for $j^q \in \{1, \ldots, J_q\}$.

In this way we have sampled all unknowns (missing values and parameters) from their full conditional distributions. The details on deriving the full conditional distributions of the missing values in the "I-steps", the full conditional distribution of $\boldsymbol{\theta}^{(1)}$ where $\mathbf{x}_1$ follows a multinomial distribution, as well as the Metropolis-Hastings update for the multinomial logistic regression model in the "P-steps" have been presented in the RFRM strategy. For details on deriving the full conditional distributions of the remaining parameters in the "P-steps", such as $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \Phi_q)$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is continuous, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q)$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is binary) and $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is ordinal), please see Appendix C.

5.1.1 Simulation study

We now illustrate the performances of FRM, RFRM and MICE through a simulation study. In the MICE approach, we use the model and prior specifications described in the RFRM strategy to generate missing values. We are interested to compare the results between a non-robust modelling strategy (FRM) and a robust modelling strategy (RFRM), as well as the results between the joint modelling specification with t-errors (RFRM) with the equivalent fully conditional specification (MICE).

We simulate data sets that contain variables measured on binary, ordinal, continuous

and nominal (greater than two levels) scales. Specifically each data set contains one binary variable, two ordinal variables, four continuous variables, and two nominal variables. Variables are simulated in a sequential manner with each variable conditional on a subset of variables already generated. We then introduce missing values into all but one of the variables using the MAR mechanism, so that each incomplete variable has approximately 30% missing values. Specific details of how we simulated the incomplete dataset are given in Appendix D. We replicate this data generation process 500 times to generate 500 incomplete data sets.

Using the RFRM and MICE imputation strategies, we multiply impute the missing values in each incomplete data set $m = 10$ times. To generate m independent imputed data sets, we run m Metropolis within Gibbs samplers, each with a different starting value, with each sampler resulting in an imputed data set after convergence. We assume analysts may be interested in making inferences about various types of estimands arising from both univariate analyses, e.g. population means of variables or the proportions in the population taking a particular level of a categorical variable, as well as multivariate analyses, e.g. the coefficients from a regression model. See Appendix D for a full list of the estimands considered. Using the m imputed data sets, we apply the combining rules described in Equation 2.1 to construct point and interval estimates for these estimands. We also construct estimates for the same estimands when using FRM to multiply impute the missing values.

When implementing RFRM, MICE or FRM for imputations, we assume the imputer knows the data generating process and so decomposes the joint distribution of the imputation model as described in Equation 3.3. The analysis of the imputed data sets will also respect the ordering of the variables used to impute the missing values, the analysis model is thus congenial to the imputation model (Meng (1994)).

Figure 5.1 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by RFRM and FRM over the 500 data sets. These estimands include those arising from both univariate analyses and regression analyses. The first boxplot presents coverages when there are no missing data in the data sets. These coverages are as expected the closest to 0.95. The second boxplot shows the coverages obtained from using FRM while the third boxplot shows the coverages obtained from using RFRM. We see that the RFRM obtains coverages much closer to 0.95 than the coverages obtained from using FRM. The estimands that showed low coverages in the second boxplot are the regression coefficients estimates from the regression model $p(\mathbf{x}_7|\mathbf{x}_6, \mathbf{x}_3)$.

To determine whether these low coverages seen from using FRM are due to a result in biases in the estimates, we plot the absolute biases of estimands obtained from using FRM against the biases obtained from using RFRM. We distinguish between estimates arising from univariate analyses and estimates arising from multivariate analyses, i.e. regression coefficient estimates. In Figure 5.2, we see that the majority of the large biases are those arising from multivariate analyses, and are above the $y = x$ line, which indicates that FRM tends to obtain regression coefficient estimates further from the true values than RFRM. We noticed that RFRM shows a larger absolute bias of a regression coefficient, denoted by "R" (on the right hand side of Figure 5.2) than FRM. This coefficient is a coefficient estimate from the regression model $p(\mathbf{x}_9|\mathbf{x}_2, \mathbf{x}_7)$.

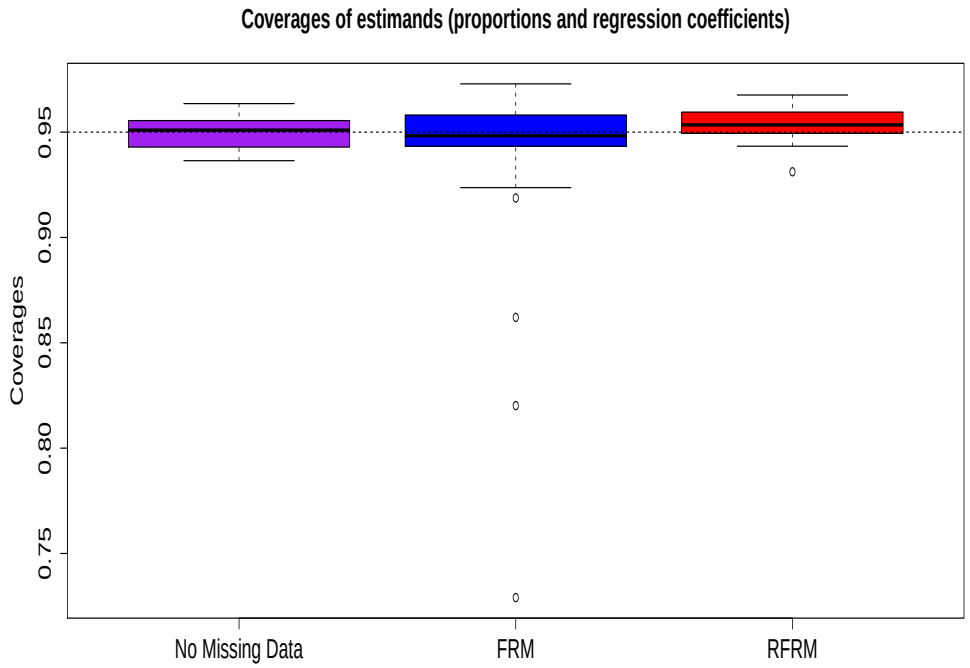


Figure 5.1: The coverages of estimands.

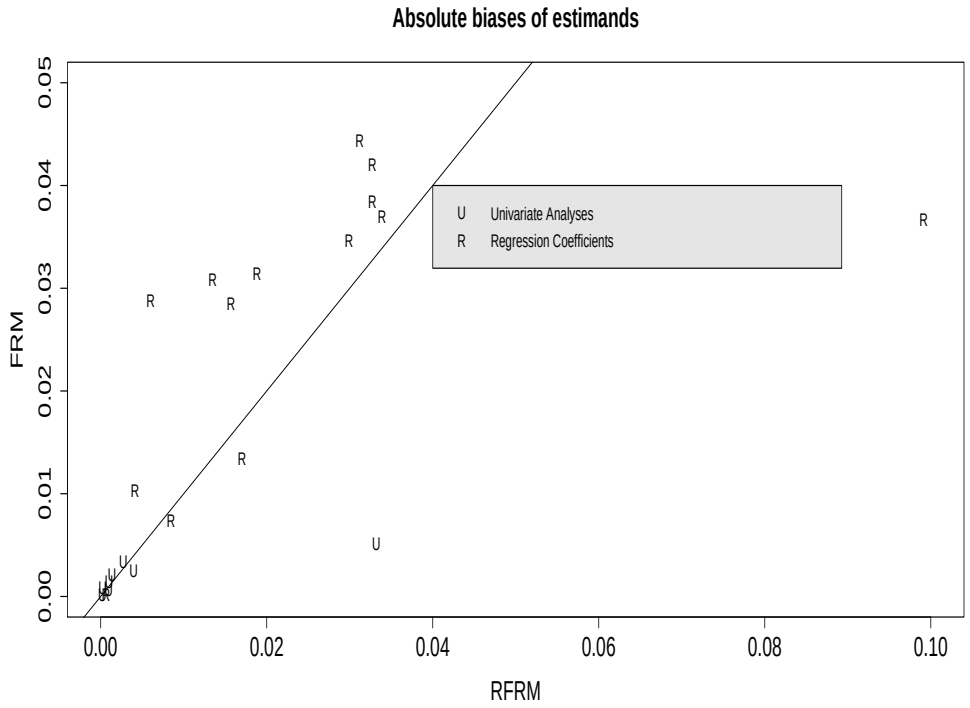


Figure 5.2: Absolute biases of estimands.

We now compare the results between RFRM and MICE. Figure 5.3 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by RFRM and MICE over the 500 data sets. These estimands include those arising from both univariate analyses and regression analyses. The first boxplot presents coverages when there are no missing data in the data sets. These coverages are as expected the closest to 0.95. The second boxplot shows the coverages obtained from using MICE while the third boxplot shows the coverages obtained from using RFRM. We see that the RFRM obtains coverages much closer to 0.95 than the coverages obtained from using MICE.

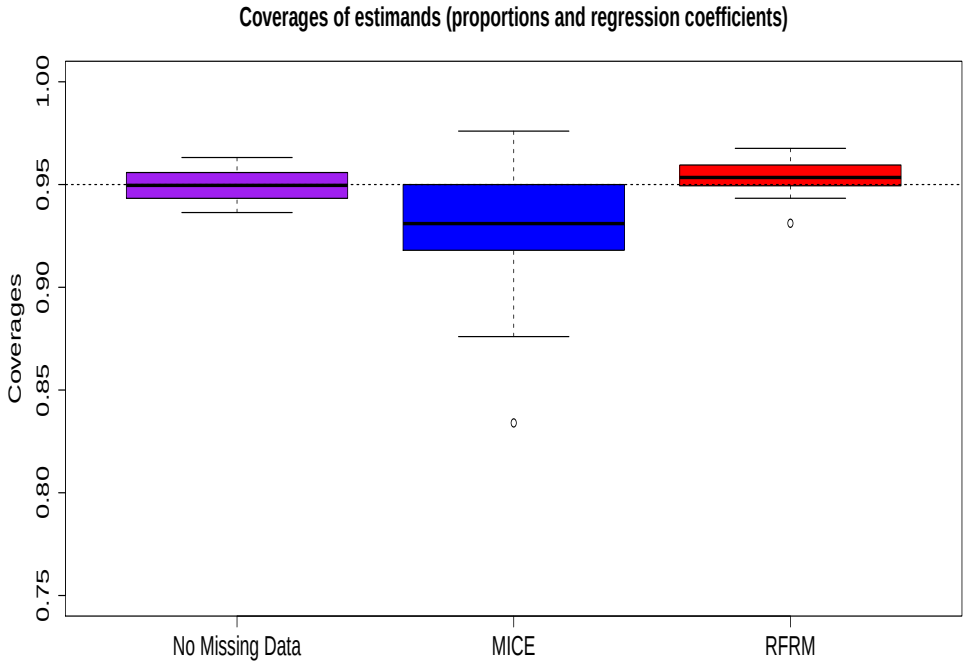


Figure 5.3: The coverages of estimands.

To determine whether these low coverages seen from using MICE are due to a result in biases in the estimates, we plot the absolute biases of estimands obtained from using MICE against the biases obtained from using RFRM. We distinguish between estimates arising from univariate analyses and estimates arising from multivariate analyses, i.e. regression coefficient estimates. In Figure 5.4, we see that the majority of the large biases are those arising from multivariate analyses, and are above the $y = x$ line, which indicates that MICE tends to obtain regression coefficient estimates further from the true values than RFRM.

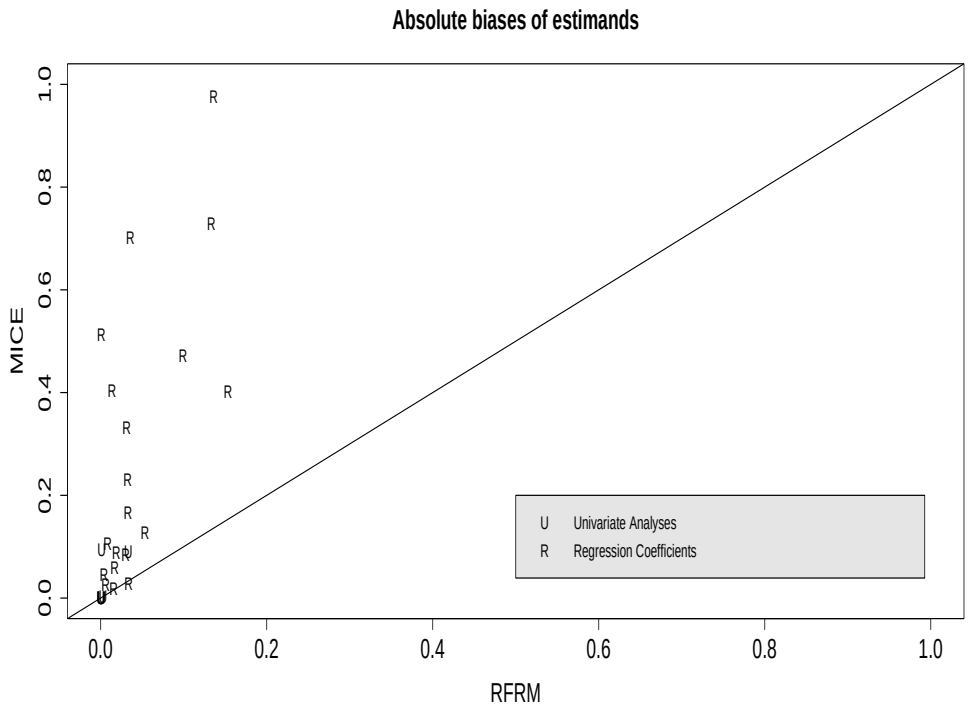


Figure 5.4: Absolute biases of estimands.

5.1.2 Simulation involving the complete cases using FRM, RFRM and MICE

We now apply FRM, RFRM and MICE to impute missing values in the breast-feeding data set taken from Chapter 3. Before applying FRM, RFRM and MICE to impute missing values in the full sample, we apply the imputation models to a simulation involving the complete case subsample. We reintroduce missing data patterns in the complete case subsample using the same fractions that appeared in the full sample. We can then run FRM, RFRM and MICE to multiply impute the missing values in this incomplete subsample, perform the analysis using the imputed data, and compare the results to the analysis performed on the subsample prior to introducing missing values.

Table 5.1 presents estimates of the regression coefficients and variances obtained from fitting the regression model of Peabody individual assessment math score (PI-ATM) on treatment and other covariates. The first column presents estimates prior to introducing missing values, the second column presents results based on the FRM, the third column presents results from using RFRM, while the final column presents results from using MICE. From an analysts point of view, the key estimate is the coefficient on treatment, which gives an estimate of the treatment effect; ideally the estimate from the imputed data should be close to that obtained from the fully observed data. We see that the bias in the treatment effect estimate from using RFRM is about 9.63% when compared to the fully observed estimate, while the bias from using FRM is 20.7% and from using MICE is 59.9%. Hence, in terms of estimating the treatment effect here, there is a potential benefit from using RFRM over FRM and MICE.

Table 5.1: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	No missing data	FRM	RFRM	MICE
Intercept.	86.344 (3.386)	85.430 (6.061)	85.489(5.160)	86.754 (5.167)
Treatment effect.	1.609 (0.948)	1.275 (1.586)	1.454(1.698)	2.573 (1.248)
The number of years between				
1979 and when the mother gave birth.	-0.025 (0.099)	0.001 (0.182)	-0.012 (0.143)	-0.062 (0.127)
The child's race-Black	-1.244 (1.117)	-0.933 (2.078)	-0.688 (1.481)	-0.595 (1.357)
The child's race-Other	2.346 (0.971)	4.159 (1.736)	4.261 (1.447)	4.252 (1.269)
Spouse present at birth.	-1.595 (1.438)	-1.860 (2.879)	-1.645 (1.970)	-1.384 (1.825)
Partner present at birth.	-0.159 (1.104)	0.177 (1.818)	0.114 (1.492)	-0.006 (1.752)
Family income.	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Child's sex.	0.811 (0.676)	0.130 (1.176)	-0.134(0.983)	0.049 (0.934)
Grandparents were present at birth.	-1.078 (1.151)	-1.175 (1.712)	-1.865(1.712)	-0.188 (1.465)
Mother's intelligence.	0.081 (0.018)	0.057 (0.032)	0.051(0.026)	0.048 (0.026)
Mother's highest educational.	0.606 (0.212)	0.628 (0.355)	0.636 (0.304)	0.610 (0.353)
Child's birth weight.	0.023 (0.021)	0.026 (0.032)	0.026(0.032)	0.018 (0.025)
Days that the child spent in hospital.	-0.045 (0.052)	-0.047 (0.089)	-0.058(0.065)	-0.055(0.063)
Days that the mother spent in hospital.	-0.283 (0.119)	-0.387 (0.214)	-0.301(0.167)	-0.431(0.156)
Weeks that the mother worked-level 2.	1.190 (0.984)	2.194 (1.556)	1.681(1.321)	2.250 (1.483)
Weeks that the mother worked-level 3.	0.038 (1.261)	1.653 (2.083)	1.679(1.803)	1.437(2.275)
Weeks that the mother worked-level 4.	2.156 (1.102)	3.392 (2.003)	3.779(1.855)	4.708(1.685)
Child was born premature-level 2.	1.265 (0.919)	-0.233 (1.531)	-0.032(1.282)	-0.155(1.180)
Child was born premature-level 3.	2.009 (2.248)	-0.226 (3.718)	0.361(3.000)	-0.446(2.605)

Figure 5.5 shows the plot of the absolute biases of the estimands (regression coefficients obtained from fitting the regression model of PI-ATM on treatment and other covariates) obtained using FRM, again when compared to the fully observed estimates, against similar biases obtained using RFRM. We see that the treatment effect, labelled as "T" lies above the $y = x$ line, indicating that FRM showed larger absolute bias in estimating the treatment effect than the one obtained using RFRM.

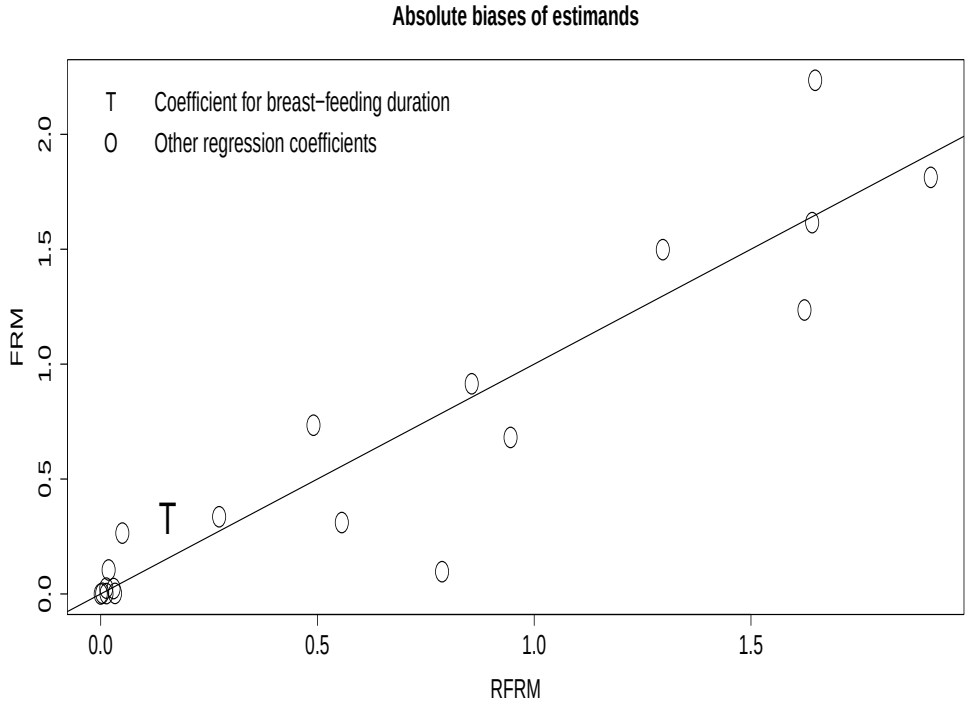


Figure 5.5: Absolute biases of estimands.

Figure 5.6 shows the plot of the absolute biases of the estimands (regression coefficients obtained from fitting the regression model of PI-ATM on treatment and other covariates) obtained using MICE, again when compared to the fully observed estimates, against similar biases obtained using RFRM. We see that the treatment effect, labelled as "T" lies above the $y = x$ line, indicating that MICE showed larger absolute bias in estimating the treatment effect than the one obtained using RFRM.

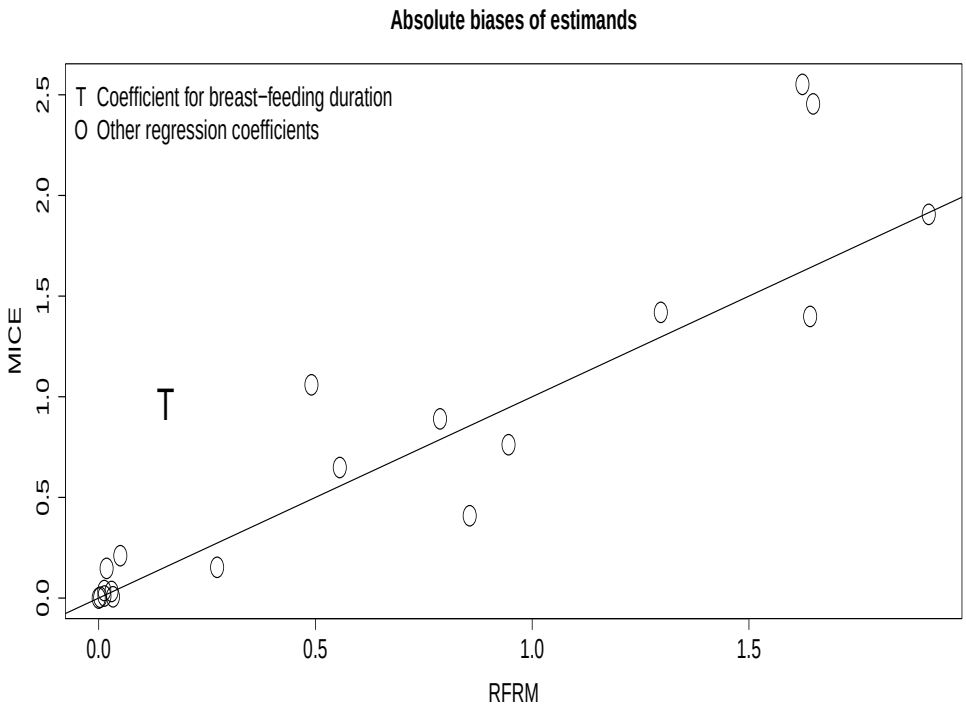


Figure 5.6: Absolute biases of estimands.

5.1.3 Application to the full data sample using FRM, RFRM and MICE

We now apply FRM, RFRM and MICE to the full data sample. For FRM we run $m = 50$ Gibbs samplers at different starting points, each for 20000 iterations to generate 50 imputed data sets, we also use RFRM and MICE to generate 50 imputed data sets. Table 5.2 presents coefficient estimates from the regression model described previously using both sets of imputed data. We see that there is no real significant difference in the treatment effect estimates between FRM and RFRM as they differ by approximately 0.069 (standard errors from using FRM and RFRM are 0.727 and 0.697 respectively). While there is difference in the treatment effect estimates between MICE and RFRM as they differ by approximately 0.333 (standard errors from using MICE and RFRM are 0.850 and 0.697 respectively).

Table 5.2: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	FRM	RFRM	MICE
Intercept.	85.222 (2.578)	85.544 (2.576)	85.395 (2.609)
Treatment effect.	0.903 (0.727)	0.834 (0.697)	1.167 (0.850)
The number of years between			
1979 and when the mother gave birth.	0.041 (0.071)	0.056 (0.075)	0.029 (0.094)
The child's race-Black	-0.549 (0.776)	-0.683(0.855)	-0.527 (0.710)
The child's race-Other	3.093 (0.726)	2.925 (0.749)	2.885 (0.691)
Spouse present at birth.	-0.101 (1.227)	-0.193 (1.139)	-0.078 (1.075)
Partner present at birth.	0.294 (0.824)	0.769 (0.785)	0.172 (0.747)
Family income.	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Child's sex.	0.740 (0.468)	0.868 (0.502)	0.867 (0.531)
Grandparents were present at birth.	-0.537 (0.806)	-1.142 (0.757)	-0.276 (0.581)
Mother's intelligence.	0.108 (0.014)	0.111 (0.014)	0.105 (0.012)
Mother's highest educational.	0.580 (0.162)	0.570 (0.154)	0.562 (0.144)
Child's birth weight.	0.013 (0.016)	0.014(0.016)	0.016 (0.014)
Days that the child spent in hospital.	-0.055 (0.046)	-0.058(0.044)	-0.062 (0.049)
Days that the mother spent in hospital.	-0.104 (0.079)	-0.101(0.085)	-0.155 (0.084)
Weeks that the mother worked-level 2.	0.960 (0.704)	0.403(0.670)	1.055 (0.806)
Weeks that the mother worked-level 3.	1.040 (0.925)	0.435 (0.867)	1.138 (1.080)
Weeks that the mother worked-level 4.	1.898 (0.871)	1.167(0.799)	2.182 (1.015)
Child was born premature-level 2.	0.811 (0.659)	0.782(0.636)	0.913 (0.674)
Child was born premature-level 3.	0.958 (1.372)	0.834(1.514)	1.176 (1.804)

5.1.4 Summary

The robust model RFRM seems to perform better than FRM and MICE in the simulation study and the complete cases analysis. In both the simulation study and real data application, we set the shape and rate values for the gamma prior equal to 10, which results in a marginal t distribution with degrees of freedom $\nu = 20$ for the priors and the errors of the linear regression models. We have increased the degrees of freedom ν to 100 in the RFRM approach and the results obtained were similar to the FRM approach. The reason is that assumptions of normally distributed errors can be viewed as a special case within this framework by letting ν large enough. In the latent variable representation, instead of modelling the variance using a t-prior, we set the variance equal to 1 for identifiability. We are not sure whether this will affect the result of our analysis, as by setting the variance equal to 1, the latent variable model is not capturing the heavy tailed features of the variable.

5.2 Lasso factored regression model (LFRM)

We now present the full conditional distributions in the Metropolis within Gibbs sampler required to generate imputed datasets using LFRM. We first express the joint posterior distribution of all unknowns; these comprise model parameters Θ, and the missing and latent values introduced through the data augmentation, denote the set of all unobserved data as $\mathbf{X}_{unobs}$. The joint posterior can then be expressed as,

$$p(\Theta, \mathbf{X}_{unobs}|\mathbf{X}_{obs}) \propto \prod_{i=1}^{n} \left\{ p(x_{i,1}|\boldsymbol{\theta}^{(1)})p(\boldsymbol{\theta}^{(1)}) \prod_{j=2}^{p} p(x_{i,j}|x_{i,1}, \ldots, x_{i,j-1}, \boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)}) \right\}.$$

We assume that there are k nominal variables and we then order the variables so that $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are nominal, each $\mathbf{x}_q, \quad q = 1, \ldots, k$ taking a set of possible values $1, \ldots, L_q$. Variables $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_p$ could then be measured on a binary, ordinal or continuous scale. We provide expressions for each conditional regression model $p(x_{i,q}|x_{i,1}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)})$. First if q = 1 then,

$$p(x_{i,1}|\boldsymbol{\theta}^{(1)}) = \text{Multinomial}(\pi_1^{(1)}, \ldots, \pi_{L_1}^{(1)}),$$

where $\pi_{j^1}^{(1)} = p(x_{i,1} = j^1|\boldsymbol{\theta}^{(1)}) = \frac{\exp(\beta_{0,j^1}^{(1)})}{\sum_{s=1}^{L_1} \exp(\beta_{0,s}^{(1)})}, \quad j^1 = 1, \ldots, L_1$. We place a Dirichlet prior on $\boldsymbol{\pi}^{(1)}$, i.e.

$$p(\boldsymbol{\pi}^{(1)}) \propto \text{Dirichlet}(\alpha_1, \ldots, \alpha_{L_1}),$$

where we set all $\alpha_{j^1} = 0.5$, $j^1 = 1, \ldots, L_1$ (corresponding to the Jeffreys prior), another

common choice could be to set all $\alpha_{j^1} = 0$. For $q = 2, \ldots, k$,

$$p(x_{i,q}|x_{i,1}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \text{Multinomial}(\pi_{i,1}^{(q)}, \ldots, \pi_{i,L_q}^{(q)}),$$

where

$$\pi_{i,j^q}^{(q)} = \frac{\exp\left(\beta_{0,j^q}^{(q)} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b} \beta_{b,j^q}^{(q),j^b} I(x_{i,b} = j^b)\right)}{\sum_{u=1}^{L_q} \exp\left(\beta_{0,u}^{(q)} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b} \beta_{b,u}^{(q),j^b} I(x_{i,b} = j^b)\right)},$$

where $j^q = 1, \ldots, L_q$. Let $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}_1^{(q)}, \ldots, \boldsymbol{\theta}_{L_q}^{(q)})$
with $\boldsymbol{\theta}_{j^q}^{(q)} = (\beta_{0,j^q}^{(q)}, \beta_{1,j^q}^{(q),2}, \ldots, \beta_{1,j^q}^{(q),L_2}, \ldots, \beta_{q-1,j^q}^{(q),2}, \ldots, \beta_{q-1,j^q}^{(q),L_{q-1}})$ for $j^q = 1, \ldots, L_q$ and we set $\boldsymbol{\theta}_1^{(q)} = \mathbf{0}$ for identifiability . We specify the prior for each $\boldsymbol{\theta}_{j^q}^{(q)}, j^q = 2, \ldots, L_q$ by

$$p(\boldsymbol{\theta}^{(q)}) = \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}).$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with large entries.

For $\mathbf{x}_q,\ q \in \{k+1, \ldots, p\}$ is continuous (for now assuming $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{q-1}$ are continuous),

$$p(x_{i,q}|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = N(\beta_0^{(q)} + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b} \beta_b^{(q),j^b} I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1} \beta_b^{(q)} x_{i,b}, \phi_q^{-1}).$$

We specify a prior for $\boldsymbol{\theta}^{(q)}$,

$$\begin{aligned} p(\boldsymbol{\theta}^{(q)}) &= p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)}) \\ &= p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)}) p(\boldsymbol{\eta}_q|\lambda_q^2, \phi_q, \beta_0^{(q)}) p(\lambda_q^2|\phi_q, \beta_0^{(q)}) p(\phi_q|\beta_0^{(q)}) p(\beta_0^{(q)}), \end{aligned}$$

where $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$ and
$\boldsymbol{\eta}_q = (\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2)'$. The prior distribution for $\boldsymbol{\beta}^{(q)}$ follows a multivariate normal distibution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_q$,

$$p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \phi_q) \sim \text{MVN}(\mathbf{0}, \frac{1}{\phi_q}\mathbf{D}_\eta^{(q)}), \text{ with } \mathbf{D}_\eta^{(q)} = \text{diag}(\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2), \tag{5.7}$$

with the distribution of $\eta_{j,q}^2,\ j = 1, \ldots, q-1$ is given by

$$p(\eta_{j,q}^2|\lambda_q) = \frac{\lambda_q^2}{2} \exp\left(\frac{-\lambda_q^2 \eta_{j,q}^2}{2}\right),\ j = 1, \ldots, q-1, \tag{5.8}$$

and this is the conditional Laplace prior for $\boldsymbol{\beta}^{(q)}$ suggested by Park and Casella (2008).

We consider a diffuse hyperprior on λ_q^2 of the form

$$p(\lambda_q^2) = \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right),\ \lambda_q^2 > 0,\ r > 0,\ \delta > 0, \tag{5.9}$$

which is a gamma distribution with shape and rate r and δ respectively. We also specify the prior distributions for ϕ_q and $\beta_0^{(q)}$ as follows:

$$\begin{aligned} p(\phi_q) &\propto \frac{1}{\phi_q}, \\ p(\beta_0^{(q)}) &\propto 1. \end{aligned}$$

For $\mathbf{x}_q,\ q \in \{k+1,\ldots,p\}$ is binary (for now assuming $\mathbf{x}_{k+1},\ldots,\mathbf{x}_{q-1}$ are continuous),

$$\begin{aligned} &x_{i,q} = I(x_{i,q}^* > 0), \quad \text{where} \\ &p(x_{i,q}^*|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \mathrm{N}(\beta_0^{(q)} + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b}\beta_b^{(q),j^b} I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1}\beta_b^{(q)} x_{i,b}, 1). \end{aligned}$$

We specify a prior for $\boldsymbol{\theta}^{(q)}$ by,

$$p(\boldsymbol{\theta}^{(q)}) = p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}) = p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)})p(\boldsymbol{\eta}_q|\lambda_q^2, \beta_0^{(q)})p(\lambda_q^2|\beta_0^{(q)})p(\beta_0^{(q)}),$$

where the prior distributions for $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2$ are given by Equation 5.7, Equation 5.8 and Equation 5.9 respectively. This is the Laplace prior for $\boldsymbol{\beta}^{(q)}$ suggested by Park and Casella (2008). We also specify a prior distribution for $\beta_0^{(q)}$,

$$p(\beta_0^{(q)}) \propto 1.$$

For $\mathbf{x}_q,\ q \in \{k+1,\ldots,p\}$ is ordinal taking values in $\{1,\ldots,J_q\}$ (for now assuming $\mathbf{x}_{k+1},\ldots,\mathbf{x}_{q-1}$ are continuous),

$$\begin{aligned} &x_{i,q} = j^q\ I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}), \quad \text{where} \\ &p(x_{i,q}^*|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = N(\beta_0^{(q)} + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b}\beta_b^{(q),j^b} I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1}\beta_b^{(q)} x_{i,b}, 1), \end{aligned}$$

where the $\gamma_{j^q}^{(q)}$ are threshold values, with $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_{J_q}^{(q)} = \infty$, for $j^q \in \{1, ..., J_q\}$.

We also specify a prior for $\boldsymbol{\theta}^{(q)}$ by,

$$\begin{aligned} p(\boldsymbol{\theta}^{(q)}) &= p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}) \\ &= p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})p(\boldsymbol{\eta}_q|\lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})p(\lambda_q^2|\beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})p(\beta_0^{(q)}|\boldsymbol{\gamma}^{(q)})p(\boldsymbol{\gamma}^{(q)}), \end{aligned}$$

where $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1, \ldots, J_q\}\right\}$. The prior distributions for $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2$ are given by Equation 5.7, Equation 5.8 and Equation 5.9 respectively. This is the Laplace prior for $\boldsymbol{\beta}^{(q)}$ (Park and Casella (2008)). We also specify a prior for $\beta_0^{(q)}$ as follows:

$$p(\beta_0^{(q)}) \propto 1,$$

and we place an improper uniform prior on $\boldsymbol{\gamma}^{(q)}$ i.e.

$$p(\boldsymbol{\gamma}^{(q)}) \propto I(\boldsymbol{\gamma}^{(q)} \in \Omega^{(q)}),$$

where $\Omega^{(q)} = \left\{\gamma_{j^q}^{(q)} : \gamma_0^{(q)} = -\infty < \gamma_1^{(q)} = 0 < \gamma_2^{(q)} < \ldots < \gamma_{J_{q-1}}^{(q)} < \gamma_{J_q}^{(q)} = \infty\right\}$.

If $x_{i,q},\ q \in \{k+1, \ldots, q-1\}$ is not continuous then we replace $x_{i,q}$ with $x_{i,q}^*$ in the model for $p(x_{i,q}|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)})$. For notational convenience when $x_{i,q}$ is nominal or continuous we assume $x_{i,q} = x_{i,q}^*,\ q \in \{1, \ldots, p\}$.

With this model and prior specification we can present the full conditional distribution required to impute missing values in the Metropolis within Gibbs sampler. Following Schafer (1997) we present this as a data augmentation scheme and so first present the "I-steps" followed by the "P-steps".

In the "I-steps", conditional on parameter values, first impute missing nominal values $x_{i,q}^*$ from the following discrete distribution

$$\begin{aligned} &p(x_{i,q}^* = j^q | x_{i,1}^* = j^1, \ldots, x_{i,q-1}^* = j^{q-1}, x_{i,q+1}^* = j^{q+1}, \ldots, x_{i,k}^* = j^k, x_{i,k+1}^*, \ldots, x_{i,p}^*, \Theta) \\ &= \tilde{\pi}_{i,j^q}^{(q)}, \end{aligned}$$

where $j^q \in \{1, \ldots, L_q\}$, and

$$\tilde{\pi}_{i,j^q}^{(q)} = \frac{\displaystyle\prod_{b=q}^{k} \pi_{i,j^b}^{(b)} \prod_{b=k+1}^{p} \exp\left(\frac{-\phi_b}{2}(\tilde{x}_{i,b}^* - \beta_q^{(b),j^q} I(x_{i,q}^* = j^q) - \sum_{t=k+1}^{(b-1)} \beta_t^{(b)} x_{i,t}^*)^2\right)}{\displaystyle\sum_{u=1}^{L_q} \left\{\pi_{i,u}^{(q)} \prod_{b=q+1}^{k} \pi_{i,j^b}^{(b)}\right\} \prod_{b=k+1}^{p} \exp\left(\frac{-\phi_b}{2}(\tilde{x}_{i,b}^* - \beta_q^{(b),j^q} I(x_{i,q}^* = u) - \sum_{t=k+1}^{(b-1)} \beta_t^{(b)} x_{i,t}^*)^2\right)},$$

where

$$\tilde{x}_{i,b}^* = x_{i,b}^* - \beta_0^{(b)} - \sum_{s=1}^{q-1} \sum_{j^s=2}^{L_s} \beta_s^{(b),j^s} I(x_{i,s}^* = j^s) - \sum_{s=q+1}^{k} \sum_{j^s=2}^{L_s} \beta_s^{(b),j^s} I(x_{i,s}^* = j^s).$$

Next impute missing continuous values $x_{i,q}^*$ or latent values $x_{i,q}^*$ with missing $x_{i,q}$ from,

$$p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, x_{i,q+1}^*, \ldots, x_{i,k}^*, x_{i,k+1}^*, \ldots, x_{i,p}^*, \Theta) = \mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_q^{-1}),$$

where

$$\widetilde{\mu}_{i,q} = \widetilde{\phi}_q^{-1} \left\{ \frac{\mu_{i,q}}{\phi_q^{-1}} + \sum_{s=q+1}^{p} \frac{\beta_q^{(s)}}{\phi_s^{-1}} \left[x_{i,s}^* - (\mu_{i,s} - \beta_q^{(s)} x_{i,q}^*) \right] \right\},$$

and

$$\widetilde{\phi}_q^{-1} = \left(\frac{1}{\phi_q^{-1}} + \sum_{s=q+1}^{p} \frac{(\beta_q^{(s)})^2}{\phi_s^{-1}} \right)^{-1}$$

where $\mu_{i,q} = \beta_0^{(q)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \beta_b^{(q),j^b} I(x_{i,b}^* = j^b) + \sum_{b=k+1}^{q-1} \beta_b^{(q)} x_{i,b}^*$ and $\mu_{i,s} = \beta_0^{(s)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \beta_b^{(s),j^b} I(x_{i,b}^* = j^b) + \sum_{b=k+1}^{s-1} \beta_b^{(s)} x_{i,b}^*$.

Next impute latent values $x_{i,q}^*$ with observed ordinal variable $x_{i,q}$ from,

$$p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, x_{i,q+1}^*, \ldots, x_{i,p}^*, \Theta) = \frac{I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) \mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_q^{-1})}{\Phi(\frac{\gamma_{j^q}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_q^{-1}}) - \Phi(\frac{\gamma_{j^q-1}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_q^{-1}})},$$

which is a normal distribution $\mathrm{N}(\widetilde{\mu}_{i,q}, \phi_q^{-1})$ truncated to $(\gamma_{j^q-1}^{(q)}, \gamma_{j^q}^{(q)})$.

Next impute latent values $x_{i,q}^*$ with observed binary variable $x_{i,q}$ from,

$$p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, x_{i,q+1}^*, \ldots, x_{i,p}^*, \Theta) = \frac{I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) \mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_q^{-1})}{\Phi(\frac{\gamma_{j^q}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_q^{-1}}) - \Phi(\frac{\gamma_{j^q-1}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_q^{-1}})},$$

which is a normal distribution $\mathrm{N}(\widetilde{\mu}_{i,q}, \phi_q^{-1})$ truncated to $(\gamma_{j^q-1}^{(q)}, \gamma_{j^q}^{(q)})$ where the $\gamma_{j^q}^{(q)}$ are threshold values, with $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_2^{(q)} = \infty$.

Once we have imputed values for $x_{i,q}^*$ (with missing $x_{i,q}$), we define a function $g(x_{i,q}^*)$ to map each $x_{i,q}^*$ back to its original measurement scale, and thus create an imputed data set. The mapping function is defined as follows,

$$g(x_{i,q}^*) = \begin{cases} I(x_{i,q}^* > 0) & \text{if } x_{i,q}^* \text{ is binary,} \\ j^q \; I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) & \text{if } x_{i,q}^* \text{ is ordinal, } j^q \in \{1, \ldots, J_q\}, \\ x_{i,q}^* & \text{if } x_{i,q}^* \text{ is continuous,} \\ x_{i,q}^* & \text{if } x_{i,q}^* \text{ is nominal, } j^q \in \{1, \ldots, L_q\}, \end{cases}$$

Denote an imputed value for $x_{i,q}$ by $x_{i,q}^{(t)}$ at iteration t, and an imputed data set at iteration t by $\mathbf{X}_{com}^{(t)}$. In the "P-steps", conditional on an imputed data set, first sample values for

$\boldsymbol{\theta}^{(1)}$ from a Dirichlet ($\boldsymbol{\alpha}$) distribution with parameters

$$\boldsymbol{\alpha} = \left(\alpha_1 + \sum_{i=1}^{n} I(x_{i,1} = 1),\ \alpha_2 + \sum_{i=1}^{n} I(x_{i,1} = 2), \ldots, \alpha_{L_1} + \sum_{i=1}^{n} I(x_{i,1} = L_1) \right)'.$$

Next propose values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}_1^{(q)}, \ldots, \boldsymbol{\theta}_{L_q}^{(q)})$
with $\boldsymbol{\theta}_{j^q}^{(q)} = (\beta_{0,j^q}^{(q)}, \beta_{1,j^q}^{(q),2}, \ldots, \beta_{1,j^q}^{(q),L_2}, \ldots, \beta_{q-1,j^q}^{(q),2}, \ldots, \beta_{q-1,j^q}^{(q),L_{q-1}})$, for $j^q = 1, \ldots, L_q$ using a Metropolis-Hastings step, where we set $\boldsymbol{\theta}_1^{(q)} = \mathbf{0}$ for identifiability. We consider a proposal distribution based on the imputed data log-likelihood at each iteration t, $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)})$, which can be expressed as

$$l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)}) = \sum_{w=1}^{L_q} \sum_{i=1}^{n} \ln \left(\frac{\exp\left(\beta_{0,w}^{(q)} + \sum_{b=1}^{q-1} \sum_{j^b=2}^{L_b} \beta_{b,w}^{(q),j^b} f(x_{i,b}^{(t)}, x_{i,b}, m_{i,b}) \right)}{\sum_{u=1}^{L_q} \exp\left(\beta_{0,u}^{(q)} + \sum_{b=1}^{q-1} \sum_{j^b=2}^{L_b} \beta_{b,u}^{(q),j^b} f(x_{i,b}^{(t)}, x_{i,b}, m_{i,b}) \right)} \right)^{I(x_{i,q}=w)},$$

where $f(x_{i,b}^{(t)}, x_{i,b}, m_{i,b}) = (I(x_{i,b}^{(t)} = j^b) m_{i,b} + I(x_{i,b} = j^b)(1 - m_{i,b}))$, with $m_{i,b}$ is a missing data indicator with $m_{i,b} = 1$ indicating $x_{i,b}$ is missing and $m_{i,b} = 0$ indicating $x_{i,b}$ is observed. We generate proposals for $\boldsymbol{\theta}^{(q)}$ from a $\mathrm{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ distribution, and $\boldsymbol{\mu}^{(t)}$ is the value that maximises $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)})$ and $\boldsymbol{\Sigma}^{(t)}$ is given by $-E[\frac{\delta^2}{\delta\beta_{j,w}^{(q)} \delta\beta_{\tilde{j},\tilde{w}}^{(q)}} l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)})]^{-1}_{\boldsymbol{\theta}^{(q)} = \boldsymbol{\mu}^{(t)}}$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)})$, $q \in \{k+1, \ldots, p\}$ when $\mathbf{x}_q^*$ is continuous from the joint posterior distribution of $\boldsymbol{\beta}^{(q)} =$
$(\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$, $\boldsymbol{\eta}_q = (\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2)'$, λ_q^2, ϕ_q and $\beta_0^{(q)}$, which is given by

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)} | \mathbf{X}_{com}^{(t)}) &= p(\boldsymbol{\beta}^{(q)} | \boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) p(\boldsymbol{\eta}_q | \lambda_q^2, \phi_q, \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) \\ &\quad p(\lambda_q^2 | \phi_q, \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) p(\phi_q | \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) p(\beta_0^{(q)} | \mathbf{X}_{com}^{(t)}). \end{aligned}$$

Now we sample $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)}$ from their full conditional distributions in the following way:

1. We first sample $\boldsymbol{\beta}^{(q)}$ from

$$p(\boldsymbol{\beta}^{(q)} | \phi_q, \boldsymbol{\eta}_q, \widetilde{\mathbf{X}}_q', \widetilde{\mathbf{x}}_q^*) = \mathrm{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}),$$

with

$$\hat{\boldsymbol{\beta}}^{(q)} = \phi_q \mathbf{V}^{(q)} \widetilde{\mathbf{X}}_q' \widetilde{\mathbf{x}}_q^*,$$

and

$$\mathbf{V}^{(q)} = \frac{1}{\phi_q}\left(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q + \mathbf{D}_\eta^{(q)}\right)^{-1},$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*, \ldots, \widetilde{x}_{n,q}^*)'$, with $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\eta^{(q)}$ is a $(q-1) \times (q-1)$ diagonal matrix with entries $\frac{1}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

2. We then sample values for $\eta_{j,q}^2,\ j = 1, \ldots, q-1$ from

$$p(\eta_{j,q}^2 | \boldsymbol{\beta}^{(q)}, \lambda_q^2, \phi_q), \quad j = 1, \ldots, q-1$$

which is an inverse-Gaussian distribution with parameters

$$\mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2 \phi_q}} \ \text{ and } \ \lambda_q' = \lambda_q^2, \quad j = 1, \ldots, q-1$$

in the parameterization of the inverse-Gaussian density given by

$$f(g) = \sqrt{\frac{\lambda_q'}{2\pi}} g^{-\frac{3}{2}} \exp\left\{-\frac{\lambda_q'(g-\mu_q')^2}{2(\mu_q')^2 g}\right\},\ g = \frac{1}{\eta_{j,q}^2},\ g > 0.$$

3. Next sample a value for λ_q^2 from

$$p(\lambda_q^2 | \boldsymbol{\eta}_q) \sim \text{Gamma}\left((q-1) + r, \sum_{j=1}^{q-1} \frac{\eta_{j,q}^2}{2} + \delta\right).$$

4. We then sample a value for ϕ_q from

$$p(\phi_q | \boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \widetilde{\mathbf{x}}_q^*, \widetilde{\mathbf{X}}_q) = \text{Gamma}\left(\frac{n + (q-1)}{2}, \frac{(\widetilde{\mathbf{x}}_q^* - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{x}}_q^* - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)}) + (\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right),$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*, \ldots, \widetilde{x}_{n,q}^*)'$, with $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\eta^{(q)}$ is a $(q-1) \times (q-1)$ diagonal matrix with entries $\frac{1}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

5. We sample a value for $\beta_0^{(q)}$ from

$$p(\beta_0^{(q)} | \phi_q, \boldsymbol{\beta}^{(q)}, x_{i,q}^*, \widetilde{\mathbf{x}}_{i,q}) \sim \text{N}(\frac{\sum_{i=1}^n (x_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}, \frac{\phi_q^{-1}}{n}).$$

where
$\widetilde{\mathbf{x}}_{i,q} = (I(x^*_{i,1} = 2), \ldots, I(x^*_{i,1} = L_1), \ldots, I(x^*_{i,k} = 2), \ldots, I(x^*_{i,k} = L_k), x^*_{i,k+1}, \ldots, x^*_{i,q-1})$ for $i = 1, \ldots, n$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \beta^{(q)}_0),\ q \in \{k+1, \ldots, p\}$ when $\mathbf{x}_q$ is binary from the joint posterior distribution of $\boldsymbol{\beta}^{(q)} = (\beta^{(q),2}_1, \ldots, \beta^{(q),L_1}_1, \ldots, \beta^{(q),2}_k, \ldots, \beta^{(q),L_k}_k, \beta^{(q)}_{k+1}, \ldots, \beta^{(q)}_{q-1})'$, $\boldsymbol{\eta}_q = (\eta^2_{1,q}, \ldots, \eta^2_{q-1,q})'$, λ^2_q, and $\beta^{(q)}_0$, which is given by

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \beta^{(q)}_0 | \mathbf{X}^{(t)}_{com}) &= p(\boldsymbol{\beta}^{(q)} | \boldsymbol{\eta}_q, \lambda^2_q, \beta^{(q)}_0, \mathbf{X}^{(t)}_{com}) p(\boldsymbol{\eta}_q | \lambda^2_q, \beta^{(q)}_0, \mathbf{X}^{(t)}_{com}) p(\lambda^2_q | \beta^{(q)}_0, \mathbf{X}^{(t)}_{com}) \\ &\quad p(\beta^{(q)}_0 | \mathbf{X}^{(t)}_{com}), \end{aligned}$$

Now we sample $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \beta^{(q)}_0$ from their full conditional distributions in the following way:

1. We first sample $\boldsymbol{\beta}^{(q)}$ from

$$p(\boldsymbol{\beta}^{(q)} | \boldsymbol{\eta}_q, \widetilde{\mathbf{X}}'_q, \widetilde{\mathbf{x}}^*_q) = \text{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}), \tag{5.10}$$

with

$$\hat{\boldsymbol{\beta}}^{(q)} = \mathbf{V}^{(q)} \widetilde{\mathbf{X}}'_q \widetilde{\mathbf{x}}^*_q,$$

and

$$\mathbf{V}^{(q)} = \left(\widetilde{\mathbf{X}}'_q \widetilde{\mathbf{X}}_q + \mathbf{D}^{(q)}_\eta \right)^{-1},$$

where $\widetilde{\mathbf{x}}^*_q = (\widetilde{x}^*_{1,q}, \ldots, \widetilde{x}^*_{n,q})'$, with $\widetilde{x}^*_{i,q} = x^*_{i,q} - \beta^{(q)}_0$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x^*_{i,1} = 2), \ldots, I(x^*_{i,1} = L_1), \ldots, I(x^*_{i,k} = 2), \ldots, I(x^*_{i,k} = L_k), x^*_{i,k+1}, \ldots, x^*_{i,q-1})$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}^{(q)}_\eta$ is a $(q-1) \times (q-1)$ diagonal matrix with entries $\frac{1}{\eta^2_{j,q}}$ for $j = 1, \ldots, q-1$.

2. We then sample values for $\eta^2_{j,q},\ j = 1, \ldots, q-1$ from

$$p(\eta^2_{j,q} | \boldsymbol{\beta}^{(q)}, \lambda^2_q), \quad j = 1, \ldots, q-1$$

which is an inverse-Gaussian distribution with parameters

$$\mu'_q = \sqrt{\frac{\lambda^2_q}{(\beta^{(q)}_j)^2}} \ \text{ and } \ \lambda'_q = \lambda^2_q, \quad j = 1, \ldots, q-1$$

in the parameterization of the inverse-Gaussian density given by

$$f(g) = \sqrt{\frac{\lambda'_q}{2\pi}} g^{-\frac{3}{2}} \exp\left\{ -\frac{\lambda'_q (g - \mu'_q)^2}{2(\mu'_q)^2 g} \right\},\ g = \frac{1}{\eta^2_{j,q}},\ g > 0. \tag{5.11}$$

3. Next sample a value for λ_q^2 from

$$p(\lambda_q^2|\boldsymbol{\eta}_q) \sim \text{Gamma}\left((q-1)+r, \sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2}+\delta\right). \tag{5.12}$$

4. We sample a value for $\beta_0^{(q)}$ from

$$p(\beta_0^{(q)}|\boldsymbol{\beta}^{(q)}, x_{i,q}^*, \widetilde{\mathbf{x}}_{i,q}) \sim \text{N}(\frac{\sum_{i=1}^n (x_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}, \frac{1}{n}), \tag{5.13}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(I(x_{i,1}^*=2),\ldots,I(x_{i,1}^*=L_1),\ldots,I(x_{i,k}^*=2),\ldots,I(x_{i,k}^*=L_k),x_{i,k+1}^*,\ldots,x_{i,q-1}^*)$, for $i=1,\ldots,n$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})$, $q \in \{k+1,\ldots,p\}$ when $\mathbf{x}_q$ is ordinal from the joint posterior distribution of $\boldsymbol{\beta}^{(q)} =$
$(\beta_1^{(q),2},\ldots,\beta_1^{(q),L_1},\ldots,\beta_k^{(q),2},\ldots,\beta_k^{(q),L_k},\beta_{k+1}^{(q)},\ldots,\beta_{q-1}^{(q)})'$, $\boldsymbol{\eta}_q = (\eta_{1,q}^2,\ldots,\eta_{q-1,q}^2)'$, λ_q^2, $\beta_0^{(q)}$ and $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1,\ldots,J_q\}\right\}$, which is given by

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}|\mathbf{X}_{com}^{(t)}) &= p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) p(\boldsymbol{\eta}_q|\lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) \\ &\quad p(\lambda_q^2|\beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) p(\beta_0^{(q)}|\boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) p(\boldsymbol{\gamma}^{(q)}|\mathbf{X}_{com}^{(t)}), \end{aligned}$$

Now we sample $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}$ from their full conditional distributions in the following way:

1. We first sample $\boldsymbol{\beta}^{(q)}$ from its full conditional distribution which is given by Equation 5.10

2. We then sample values for $\eta_{j,q}^2$, $j=1,\ldots,q-1$. from its full conditional distribution which is given by Equation 5.11.

3. Next sample a value for λ_q^2 from its full conditional distribution which is given by Equation 5.12.

4. We sample a value for $\beta_0^{(q)}$ from its full conditional distribution which is given by Equation 5.13.

5. We then update the threshold values $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1,\ldots,J_q\}\right\}$. The full conditional distribution of the threshold values $\gamma_{j^q}^{(q)}$ for $j^q \in \{1,\ldots,J_q\}$ is uniformly distributed on the interval

$$\left[\max\left\{\max\left\{x_{i,q}^* : x_{i,q} = j^q\right\}, \gamma_{j^q-1}^{(q)}\right\}, \min\left\{\min\left\{x_{i,q}^* : x_{i,q} = j^q+1\right\}, \gamma_{j^q+1}^{(q)}\right\}\right],$$

where $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_{J_q}^{(q)} = \infty$, for $j^q \in \{1, \ldots, J_q\}$.

In this way we have sampled all unknowns (missing values and parameters) from their full conditional distributions. The details on deriving the full conditional distributions of the missing values in the "I-steps", the full conditional distribution of $\boldsymbol{\theta}^{(1)}$ where $\mathbf{x}_1$ follows a multinomial distribution, as well as the Metropolis-Hastings update for the multinomial logistic regression model in the "P-steps" have been presented in the LFRM modelling strategy. For details on deriving the full conditional distributions of the remaining parameters in the "P-steps", such as $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q, \phi_q, \beta_0^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is continuous, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q, \beta_0^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is binary) and $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is ordinal), please see Appendix E.

5.2.1 Simulation study

We now illustrate the performances of FRM, LFRM and MICE through a simulation study. In the MICE approach, we use the model and prior specifications described in the LFRM strategy to generate missing values. We are interested to compare the results between FRM and the Bayesian Lasso modelling strategy (LFRM), as well as the results between the joint modelling specification with Bayesian Lasso (LFRM) with the equivalent fully conditional specification (MICE).

We simulate data sets that contain variables measured on binary, ordinal, continuous and nominal (greater than two levels) scales. Specifically each data set contains 5 binary variables, 5 ordinal variables, 10 continuous variables, and 5 nominal variables. Variables are simulated in a sequential manner with each variable conditional on a subset of variables already generated. We then introduce missing values into nine of the variables using the MAR mechanism, so that each incomplete variable has approximately 30% missing values. Specific details of how we simulated the incomplete dataset are given in Appendix F. We replicate this data generation process 500 times to generate 500 incomplete data sets.

Using the LFRM and MICE imputation strategies, we multiply impute the missing values in each incomplete data set $m = 10$ times. To generate m independent imputed data sets, we run m Metropolis within Gibbs samplers, each with a different starting value, with each sampler resulting in an imputed data set after convergence. We assume analysts may be interested in making inferences about various types of estimands arising from both univariate analyses, e.g. population means of variables or the proportions in the population taking a particular level of a categorical variable, as well as multivariate analyses, e.g. the coefficients from a regression model. Using the m imputed data sets, we apply the combining rules described in Equation 2.1 to construct point and interval estimates for these estimands. We also construct estimates for the same estimands when using FRM to multiply impute the missing values.

When implementing MICE, LFRM or FRM for imputations, we assume the imputer knows the data generating process and so decomposes the joint distribution of the imputation model as described in Equation 3.3. We then performed analysis on the imputed

data sets and obtained estimands arising from both univariate analyses and multivariate analyses. See Appendix F for a full list of the estimands considered.

Figure 5.7 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by FRM and LFRM over the 500 data sets. These estimands include those arising from both univariate analyses and regression analyses. The first box plot presents coverages when there are no missing data in the datasets. These coverage are as expected the closest to 0.95. The second box plot shows the coverages from FRM while the third box plot shows the coverages obtained from LFRM. We noticed that LFRM showed some low coverages in the plot. These low coverages are the regression coefficients estimates from the regression model $p(\mathbf{x}_{25}|\mathbf{x}_1, \ldots, \mathbf{x}_{24})$ and the lowest coverage is the intercept estimate of this regression model.

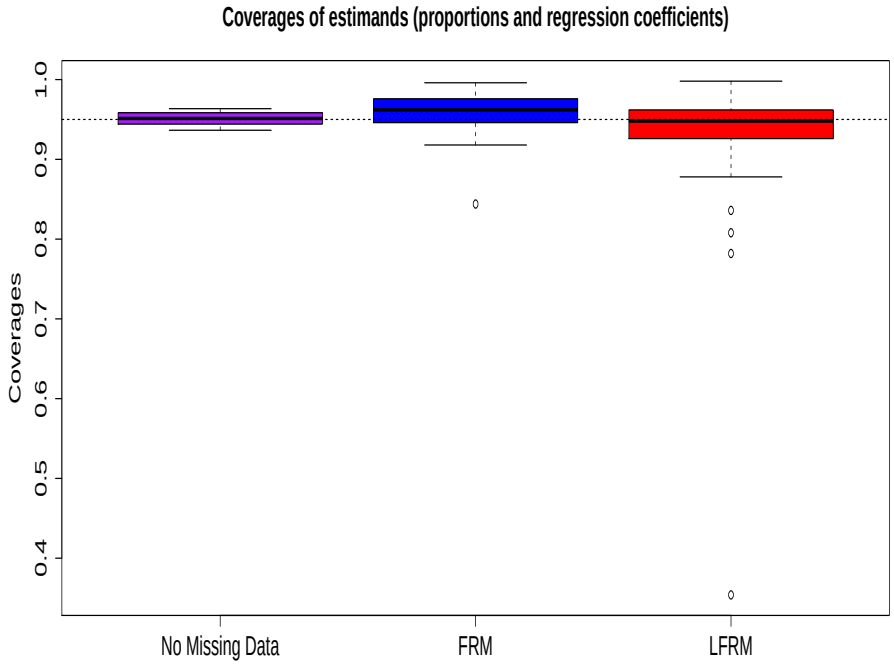


Figure 5.7: The coverages of estimands.

To determine whether these low coverages seen from using LFRM are due to a result in biases in the estimates, we plot the biases in the estimates obtained from using FRM against the biases obtained from using LFRM. We distinguish between estimates arising from univariate analyses and estimates arising from multivariate analyses, i.e. regression coefficient estimates. We see that the majority of the large biases are those arising from multivariate analyses, and are below the $y = x$ line, which indicates that LFRM tends to obtain regression coefficient estimates further from the true values than FRM. We noticed that LFRM shows a large absolute bias of a regression coefficient, denoted by "R" (on the right hand side of Figure 5.8). This estimand is the intercept estimate of the regression model $p(\mathbf{x}_{25}|\mathbf{x}_1, \ldots, \mathbf{x}_{24})$, which also shows the lowest coverage in Figure 5.7.

We now compare the results between LFRM and MICE. Figure 5.9 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by MICE and LFRM over the 500 data sets. These estimands

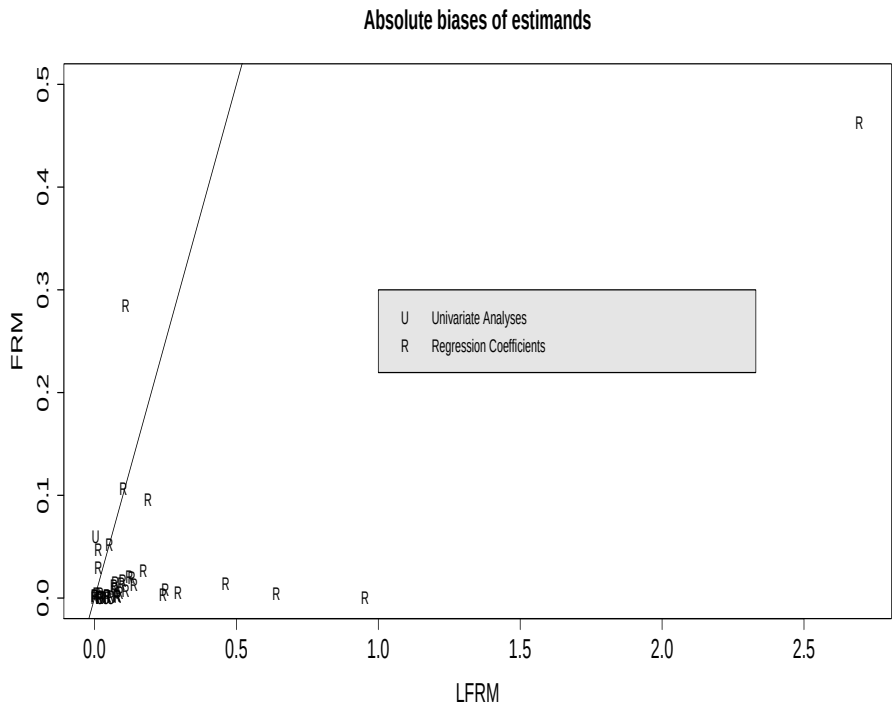


Figure 5.8: Absolute biases of estimands.

include those arising from both univariate analyses and regression analyses. The first box plot presents coverages when there are no missing data in the datasets. These coverage are as expected the closest to 0.95. The second box plot shows the coverages from MICE while the third box plot shows the coverages obtained from LFRM. We noticed that both MICE and LFRM dont't show any significant differences in the coverages, as they both have some low coverages in the plot. The lowest coverage from both plot is the intercept estimate of the regression model $p(\mathbf{x}_{25}|\mathbf{x}_1, \ldots, \mathbf{x}_{24})$.

We now plot the biases in the estimates obtained from using MICE against the biases obtained from using LFRM. We distinguish between estimates arising from univariate analyses and estimates arising from multivariate analyses, i.e. regression coefficient estimates. We see that the majority of the large biases are those arising from multivariate analyses, and are below the $y = x$ line, which indicates that LFRM tends to obtain regression coefficient estimates further from the true values than MICE. We noticed that both MICE and LFRM show a large absolute bias of a regression coefficient, denoted by "R" (on the right hand side of Figure 5.10). This estimand is the intercept estimate of the regression model $p(\mathbf{x}_{25}|\mathbf{x}_1, \ldots, \mathbf{x}_{24})$, which also shows the lowest coverage in Figure 5.9.

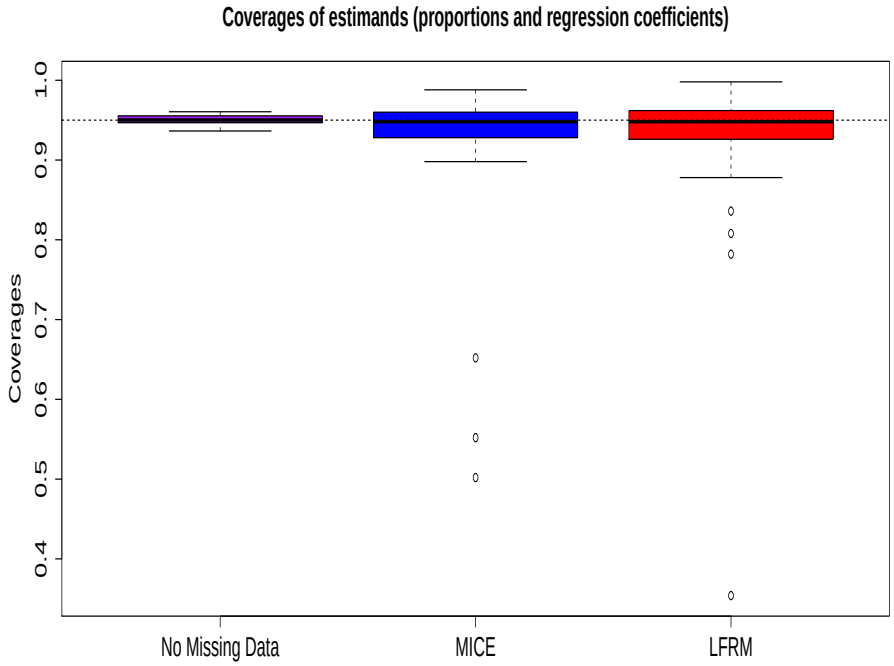


Figure 5.9: The coverages of estimands.

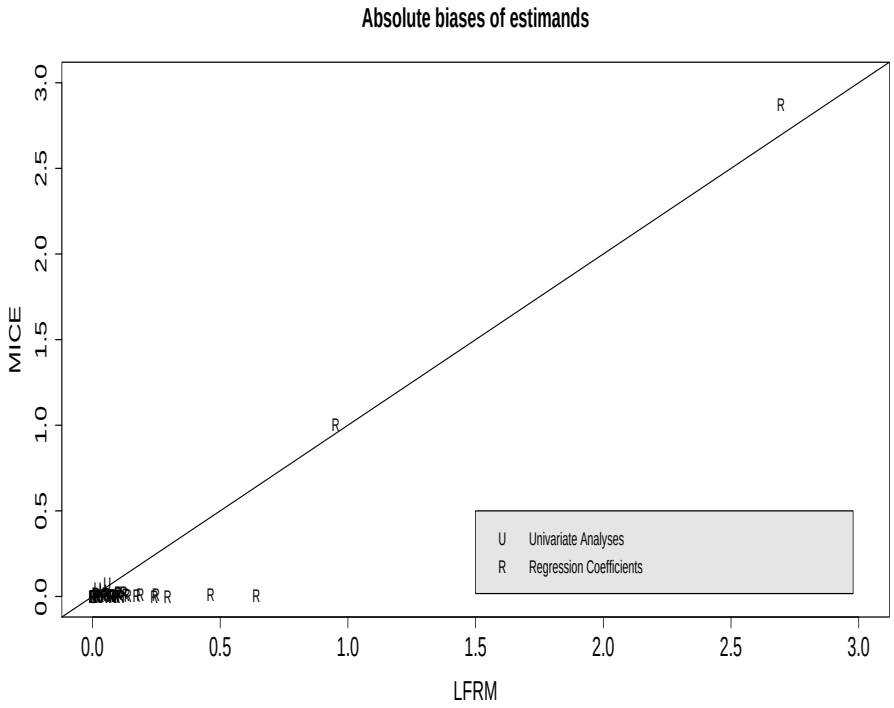


Figure 5.10: Absolute biases of estimands.

5.2.2 Simulation involving the complete cases using FRM, LFRM and MICE

We now apply FRM, LFRM and MICE to impute missing values in the breast-feeding data set taken from Chapter 3. Before applying FRM, LFRM and MICE to impute missing values in the full sample, we apply the imputation models to a simulation involving the complete case subsample. We reintroduce missing data patterns in the complete case subsample using the same fractions that appeared in the full sample. We can then run FRM, LFRM and MICE to multiply impute the missing values in this incomplete subsample, perform the analysis using the imputed data, and compare the results to the analysis performed on the subsample prior to introducing missing values.

Table 5.3 presents estimates of the regression coefficients and variances obtained from fitting the regression model of Peabody individual assessment math score (PI-ATM) on treatment and other covariates. The first column presents estimates prior to introducing missing values, the second column presents results based on FRM, the third column presents results from LFRM while the final column presents results from MICE. From an analysts point of view, the key estimate is the coefficient on treatment, which gives an estimate of the treatment effect; ideally the estimate from the incomplete data should be close to that obtained from the fully observed data. We see that the bias in the treatment effect estimate from using LFRM is about 13.4% when compared to the fully observed estimate, while the bias from using FRM is 20.7% and from using MICE is 3.418%. Hence, in terms of estimating the treatment effect here, there is potential benefit from using MICE over LFRM and FRM.

In Figure 5.11 we also plot the absolute biases in the regression coefficient estimates from using FRM, again when compared to the fully observed estimates, against similar biases from using LFRM. We see that in general using FRM for imputation has no significant potential to obtain estimates closer to those obtained from the fully observed data than using LFRM. We noticed that LFRM shows a large absolute bias of a regression coefficient (on the right hand side of Figure 5.11), this estimand is the intercept estimate of the regression model of PI-ATM.

In Figure 5.12 we also plot the absolute biases in the regression coefficient estimates from using MICE, again when compared to the fully observed estimates, against similar biases from using LFRM. We see that in general using LFRM for imputation has no significant potential to obtain estimates closer to those obtained from the fully observed data than using MICE. We noticed that LFRM shows a large absolute bias of a regression coefficient (on the right hand side of Figure 5.11), this estimand is the intercept estimate of the regression model of PI-ATM.

Table 5.3: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	No missing data	FRM	LFRM	MICE
Intercept.	86.344 (3.386)	85.430 (6.061)	90.535 (5.182)	89.836 (4.769)
Treatment effect.	1.609 (0.948)	1.275 (1.586)	1.393 (1.254)	1.664 (0.982)
The number of years between				
1979 and when the mother gave birth.	-0.025 (0.099)	0.001 (0.182)	0.006 (0.130)	-0.068 (0.113)
The child's race-Black	-1.244 (1.117)	-0.933 (2.078)	-0.876 (1.433)	-0.819 (1.188)
The child's race-Other	2.346 (0.971)	4.159 (1.736)	3.169 (1.443)	3.843 (1.072)
Spouse present at birth.	-1.595 (1.438)	-1.860 (2.879)	-1.069 (2.118)	-1.270 (1.552)
Partner present at birth.	-0.159 (1.104)	0.177 (1.818)	-0.589 (1.458)	0.009(1.220)
Family income.	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Child's sex.	0.811 (0.676)	0.130 (1.176)	0.085 (0.900)	0.171 (0.732)
Grandparents were present at birth.	-1.078 (1.151)	-1.175 (1.712)	0.081 (1.443)	-0.523(1.272)
Mother's intelligence.	0.081 (0.018)	0.057 (0.032)	0.041 (0.025)	-0.057 (0.020)
Mother's highest educational.	0.606 (0.212)	0.628 (0.355)	0.459 (0.266)	0.572 (0.236)
Child's birth weight.	0.023 (0.021)	0.026 (0.032)	0.012 (0.028)	0.026 (0.022)
Days that the child spent in hospital.	-0.045 (0.052)	-0.047 (0.089)	-0.061 (0.066)	-0.054 (0.055)
Days that the mother spent in hospital.	-0.283 (0.119)	-0.387 (0.214)	-0.309 (0.181)	-0.366 (0.128)
Weeks that the mother worked-level 2	1.190 (0.984)	2.194 (1.556)	1.692 (1.475)	2.526 (1.287)
Weeks that the mother worked-level 3	0.038 (1.261)	1.653 (2.083)	1.400 (1.933)	2.256 (1.635)
Weeks that the mother worked-level 4	2.156 (1.102)	3.392 (2.003)	3.553 (2.003)	5.419 (1.287)
Child was born premature-level 2	1.265 (0.919)	-0.233 (1.531)	-0.063 (1.212)	-0.503 (0.986)
Child was born premature-level 3	2.009 (2.248)	-0.226 (3.718)	-0.353 (2.820)	-0.222 (2.442)

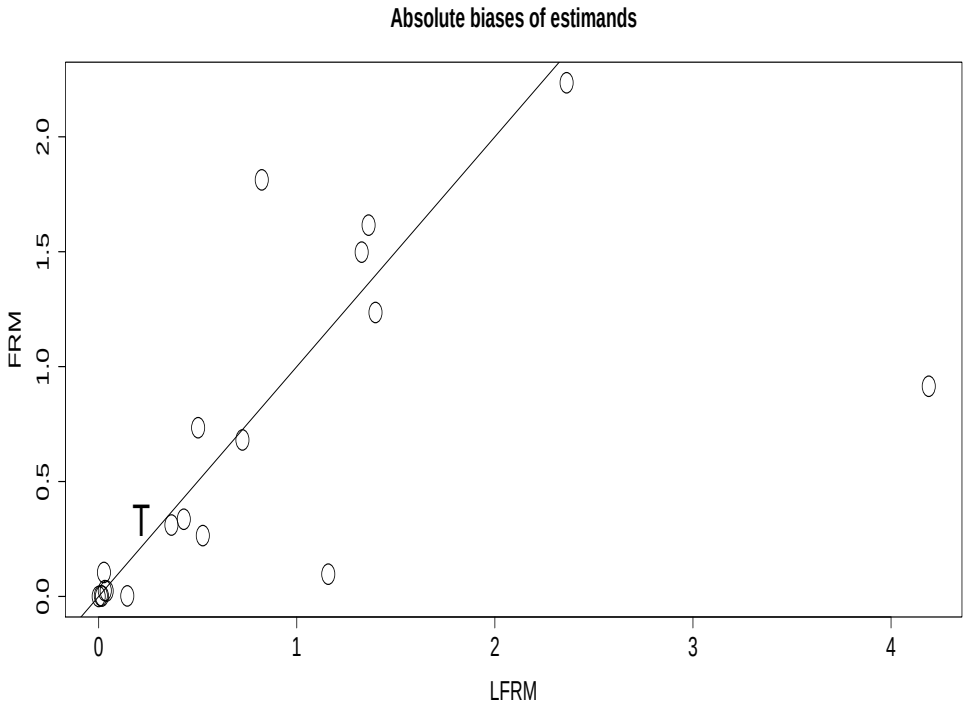


Figure 5.11: Absolute biases of estimands.

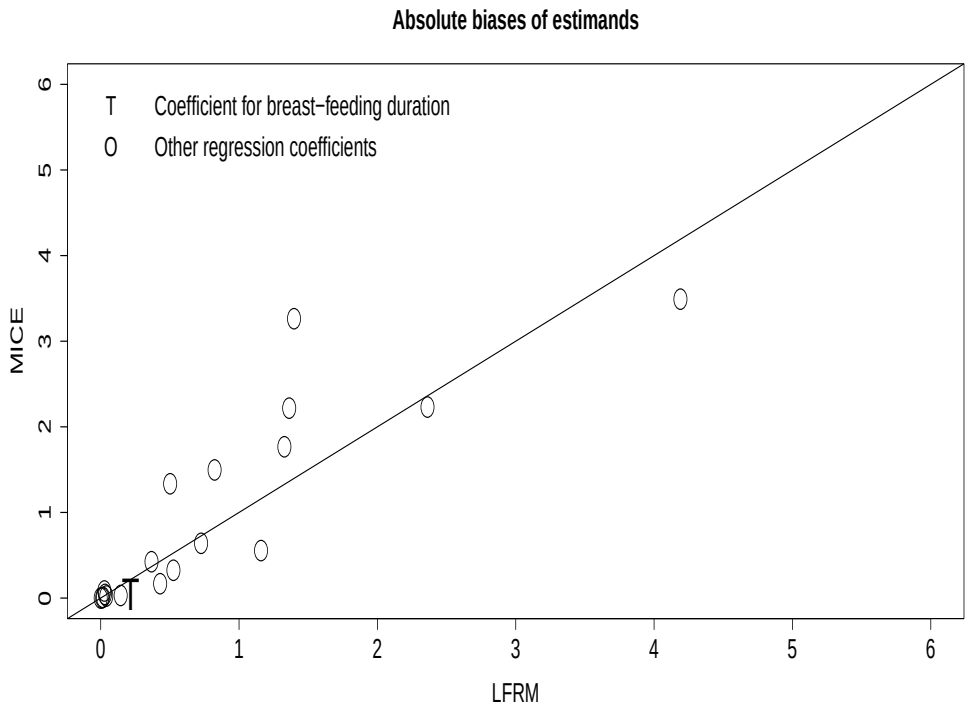


Figure 5.12: Absolute biases of estimands.

5.2.3 Application to the full data sample using FRM, LFRM and MICE

Simirlay to before, we now apply FRM, LFRM and MICE to the full data sample. For FRM we run $m = 50$ Gibbs samplers at different starting points each for 20000 iterations to generate 50 imputed data sets, we also use LFRM and MICE to generate 50 imputed data sets. Table 5.4 presents coefficient estimates from the regression model described above using both sets of imputed data. We see that there is some significant difference in the treatment effect estimates using FRM and LFRM; they differ by approximately 0.813 (standard errors from using FRM and LFRM are 0.727 and 0.709 respectively). Also, there is some significant difference in the treatment effect estimates using MICE and LFRM; they differ by approximately 0.69 (standard errors from using MICE and LFRM are 0.610 and 0.709 respectively).

Table 5.4: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	FRM	LFRM	MICE
Intercept.	85.222 (2.578)	86.897 (2.375)	86.475(2.038)
Treatment effect.	0.903 (0.727)	1.716 (0.709)	1.026(0.610)
The number of years between			
1979 and when the mother gave birth.	0.041 (0.071)	0.079 (0.075)	0.035(0.048)
The child's race-Black	-0.549 (0.776)	-0.607 (0.848)	-0.556(0.678)
The child's race-Other	3.093 (0.726)	2.908 (0.718)	2.879(0.594)
Spouse present at birth.	-0.101 (1.227)	0.095 (1.153)	-0.001 (1.000)
Partner present at birth.	0.294 (0.824)	-0.129 (0.746)	0.141 (0.606)
Family income.	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Child's sex.	0.740 (0.468)	0.709 (0.485)	0.820 (0.379)
Grandparents were present at birth.	-0.537 (0.806)	0.221(0.832)	-0.401(0.637)
Mother's intelligence.	0.108 (0.014)	0.109 (0.014)	0.104 (0.011)
Mother's highest educational.	0.580 (0.162)	0.461 (0.153)	0.517(0.122)
Child's birth weight.	0.013 (0.016)	0.009(0.016)	0.013 (0.013)
Days that the child spent in hospital.	-0.055 (0.046)	-0.068(0.042)	-0.081 (0.035)
Days that the mother spent in hospital.	-0.104 (0.079)	-0.131 (0.092)	-0.151 (0.072)
Weeks that the mother worked-level 2.	0.960 (0.704)	1.186 (0.752)	1.200 (0.592)
Weeks that the mother worked-level 3.	1.040 (0.925)	1.397 (1.014)	1.039 (0.765)
Weeks that the mother worked-level 4.	1.898 (0.871)	2.198(0.935)	2.157 (0.667)
Child was born premature-level 2.	0.811 (0.659)	0.702 (0.693)	0.695 (0.539)
Child was born premature-level 3.	0.958 (1.372)	0.794 (1.569)	0.837(1.183)

5.2.4 Summary

The model FRM seems to perform better than LFRM in terms of coverages in the simulation and the complete case analysis. However, there is no significant difference in the results obtained using LFRM and using MICE in the simulation study. For both simulation and real data application, we set the values of shape (denoted by r) and rate (denoted by δ) for the diffuse hyperprior on the shrinkage parameter, λ^2 equal to 2. Other values such as 1,5 and 10 have been used and the results do not show any significant differences. An alternative to a diffuse hyperprior on the shrinkage parameter λ^2 is to consider the empirical Bayes by marginal maximum likelihood method suggested by Park and Casella (2008) to determine the value of λ^2. We are interested to see how this empirical method will affect our imputations.

5.3 Modified factored regression model (MFRM)

In this section, we will be considering combining elements of RFRM and LFRM into one modelling strategy. We refer to this this modelling strategy as the modified factored regression model (MFRM). We now present the full conditional distributions in the Metropolis within Gibbs sampler required to generate imputed datasets using MFRM. We first express the joint posterior distribution of all unknowns; these comprise model parameters Θ, and the missing and latent values introduced through the data augmentation, denote the set of all unobserved data as $\mathbf{X}_{unobs}$. The joint posterior can then be expressed as,

$$p(\Theta, \mathbf{X}_{unobs}|\mathbf{X}_{obs}) \propto \prod_{i=1}^{n} \left\{ p(x_{i,1}|\boldsymbol{\theta}^{(1)})p(\boldsymbol{\theta}^{(1)}) \prod_{j=2}^{p} p(x_{i,j}|x_{i,1}, \ldots, x_{i,j-1}, \boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)}) \right\}.$$

We assume that there are k nominal variables and we then order the variables so that $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are nominal, each $\mathbf{x}_q, \quad q = 1, \ldots, k$ taking a set of possible values $1, \ldots, L_q$. Variables $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_p$ could then be measured on a binary, ordinal or continuous scale. We provide expressions for each conditional regression model $p(x_{i,q}|x_{i,1}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)})$. First if q = 1 then,

$$p(x_{i,1}|\boldsymbol{\theta}^{(1)}) = \text{Multinomial}(\pi_1^{(1)}, \ldots, \pi_{L_1}^{(1)}),$$

where $\pi_{j^1}^{(1)} = p(x_{i,1} = j^1|\boldsymbol{\theta}^{(1)}) = \frac{\exp(\beta_{0,j^1}^{(1)})}{\sum_{s=1}^{L_1} \exp(\beta_{0,s}^{(1)})}, \quad j^1 = 1, \ldots, L_1$. We place a Dirichlet prior on $\boldsymbol{\pi}^{(1)}$, i.e.

$$p(\boldsymbol{\pi}^{(1)}) \propto \text{Dirichlet}(\alpha_1, \ldots, \alpha_{L_1}),$$

we set all $\alpha_{j^1} = 0.5$, $j^1 = 1, \ldots, L_1$ (corresponding to the Jeffreys prior), another common

choice could be to set all $\alpha_{j^1} = 0$. For $q = 2, \ldots, k$,

$$p(x_{i,q}|x_{i,1}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \text{Multinomial}(\pi^{(q)}_{i,1}, \ldots, \pi^{(q)}_{i,L_q}),$$

where

$$\pi^{(q)}_{i,j^q} = \frac{\exp\left(\beta^{(q)}_{0,j^q} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b} \beta^{(q),j^b}_{b,j^q} I(x_{i,b} = j^b)\right)}{\sum_{u=1}^{L_q} \exp\left(\beta^{(q)}_{0,u} + \sum_{b=1}^{q-1}\sum_{j^b=2}^{L_b} \beta^{(q),j^b}_{b,u} I(x_{i,b} = j^b)\right)},$$

where $j^q = 1, \ldots, L_q$. Let $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}^{(q)}_1, \ldots, \boldsymbol{\theta}^{(q)}_{L_q})$
with $\boldsymbol{\theta}^{(q)}_{j^q} = (\beta^{(q)}_{0,j^q}, \beta^{(q),2}_{1,j^q}, \ldots, \beta^{(q),L_2}_{1,j^q}, \ldots, \beta^{(q),2}_{q-1,j^q}, \ldots, \beta^{(q),L_{q-1}}_{q-1,j^q})$ for $j^q = 1, \ldots, L_q$ and we set $\boldsymbol{\theta}^{(q)}_1 = \mathbf{0}$ for identifiability . We specify the prior for each $\boldsymbol{\theta}^{(q)}_{j^q}, j^q = 2, \ldots, L_q$ by

$$p(\boldsymbol{\theta}^{(q)}) = \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}).$$

where $\boldsymbol{\Sigma}$ is diagonal matrix with large entries.

For $\mathbf{x}_q,\ q \in \{k+1, \ldots, p\}$ is continuous (for now assuming $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{q-1}$ are continuous),

$$\begin{aligned} &p(x_{i,q}|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \\ &\text{N}(\beta^{(q)}_0 + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b} \beta^{(q),j^b}_b I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1} \beta^{(q)}_b x_{i,b}, (\tau_q \phi_{i,q})^{-1}), \end{aligned}$$

We specify a prior for $\boldsymbol{\theta}^{(q)}$,

$$\begin{aligned} p(\boldsymbol{\theta}^{(q)}) &= p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \psi_q, \beta^{(q)}_0, \tau_q, \Phi_q) \\ &= p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \lambda^2_q, \psi_q, \beta^{(q)}_0, \tau_q, \Phi_q) p(\boldsymbol{\eta}_q|\lambda^2_q, \psi_q, \beta^{(q)}_0, \tau_q, \Phi_q) p(\lambda^2_q|\psi_q, \beta^{(q)}_0, \tau_q, \Phi_q) \\ &\quad p(\psi_q|\beta^{(q)}_0, \tau_q, \Phi_q) p(\beta^{(q)}_0|\tau_q, \Phi_q) p(\tau_q|\Phi_q) p(\Phi_q), \end{aligned}$$

where $\boldsymbol{\beta}^{(q)} = (\beta^{(q),2}_1, \ldots, \beta^{(q),L_1}_1, \ldots, \beta^{(q),2}_k, \ldots, \beta^{(q),L_k}_k, \beta^{(q)}_{k+1}, \ldots, \beta^{(q)}_{q-1})'$,
$\boldsymbol{\eta}_q = (\eta^2_{1,q}, \ldots, \eta^2_{q-1,q})'$ and $\Phi_q = (\phi_{i,q}, \ldots, \phi_{n,q})'$. The prior distribution for $\boldsymbol{\beta}^{(q)}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_q$,

$$p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \psi_q) \sim \text{MVN}(\mathbf{0}, \frac{1}{\psi_q}\mathbf{D}^{(q)}_\eta),\ \text{with } \mathbf{D}^{(q)}_\eta = \text{diag}(\eta^2_{1,q}, \ldots, \eta^2_{q-1,q}), \tag{5.14}$$

with the distribution of $\eta^2_{j,q},\ j = 1, \ldots, q-1$ is given by

$$p(\eta^2_{j,q}|\lambda^2_q) = \frac{\lambda^2_q}{2} \exp\left(\frac{-\lambda^2_q \eta^2_{j,q}}{2}\right),\ j = 1, \ldots, q-1, \tag{5.15}$$

and this is the conditional Laplace prior for $\boldsymbol{\beta}^{(q)}$ suggested by Park and Casella (2008). We consider a diffuse hyperprior on λ_q^2 of the form

$$p(\lambda_q^2) = \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right),\ \lambda_q^2 > 0,\ r > 0,\ \delta > 0, \tag{5.16}$$

which is a gamma distribution with shape and rate r and δ respectively. The prior distribution for $p(\Phi_q)$ follows a gamma distribution with shape and rate parameters equal to $\frac{\nu}{2}$,

$$p(\phi_{i,q}) \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}),\ i = 1, \ldots, n,$$

and the prior distribution for τ_q is given by

$$p(\tau_q) \ \propto \ \frac{1}{\tau_q}.$$

and this results in the errors of the linear regression model follow a marginal t distribution with degrees of freedom ν. We also specify the prior distributions for ψ_q and $\beta_0^{(q)}$ as follows:

$$p(\psi_q) \ \propto \ \frac{1}{\psi_q}, \tag{5.17}$$

and

$$p(\beta_0^{(q)}) \ \propto \ 1. \tag{5.18}$$

For $\mathbf{x}_q,\ q \in \{k+1, \ldots, p\}$ is binary (for now assuming $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{q-1}$ are continuous),

$$x_{i,q} = I(x_{i,q}^* > 0), \quad \text{where}$$
$$p(x_{i,q}^*|x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\theta}^{(q)}) = \text{N}(\beta_0^{(q)} + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b}\beta_b^{(q),j^b}I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1}\beta_b^{(q)}x_{i,b}, 1).$$

We specify a prior for $\boldsymbol{\theta}^{(q)}$,

$$\begin{aligned} p(\boldsymbol{\theta}^{(q)}) &= p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}) \\ &= p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)})p(\boldsymbol{\eta}_q|\lambda_q^2, \psi_q, \beta_0^{(q)})p(\lambda_q^2|\psi_q, \beta_0^{(q)})p(\psi_q|\beta_0^{(q)})p(\beta_0^{(q)}), \end{aligned}$$

where $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$ and $\boldsymbol{\eta}_q = (\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2)'$. The prior distributions for $\boldsymbol{\beta}^{(q)}$ and $\boldsymbol{\eta}_q$ are given by Equation 5.14 and Equation 5.15 respectively and this is the conditional Laplace prior for $\boldsymbol{\beta}^{(q)}$ suggested by Park and Casella (2008). We specify prior distributions for λ_q^2, ψ_q and $\beta_0^{(q)}$ which are given by Equation 5.16, Equation 5.17 and Equation 5.18 respectively.

For $\mathbf{x}_q,\ q \in \{k+1,\ldots,p\}$ is ordinal taking values in $\{1,\ldots,J_q\}$ (for now assuming $\mathbf{x}_{k+1},\ldots,\mathbf{x}_{q-1}$ are continuous),

$$x_{i,q} = j^q\ I(\gamma^{(q)}_{j^q-1} < x^*_{i,q} < \gamma^{(q)}_{j^q}), \quad \text{where}$$
$$p(x^*_{i,q}|x_{i,1},x_{i,2},\ldots,x_{i,q-1},\boldsymbol{\theta}^{(q)}) = \mathrm{N}(\beta^{(q)}_0 + \sum_{b=1}^{k}\sum_{j^b=2}^{L_b}\beta^{(q),j^b}_b I(x_{i,b}=j^b) + \sum_{b=k+1}^{q-1}\beta^{(q)}_b x_{i,b}, 1),$$

where the $\gamma^{(q)}_{j^q}$ are threshold values, with $\gamma^{(q)}_0 = -\infty$, $\gamma^{(q)}_1 = 0$ and $\gamma^{(q)}_{J_q} = \infty$, for $j^q \in \{1,...,J_q\}$.

We also specify a prior for $\boldsymbol{\theta}^{(q)}$,

$$\begin{aligned} p(\boldsymbol{\theta}^{(q)}) &= p(\boldsymbol{\beta}^{(q)},\boldsymbol{\eta}_q,\lambda^2_q,\psi_q,\beta^{(q)}_0,\boldsymbol{\gamma}^{(q)}) \\ &= p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q,\lambda^2_q,\psi_q,\beta^{(q)}_0,\boldsymbol{\gamma}^{(q)})p(\boldsymbol{\eta}_q|\lambda^2_q,\psi_q,\beta^{(q)}_0,\boldsymbol{\gamma}^{(q)})p(\lambda^2_q|\psi_q,\beta^{(q)}_0,\boldsymbol{\gamma}^{(q)}) \\ &\quad p(\psi_q|\beta^{(q)}_0,\boldsymbol{\gamma}^{(q)})p(\beta^{(q)}_0|\boldsymbol{\gamma}^{(q)})p(\boldsymbol{\gamma}^{(q)}), \end{aligned}$$

where $\boldsymbol{\beta}^{(q)} = (\beta^{(q),2}_1,\ldots,\beta^{(q),L_1}_1,\ldots,\beta^{(q),2}_k,\ldots,\beta^{(q),L_k}_k,\beta^{(q)}_{k+1},\ldots,\beta^{(q)}_{q-1})'$, $\boldsymbol{\eta}_q = (\eta^2_{1,q},\ldots,\eta^2_{q-1,q})'$ and $\boldsymbol{\gamma}^{(q)} = \left\{\gamma^{(q)}_{j^q} : j^q \in \{1,\ldots,J_q\}\right\}$. The prior distributions for $\boldsymbol{\beta}^{(q)}$ and $\boldsymbol{\eta}_q$ are given by Equation 5.14 and Equation 5.15 respectively and this is the conditional Laplace prior for $\boldsymbol{\beta}^{(q)}$. We specify prior distributions for λ^2_q, ψ_q and $\beta^{(q)}_0$ which are given by Equation 5.16, Equation 5.17 and Equation 5.18 respectively. We place an improper uniform prior on $\boldsymbol{\gamma}^{(q)}$ i.e.

$$p(\boldsymbol{\gamma}^{(q)}) \propto I(\boldsymbol{\gamma}^{(q)} \in \Omega^{(q)}),$$

where $\Omega^{(q)} = \left\{\gamma^{(q)}_{j^q} : \gamma^{(q)}_0 = -\infty < \gamma^{(q)}_1 = 0 < \gamma^{(q)}_2 < \ldots < \gamma^{(q)}_{J_{q-1}} < \gamma^{(q)}_{J_q} = \infty\right\}$.

If $x_{i,q},\ q \in \{k+1,\ldots,q-1\}$ is not continuous then we replace $x_{i,q}$ with $x^*_{i,q}$ in the model for $p(x_{i,q}|x_{i,1},x_{i,2},\ldots,x_{i,q-1},\boldsymbol{\theta}^{(q)})$. For notational convenience when $x_{i,q}$ is nominal or continuous we assume $x_{i,q} = x^*_{i,q},\ q \in \{1,\ldots,p\}$.

With this model and prior specification we can present the full conditional distribution required to impute missing values in the Metropolis within Gibbs sampler. Following Schafer (1997) we present this a data augmentation scheme and so first present the "I-steps" followed by the "P-steps".

In the "I-steps", conditional on parameter values, first impute missing nominal values $x^*_{i,q}$ from the following discrete distribution

$$\begin{aligned} &p(x^*_{i,q} = j^q | x^*_{i,1} = j^1,\ldots,x^*_{i,q-1} = j^{q-1}, x^*_{i,q+1} = j^{q+1},\ldots,x^*_{i,k} = j^k, x^*_{i,k+1},\ldots,x^*_{i,p},\Theta) \\ &= \tilde{\pi}^{(q)}_{i,j^q}, \end{aligned}$$

where $j^q \in \{1, \ldots, L_q\}$, and

$$\tilde{\pi}_{i,j^q}^{(q)} = \frac{\prod_{b=q}^{k} \pi_{i,j^b}^{(b)} \prod_{b=k+1}^{p} \exp\left(\frac{-\tau_b \phi_{i,b}}{2} (\tilde{x}_{i,b}^* - \beta_q^{(b),j^q} I(x_{i,q}^* = j^q) - \sum_{t=k+1}^{(b-1)} \beta_t^{(b)} x_{i,t}^*)^2 \right)}{\sum_{u=1}^{L_q} \left\{ \pi_{i,u}^{(q)} \prod_{b=q+1}^{k} \pi_{i,j^b}^{(b)} \right\} \prod_{b=k+1}^{p} \exp\left(\frac{-\tau_b \phi_{i,b}}{2} (\tilde{x}_{i,b}^* - \beta_q^{(b),j^q} I(x_{i,q}^* = u) - \sum_{t=k+1}^{(b-1)} \beta_t^{(b)} x_{i,t}^*)^2 \right)},$$

where

$$\tilde{x}_{i,b}^* = x_{i,b}^* - \beta_0^{(b)} - \sum_{s=1}^{q-1} \sum_{j^s=2}^{L_s} \beta_s^{(b),j^s} I(x_{i,s}^* = j^s) - \sum_{s=q+1}^{k} \sum_{j^s=2}^{L_s} \beta_s^{(b),j^s} I(x_{i,s}^* = j^s).$$

Next impute missing continuous values $x_{i,q}^*$ or latent values $x_{i,q}^*$ with missing $x_{i,q}$ from,

$$p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, x_{i,q+1}^*, \ldots, x_{i,k}^*, x_{i,k+1}^*, \ldots, x_{i,p}^*, \Theta) = \mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_{i,q}^{-1}),$$

where

$$\widetilde{\mu}_{i,q} = \widetilde{\phi}_{i,q}^{-1} \left\{ \frac{\mu_{i,q}}{(\tau_q \phi_{i,q})^{-1}} + \sum_{s=q+1}^{p} \frac{\beta_q^{(s)}}{(\tau_s \phi_{i,s})^{-1}} \left[x_{i,s}^* - (\mu_{i,s} - \beta_q^{(s)} x_{i,q}^*) \right] \right\},$$

and

$$\widetilde{\phi}_{i,q}^{-1} = \left(\frac{1}{(\tau_q \phi_{i,q})^{-1}} + \sum_{s=q+1}^{p} \frac{(\beta_q^{(s)})^2}{(\tau_s \phi_{i,s})^{-1}} \right)^{-1},$$

where $\mu_{i,q} = \beta_0^{(q)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \beta_b^{(q),j^b} I(x_{i,b}^* = j^b) + \sum_{b=k+1}^{q-1} \beta_b^{(q)} x_{i,b}^*$ and
$\mu_{i,s} = \beta_0^{(s)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \beta_b^{(s),j^b} I(x_{i,b}^* = j^b) + \sum_{b=k+1}^{s-1} \beta_b^{(s)} x_{i,b}^*$.

Next impute latent values $x_{i,q}^*$ with observed ordinal variable $x_{i,q}$ from,

$$p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, x_{i,q+1}^*, \ldots, x_{i,p}^*, \Theta) = \frac{I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) \mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_{i,q}^{-1})}{\Phi(\frac{\gamma_{j^q}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_{i,q}^{-1}}) - \Phi(\frac{\gamma_{j^q-1}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_{i,q}^{-1}})},$$

which is a normal distribution $\mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_{i,q}^{-1})$ truncated to $(\gamma_{j^q-1}^{(q)}, \gamma_{j^q}^{(q)})$.

Next impute latent values $x_{i,q}^*$ with observed binary variable $x_{i,q}$ from,

$$p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, x_{i,q+1}^*, \ldots, x_{i,p}^*, \Theta) = \frac{I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) \mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_{i,q}^{-1})}{\Phi(\frac{\gamma_{j^q}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_{i,q}^{-1}}) - \Phi(\frac{\gamma_{j^q-1}^{(q)} - \widetilde{\mu}_{i,q}}{\widetilde{\phi}_{i,q}^{-1}})},$$

which is a normal distribution $\mathrm{N}(\widetilde{\mu}_{i,q}, \widetilde{\phi}_{i,q}^{-1})$ truncated to $(\gamma_{j^q-1}^{(q)}, \gamma_{j^q}^{(q)})$ where the $\gamma_{j^q}^{(q)}$ are threshold values, with $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_2^{(q)} = \infty$.

Once we have imputed values for $x_{i,q}^*$ (with missing $x_{i,q}$), we define a function $g(x_{i,q}^*)$ to map each $x_{i,q}^*$ back to its original measurement scale, and thus create an imputed data set. The mapping function is defined as follows,

$$g(x_{i,q}^*) = \begin{cases} I(x_{i,q}^* > 0) & \text{if } x_{i,q}^* \text{ is binary,} \\ j^q \; I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) & \text{if } x_{i,q}^* \text{ is ordinal, } j^q \in \{1, \ldots, J_q\}, \\ x_{i,q}^* & \text{if } x_{i,q}^* \text{ is continuous,} \\ x_{i,q}^* & \text{if } x_{i,q}^* \text{ is nominal, } j^q \in \{1, \ldots, L_q\}, \end{cases}$$

Denote an imputed value for $x_{i,q}$ by $x_{i,q}^{(t)}$ at iteration t, and an imputed data set at iteration t by $\mathbf{X}_{com}^{(t)}$. In the "P-steps", conditional on an imputed data set, first sample values for $\boldsymbol{\theta}^{(1)}$ from a Dirichlet $(\boldsymbol{\alpha})$ distribution with parameters

$$\boldsymbol{\alpha} = \left(\alpha_1 + \sum_{i=1}^{n} I(x_{i,1} = 1), \; \alpha_2 + \sum_{i=1}^{n} I(x_{i,1} = 2), \ldots, \alpha_{L_1} + \sum_{i=1}^{n} I(x_{i,1} = L_1) \right)'.$$

Next propose values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}_1^{(q)}, \ldots, \boldsymbol{\theta}_{L_q}^{(q)})$
with $\boldsymbol{\theta}_{j^q}^{(q)} = (\beta_{0,j^q}^{(q)}, \beta_{1,j^q}^{(q),2}, \ldots, \beta_{1,j^q}^{(q),L_2}, \ldots, \beta_{q-1,j^q}^{(q),2}, \ldots, \beta_{q-1,j^q}^{(q),L_{q-1}})$, for $j^q = 1, \ldots, L_q$ using a Metropolis-Hastings step, where we set $\boldsymbol{\theta}_1^{(q)} = \mathbf{0}$ for identifiability. We consider a proposal distribution based on the imputed data log-likelihood at each iteration t, $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)})$, which can be expressed as

$$l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)}) = \sum_{w=1}^{L_q} \sum_{i=1}^{n} \ln \left(\frac{\exp\left(\beta_{0,w}^{(q)} + \sum_{b=1}^{q-1} \sum_{j^b=2}^{L_b} \beta_{b,w}^{(q),j^b} f(x_{i,b}^{(t)}, x_{i,b}, m_{i,b}) \right)}{\sum_{u=1}^{L_q} \exp\left(\beta_{0,u}^{(q)} + \sum_{b=1}^{q-1} \sum_{j^b=2}^{L_b} \beta_{b,u}^{(q),j^b} f(x_{i,b}^{(t)}, x_{i,b}, m_{i,b}) \right)} \right)^{I(x_{i,q}=w)},$$

where $f(x_{i,b}^{(t)}, x_{i,b}, m_{i,b}) = (I(x_{i,b}^{(t)} = j^b) m_{i,b} + I(x_{i,b} = j^b)(1 - m_{i,b}))$, with $m_{i,b}$ is a missing data indicator with $m_{i,b} = 1$ indicating $x_{i,b}$ is missing and $m_{i,b} = 0$ indicating $x_{i,b}$ is observed. We generate proposals for $\boldsymbol{\theta}^{(q)}$ from a $N(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ distribution, and $\boldsymbol{\mu}^{(t)}$ is the value that maximises $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)})$ and $\boldsymbol{\Sigma}^{(t)}$ is given by $-E[\frac{\delta^2}{\delta \beta_{j,w}^{(q)} \delta \beta_{\tilde{j},\tilde{w}}^{(q)}} l(\boldsymbol{\theta}^{(q)}; \mathbf{X}_{com}^{(t)})]_{\boldsymbol{\theta}^{(q)}=\boldsymbol{\mu}^{(t)}}^{-1}$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q)$ for $q \in \{k+1, \ldots, p\}$, when $\mathbf{x}_q^*$ is continuous from the joint posterior distribution of
$\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$, $\boldsymbol{\eta}_q = (\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2)'$, λ_q^2,

ψ_q, $\beta_0^{(q)}$, τ_q, and $\Phi_q = (\phi_{1,q}, \ldots, \phi_{n,q})'$, which is given by

$$
\begin{aligned}
&p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q | \mathbf{X}_{com}^{(t)}) = \\
&p(\boldsymbol{\beta}^{(q)} | \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q, \mathbf{X}_{com}^{(t)}) p(\boldsymbol{\eta}_q | \lambda_q^2, \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q, \mathbf{X}_{com}^{(t)}) \\
&p(\lambda_q^2 | \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q, \mathbf{X}_{com}^{(t)}) p(\psi_q | \beta_0^{(q)}, \tau_q, \Phi_q, \mathbf{X}_{com}^{(t)}) p(\beta_0^{(q)} | \tau_q, \Phi_q, \mathbf{X}_{com}^{(t)}) \\
&p(\tau_q | \Phi_q, \mathbf{X}_{com}^{(t)}) p(\Phi_q | \mathbf{X}_{com}^{(t)})
\end{aligned}
$$

Now we sample values for $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q$ from their full conditional distributions in the following way:

1. We first sample $\boldsymbol{\beta}^{(q)}$ from

$$p(\boldsymbol{\beta}^{(q)} | \phi_q, \psi_q, \boldsymbol{\eta}_q, \tau_q, \widetilde{\mathbf{X}}_q, \widetilde{\mathbf{x}}_q^*) = \text{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}),$$

with

$$\hat{\boldsymbol{\beta}}^{(q)} = \mathbf{V}^{(q)} \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_\phi^{(q)} \widetilde{\mathbf{x}}_q^*,$$

and

$$\mathbf{V}^{(q)} = (\widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_\phi^{(q)} \widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)})^{-1},$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*, \ldots, \widetilde{x}_{n,q}^*)'$, with $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n \times n$ diagonal matrix with entries $\phi_{i,q}$ for $i = 1, \ldots, n$ and $\mathbf{D}_{\psi,\eta}^{(q)}$ is a $(q-1) \times (q-1)$ diagonal matrix with entries $\frac{\psi_q}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

2. We then sample values for $\eta_{j,q}^2,\ j = 1, \ldots, q-1$ from

$$p(\eta_{j,q}^2 | \boldsymbol{\beta}^{(q)}, \lambda_q^2, \psi_q), \quad j = 1, \ldots, q-1$$

which is an inverse-Gaussian distribution with parameters

$$\mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2 \psi_q}} \quad \text{and} \quad \lambda_q' = \lambda_q^2, \quad j = 1, \ldots, q-1$$

in the parameterization of the inverse-Gaussian density given by

$$f(g) = \sqrt{\frac{\lambda_q'}{2\pi}} g^{-\frac{3}{2}} \exp\left\{ -\frac{\lambda_q'(g - \mu_q')^2}{2(\mu_q')^2 g} \right\},\ g = \frac{1}{\eta_{j,q}^2},\ g > 0.$$

3. Next sample a value for λ_q^2 from

$$p(\lambda_q^2|\boldsymbol{\eta}_q) \sim \text{Gamma}\left((q-1)+r, \sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2}+\delta\right).$$

4. Next sample a value for ψ_q from

$$p(\psi_q|\boldsymbol{\eta}_q, \boldsymbol{\beta}^{(q)}) = \text{Gamma}\left(\frac{q-1}{2}, \frac{(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right),$$

where $\mathbf{D}_\eta^{(q)}$ is a $(q-1)\times(q-1)$ diagonal matrix with entries $\frac{1}{\eta_{j,q}^2}$ for $j=1,\ldots,q-1$.

5. Next sample a value for $\beta_0^{(q)}$ from

$$p(\beta_0^{(q)}|\tau_q, \phi_q, x_{i,q}^*, \widetilde{\mathbf{x}}_{i,q}) \sim \text{N}(\frac{\sum_{i=1}^n \phi_{i,q}(x_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{\sum_{i=1}^n \phi_{i,q}}, (\tau_q\sum_{i=1}^n \phi_{i,q})^{-1}),$$

where
$\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^*=2),\ldots,I(x_{i,1}^*=L_1),\ldots,I(x_{i,k}^*=2),\ldots,I(x_{i,k}^*=L_k),x_{i,k+1}^*,\ldots,x_{i,q-1}^*)$ for $i=1,\ldots,n$.

6. We then sample a value for τ_q from

$$p(\tau_q|\phi_q, \boldsymbol{\beta}^{(q)}, \widetilde{\mathbf{x}}_q^*, \widetilde{\mathbf{X}}_q) \sim \text{Gamma}\left(\frac{n}{2}, \frac{(\widetilde{\mathbf{x}}_q^* - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'\mathbf{D}_\phi^{(q)}(\widetilde{\mathbf{x}}_q^* - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})}{2}\right),$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*,\ldots,\widetilde{x}_{n,q}^*)'$, with $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q},\ldots,\widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^*=2),\ldots,I(x_{i,1}^*=L_1),\ldots,I(x_{i,k}^*=2),\ldots,I(x_{i,k}^*=L_k),x_{i,k+1}^*,\ldots,x_{i,q-1}^*)$ for $i=1,\ldots,n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n\times n$ diagonal matrix with entries $\phi_{i,q}$ for $i=1,\ldots,n$.

7. We then sample values for $\phi_{i,q}$ from

$$p(\phi_{i,q}|\boldsymbol{\eta}_q, \tau_q, \widetilde{x}_{i,q}^*, \widetilde{\mathbf{x}}_{i,q}) \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu+\tau_q(\widetilde{x}_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2}{2}\right), \quad i=1,\ldots,n.$$

where $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{x}}_{i,q} =$
$(I(x_{i,1}^*=2),\ldots,I(x_{i,1}^*=L_1),\ldots,I(x_{i,k}^*=2),\ldots,I(x_{i,k}^*=L_k),x_{i,k+1}^*,\ldots,x_{i,q-1}^*)$ for $i=1,\ldots,n$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)})$ for $q\in\{k+1,\ldots,p\}$, when $\mathbf{x}_q$ is binary from the joint posterior distribution of
$\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2},\ldots,\beta_1^{(q),L_1},\ldots,\beta_k^{(q),2},\ldots,\beta_k^{(q),L_k},\beta_{k+1}^{(q)},\ldots,\beta_{q-1}^{(q)})'$, $\boldsymbol{\eta}_q = (\eta_{1,q}^2,\ldots,\eta_{q-1,q}^2)'$, λ_q^2,

ψ_q and $\beta_0^{(q)}$, which is given by

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)} | \mathbf{X}_{com}^{(t)}) &= p(\boldsymbol{\beta}^{(q)} | \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) p(\boldsymbol{\eta}_q | \lambda_q^2, \psi_q, \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) \\ & \quad p(\lambda_q^2 | \psi_q, \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) p(\psi_q | \beta_0^{(q)}, \mathbf{X}_{com}^{(t)}) p(\beta_0^{(q)} | \mathbf{X}_{com}^{(t)}) \end{aligned}$$

We sample values for $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}$ from their full conditional distributions in the following way:

1. We first $\boldsymbol{\beta}^{(q)}$ from

$$p(\boldsymbol{\beta}^{(q)} | \psi_q, \boldsymbol{\eta}_q, \widetilde{\mathbf{X}}_q', \widetilde{\mathbf{x}}_q^*) = \text{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}), \tag{5.19}$$

with

$$\hat{\boldsymbol{\beta}}^{(q)} = \mathbf{V}^{(q)} \widetilde{\mathbf{X}}_q' \widetilde{\mathbf{x}}_q^*,$$

and

$$\mathbf{V}^{(q)} = (\widetilde{\mathbf{X}}_q' \widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)})^{-1},$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*, \ldots, \widetilde{x}_{n,q}^*)'$, with $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_{\psi,\eta}^{(q)}$ is a $(q-1) \times (q-1)$ diagonal matrix with entries $\frac{\psi_q}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

2. We then sample values for $\eta_{j,q}^2,\ j = 1, \ldots, q-1$ from

$$p(\eta_{j,q}^2 | \boldsymbol{\beta}^{(q)}, \lambda_q^2, \psi_q), \quad j = 1, \ldots, q-1$$

which is an inverse-Gaussian distribution with parameters

$$\mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2 \psi_q}} \text{ and } \lambda_q' = \lambda_q^2, \quad j = 1, \ldots, q-1$$

in the parameterization of the inverse-Gaussian density given by

$$f(g) = \sqrt{\frac{\lambda_q'}{2\pi}} g^{-\frac{3}{2}} \exp\left\{ -\frac{\lambda_q'(g - \mu_q')^2}{2(\mu_q')^2 g} \right\},\ g = \frac{1}{\eta_{j,q}^2}\ g > 0. \tag{5.20}$$

3. Next sample a value for λ_q^2 from

$$p(\lambda_q^2 | \boldsymbol{\eta}_q) \sim \text{Gamma}\left((q-1) + r, \sum_{j=1}^{q-1} \frac{\eta_{j,q}^2}{2} + \delta \right). \tag{5.21}$$

4. We then sample a value for ψ_q from

$$p(\psi_q|\boldsymbol{\eta}_q, \boldsymbol{\beta}^{(q)}) = \text{Gamma}\left(\frac{q-1}{2}, \frac{(\boldsymbol{\beta}^{(q)})'\mathbf{D}_{\eta}^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right), \tag{5.22}$$

where $\mathbf{D}_{\eta}^{(q)}$ is a $(q-1)\times(q-1)$ diagonal matrix with entries $\frac{1}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

5. We then sample a value for $\beta_0^{(q)}$ from

$$p(\beta_0^{(q)}|\boldsymbol{\beta}^{(q)}, x_{i,q}^*, \widetilde{\mathbf{x}}_{i,q}) \sim \text{N}(\frac{\sum_{i=1}^{n}(x_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}, \frac{1}{n}), \tag{5.23}$$

where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})$ for $q \in \{k+1, \ldots, p\}$, when $\mathbf{x}_q$ is ordinal from the joint posterior distribution of $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$, $\boldsymbol{\eta}_q = (\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2)'$, λ_q^2, ψ_q, $\beta_0^{(q)}$ and $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1, \ldots, J_q\}\right\}$, which is given by

$$\begin{aligned}
&p(\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}|\mathbf{X}_{com}^{(t)}) = \\
&p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)})p(\boldsymbol{\eta}_q|\lambda_q^2, \psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)}) \\
&p(\lambda_q^2|\psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)})p(\psi_q|\beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)})p(\beta_0^{(q)}|\boldsymbol{\gamma}^{(q)}, \mathbf{X}_{com}^{(t)})p(\boldsymbol{\gamma}^{(q)}|\mathbf{X}_{com}^{(t)})
\end{aligned}$$

We sample values for $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}$ from their full conditional distributions in the following way:

1. We first $\boldsymbol{\beta}^{(q)}$ from its full conditional distribution which is given by Equation 5.19.

2. We then sample values for $\eta_{j,q}^2$, for $j = 1, \ldots, q-1$ from its full conditional distribution which is given by Equation 5.20.

3. Next sample a value for λ_q^2 from its full conditional distribution which is given by Equation 5.21.

4. We then sample a value for ψ_q from its full conditional distribution which is given by Equation 5.22.

5. We then sample a value for $\beta_0^{(q)}$ from its full conditional distribution which is given by Equation 5.23.

6. We then update the threshold values $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1, \ldots, J_q\}\right\}$. The full conditional distribution of the threshold values $\gamma_{j^q}^{(q)}$ for $j^q \in \{1, \ldots, J_q\}$ is uniformly distributed

on the interval

$$\left[\max\left\{\max\left\{x^*_{i,q} : x_{i,q} = j^q\right\}, \gamma^{(q)}_{j^q-1}\right\}, \min\left\{\min\left\{x^*_{i,q} : x_{i,q} = j^q + 1\right\}, \gamma^{(q)}_{j^q+1}\right\}\right],$$

where $\gamma^{(q)}_0 = -\infty$, $\gamma^{(q)}_1 = 0$ and $\gamma^{(q)}_{J_q} = \infty$, for $j^q \in \{1, \ldots, J_q\}$.

In this way we have sampled all unknowns (missing values and parameters) from their full conditional distributions. The details on deriving the full conditional distributions of the missing values in the "I-steps", the full conditional distribution of $\boldsymbol{\theta}^{(1)}$ where $\mathbf{x}_1$ follows a multinomial distribution, as well as the Metropolis-Hastings update for the multinomial logistic regression model in the "P-steps" have been presented in the MFRM modelling strategy. For details on deriving the full conditional distributions of the remaining parameters in the "P-steps", such as $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \psi_q, \beta^{(q)}_0, \tau_q, \Phi_q)$ for $\mathbf{x}^*_q$ when $\mathbf{x}^*_q$ is continuous, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \psi_q, \beta^{(q)}_0)$ for $\mathbf{x}^*_q$ when $\mathbf{x}^*_q$ is a latent variable ($\mathbf{x}_q$ is binary) and $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \psi_q, \beta^{(q)}_0, \boldsymbol{\gamma}^{(q)})$ for $\mathbf{x}^*_q$ when $\mathbf{x}^*_q$ is a latent variable ($\mathbf{x}_q$ is ordinal), please see Appendix G.

5.3.1 Simulation study

We now illustrate the performances of FRM, MICE and MFRM through a simulation study. In the MICE approach, we use the model and prior specifications described in the MFRM strategy to generate missing values. We are interested to compare the results between a FRM and MFRM, as well as the results between the joint modelling specification with t-error for the regression model and Bayesian lasso for the regression parameters (MFRM) with the equivalent fully conditional specification (MICE).

We simulate data sets that contain variables measured on binary, ordinal, continuous and nominal (greater than two levels) scales. Specifically each data set contains 5 binary variables, 5 ordinal variables, 10 continuous variables, and 5 nominal variables. Variables are simulated in a sequential manner with each variable conditional on a subset of variables already generated. We then introduce missing values into nine of the variables using the MAR mechanism, so that each incomplete variable has approximately 30% missing values. Specific details of how we simulated the incomplete dataset are given in Appendix H. We replicate this data generation process 500 times to generate 500 incomplete data sets.

Using the MFRM and MICE imputation strategies, we multiply impute the missing values in each incomplete data set $m = 10$ times. To generate m independent imputed data sets, we run m Metropolis within Gibbs samplers, each with a different starting value, with each sampler resulting in an imputed data set after convergence. We assume analysts may be interested in making inferences about various types of estimands arising from both univariate analyses, e.g. population means of variables or the proportions in the population taking a particular level of a categorical variable, as well as multivariate analyses, e.g. the coefficients from a regression model. Using the m imputed data sets, we apply the combining rules described in Equation 2.1 to construct point and interval estimates for these estimands. We also construct estimates for the same estimands when using FRM to multiply impute the missing values.

When implementing MICE, MFRM or FRM for imputations, we assume the imputer knows the data generating process and so decomposes the joint distribution of the imputation model as described in Equation 3.3. We then performed analysis on the imputed data sets and obtained estimands arising from both univariate analyses and multivariate analyses. See Appendix H for a full list of the estimands considered.

Figure 5.13 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by FRM and MFRM over the 500 datasets. These estimands include those arising from both univariate analyses and regression analyses. The first box plot presents coverages when there are no missing data in the datasets. These coverage are as expected the closest to 0.95. The second box plot shows the coverages from FRM while the third box plot shows the coverages obtained from MFRM. We noticed that MFRM obtains coverages much closer to 0.95 than the coverages obtained from using FRM. The parameter that show low coverage in the second boxplot is the univariate analysis of proportion of $\mathbf{x}_{15} = 1$.

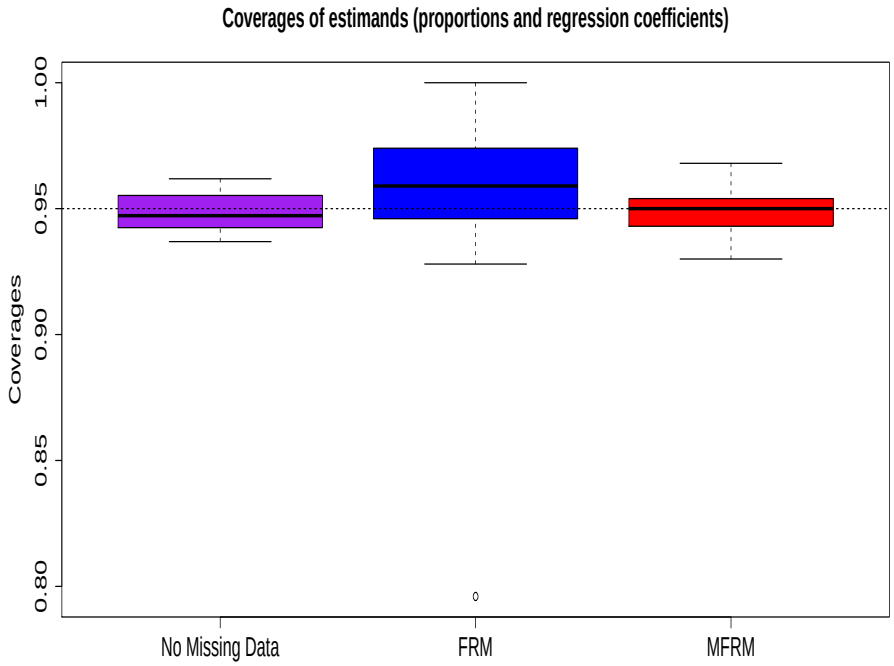


Figure 5.13: The coverages of estimands.

We now plot the biases in the estimates obtained from using FRM against the biases obtained from using MFRM. We distinguish between estimates arising from univariate analyses and estimates arising from multivariate analyses, i.e. regression coefficient estimates. We see that the majority of the large biases are those arising from multivariate analyses, and are above the $y = x$ line, which indicates that FRM tends to obtain regression coefficient estimates further from the true values than MFRM.

We now compare the results between MFRM and MICE. Figure 5.15 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by MICE and MFRM over the 500 datasets. These estimands include those arising from both univariate analyses and regression analyses. The first box

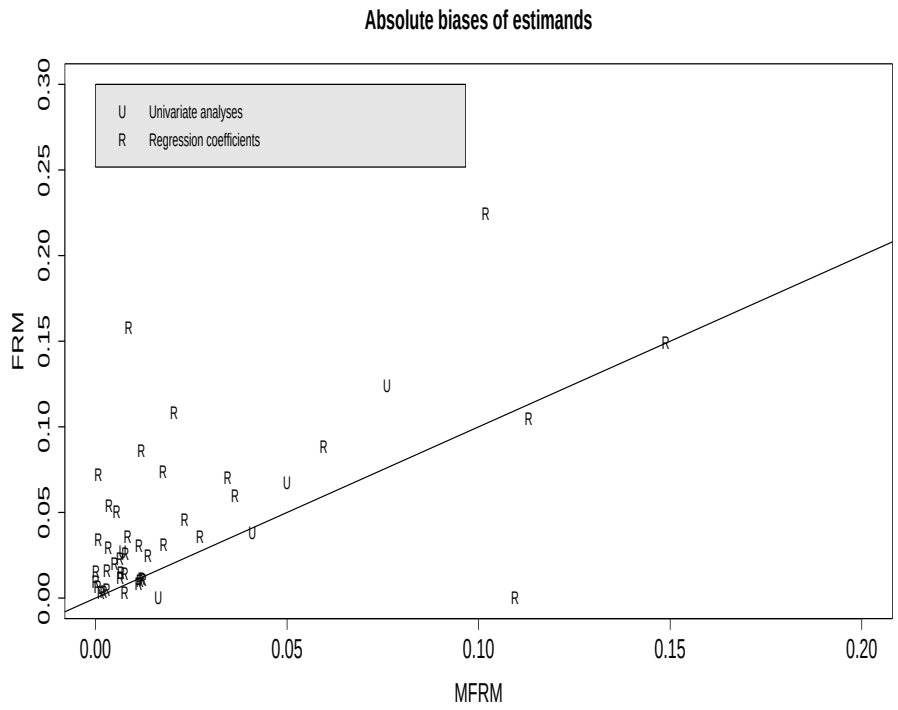


Figure 5.14: Absolute biases of estimands.

plot presents coverages when there are no missing data in the datasets. These coverage are as expected the closest to 0.95. The second box plot shows the coverages from MICE while the third box plot shows the coverages obtained from MFRM. We noticed that MFRM obtains coverages much closer to 0.95 than the coverages obtained from using MICE.

We now plot the biases in the estimates obtained from using MICE against the biases obtained from using MFRM. We distinguish between estimates arising from univariate analyses and estimates arising from multivariate analyses, i.e. regression coefficient estimates. We see that the majority of the large biases are those arising from multivariate analyses, and are above the $y = x$ line, which indicates that MICE tends to obtain regression coefficient estimates further from the true values than MFRM.

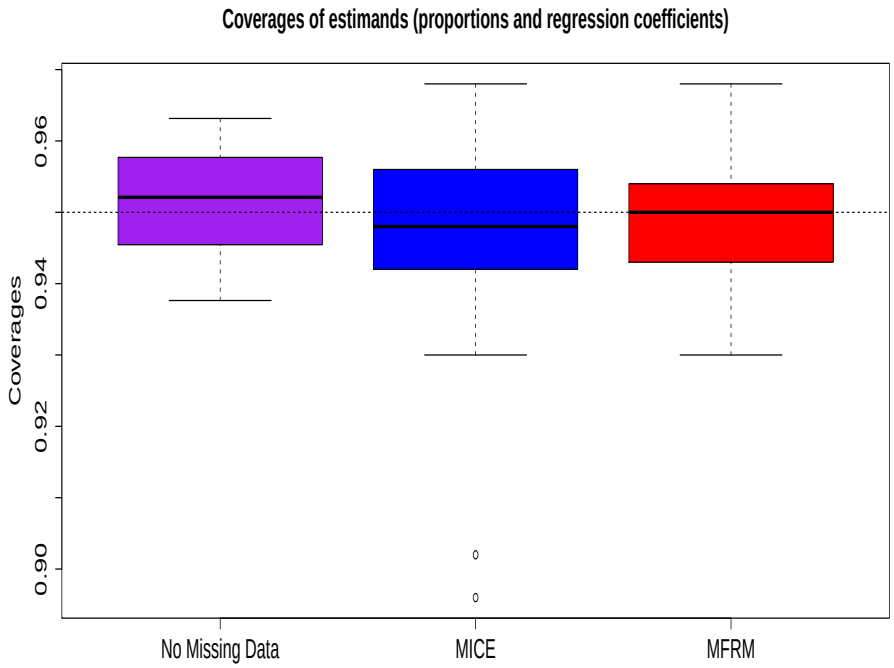


Figure 5.15: The coverages of estimands.

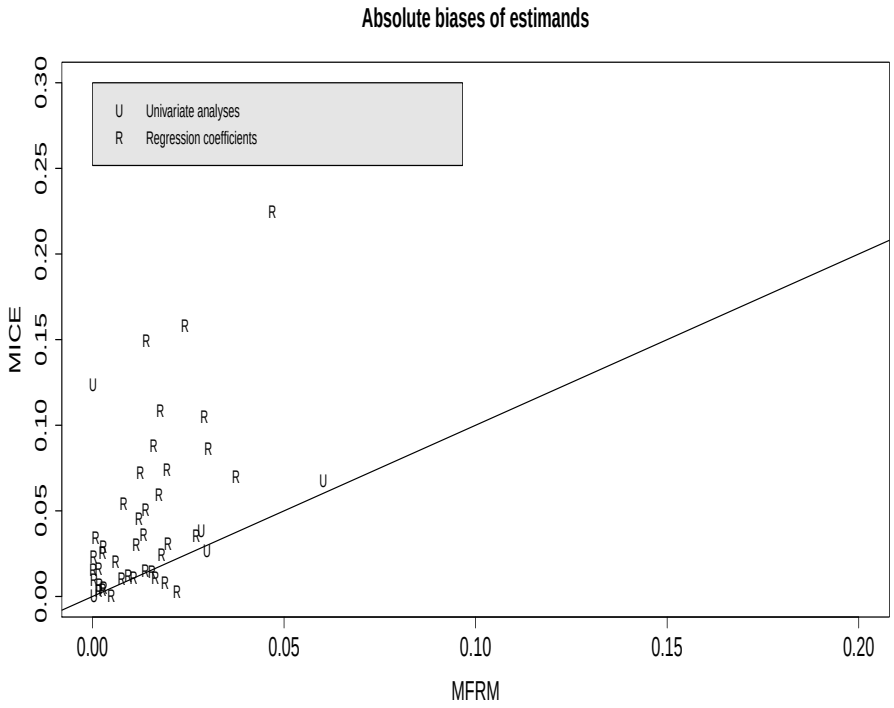


Figure 5.16: Absolute biases of estimands.

5.3.2 Simulation involving the complete cases using FRM, MFRM and MICE

We now apply the FRM, MFRM and MICE to impute missing values in the breast-feeding data set taken from Chapter 3. Before applying FRM, MFRM and MICE to impute missing values in the full sample, we apply the imputation model to a simulation involving the complete case subsample. We reintroduce missing data patterns in the complete case subsample using the same fractions that appeared in the full sample. We can then run FRM, MFRM and MICE to multiply impute the missing values in this incomplete subsample, perform the analysis using the imputed data, and compare the results to the analysis performed on the subsample prior to introducing missing values.

Table 5.5 presents estimates of the regression coefficients and variances obtained from fitting the regression model of PI-ATM on treatment and other covariates. The first column presents estimates prior to introducing missing values, the second column presents results based on the FRM, third column presents results from MFRM while the final column presents results from MICE . From an analysts point of view, the key estimate is the coefficient on treatment, which gives an estimate of the treatment effect; ideally the estimate from the incomplete data should be close to that obtained from the fully observed data. We see that the bias in the treatment effect estimate from using MFRM is about 25.5% when compared to the fully observed estimate, while the bias from using FRM is 20.7% and from using MICE is 19.2%. Hence, in terms of estimating the treatment effect here, there is no potential benefit from using MFRM over FRM and MICE.

In Figure 5.17 we also plot the absolute biases in the regression coefficient estimates from using FRM, again when compared to the fully observed estimates, against similar biases from using MFRM. We see that in this case using FRM for imputation has no potential to obtain estimates closer to those obtained from the fully observed data than using MFRM.

In Figure 5.18 we also plot the absolute biases in the regression coefficient estimates from using MICE, again when compared to the fully observed estimates, against similar biases from using MFRM. We see that in this case using MFRM for imputation has no potential to obtain estimates closer to those obtained from the fully observed data than using MICE.

Table 5.5: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	No missing data	FRM	MFRM	MICE
Intercept.	86.344 (3.386)	85.430 (6.061)	87.095 (7.655)	86.341 (5.923)
Treatment effect.	1.609 (0.948)	1.275 (1.586)	2.020(1.897)	1.918 (1.835)
The number of years between				
1979 and when the mother gave birth.	-0.025 (0.099)	0.001 (0.182)	-0.036 (0.174)	-0.046 (0.110)
The child's race-Black	-1.244 (1.117)	-0.933 (2.078)	-1.005 (1.924)	-0.670 (1.943)
The child's race-Other	2.346 (0.971)	4.159 (1.736)	3.887 (1.819)	4.245 (1.605)
Spouse present at birth.	-1.595 (1.438)	-1.860 (2.879)	-1.769 (2.802)	-1.526 (2.552)
Partner present at birth.	-0.159 (1.104)	0.177 (1.818)	-0.061 (2.238)	0.195 (1.952)
Family income.	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Child's sex.	0.811 (0.676)	0.130 (1.176)	0.142 (1.391)	0.109 (1.201)
Grandparents were present at birth.	-1.078 (1.151)	-1.175 (1.712)	-0.901 (2.243)	-0.739 (1.903)
Mother's intelligence.	0.081 (0.018)	0.057 (0.032)	0.056 (0.035)	0.049 (0.022)
Mother's highest educational.	0.606 (0.212)	0.628 (0.355)	0.483 (0.444)	0.583 (0.249)
Child's birth weight.	0.023 (0.021)	0.026 (0.032)	0.018 (0.055)	0.032 (0.023)
Days that the child spent in hospital.	-0.045 (0.052)	-0.047 (0.089)	-0.028 (0.095)	-0.032 (0.054)
Days that the mother spent in hospital.	-0.283 (0.119)	-0.387 (0.214)	-0.348 (0.207)	-0.456 (0.126)
Weeks that the mother worked-level 2	1.190 (0.984)	2.194 (1.556)	2.304 (1.751)	2.209 (1.414)
Weeks that the mother worked-level 3	0.038 (1.261)	1.653 (2.083)	2.701 (2.467)	0.793 (1.910)
Weeks that the mother worked-level 4	2.156 (1.102)	3.392 (2.003)	4.876 (2.468)	4.514 (2.410)
Child was born premature-level 2	1.265 (0.919)	-0.233 (1.531)	-0.017 (1.874)	0.034 (1.370)
Child was born premature-level 3	2.009 (2.248)	-0.226 (3.718)	-0.670 (5.879)	0.263 (4.527)

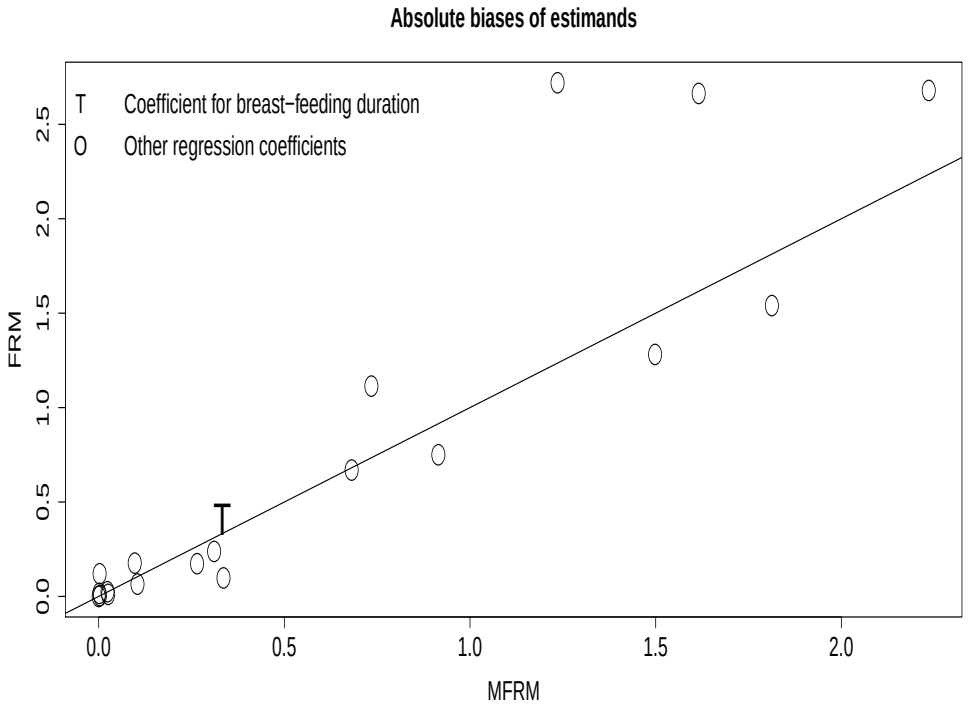


Figure 5.17: Absolute biases of estimands.

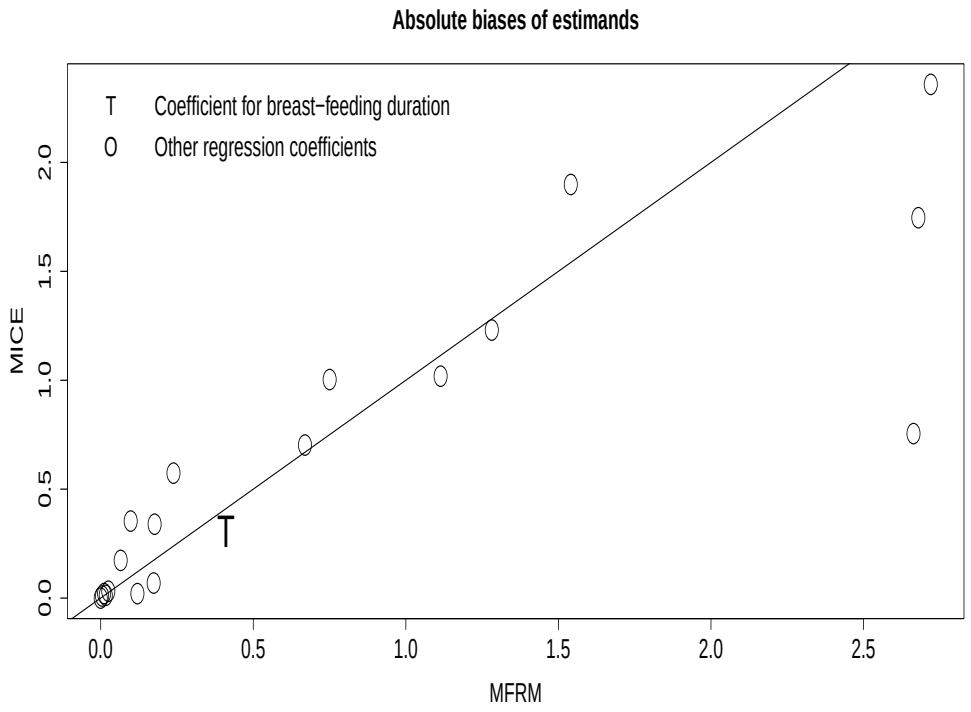


Figure 5.18: Absolute biases of estimands.

5.3.3 Application to the full data sample using FRM, MFRM and MICE

We now apply FRM, MFRM and MICE to the full data sample. For FRM we run $m = 50$ Gibbs samplers at different starting points each for 20000 iterations to generate 50 imputed data sets, we also use MFRM and MICE to generate 50 imputed data sets. Table 5.6 presents coefficient estimates from the regression model described above using both sets of imputed data. We see that there is no real significant difference in the treatment effect estimates using FRM and MFRM; they differ by approximately 0.003 (standard errors from using FRM and MFRM are 0.727 and 1.153 respectively). Also, we notice that there is no real significant difference in the treatment effect estimates using MFRM and MICE; they differ by approximately 0.032 (standard errors from using MFRM and MICE are 1.153 and 0.639 respectively).

Table 5.6: Estimates of regression coefficients and standard errors (standard errors are in parentheses)

	FRM	MFRM	MICE
Intercept.	85.222 (2.578)	88.139(4.101)	86.188 (2.175)
Treatment effect.	0.903 (0.727)	0.906(1.153)	0.938 (0.639)
The number of years between			
1979 and when the mother gave birth.	0.041 (0.071)	0.162(0.083)	0.028 (0.055)
The child's race-Black	-0.549 (0.776)	-0.264(1.174)	-0.649 (0.716)
The child's race-Other	3.093 (0.726)	3.144(1.043)	2.749 (0.577)
Spouse present at birth.	-0.101 (1.227)	-0.317(1.406)	-0.258 (0.935)
Partner present at birth.	0.294 (0.824)	0.122(1.510)	0.130 (0.662)
Family income.	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Child's sex.	0.740 (0.468)	0.567(0.659)	0.851 (0.393)
Grandparents were present at birth.	-0.537 (0.806)	-0.403(0.991)	-0.205 (0.682)
Mother's intelligence.	0.108 (0.014)	0.118(0.017)	0.107 (0.011)
Mother's highest educational.	0.580 (0.162)	0.361(0.285)	0.527 (0.130)
Child's birth weight.	0.013 (0.016)	-0.003(0.028)	0.015 (0.013)
Days that the child spent in hospital.	-0.055 (0.046)	-0.015 (0.037)	-0.064 (0.039)
Days that the mother spent in hospital.	-0.104 (0.079)	0.000(0.000)	-0.182(0.068)
Weeks that the mother worked-level 2	0.960 (0.704)	0.958(0.908)	1.121 (0.562)
Weeks that the mother worked-level 3	1.040 (0.925)	0.745(1.237)	1.009 (0.883)
Weeks that the mother worked-level 4	1.898 (0.871)	1.700(1.054)	1.933 (0.691)
Child was born premature-level 2	0.811 (0.659)	0.361(0.899)	0.783 (0.564)
Child was born premature-level 3	0.958 (1.372)	-0.768(1.996)	0.488 (1.433)

5.3.4 Summary

The MFRM seems to perform better than FRM and MICE in terms of coverages in the simulation study. For both the simulation study and real data application, we set the values of shape (denoted by r) and rate (denoted by δ) for the hyperprior on the shrinkage parameter, λ^2 equal to 2, and we set the degrees of freedom $\nu = 20$ for the t-priors. This results in a ratio value of 10 ($\frac{\nu}{\delta}$). Other ratios have been considered such as 2,5 and 20 where we fix the degrees of freedom $\nu = 20$ and vary the values of $r = \delta = 10$, $r = \delta = 4$ and $r = \delta = 1$. The results obtained do not show any significant differences. Future work includes investigating the ratio of $\frac{\nu}{\delta}$, where we try to find the optimal ratio that may potentially lead to more plausible imputations for our model MFRM.

5.4 Comparison of results from the real data complete cases

We now compare the results of our modelling strategies: FRM, RFRM, LFRM and MFRM through the real data complete cases. The coefficients estimates from each method are tabulated in Table 5.7. We then calculate the absolute biases of coefficients estimates of each method, compare to the estimates obtained from the fully observed complete case data. The results are then tabulated in Table 5.8. We want to find which model will give us the most numbers of estimates that show the smallest absolute biases. Model FRM has 5 estimates that show the smallest absolute biases, model RFRM has 5 estimates, model LFRM has only 3 estimate and model MFRM has 6 estimates that show the smallest absolute bias. There is no significant difference in estimating the coefficient for family income. However, we cannot be certain if the coefficients estimates obtained from the MFRM are reliable estimates of the true estimates. We only use this complete case scenario simply to illustrate the different modelling strategies.

Table 5.7: Estimates of regression coefficients

	No missing data	FRM	RFRM	LFRM	MFRM
Intercept.	86.344	85.430	85.489	90.535	87.095
Treatment effect.	1.609	1.275	1.454	1.393	2.020
The number of years between					
1979 and when the mother gave birth.	-0.025	0.001	-0.012	0.006	-0.036
The child's race-Black	-1.244	-0.933	-0.688	-0.876	-1.005
The child's race-Other	2.346	4.159	4.261	3.169	3.887
Spouse present at birth.	-1.595	-1.860	-1.645	-1.069	-1.769
Partner present at birth.	-0.159	0.177	0.114	-0.589	-0.061
Family income.	0.000	0.000	0.000	0.000	0.000
Child's sex.	0.811	0.130	-0.134	0.085	0.142
Grandparents were present at birth.	-1.078	-1.175	-1.865	0.081	-0.901
Mother's intelligence.	0.081	0.057	0.051	0.041	0.056
Mother's highest educational.	0.606	0.628	0.636	0.459	0.483
Child's birth weight.	0.023	0.026	0.026	0.012	0.018
Days that the child spent in hospital.	-0.045	-0.047	-0.058	-0.061	-0.028
Days that the mother spent in hospital.	-0.283	-0.387	-0.301	-0.309	-0.348
Weeks that the mother worked-level 2	1.190	2.194	1.681	1.692	2.304
Weeks that the mother worked-level 3	0.038	1.653	1.679	1.400	2.701
Weeks that the mother worked-level 4	2.156	3.392	3.779	3.553	4.876
Child was born premature-level 2	1.265	-0.233	-0.032	-0.063	-0.017
Child was born premature-level 3	2.009	-0.226	0.361	-0.353	-0.670

Table 5.8: Absolute biases of regression coefficients estimates

	FRM	RFRM	LFRM	MFRM
Intercept.	0.91470	0.85585	4.19035	0.75030
Treatment effect.	0.3361	0.15478	0.21639	0.41108
The number of years between				
1979 and when the mother gave birth.	0.02546	0.01291	0.03090	0.01118
The child's race-Black	0.31086	0.55615	0.36725	0.23846
The child's race-Other	1.81284	1.91479	0.82323	1.54060
Spouse present at birth.	0.26494	0.04990	0.52563	0.17356
Partner present at birth.	0.33607	0.27320	0.42966	0.09825
Family income.	0.00000	0.00000	0.00000	0.00000
Child's sex.	0.68117	0.94542	0.72568	0.66939
Grandparents were present at birth.	0.09706	0.78743	1.15896	0.17665
Mother's intelligence.	0.02371	0.02965	0.04035	0.02493
Mother's highest educational.	0.00251	0.03301	0.14442	0.12031
Child's birth weight.	0.00340	0.00349	0.01087	0.00459
Days that the child spent in hospital.	0.00181	0.01334	0.01659	0.01659
Days that the mother spent in hospital.	0.10447	0.01826	0.02645	0.06547
Weeks that the mother worked-level 2.	0.73436	0.49079	0.50201	1.11401
Weeks that the mother worked-level 3.	1.61583	1.64119	1.36248	2.66374
Weeks that the mother worked-level 4.	1.23567	1.62340	1.39734	2.66374
Child was born premature-level 2.	1.49833	1.29667	1.32767	1.28171
Child was born premature-level 3.	2.23535	1.64814	2.36216	2.67962

Chapter 6

Statistical disclosure control

When releasing data for public use, statistical agencies seek to reduce the risk of disclosure, at the same time trying to preserve the utility of the release data. A common approach is to alter the original data for public release by applying statistical disclosure limitation (SDL) methods, such as adding random noise to data values (Fuller (1993)), top coding variables (Willenborg and de Waal (2001)), swapping data values (Fienberg and McIntyre (2004)) and among others to those variables that will potentially be identified. However, these methods can distort relationships between variables in the data set. An alternative approach, called partially synthetic data was proposed by Little (1993b). This approach is to release partially synthetic data sets, which comprise some original data values with values at high risk of disclosure being replaced by multiply imputed synthetic values. Valid inference can be made using the combining rules described by Reiter (2003). These combining rules are not the same as those for multiple imputation of missing data (see Chapter 2 for more details).

In this chapter, we will be applying a version of FRM to protect confidentiality of a genuine data set taken from the current population survey (CPS) by generating multiply imputed partially synthetic data sets. Before applying FRM, we will decide on the procedure for selecting the data points that need to be replaced with synthetic data. We then compute and analyze risk and utility measures of the synthetic data sets.

6.1 Selecting the values

In this section, we will discuss the process of selecting the data points that need to be replaced with synthetic values. We will be focusing on the categorical variables because in most scenarios, the key variables which can lead to identification are assumed to be categorical (Drechsler and Reiter (2008)). Suppose we have categorical variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, where $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n,j})'$ for variables $j = 1, \ldots, p$ and units $i = 1, \ldots, n$. Each $x_{i,j}$ for can take values $1, 2, \ldots, w_j$ where w_j is the number of levels in variable $\mathbf{x}_j$. We then order the variables in an anti-lexicographical way, meaning that the first subscript $\mathbf{x}_1$ will

vary the fastest, the second subscript $\mathbf{x}_2$ will vary the second fastest, and so on (see Schafer (1997) for more details). An illustration of anti-lexicographical order is shown below:

Table 6.1: Example of anti-lexicographical ordering

$\mathbf{x}_1$	$\mathbf{x}_2$	$\cdots$	$\mathbf{x}_p$
1	1	$\cdots$	1
2	1	$\cdots$	1
$\vdots$	$\vdots$		$\vdots$
d_1	1	$\cdots$	1
1	2	$\cdots$	1
2	2	$\cdots$	1
$\vdots$	$\vdots$		$\vdots$
d_1	d_2	$\cdots$	d_p

We cross-classify all the variables and we then denote the maximum number of possible combinations of all the variables by D, where $D = \prod_{j=1}^{p} w_j$. We can denote a variable $\mathbf{h} = (h_1, \ldots, h_n)'$ for each unit i where $h_i \in \{1, \ldots, D\}$ and corresponds to the particular combination $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$. Now let the frequencies for the different values of h_i be denoted by

$$f_g = \sum_{i=1}^{n} I(h_i = g), \quad g = 1, \ldots, D,$$

where $I(\cdot)$ is the indicator function: $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. The combination value g is unique if $f_g = 1$. Similarly, $f_g = 2$ means combination value g appears twice in the data set and so on. Let the frequencies of frequencies be denoted

$$n_r = \sum_{g=1}^{D} I(f_g = r), \quad r = 1, 2, \ldots$$

where n_1 represents the number of unique combinations in the data set.

Now suppose that subject i is unique, i.e. $h_i = g, f_g = 1, g \in \{1, \ldots, D\}$. Instead of synthesizing the whole row of data points $(x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ corresponding to this subject i, we only synthesize the data points that make this subject i unique. Table 6.2 shows a simple illustration of this idea.

Table 6.2: Illustration of choosing one data point that need to be synthesized

subject	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	Combination value ($\mathbf{h}$)	Frequency (f_g)
1	2	2	1	2	12	2
2	2	2	1	2	12	2
i	2	2	1	(1)	4	1

Table 6.2 shows a data set consists of 3 units and 4 variables, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, where

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are binary variables and $\mathbf{x}_4$ is a categorical variable with 3 levels. The column $\mathbf{h}$ represents the combination values for each unit and the frequencies for the combination values are denoted by column f_g. All the values in columns $\mathbf{h}$ and f_g are derived using the method discussed previously. Unit 1 and 2 have the same set of data points (2,2,1,2), which is corresponding to the combination value 12 ($h_1 = h_2 = 12$) and this combination appears twice, hence the frequency $f_{12} = 2$. We noticed that unit i is unique, i.e. $f_4 = 1$, with $g = 4$ being the combination value corresponds to unit i. We now need to find the data points that make this subject i unique. It is obvious that the only data point that we need to replace with synthetic data in this example is $x_{i,4}$ (the circled number). Lets formalise this idea of selecting the data points that need to be replaced with synthetic data. We first denote the function $h = h(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ where $h_{i\setminus\{\mathbf{x}_j\}}$ means the element $\mathbf{x}_j$ is being excluded from the set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\}$, i.e. $h_{i\setminus\{\mathbf{x}_j\}} = h(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p)$. The main idea behind this selection procedure is that we try to find a data point $x_{i,j}$ such that when we recompute $h_{i\setminus\{\mathbf{x}_j\}}$, for unit $i = 1, \ldots, n$ with a new set of cross-classified values for $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p$, i.e. aggregated over variable $\mathbf{x}_j$ that takes values $g' \in \{1, \ldots, D'\}$, then for $g' = h_{i\setminus\{\mathbf{x}_j\}}$, we would like $f_{g'} > C_s$, where C_s is some threshold value. If such a variable $\mathbf{x}_j$ can be found, we then replace $x_{i,j}$ with a synthetic value. So if we now aggregate over variable $\mathbf{x}_4$ in the data set from Table 6.2, we have:

Table 6.3: Data set from Table 6.2 when aggregating over variable $\mathbf{x}_4$

subject	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	Combination value ($\mathbf{h}$)	Frequency ($f_{g'}$)
1	2	2	1	4	3
2	2	2	1	4	3
i	2	2	1	4	3

From Table 6.3 we noticed that after aggregating over variable $\mathbf{x}_4$, each unit has a new combination value as well as a new frequency value $f_{g'}$. Unit i is no longer unique since $f_4 = 3$ where $g' = 4$. Hence, instead of replacing $x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}$ with synthetic values, we only need to synthesize $x_{i,4}$. In this simple case, we set the threshold value $C_s = 1$ such that $f_{g'} > 1$ in order to select the data points that need to be replaced with synthetic data for units which are unique, i.e. $f_g = 1$. By setting $C_s = 2$ such that $f_{g'} > 2$, we are selecting the data points that need to be replaced with synthetic data for units which are unique, i.e. $f_g = 1$ as well as for units which have combination value that repeat twice in the data sets as $f_g = 2$. In general, by setting $C_s = r,\ r = 1, 2, \ldots$, we are selecting the data points for units $i = 1, \ldots, n$ such that $f_g \leq r$.

If it is not possible to find a $x_{i,j}$ such that $f_{g'} > C_s$ for $g' = h_{i\setminus\{\mathbf{x}_j\}}$, we can consider a set of variables $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ where $j' \neq j$ and we try to find data points $x_{i,j}, x_{i,j'}$ such that for $g' = h_{i\setminus\{\mathbf{x}_j, \mathbf{x}_{j'}\}}$, where we aggregate over pairs of variables in the data set, we would like $f_{g'} > C_s$, where C_s is the threshold value. If such variables $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ can be found, we then replace $x_{i,j}, x_{i,j'}$ with synthetic values. Table 6.4 shows an example where there are 2 data points that we need to replace with synthetic data, which are $x_{i,2}$ and $x_{i,4}$.

Table 6.4 shows a data set consists of 3 units and 4 variables, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are binary variables and $\mathbf{x}_4$ is a categorical variable with 3 levels. The column $\mathbf{h}$ represents the combination values for each unit and the frequencies for the combination values are denoted by column f_g. We noticed that unit i is unique, i.e. $f_{18} = 1$, with

Table 6.4: Illustration of choosing two data points that need to be synthesized

subject	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	Combination value ($\mathbf{h}$)	Frequency (f_g)
1	2	2	1	2	12	2
2	2	2	1	2	12	2
i	2	(1)	1	(3)	18	1

$g = 18$ being the combination value corresponds to unit i. We now need to find the data points that make this subject i unique and we first aggregate over variable $\mathbf{x}_4$ and we have:

Table 6.5: Data set from Table 6.4 when aggregating over variable $\mathbf{x}_4$

subject	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	Combination value ($\mathbf{h}$)	Frequency ($f_{g'}$)
1	2	2	1	4	2
2	2	2	1	4	2
i	2	(1)	1	2	1

So similar to before, we notice that after aggregating over variable $\mathbf{x}_4$, each unit has a new combination value as well as a new frequency value $f_{g'}$. Unit i is still unique after aggregating over variable $\mathbf{x}_4$ since $f_2 = 1$ where $g' = 2$. Hence, we need to consider aggregating over pairs of variables so that unit i will be unique afterwards, in this case, we see that aggregating over variable $\mathbf{x}_2$ results in:

Table 6.6: Data set from Table 6.4 by aggregating over variables $\mathbf{x}_2$ and $\mathbf{x}_4$

subject	$\mathbf{x}_1$	$\mathbf{x}_3$	Combination value ($\mathbf{h}$)	Frequency ($f_{g'}$)
1	2	1	2	3
2	2	1	2	3
i	2	1	2	3

Hence unit i is no longer unique after we aggregate over variable $\mathbf{x}_2$ and $\mathbf{x}_4$ since $f_2 = 3$ where $g' = 2$. Hence, we found the data points that need to be replaced with synthetic values, which are $x_{i,2}$ and $x_{i,4}$. In general, if we can't satisfy the criteria $h_{i\setminus\{\mathbf{x}_j,\mathbf{x}_{j'}\}} > C_s$ by aggregating over variables $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$, we will keep increasing the number of variables $\{\mathbf{x}_j, \mathbf{x}_{j'}, \mathbf{x}_{j''}, \ldots\}$, where $j \neq j' \neq j''$ until the criteria $h_{i\setminus\{\mathbf{x}_j,\mathbf{x}_{j'},\mathbf{x}_{j''},\ldots\}} > C_s$ is satisfied.

If it is possible to find a $x_{i,j}$ such that $f_{g'} > C_s$ for $g' = h_{i\setminus\{\mathbf{x}_j\}}$ and a $x_{i,\tilde{j}}$ where $\tilde{j} \neq j$ such that $f_{\tilde{g}} > C_s$ for $\tilde{g} = h_{i\setminus\{\mathbf{x}_{\tilde{j}}\}}$, we will select both the data points $x_{i,j}$ and $x_{i,\tilde{j}}$. Table 6.7 shows an illustration of this case where we select both $x_{i,2}$ and $x_{i,4}$ given we can make unit i no longer unique by aggregating over either $\mathbf{x}_2$ or $\mathbf{x}_4$.

From Table 6.7, we noticed that unit i is unique, i.e. $f_6 = 1$, with $g = 6$ being the combination value corresponds to unit i. We now need to find the data points that make this subject i unique. We first aggregate over variable $\mathbf{x}_4$ and the data set is tabulated in Table 6.8. We noticed that unit i is no longer unique after we aggregate over variable $\mathbf{x}_4$

Table 6.7: Illustration of choosing both data points that need to be synthesized

subject	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	Combination value ($\mathbf{h}$)	Frequency (f_g)
1	2	2	1	2	12	2
2	2	2	1	2	12	2
i	2	(1)	2	(1)	6	1
4	2	1	2	3	22	2
5	2	1	2	3	22	2
6	2	2	2	1	8	3
7	2	2	2	1	8	3
8	2	2	2	1	8	3

since $f_6 = 3$ where $g' = 6$. Now suppose instead of aggregating over $\mathbf{x}_4$, we aggregate over variable $\mathbf{x}_2$ and the data set is shown in Table 6.9.

Hence unit i is no longer unique after we aggregate over variable $\mathbf{x}_2$ since $f_4 = 4$ where $g' = 4$. This shows that it is possible to find a $x_{i,j}$ such that $f_{g'} > C_s$ for $g' = h_{i\setminus\{\mathbf{x}_j\}}$ and a $x_{i,\tilde{j}}$ where $\tilde{j} \neq j$ such that $f_{\tilde{g}} > C_s$ for $\tilde{g} = h_{i\setminus\{\mathbf{x}_{\tilde{j}}\}}$, in this case, we will select both the data points $x_{i,2}$ and $x_{i,4}$. In the general case of we have two subsets of variables $\mathbf{A}_i, \mathbf{B}_i \subseteq \{x_{i,1}, x_{i,2}, \ldots, x_{i,p}\}$ and $f_{g'} > C_s$ for $g' = h_{i\setminus\{\mathbf{A}_i\}}$ and $f_{\tilde{j}} > C_s$ for $\tilde{j} = h_{i\setminus\{\mathbf{B}_i\}}$, then we select the union $\mathbf{A}_i \cup \mathbf{B}_i$ to synthesize.

Table 6.8: Data set from Table 6.7 by aggregating over variables $\mathbf{x}_4$

subject	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	Combination value ($\mathbf{h}$)	Frequency ($f_{g'}$)
1	2	2	1	4	2
2	2	2	1	4	2
i	2	1	2	6	3
4	2	1	2	6	3
5	2	1	2	6	3
6	2	2	2	8	3
7	2	2	2	8	3
8	2	2	2	8	3

Table 6.9: Data set from Table 6.7 by aggregating over variables $\mathbf{x}_2$

subject	$\mathbf{x}_1$	$\mathbf{x}_3$	$\mathbf{x}_4$	Combination value ($\mathbf{h}$)	Frequency ($f_{g'}$)
1	2	1	2	6	2
2	2	1	2	6	2
i	2	2	1	4	4
4	2	2	3	12	2
5	2	2	3	12	2
6	2	2	1	4	4
7	2	2	1	4	4
8	2	2	1	4	4

6.2 Risk of identification

We now briefly describe how to evaluate the risk of identification in the partially synthetic data sets. We assume the variable that the intruder possesses are categorical variables. Now suppose we have a data set with only key categorical variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, where $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n,j})'$ for $j = 1, \ldots, p$. Each $x_{i,j}$ for can take values $1, 2, \ldots, w_j$ where w_j is the number of levels in variable $\mathbf{x}_j$. From Section 6.1, each subject i is assigned a value $h_i \in \{1, \ldots, D\}$ depending on its value $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$ with D as defined previously. We denote the d multiply imputed data sets by $\mathbf{D}_{syn}^{(l)}$, for $l = 1, \ldots, d$. For each of the partially synthetic data $\mathbf{D}_{syn}^{(l)}$, we can repeat the process in Section 6.1 where each subject i is assigned a value $h_i^{(l)} \in \{1, \ldots, D\}$ depending on its partially synthetic values $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$. The risk analysis proceeds as follows:

Suppose the intruder possesses categorical variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, where $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n,j})'$ for $j = 1, \ldots, p$. From Section 6.1, we can express this set of variables as n records of target information $t_1, t_2, \ldots, t_n$ depending on its value $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$. We start by matching the first target record t_1 to $h_i^{(1)}$ from the first partially synthetic data $\mathbf{D}_{syn}^{(1)}$. We then create a vector $\mathbf{m}_1^{(1)} = (m_{1,1}^{(1)}, m_{2,1}^{(1)}, \ldots, m_{n,1}^{(1)})'$ where $m_{i,1}^{(1)} = 1$ for unit $i = 1, \ldots, n$ if the target record t_1 matches record $h_i^{(1)}$, and zero otherwise. We can then assign a probability vector $\mathbf{p}_1^{(1)} = (p_{1,1}^{(1)}, p_{2,1}^{(1)}, \ldots, p_{n,1}^{(1)})'$, where $p_{i,1}^{(1)} = \frac{m_{i,1}^{(1)}}{\sum_{i=1}^{n} m_{i,1}^{(1)}}$ for units $i = 1, \ldots, n$. If $\sum_{i=1}^{n} m_{i,1}^{(1)} = 0$, then $\mathbf{p}_1^{(1)} = \mathbf{0}$. For the first partially synthetic data $\mathbf{D}_{syn}^{(1)}$, this process is repeated for all target records t_k, for $k = 1, \ldots, n$, resulting in n sets of probability vectors $\mathbf{p}_k^{(1)}$, for $k = 1, \ldots, n$. We then repeat the matching process again for the d imputed synthetic data sets, where we have $\mathbf{p}_k^{(l)} = (p_{1,k}^{(l)}, p_{2,k}^{(l)}, \ldots, p_{n,k}^{(l)})'$, for $k = 1, \ldots, n$ and $l = 1, \ldots, d$. We then average the probability vectors across d imputed data sets and we have $\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2, \ldots, \bar{\mathbf{p}}_n$ where $\bar{\mathbf{p}}_k = (\bar{p}_{1,k}, \bar{p}_{2,k}, \ldots, \bar{p}_{n,k})'$ for $k = 1, \ldots, n$, with each $\bar{p}_{i,k} = \sum_{l=1}^{d} \frac{p_{i,k}^{(l)}}{d}$ for target records $k = 1, \ldots, n$ and units $i = 1, \ldots, n$.

We then denote a $n \times n$ matrix $\mathbf{P}$ in the following way:

$$\mathbf{P} = \begin{pmatrix} \bar{p}_{1,1} & \bar{p}_{1,2} & \cdots & \cdots & \bar{p}_{1,n} \\ \bar{p}_{2,1} & \bar{p}_{2,2} & \cdots & \cdots & \bar{p}_{2,n} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \bar{p}_{n,1} & \bar{p}_{n,2} & \cdots & \cdots & \bar{p}_{n,n} \end{pmatrix}$$

with $\bar{p}_{i,j}$ represent the i-th row and j-th column entries in the matrix $\mathbf{P}$,, which is the probability that the j-th target record is being matched to the i-th unit in the data set.

We now describe our risk measures. The first risk measure, which we call perceived match risk, is the number of units with maximum probability of being identified that exceed some threshold value c that we consider risky, i.e. we compute $\sum_{j=1}^{n} I(\max_{1 \le i \le n}(\bar{p}_{i,j}) > c)$. In our case, we set this threshold value to be 0.20. This represents the number of

records that have a high chance of being identified, either correctly or incorrectly. The second risk measure, which we call expected match risk, represents the number of records that the intruder would guess correctly from the synthetic data. This can be calculated from the trace of $\mathbf{P}$, i.e. $\sum_{i=1}^{n} \bar{p}_{i,i}$. The third risk measure, which we call true match risk, equals $\sum_{i=1}^{n} I(\bar{p}_{i,i} = 1)$, i.e. the number of units the intruder will match correctly with probability 1. For more details on this framework, see Reiter and Mitra (2009) and Drechsler and Reiter (2008).

6.3 Current population survey (CPS)

We now proceed to generate partially synthetic data sets for a real data set. This real data set is drawn from the March 2000 US current population survey (CPS). This data set consists of 49436 units and 9 variables. Similar data set has also been used by Reiter (2005b) and Drechsler and Reiter (2008). With the generated synthetic data sets, we wish to compare the different risk measures calculated from the synthetic data sets with different amount of synthetic data.

6.3.1 Description of variables

The are 9 variables in the data set. These variables are:

- $\mathbf{x}_1$ - Race. (Nominal: 1-White, 2-Black, 3-Others)
- $\mathbf{x}_2$ - Marital Status. (Nominal: 7 levels)
- $\mathbf{x}_3$ - Age. (Continuous: 15-90)
- $\mathbf{x}_4$ - Income. (Continuous: 1-768700)
- $\mathbf{x}_5$ - Social security payment. (Continuous: 0,1-50000)
- $\mathbf{x}_6$ - Tax. (Continuous: 0,1-100000)
- $\mathbf{x}_7$ - Child support payment. (Continuous: 0,1-23920)
- $\mathbf{x}_8$ - Sex. (Binary: 1-Male, 2-Female)
- $\mathbf{x}_9$ - Highest education level attained. (Ordinal: 16 levels)

where race has three levels: 1-White, 2-Black, 3-Others. Marital status has seven levels: 1 for married civilians with both spouses present at the home; 2 for married people in the armed forces with both spouses present at the home; 3 for married people with one spouse not present at the home; 4 for widowers; 5 for divorced people; 6 for separated people and 7 for people who never have been married. The variable $\mathbf{x}_9$ (highest attained education level) has 16 levels in correspondence with years of schooling. For examples, 31 represents

highest educational attainments of less than first grade, 32 represents 1st to 4th grade, 33 represents 5th or 6th grade, 34 represents 7th and 8th grade, 35 represents highest educational attainments of 9th grade, 36 represents 10th grade, 37 represents 11th grade, 38 represents 12th grade, 39 represents high school graduate, 40 represents some college but no degree, 41 represents associate degree in college - occupation/vocation program, 42 represents associate degree in college - academic program, 43 represents Bachelor's degree, 44 represents Master's degree, 45 represents professional school degree and 46 represents doctoral degree. There are 5 continuous variables in the data sets, e.g. age, income, social security payment, tax and child support payment. We noticed that out of the 5 continuous variables, 3 continuous variables such as social security, tax and child support payment show spikes at zero (see Appendix I for density plots of these variables). In order to deal with the huge spikes at zero, we will be introducing a new modelling strategies in the next section.

6.3.2 Skip factored regression model (SFRM)

Suppose we have a fully observed $n \times p$ data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})'$, $j = 1, \ldots, p$ is the j-th variable in $\mathbf{X}$. We assume that there are k nominal variables and following Chapter 3, we order the variables so that $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are nominal, each $\mathbf{x}_q$, $q = 1, \ldots, k$ taking a set of possible values $1, \ldots, L_q$. Variables $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_p$ could then be measured on a binary, ordinal, continuous scale or continuous variables with spikes at zero. Let $\mathbf{x}_q$, $q \in \{k+1, \ldots, p\}$ be the continuous variable with a huge spike of zero. We now extend the FRM strategy to accommodate the huge spike at zero in the variable $\mathbf{x}_q$. First, we generate a latent variable $\mathbf{z}_q^*$ where $\mathbf{z}_q^* = (z_{1,q}^*, z_{2,q}^*, \ldots, z_{n,q}^*)'$ from the linear regression model represented by,

$$p(z_{i,q}^* | x_{i,1}, x_{i,2}, \ldots, x_{i,q-1}, \boldsymbol{\varphi}^{(q)}) = N(\varphi_0^{(q)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \varphi_b^{(q),j^b} I(x_{i,b} = j^b) + \sum_{b=k+1}^{q-1} \varphi_b^{(q)} x_{i,b}, 1).$$

We then denote an indicator variable, $\mathbf{z}_q$, such that $\mathbf{z}_q = (z_{1,q}, \ldots, z_{n,q})'$ where $z_{i,q} = I(z_{i,q}^* > 0)$. We can generate the synthetic values $x_{i,q}^*$ from the following way (for notational convenience we assume that if $x_{i,k}$, $k = 1, \ldots, p$ is continuous, latent or nominal then $x_{i,k} = x_{i,k}^*$):

$$\begin{aligned} x_{i,q}^* &= 0 \text{ for } z_{i,q} = 0, \\ x_{i,q}^* &\sim p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, x_{i,q+1}^*, \ldots, x_{i,k}^*, x_{i,k+1}^*, \ldots, x_{i,p}^*, \Theta) \text{ for } z_{i,q} = 1, \quad (6.1) \end{aligned}$$

where $x_{i,q}^*$ for $z_{i,q} = 1$ is generated from its full conditional distribution (Equation 6.1) which follows a normal distribution with mean

$$\widetilde{\mu}_{i,q} = \widetilde{\phi}_q^{-1} \left\{ \frac{\mu_{i,q}}{\phi_q^{-1}} + \sum_{s=q+1}^{p} \frac{\beta_q^{(s)}}{\phi_s^{-1}} \left[x_{i,s}^* - (\mu_{i,s} - \beta_q^{(s)} x_{i,q}^*) \right] \right\}, \text{ for } z_{i,q} = 1,$$

and variance

$$\widetilde{\phi}_q^{-1} = \left(\frac{1}{\phi_q^{-1}} + \sum_{s=q+1}^{p} \frac{(\beta_q^{(s)})^2}{\phi_s^{-1}} \right)^{-1},$$

where $\mu_{i,q} = \beta_0^{(q)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \beta_b^{(q),j^b} I(x_{i,b}^* = j^b) + \sum_{b=k+1}^{q-1} \beta_b^{(q)} x_{i,b}^*$ and $\mu_{i,s} = \beta_0^{(s)} + \sum_{b=1}^{k} \sum_{j^b=2}^{L_b} \beta_b^{(s),j^b} I(x_{i,b}^* = j^b) + \sum_{b=k+1}^{s-1} \beta_b^{(s)} x_{i,b}^*$. The full conditional distributions of the rest of the parameters can be derived using the same strategies from Chapter 3. We name this modelling strategy as skip factored regression model (SFRM).

6.3.3 Simulation study

We now compare the performances of FRM and SFRM using a simulation study. We simulate $x_{i,1}, x_{i,2}, \ldots, x_{i,5},\ i = 1, \ldots, 1000$ in the following way:

$$\begin{aligned}
x_{i,1} &\quad \text{simulated from a discrete distribution taking values} \in \{1,2,3\} \\
&\quad \text{with probabilities 0.3, 0.4 and 0.3 respectively.} \\
x_{i,2} &\sim p(x_{i,2}|x_{i,1}, \boldsymbol{\theta}^{(2)}), \\
x_{i,3} &\sim \mathrm{N}(11, 4) \\
x_{i,4}^* &\sim \mathrm{N}(5 + 2x_{i,3}, 9), \\
x_{i,4} &= I(x_{i,4}^* > 27) \\
z_{i,5}^* &\sim \mathrm{N}(6 + 3 * I(x_{i,1} = 2) + 4 * I(x_{i,1} = 3) - x_{i,3}, 9), \\
z_{i,5} &= I(z_{i,5}^* > 0)
\end{aligned}$$

We then simulate $x_{i,5}$ as below:

$$\begin{aligned}
x_{i,5} &= 0 \ \text{ for } \ z_{i,5} = 0 \\
x_{i,5} &\sim \mathrm{N}(1 + 3 * I(x_{i,1} = 2) * I(z_{i,5} = 1) + I(x_{i,1} = 3) * I(z_{i,5} = 1) - \\
&\quad x_{i,3} * I(z_{i,5} = 1), 4) \ \text{ for } \ z_{i,5} = 1
\end{aligned}$$

where

$$p(x_{i,2} = j^2 | x_{i,1}, \boldsymbol{\theta}^{(2)}) = \frac{\exp(\theta_{0,j^2}^{(2)} + \theta_{1,j^2}^{(2)} I(x_{i,1} = 2) + \theta_{2,j^2}^{(2)} I(x_{i,1} = 3))}{\sum_{s=1}^{5} \exp(\theta_{0,s}^{(2)} + \theta_{1,s}^{(2)} I(x_{i,1} = 2) + \theta_{2,s}^{(2)} I(x_{i,1} = 3))}$$

where $j^2 \in \{1, \ldots, 5\}$ and $\theta_{0,1}^{(2)} = \theta_{1,1}^{(2)} = \theta_{2,1}^{(2)} = 0$ for identifiability, $\theta_{0,2}^{(2)} = 1, \theta_{1,2}^{(2)} = -3, \theta_{2,2}^{(2)} = -1, \theta_{0,3}^{(2)} = -1, \theta_{1,3}^{(2)} = 2, \theta_{2,3}^{(2)} = 2, \theta_{0,4}^{(2)} = -1, \theta_{1,4}^{(2)} = 2, \theta_{2,4}^{(2)} = -2, \theta_{0,5}^{(2)} = -3, \theta_{1,5}^{(2)} = 1, \theta_{2,5}^{(2)} = 5$, and $I(\cdot)$ is the indicator function. Variables $\mathbf{x}_1$ and $\mathbf{x}_2$ are nominal variables, $\mathbf{x}_3$ is continuous, $\mathbf{x}_4$ is binary while $\mathbf{x}_5$ is the continuous variable with spike at zero.

We replicate this data generation process 1000 times to generate 1000 data sets. We want to generate synthetic values for $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_5$. Using the SFRM imputation strategy, we generate multiply imputed partially, synthetic data sets for each data set

$d = 10$ times. To generate d independent partially synthetic data, we run d Metropolis within Gibbs samplers, each with a different starting value, with each sampler resulting in a synthetic data set after convergence. We assume analysts may be interested in making inferences about various types of estimands arising from both univariate analyses, e.g. population means of variables or the proportion in the population taking a particular level of a categorical variable, as well as multivariate analyses, e.g. the coefficient from a regression model. Using the d imputed data sets, we apply the combining rules described in Chapter 2 (see Equation 2.3) to construct point and interval estimates for these estimands. We also construct estimates for the same estimands when using FRM to generate multiply imputed partially synthetic data sets.

When implementing SFRM or FRM for imputations, we assume the imputer knows the data generating process and so decomposes the joint distribution of the imputation model as described in Equation 3.3. The analysis of the imputed data sets will also respect the ordering of the variables used to generate the synthetic data, the analysis model is thus congenial to the imputation model. See Appendix J for a full list of the estimands considered.

Figure 6.1 presents box plots of coverages for the estimands using the 95% confidence intervals constructed from the imputations generated by FRM and SFRM over the 1000 datasets. These estimands include those arising from both univariate analyses and regression analyses. The first box plot shows the coverages from FRM while the second box plot shows the coverages obtained from SFRM. We see that the SFRM obtains coverages much closer to 0.95 than the coverages obtained from using FRM.

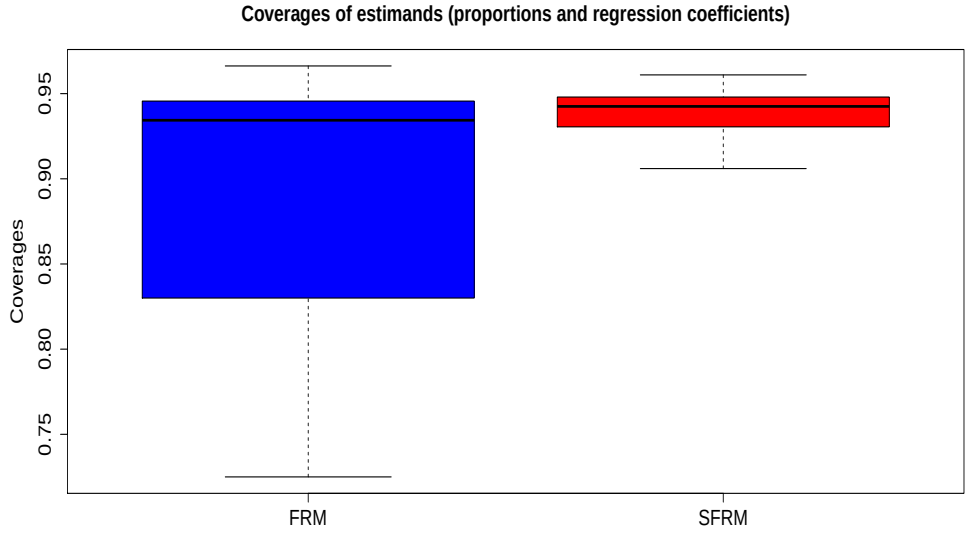


Figure 6.1: The coverages of estimands.

6.3.4 Generating partially synthetic data sets

We have proposed a method to model the data set with continuous variables with spikes at zero. We now proceed to generate partially synthetic data sets for the CPS data set.

We generate synthetic values for race, marital status, sex and highest education attained using the methods described in the previous section. Such key variables have been used by Drechsler and Reiter (2008). We select the threshold values $C_s = 0, 1, 2, 5, 10$ and 50 where 0 means we synthesize all the variables, and we also calculated the number of data points being synthesized with respect to each C_s. We repeat the process 50 times to generate 50 imputed data sets for each of the chosen threshold value C_s. We then determine the respective risk measures described in the previous section and the results are tabulated as below:

Table 6.10: Risk measures

Number of data points being synthesized (value of C_s)	Perceived (0.2)	Expected match	True match
95 (1)	271	472.32	0
545 (2)	167	419.30	0
1697 (5)	8	346.25	0
3993 (10)	0	271.88	0
19084 (50)	1	110.39	0
197744 (0)	0	2.12	0

The result shows that there is no true match risk in the multiply imputed data sets. We also noticed that the perceived risk and expected match risk reduce when we increase the number of data points being replaced. We are interested to know how the partially synthetic data sets are affecting the utility of the data sets. We will discuss the utility measure in the next section.

6.3.5 Propensity score measure

In this section, we present a quantitative measure of data utility for the partially synthetic data using the propensity score approach (Woo et al. (2009)) . We first merge the original data set (size n), with partially synthetic data set (size n) $\mathbf{D}^{(l)}_{syn}$, creating a data set $\mathbf{X}^{(l)}_{merge},\ l = 1, \ldots, d$, with size of $2n$. We denote a vector $\mathbf{T}^{(l)} = (T^{(l)}_1, T^{(l)}_2, \ldots, T^{(l)}_{2n})$ where $T^{(l)}_i = 1$ if subject i belongs to the partially synthetic data set, in our case $T^{(l)}_1, \ldots, T^{(l)}_n$ is equal to 0 and $T^{(l)}_{n+1}, \ldots, T^{(l)}_{2n}$ is equal to 1. We then regress $\mathbf{T}^{(l)}$ on $\mathbf{X}^{(l)}_{merge}$ and use the estimated probabilities from the regression model as the propensity scores. Each of the propensity score for subject i is denoted by $u^{(l)}_i$ for $i = 1, \ldots, 2n$. We then average the propensity score $u^{(l)}_i$ across 50 imputed data sets and denoted the average propensity score by $\bar{u}_i$ for $i = 1, \ldots, 2n$. We then calculate the data utility measure as below:

$$U_p = \frac{1}{2n} \sum_{i=1}^{2n} (\bar{u}_i - 0.5)^2, \tag{6.2}$$

where $2n$ is the total number of units in the merged data set. When the original and partially synthetic data have the same distribution, the propensity scores for all units

should be approximately equal to 0.5, so that U_p is close to zero. On the other hand, if $U_p = 0.25$, then the two data sets are completely distinguishable.

We now calculate the global propensity scores for our partially synthetic data sets that we synthesize in Section 6.3.4, and calculate the utility measure using Equation 6.2. The results are tabulated below:

Table 6.11: Global propensity scores

Number of data points being synthesized (value of C_s)	Global propensity score	Utility
95 (1)	4.742×10^{-7}	0.9999995
545 (2)	1.187×10^{-5}	0.9999881
1697 (5)	8.462×10^{-5}	
3993 (10)	0.0003104	0.9999154
19084 (50)	0.003523	0.996477
197744 (0)	0.007936	0.992064

The utility is represented by the values of one minus the global propensity scores, where the smaller the values, the less the utility of the data set. We notice that the utility has been reduced as we increase the number of data points being synthesized as expected. The relationship between the risk measures and utility is illustrated in Figure 6.2 and Figure 6.3.

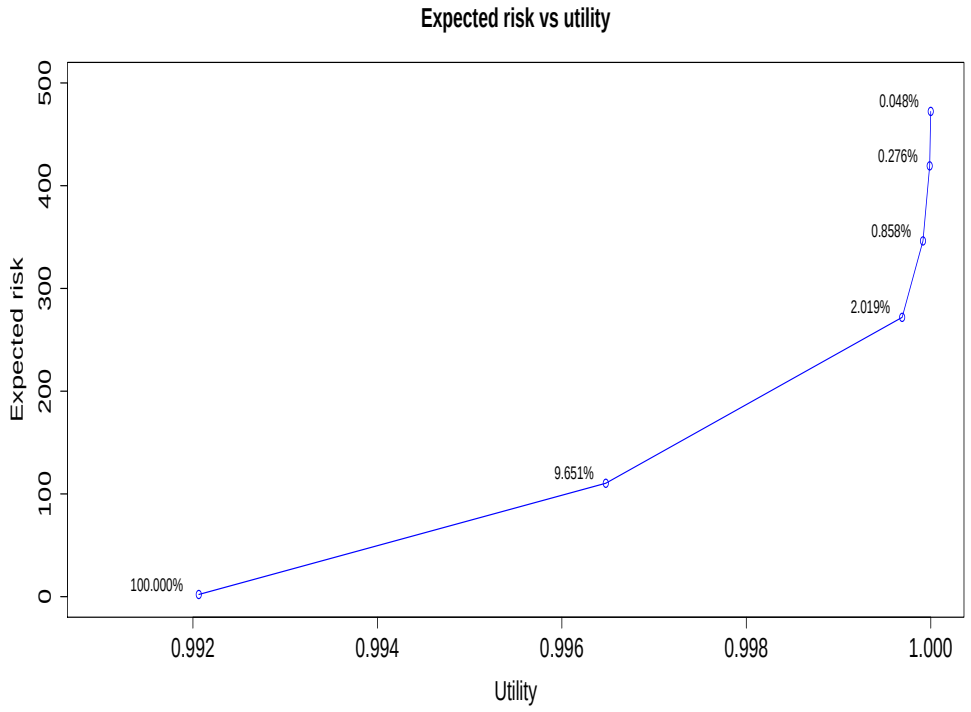


Figure 6.2: Expected risk versus utility plot.

Figure 6.2 and Figure 6.3 show the plots of expected risk and perceived risk with threshold value 0.2 against the utility respectively. Each of the points in both plots are

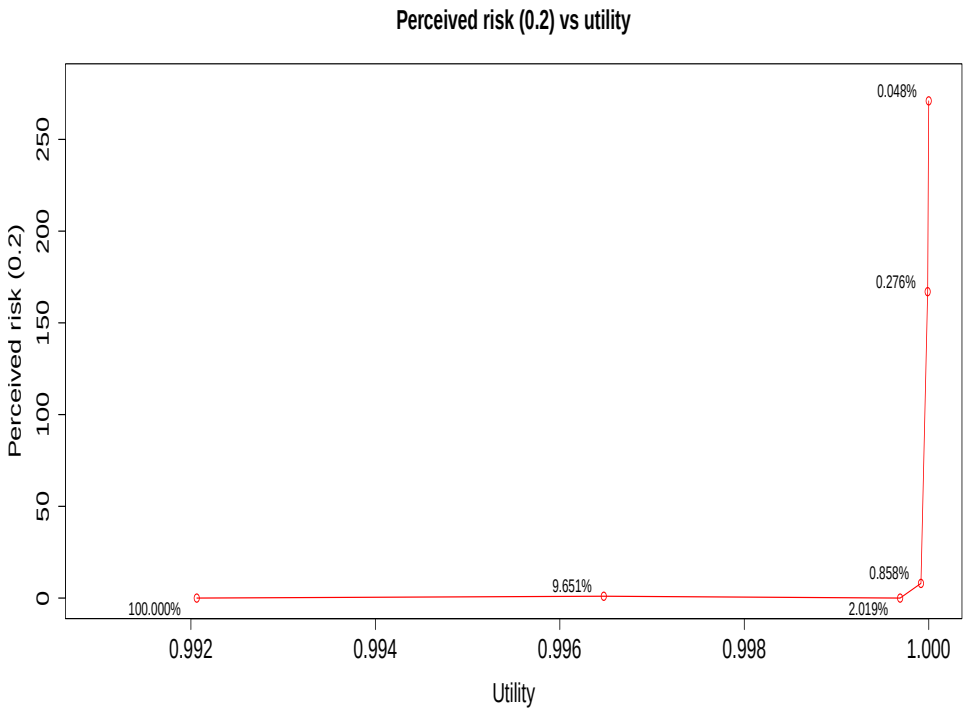


Figure 6.3: Perceived risk (0.2) versus utility plot.

labelled with the percentages of the key variables values being synthesized. Similarly to before, we noticed that the utility has been reduced (values of global propensity scores are increasing) as we increase the percentage of data points being synthesized, but this also results in reduction of the disclosure risk.

Chapter 7

Conclusions

In this thesis, we have reviewed two major applications of multiple imputation: imputing missing values and generating partially synthetic data for protecting confidentiality of public use data sets. We have proposed multiple imputation modelling strategies that can impute the missing values, as well as generating partially synthetic data sets to protect confidentiality of data sets with complex settings. In this chapter, we conclude our thesis in two parts: the first part presents concluding remarks for the missing data problems, which summarizes the modelling methods (Chapter 3), simulation study and real data application (Chapter 4) and informative priors for our model (Chapter 5). In the second part of this chapter we present some conclusions concerning contributions to the data confidentiality in Chapter 6, where we discuss our modelling strategy on generating partially synthetic data for protecting confidentiality and some wider context on the applications of synthetic data.

7.1 Missing data problems

Missing data are a common problem for many researchers and statisticians in the world. Failing to handle the missing data adequately will lead to biased estimates of parameters such as means or regression coefficients in statistical analysis. We proposed a model that can be used to impute the missing values, where the imputed values are drawn from the posterior predictive distribution in order to preserve the statistical properties in the imputed data sets. The proposed model, FRM, is a joint modelling approach which can impute missing values and generate synthetic data in data sets that contain both categorical and continuous variables with a non-monotone pattern. When the data comprise only binary or ordinal variables, then the modelling strategy allows imputations to be drawn from their posterior predictive distributions using a Gibbs sampler. When the data also comprise nominal variables then we proposed a Metropolis within Gibbs sampler with a novel Metropolis-Hastings proposal that can efficiently impute the missing values. For the fully conditional specifications, we used the modelling approach developed by Raghunathan et al. (2001), also known as MICE, to implement multiple imputation.

We used MICE as an alternative multiple imputation approach to compare to our proposed model FRM, in both the simulation study and real data application.

In the simulation study involving imputing the missing values, the results showed that FRM obtained coverages of estimands much closer to 0.95 compared to those obtained using MICE. These estimands included those arising from both univariate analyses e.g. population means of variables or the proportions in the population taking a particular level of a categorical variable, as well as multivariate analyses, e.g. the coefficients from a regression model. When implementing FRM and MICE for imputation, we explored different orderings of the predictors (to that used to generate the data) to decompose the joint distribution and impute missing values. Results showed that FRM still performed better than MICE and there aren't any significant differences in the coverages obtained using both methods. In the real data application, there is a potential benefit from using FRM over MICE in terms of estimating the treatment effect in the complete case scenario. We investigated how sensitive results were to three different ordering of variables in the decomposition of the joint distribution, and we found no significant differences in the results obtained.

We also extended our modelling strategies to incorporate different informative priors for FRM to explore robust regression modelling and the possibility of sparse relationships between the variables. In the RFRM strategy, we have assigned a marginal t-prior to the regression parameters in our model, and assumed that the errors of the linear regression models follow a marginal t distribution. The results showed that the robust model RFRM performed better than FRM in the simulation study and in the real data complete case analysis. We also compared the results between RFRM and MICE, where in the MICE approach, we used the model and prior specifications described in the RFRM strategy to generate missing values. We noticed that the RFRM obtained better coverages than the coverages obtained using MICE in both the simulation study and real data complete case analysis. Here then, we have seen that there is significant advantage using the RFRM approach, over FRM and the equivalent fully conditional specifications MICE.

For the LFRM approach, we incorporated the Bayesian Lasso (Park and Casella (2008)) into our modelling strategy but FRM seemed to perform better than LFRM in terms of coverages in the simulation study. We also compared the results between LFRM and MICE, where in the MICE approach, we use the model and prior specifications described in the LFRM strategy to generate missing values. Results showed that both MICE and LFRM did not show any significant differences in the coverages in the simulation study and real data complete case analysis. More research on methods of choosing the shrinkage parameter λ^2 can be done, and we could also consider incorporating other shrinkage methods such as the ridge regression proposed by Hoerl and Kennard (1970) into our modelling strategy.

For the MFRM approach, we assigned a marginal t-prior for the errors of the linear regression models, which are more robust to the outliers in our data, and we also assigned the Bayesian Lasso prior for the regression parameters. This strategy allows us to include robust modelling as well as exploring potential sparse relationship between variables in the data set under one modelling strategy. We also compared the results between MFRM and MICE, where in the MICE approach, we use the model and prior specifications described in the MFRM strategy to generate missing values. Results showed that MFRM performed

better than FRM and MICE in terms of coverages in the simulation study. Results showed that both MICE and RFRM did not show any significant differences in results in the real data complete case analysis. More analysis on finding the optimal ratio for $\frac{\nu}{\delta}$, where ν represents the degrees of freedom ν for the t-priors and δ represents the hyperprior on the shrinkage parameter, λ^2 can be considered in the future. We also compared the results of FRM, RFRM, LFRM and MFRM through the simulation study involving the breast-feeding data complete cases. Results showed that MFRM performed better than the rest of the modelling approaches in terms of absolute biases, however, we cannot be certain if the coefficients estimates are reliable estimates of the true estimates. Here, we only used this real data complete case scenario simply to illustrate the different modelling strategies.

In general, the FCS approach is a more flexible way to impute missing data, by imputing the data on a variable-by-variable basis by specifying an imputation model per variable. These sets of conditional distributions do not necessarily correspond to a joint model, and hence doesn't guarantee of generating imputations from a proper posterior predictive distribution. However, the proposed joint model (FRM) models the joint distribution of all the variables in the data through a sequence of generalised linear models, and allows us to draw imputations from a proper posterior distribution. It is difficult to determine whether imputations drawn from a proper posterior distribution are better than those obtained using MICE through a simulation study and a real data set. However, multiple imputation using either joint modelling or fully conditional specification normally provide a more reliable approach to handle the missing data than the complete case analysis or using the common ad-hoc methods. This is because these methods might lead to biased estimates of parameters and they don't handle the uncertainty due to the presence of missing values.

Both our joint modelling approach, FRM and the fully conditional specifications MICE assume that the missing data mechanism in the data sets is ignorable (missing at random). So what happens if the missing data mechanism is non-ignorable? The most common way to deal with this problem is to include more variables in the imputation model, thus bringing the missing data mechanism closer to missing at random. Other methods such as pattern-mixture models (Little (1993a)) or models selection approach where data imputers implicitly model the missing data mechanism can be considered. Further investigations on this area can be done in the future.

Also, multiple imputation is not the only way to handle missing data. Other approaches such as likelihood-based methods (Carpenter and Kenward (2008)) and probability weighting (Hogan and Lancaster (2004)) can be considered as alternatives in the future. We can use these methods to impute missing values and compare the results to our model FRM. It would also be interesting to explore extending this approach to deal with more complicated data structures. For example, we could consider incorporating variables that are inherently nested through the development of multi-level models. We can also consider broadening the class of models used, for example using logistic regression models for binary responses, potentially using computational techniques proposed in Kinney and Dunson (2007).

7.2 Statistical disclosure control

For statistical disclosure control, we have applied our model FRM to protect confidentiality of the current population survey (CPS) data by generating multiply imputed, partially synthetic data sets. These data sets comprised a mix of original data and the synthetic data where values chosen for synthesis are based on an approach that considers unique and sensitive units in the survey. We also extended our method to accommodate the presence of spikes at zero in some of the continuous variable of the CPS data. Our results showed that when we increased the percentage of data points being synthesized, the risks of being identified have been reduced, as well as the utility of the partially synthetic data sets.

The main advantage of using partially synthetic data approach is that the synthetic data are drawn from a plausible distribution that describes the given observed data. In our modelling strategy FRM, the synthetic values are drawn from the posterior predictive distribution that can reflect the multivariate relationship between the variables in the data set. Hence, valid inferences can be made by data users using the simple combining rules described in Chapter 2. Also, in partially synthetic data approach, only a fraction of data will be replaced by synthetic data compare to the fully synthetic approach. Hence the inferences made by the users are less sensitive to model specification.

One of the key challenges in using partially synthetic data approach for protecting confidentiality is selecting the values that will be replaced by the synthetic data. The statistical agencies need to select the units that are at high risk of disclosure such as sample unique in a data set, then replace the units' attributes with synthetic data. In Chapter 6, we have discussed a selection procedure where instead of synthesizing the whole unit, we only replaced the data points that make the unit unique. By doing so, we can minimize the number of data points needed to be synthesized for protecting confidentiality, at the same time preserving the utility of the imputed data sets. From Section 6.1, we noticed that if two or more variable sets can be found that can satisfy the criteria $f_{g'} > C_s$, we will take the union of the sets. An alternative to that is instead of taking the union of the variable sets, we take the minimum required variable set that can satisfy the criteria $f_{g'} > C_s$.

Another challenge in using the partially synthetic approach is that this is a model based imputation method. Hence the validity of the model specification has great impact on the validity of the synthetic data. The main reason is that the synthetic data only represents the relationships among the variables that have been included in the imputation model. Any relationships among the variables that are not accurately included in the imputation model will result in the data users' analysis missing those relationships.

When the data set consists of large numbers of categorical and continuous variables, it is quite difficult to specify the imputation model that can accurately capture the relationship between the variables in the data set. Hence, we can instead consider a non-parametric model such as the classification and regression tree (CART). CART is an easy to use non-parametric model that can generate partial synthetic data sets. Details on how to use CART to generate partially synthetic data can be found in Reiter (2005a). It would be interesting to compare the risks and utilities obtained using our parametric modelling method to the classification and regression tree (CART) synthesis method.

We can also explore the risk-utility trade off for our method, in order to address the decision of choosing the optimum amount of data that need to be synthesized. In order to chose an optimum amount of data for synthesis, we can use the risk-utility framework proposed by Duncan et al. (2002) which is based on quantified measures of disclosure risk and data utility. When releasing data sets, statistical agencies have to preserve the utility of the data sets as much as possible so that the data sets provided are useful to the data users, while keeping the risk of disclosure low. The partially synthetic data approach can have low disclosure risks, because the sensitive values in the data sets have been replaced by synthetic data. However, this also decreases the utility of the imputed data. Statistical agencies must decide how much synthetic data they are willing to release, at the same time striking an acceptable balance between disclosure risk and data utility. Perhaps, diagnostic plots such as Figure 6.2 and Figure 6.3 can help to make this decision, and it would be interesting to explore different criteria for how much data that we choose to synthesize.

Large survey data normally contain missing data and this will complicate the data analysis. Statistical agencies who wish to release the data for public use need to handle the missing data first before applying any statistical disclosure limitation methods on the data set or releasing synthetic data. Our modelling strategy can handle missing data and disclosure control simultaneously via a two stage imputation procedure as described in Chapter 2. First, we use multiple imputation to impute the missing values in the data set, generating m multiply imputed data sets, we then replace each value at high risk of disclosure in each imputed data set with d synthetic values using multiple imputation; thus creating $m \times d$ multiply imputed data sets. For details on how to implement this strategy and obtain valid inference, please see Reiter (2004).

One of the ethical arguments being raised when applying statistical disclosure limitations (SDL) methods or using synthetic data approach is the false disclosure risk. Suppose that intruders have a target information and wish to disclose sensitive information about certain individual in some population. If the survey data was released without any taking any confidentiality measures, the intruders can easily disclose the sensitive information about that individual. When the statistical agencies replace the sensitive values with synthetic data, the intruders might not be able to disclose the information using the target information they possess. However, after using SDL methods or synthetic data approach, there is a chance that other individuals in that survey data will have the same attributes as the target information. When the intruders attack the data sets, they might have targeted the wrong group of individuals, hence resulting in a false disclosure. Such situations will bring harm to the individuals that are falsely identified and the reputation of the statistical agencies will also be affected. These false disclosure risks can have huge impact on the health care system. In recent years, health care organizations have started to use new information technologies to reduce costs as well as improve the quality and efficiency of the health care system. With new technologies available, health care organizations can collect detailed data about the individual patient health records, which can be used to help doctors to threat their patients accordingly. Hence, health care data has became more and more valuable for scientists and big pharmaceutical companies for data mining. This prompts questions about the confidentiality measures taken by the health organizations to protect the collected health care data. If the health organizations failed to protect the confidentiality of the patients' health care information, the doctor-patient relationship will be greatly affected, and the patients will likely to withhold some sensitive information

or less likely to participate in any health care survey in the future. This might put the health organizations at risk of having less information for the health care data, which can lead to reduced quality of health care and deprive researchers of valuable data for future research. A comprehensive guideline on confidentiality in the UK health and social care can be found on http://www.hscic.gov.uk.

It is important to consider how the contributions of this thesis fit the wider contexts of statistical disclosure control. Here are some thoughts on how the work can be taken forward for relevant future contributions. One of the potential contributions is to use partially synthetic data for confidentiality protection in emerging technologies. With the development of new popular social networking sites, such as *Facebook* and *Twitter*, social network data that describe different entities and connect them can be easily acquired. Data users can then perform any social network analysis to study the data patterns that describe the connections between entities in the network. For example, internet companies such as *Google*, *Netflix* and *Amazon.com* have been making their business more efficient by analyzing such social network data, in order to study the consumers' behavior. Every time a *Facebook* user updated their info on *Facebook*, such as favourite books or movies, the user will normally see related suggestions pop up on their screens afterwards. From the business perspective, this feature is proven to be useful as this can encourage the user to spend more on buying the companies' products, as well as helping online advertisers such as *Google* to sell their adverts accurately to the targeted groups. However, confidentiality protection has become a big issue for the data users. With the accumulation of such network data as well as population survey and medical data, data intruders can use record-linkage methodologies to identify the individuals sensitive information from these data sets. Here, we can potentially use our modelling strategies to generate partially synthetic data, where our goal is to protect the confidentiality of individuals, and at the same time enable the useful analysis of social network data. Since the synthetic data are generated from a model that describes the observed data, any data patterns that are present in social network data can also be preserved. We can use the synthetic data approach to provide a reasonable approximation to the data patterns so that data users can perform any analysis, and the confidentiality of individuals' records can also be protected. The concerns on confidentiality issues with social networking data and methods to handle them have also been mentioned by Liu et al. (2009), Zhou and Pei (2008) and Hay et al. (2007).

Multiple imputation framework can be used to handle missing data and disclosure control simultaneously via a two stage imputation procedure as mentioned before. A further development in this framework is that imputations from missing data alone can potentially protect confidentiality of data sets as well. For example, survey data sets often contain missing values and it will complicate statistical analysis. Statistical agencies can then multiply impute the missing values and release the imputed data sets for public use. These imputed data sets consist of original data and imputed data, which somewhat similar to the partially synthetic data. In the partially synthetic data approach, we first select the values at high disclosure risk and replace those values with synthetic data; here, we only impute the missing values and then release the imputed data sets. The idea is that multiple imputation for missing data itself might introduce uncertainty into the imputed data, thus making it harder for intruders to make identifications or disclose any sensitive information; hence potentially protecting confidentiality of the released data. However, sample unique in the original data that do not contain any missing values will remain unique in the released data, as we only replace the missing values. Similar approach

has been mentioned by Jagannathan and Wright (2008) and it would be interesting to investigate this area further in the future.

Appendices

In the appendices we provide details that complement the material presented in the main text. Appendix A presents details of how the simulation studies in Chapter 4 were implemented. Appendix B presents some more details of the breast-feeding data analysis in Chapter 4. Appendix C presents the details of "P-steps" for RFRM in Chapter 5. Appendix D presents details of how the simulation study from Section 5.1.1 in Chapter 5 were implemented. Appendix E presents the details of "P-steps" for LFRM in Chapter 5. Appendix F presents details of how the simulation study from Section 5.2.1 in Chapter 5 were implemented. Appendix G presents the details of "P-steps" for MFRM in Chapter 5. Appendix H presents details of how the simulation study from Section 5.3.1 in Chapter 5 were implemented. Appendix I shows the density plots of social security, tax and child support payment from the CPS data in Chapter 6. Appendix J presents the estimands considered for the simulation study analysis in Chapter 6.

Appendix A - Simulation study in Chapter 4

In this section we present details of the simulation study in Chapter 4. We first describe how we simulated an incomplete data set. We then describe how we impute the missing data under the two scenarios representing the state of the imputer's knowledge, and the analyses we considered.

Data generation

We simulate $x_{i,1}, x_{i,2}, \ldots, x_{i,9},\ i = 1, \ldots, 1000$ in the following way:

$$
\begin{aligned}
x_{i,1} &\quad \text{simulated from a discrete distribution taking values} \in \{1,2,3\} \\
&\quad \text{with probabilities 0.3, 0.4 and 0.3 respectively.} \\
x_{i,2} &\sim p(x_{i,2}|x_{i,1}, \boldsymbol{\theta}^{(2)}), \\
x_{i,3} &\sim \mathrm{N}(5,1), \\
x^*_{i,4} &\sim \mathrm{N}(1+2x_{i,3},1), \\
x_{i,4} &= I(x^*_{i,4} > 11) \\
x^*_{i,5} &\sim \mathrm{N}(2+5*I(x_{i,1}=2)+I(x_{i,1}=3)+x_{i,3},2), \\
x_{i,5} &= j^5\ I(c^{(5)}_{j-1} < x^*_{i,5} < c^{(5)}_j), \quad j^5 = 1,\ldots,4 \\
x_{i,6} &\sim \mathrm{N}(8,3), \\
x_{i,7} &\sim \mathrm{N}(2+2x_{i,6}-x_{i,3},5), \\
x^*_{i,8} &\sim \mathrm{N}(3+2x_{i,7},5), \\
x_{i,8} &= j^8\ I(c^{(8)}_{j-1} < x^*_{i,8} < c^{(8)}_j), \quad j^8 = 1,\ldots,8 \\
x_{i,9} &\sim \mathrm{N}(1+4*I(x_{i,2}=2)-3*I(x_{i,2}=3)+5*I(x_{i,2}=4)+2*I(x_{i,2}=5)+5x_{i,7},3),
\end{aligned}
$$

where

$$
p(x_{i,2}=j^2|x_{i,1},\boldsymbol{\theta}^{(2)}) = \frac{\exp(\theta^{(2)}_{0,j^2}+\theta^{(2)}_{1,j^2}I(x_{i,1}=2)+\theta^{(2)}_{2,j^2}I(x_{i,1}=3))}{\sum_{s=1}^{5}\exp(\theta^{(2)}_{0,s}+\theta^{(2)}_{1,s}I(x_{i,1}=2)+\theta^{(2)}_{2,s}I(x_{i,1}=3))}
$$

where $j^2 \in \{1,\ldots,5\}$ and $\theta^{(2)}_{0,1} = \theta^{(2)}_{1,1} = \theta^{(2)}_{2,1} = 0$ for identifiability, $\theta^{(2)}_{0,2} = 1, \theta^{(2)}_{1,2} = -1, \theta^{(2)}_{2,2} = -2, \theta^{(2)}_{0,3} = -1, \theta^{(2)}_{1,3} = 3, \theta^{(2)}_{2,3} = 1, \theta^{(2)}_{0,4} = -1, \theta^{(2)}_{1,4} = 2, \theta^{(2)}_{2,4} = 2, \theta^{(2)}_{0,5} = -1, \theta^{(2)}_{1,5} = 2, \theta^{(2)}_{2,5} = 1$, and $I(\cdot)$ is the indicator function. The threshold parameters $c^{(5)}_0 = -\infty, c^{(5)}_1 = 6.962, c^{(5)}_2 = 9.292, c^{(5)}_3 = 11.59, c^{(5)}_4 = \infty$ and $c^{(8)}_0 = -\infty, c^{(8)}_1 = -2, c^{(8)}_2 = 4, c^{(8)}_3 = 9, c^{(8)}_4 = 13, c^{(8)}_5 = 17, c^{(8)}_6 = 22, c^{(8)}_7 = 28, c^{(8)}_8 = \infty$.

We introduce missing values into variables $\mathbf{x}_2, \ldots, \mathbf{x}_9$ in the following way:

$$
\begin{aligned}
p(m_{i,2}=1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,3}=1) &= \left\{ \frac{\exp(1-5*I(x_{i,2}=2)-4*I(x_{i,2}=3))}{1+\exp(1-5*I(x_{i,2}=2)-4*I(x_{i,2}=3))} \right\} (1-m_{i,2}) \text{ (MAR)}, \\
p(m_{i,4}=1) &= \left\{ \frac{\exp(-5.5+5x_{i,3})}{1+\exp(-5.5+5x_{i,3})} \right\} (1-m_{i,3}) \text{ (MAR)}, \\
p(m_{i,5}=1) &= \left\{ \frac{\exp(-5.5+5x_{i,3})}{1+\exp(-5.5+5x_{i,3})} \right\} (1-m_{i,3}) \text{ (MAR)}, \\
p(m_{i,6}=1) &= 0.3 \text{ (MCAR)} \\
p(m_{i,7}=1) &= \left\{ \frac{\exp(60-8x_{i,6})}{1+\exp(60-8x_{i,6})} \right\} (1-m_{i,6}) \text{ (MAR)}, \\
p(m_{i,8}=1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,9}=1) &= \left\{ \frac{\exp(6-x_{i,7})}{1+\exp(6-x_{i,7})} \right\} (1-m_{i,7}) \text{ (MAR)},
\end{aligned}
$$

where $m_{i,j}$ be the missing data indicator for $x_{i,j}$, where $m_{i,j}=1$ indicates $x_{i,j}$ is missing and $m_{i,j}=0$ indicates $x_{i,j}$ is observed.

Imputation and analysis

In scenario 1 we impute the missing values using the same order as the data generation process, we then obtain the following estimates from an analysis of the imputed datasets:

- Estimates of the means of $\mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_9$.
- Estimates of the proportions of units with $\mathbf{x}_2 = 1, \ldots, 5,\ \mathbf{x}_4 = 1,\ \mathbf{x}_5 = 1, \ldots, 4,\ \mathbf{x}_8 = 1, 2, \ldots, 8$
- Estimates of the regression coefficients of a linear regression from the following regression models: $p(\mathbf{x}_2|\mathbf{x}_1), p(\mathbf{x}_7|\mathbf{x}_6, \mathbf{x}_3), p(\mathbf{x}_9|\mathbf{x}_2, \mathbf{x}_7)$.

In scenario 2 we impute the missing values using a different ordering to the predictors: $\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_8, \mathbf{x}_6, \mathbf{x}_3, \mathbf{x}_9, \mathbf{x}_7$. We obtain estimates from the same analyses as above.

Appendix B - Breast-feeding data analysis

In this section we present the details of the breast-feeding data analysis mentioned in Chapter 4. We first describe the ordering of the variables we used in the imputation. As noted in Chapter 3, any variable that is fully observed will be conditioned on in every regression model and so can be placed at the beginning in the order. We then describe what transformations were applied to the variables in the study.

Description of variables

The ordering of the variables were as follows:

- $\mathbf{x}_1$ - The number of years between 1979 and when the mother gave birth. (Continuous, fully observed.)
- $\mathbf{x}_2$ - The child's race. (Nominal: 1-Hispanic, 2-Black, 3-Others, fully observed.)
- $\mathbf{x}_3$ - Whether the spouse or partner was present at birth. (Nominal: 1-Partner present, 2-Spouse present, 3-Absent, 3.54% missing values.)
- $\mathbf{x}_4$ - Family income. (Continuous, 25.69% missing values.)
- $\mathbf{x}_5$ - Breast-feeding duration. (Treatment variable, binary: 0-Control, 1-Treated, 6.24% missing values.)
- $\mathbf{x}_6$ - Child's sex. (Binary: 0-Male, 1-Female, 0.1% missing values.)
- $\mathbf{x}_7$ - Whether grandparents were present at birth. (Binary: 0-Absent, 1-Present, 3.44% missing values.)
- $\mathbf{x}_8$ - Mother's intelligence as measured by an armed forces qualification test. (Continuous, 5.16% missing values.)
- $\mathbf{x}_9$ - Mother's highest educational attainment. (Continuous, 3.70% missing values.)
- $\mathbf{x}_{10}$ - Child's birth weight. (Continuous, 3.64% missing values.)
- $\mathbf{x}_{11}$ - Number of days that the child spent in hospital. (Continuous, 10.36% missing values.)
- $\mathbf{x}_{12}$ - Number of days that the mother spent in hospital. (Continuous, 10.66% missing values.)
- $\mathbf{x}_{13}$ - Number of weeks that the mother worked in the year prior to giving birth. (Ordinal: 1-Not worked, 2-Worked for 1 to 47 weeks, 3-Worked for 48 to 51 weeks, 4-Worked for 52 weeks, 33.05% missing values.)
- $\mathbf{x}_{14}$ - Number of weeks the child was born premature. (Ordinal: 1-not preterm (zero weeks), 2- moderately preterm (one to four weeks), and very preterm (five or more weeks), 7.78% missing values.)
- $\mathbf{x}_{15}$ - Peabody individual assessment test math score (PI-ATM) administered to children at 5 or 6 years of age. (Continuous, 48.2% missing values.)

Transformations applied to the variables

We categorize the number of weeks the child was born premature into three levels: 1-not preterm (zero weeks), 2-moderately preterm (one to four weeks), and 3-very preterm (five or more weeks), with threshold values determined from guidelines of the March of Dimes (www.marchofdimes.com). The reason we categorize the variable is because it has a very large spike at zero weeks, as shown in Figure 7.1.

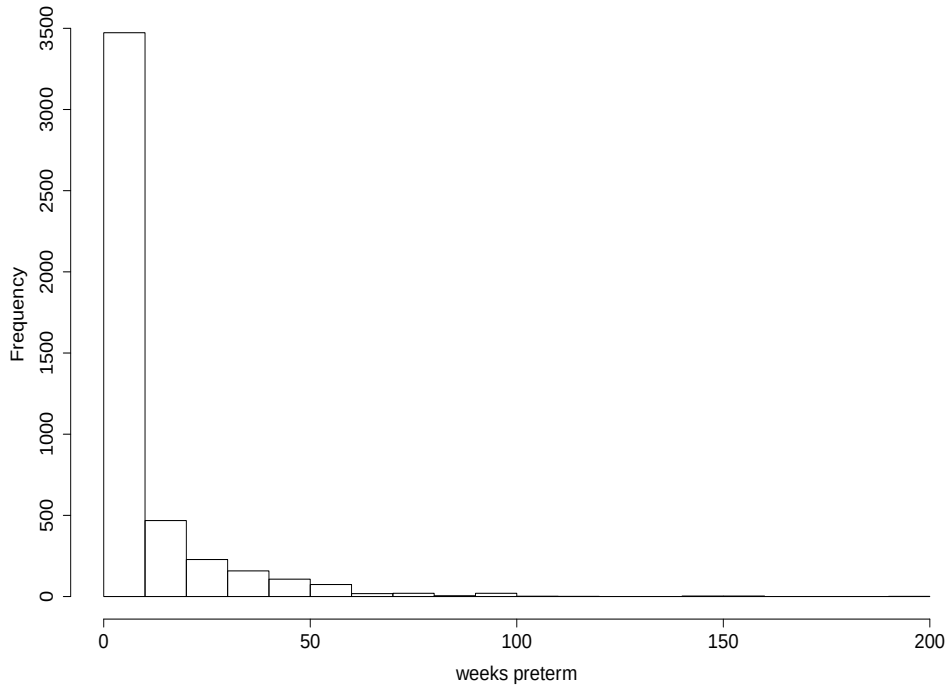


Figure 7.1: Histogram of weeks mother worked in the year before giving birth for subjects in the breast-feeding data.

From Figure 7.2 we notice that the the number of weeks that the mother worked in the year prior to giving birth has a distinct U shaped histogram, which would be difficult to capture with a normal model. Hence, we categorize the variable into four levels: 1-not worked at all, 2-worked between 1 and 47 weeks, 3-worked 48-51 weeks, and 4-worked all 52 weeks.

When implementing FRM, to ensure that the residuals of the normal linear regression models (where the response in the regression model is continuous) satisfy the normal assumption, we transform the response variable in these regression models, where the transformation is given by the Box & Cox procedure (Box and Cox (1964)). We perform the following transformations: we take the natural log of family's income ($\mathbf{x}_4$), number of days that the child spent in hospital ($\mathbf{x}_{11}$) and number of days that the mother spent in hospital ($\mathbf{x}_{12}$), we square child's birth weight ($\mathbf{x}_{10}$) and Peabody individual assessment test math score (PI-ATM) ($\mathbf{x}_{15}$), we take the square root of mother's intelligence as measured by an armed forces qualification test ($\mathbf{x}_8$).

The following diagnostic plots present normal probability plots, before and after the trans-

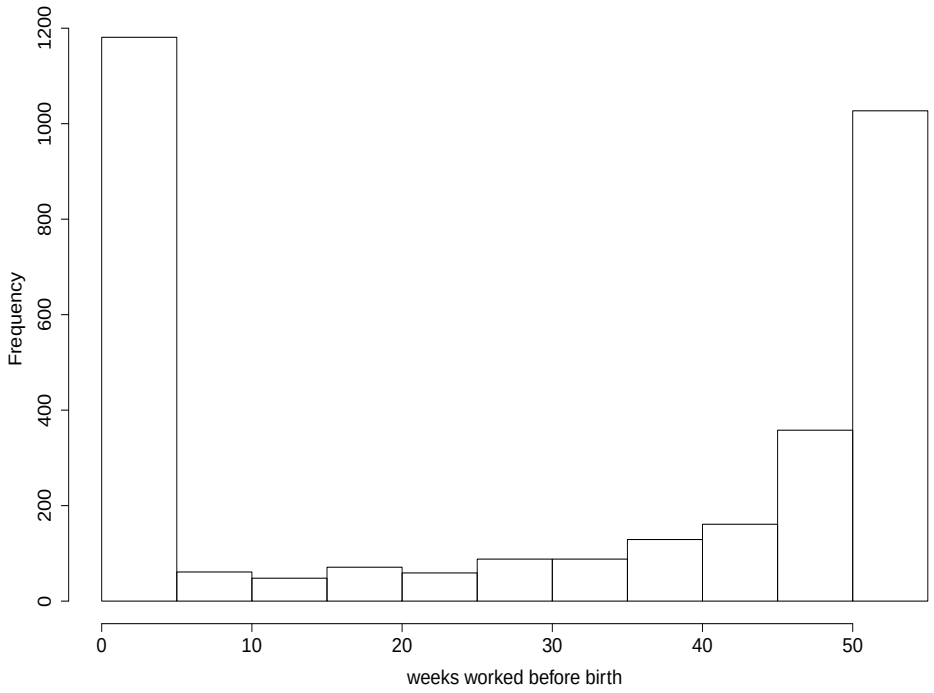


Figure 7.2: Histogram of weeks preterm for subjects in the breast feeding study.

formation, of the residuals in regression models where a transformation of the response was deemed necessary.

When implementing MICE, as each regression model to impute missing values conditions on all other variables in the data we considered potentially different Box-Cox transformations to improve normality assumptions, and hence improve the model fit. All but one variable was transformed in the same way described above. Child's birth weight ($\mathbf{x}_{10}$), which was squared when applying FRM, no longer required a transformation when applying MICE.

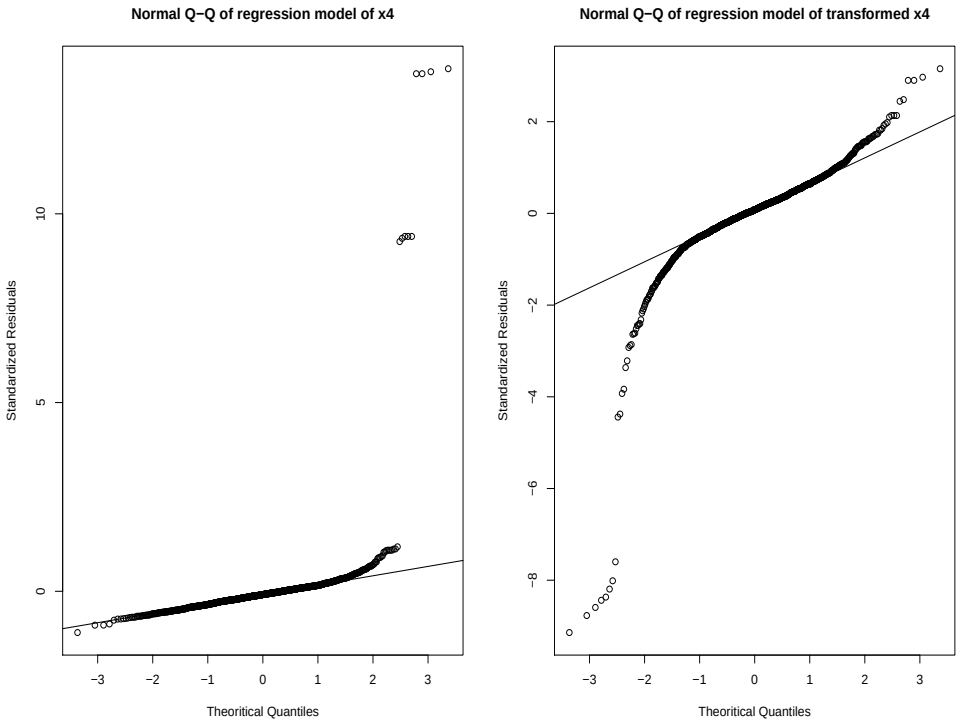


Figure 7.3: Q-Q plot of the residuals in the linear regression model of family income, $\mathbf{x}_4$ before(left) and after(right) transformation (natural log).

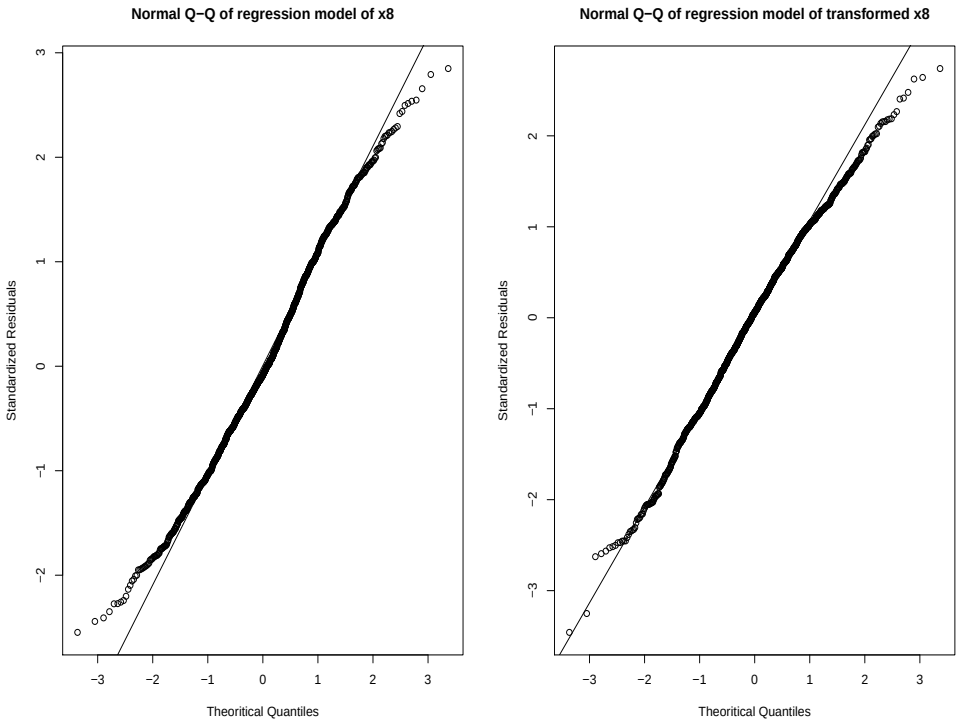


Figure 7.4: Q-Q plot of the residuals in the linear regression model of mother's intelligence as measured by an armed forces qualification test, $\mathbf{x}_8$ before(left) and after(right) transformation (square root).

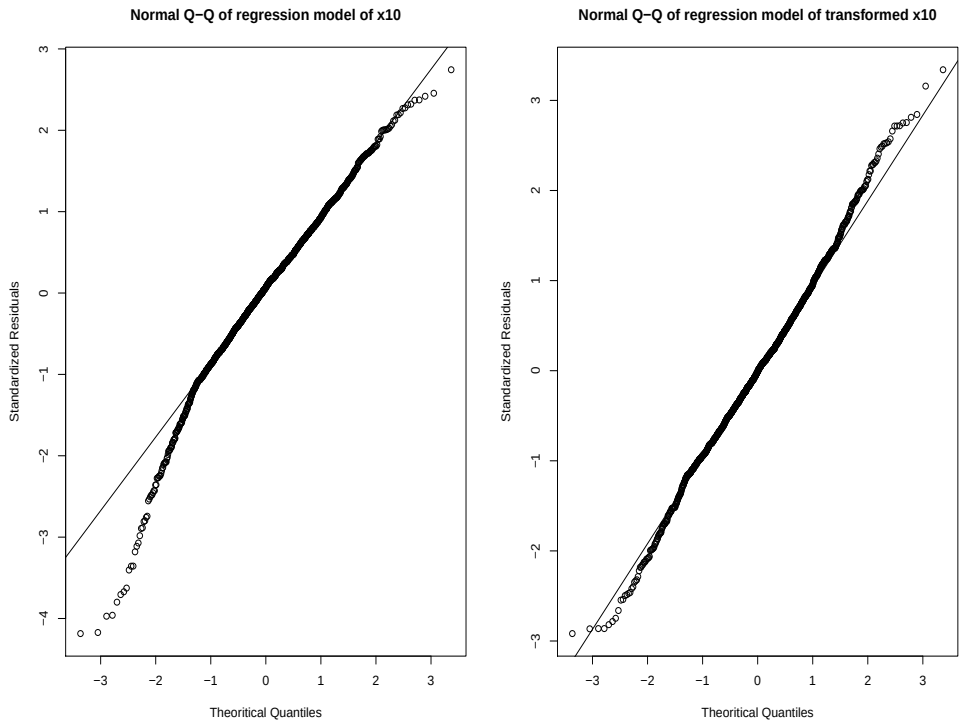


Figure 7.5: Q-Q plot of the residuals in the linear regression model of child's birth weight, $\mathbf{x}_{10}$ before(left) and after(right) transformation (square).

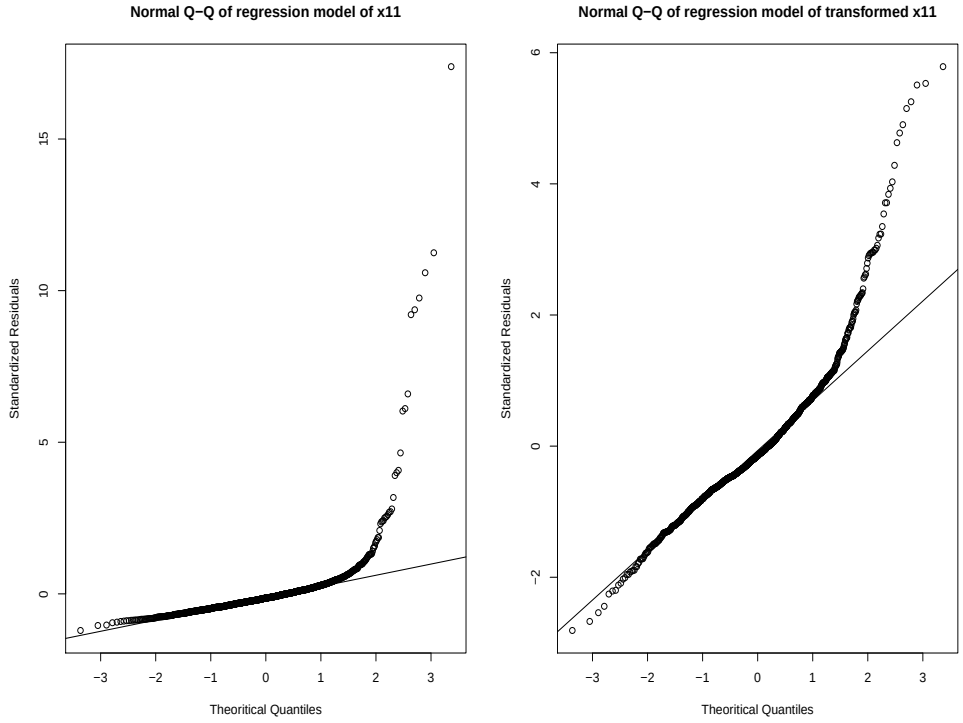


Figure 7.6: Q-Q plot of the residuals in the linear regression model of number of days that the child spent in hospital, $\mathbf{x}_{11}$ before(left) and after(right) transformation (natural log).

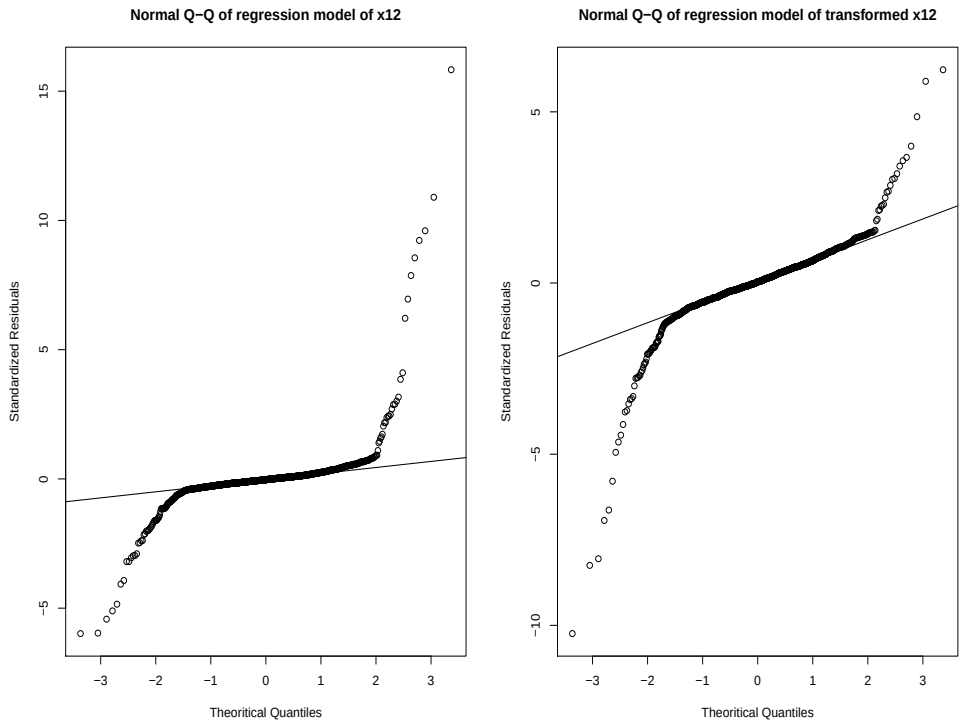


Figure 7.7: Q-Q plot of the residuals in the linear regression model of number of days that the mother spent in hospital, $\mathbf{x}_{12}$ before(left) and after(right) transformation (natural log).

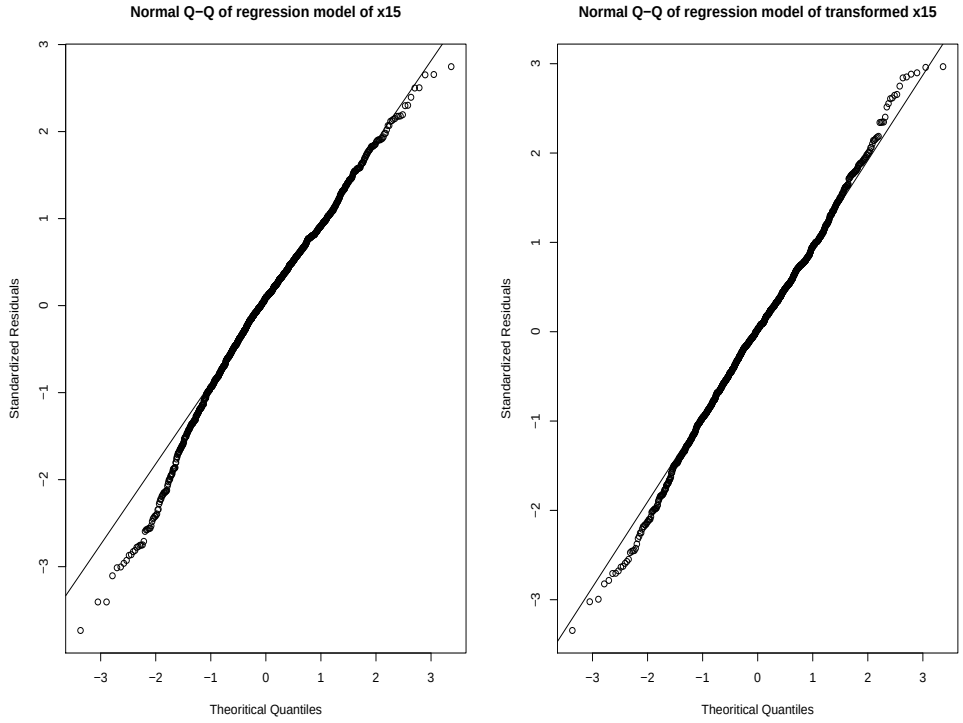


Figure 7.8: Q-Q plot of the residuals in the linear regression model of Peabody individual assessment test math score (PI-ATM), $\mathbf{x}_{15}$ before(left) and after(right) transformation (square).

Appendix C - "P-steps" for RFRM

In this section, we present the details on deriving the full conditional distributions of the parameters in the "P-steps" mentioned in the RFRM strategy in Chapter 5. The full conditional distribution of $\boldsymbol{\theta}^{(1)}$ where $\mathbf{x}_1$ follows a multinomial distribution and Metropolis-Hastings update for the multinomial logistic regression model have been presented in the RFRM section in Chapter 5. Here, we will present the details on deriving the full conditional distributions of the remaining parameters in the "P-steps", such as $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \Phi_q)$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is continuous, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q)$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is binary) and $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \boldsymbol{\gamma}^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is ordinal).

For $x_{i,q}^*$ is continuous, we use the following representation for each of the sequential regression model $p(x_{i,q}^*|x_{i,1}^*, \ldots, x_{i,q-1}^*, \boldsymbol{\theta}^{(q)})$:

$$\begin{aligned} x_{i,q}^* &= \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)} + \epsilon_{i,q}, \\ \epsilon_{i,q} &\sim \mathrm{N}(0, \frac{1}{\tau_q \phi_{i,q}}), \end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(1, I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$, for $i = 1, \ldots, n$ and $\boldsymbol{\beta}^{(q)} = (\beta_0^{(q)}, \beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$. We then propose the following priors:

$$\begin{aligned} p(\beta_j^{(q)}|\tau_q, \psi_{j,q}) &\sim \mathrm{N}(0, \frac{1}{\tau_q \psi_{j,q}}),\ j = 0, \ldots, q-1, \\ p(\tau_q) &\propto \frac{1}{\tau_q}, \\ p(\psi_{j,q}) &\sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}),\ j = 0, \ldots, q-1, \\ p(\phi_{i,q}) &\sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}),\ i = 1, \ldots, n. \end{aligned}$$

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$\begin{aligned} & \prod_{i=1}^{n} \left\{ p(x_{i,q}^*|\boldsymbol{\beta}^{(q)}, \tau_q, \phi_{i,q}) p(\phi_{i,q}) \right\} \prod_{j=0}^{q-1} \left\{ p(\beta_j^{(q)}|\tau_q, \psi_{j,q}) p(\psi_{j,q}) \right\} p(\tau_q) \\ = & \prod_{i=1}^{n} \left\{ \sqrt{\frac{\tau_q \phi_{i,q}}{2\pi}} \exp\left[-\frac{\tau_q \phi_{i,q}}{2} (x_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2 \right] \frac{1}{\Gamma(\frac{\nu}{2})} (\frac{\nu}{2})^{(\frac{\nu}{2})} \phi_{i,q}^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu \phi_{i,q}}{2} \right] \right\} \\ & \prod_{j=0}^{q-1} \left\{ \sqrt{\frac{\tau_q \psi_{j,q}}{2\pi}} \exp\left[-\frac{\tau_q \psi_{j,q}}{2} (\beta_j^{(q)})^2 \right] \frac{1}{\Gamma(\frac{\nu}{2})} (\frac{\nu}{2})^{(\frac{\nu}{2})} \psi_{j,q}^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu \psi_{j,q}}{2} \right] \right\} \frac{1}{\tau_q}. \end{aligned}$$

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q, \Phi_q$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n} \sqrt{\frac{\tau_q \phi_{i,q}}{2\pi}} \exp\left[-\frac{\tau_q \phi_{i,q}}{2}(x^*_{i,q} - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right] \prod_{j=0}^{q-1} \sqrt{\frac{\tau_q \psi_{j,q}}{2\pi}} \exp\left[-\frac{\tau_q \psi_{j,q}}{2}(\beta_j^{(q)})^2\right] \\
\propto\ & \exp\left\{-\frac{\tau_q}{2}\left[(\mathbf{x}^*_q - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\phi(\mathbf{x}^*_q - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)}) + (\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\psi\boldsymbol{\beta}^{(q)}\right]\right\} \\
\propto\ & \exp\left\{-\frac{\tau_q}{2}\left[(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)} - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\phi(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)} - \widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)}) + (\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\psi\boldsymbol{\beta}^{(q)}\right]\right\} \\
=\ & \exp\left\{-\frac{\tau_q}{2}\left[(\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^{(q)})'\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q(\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^{(q)}) + (\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\psi\boldsymbol{\beta}^{(q)}\right]\right\} \\
\propto\ & \exp\left\{-\frac{\tau_q}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q + \mathbf{D}^{(q)}_\psi)\boldsymbol{\beta}^{(q)} - 2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\},
\end{aligned}$$

Let $A_q = (\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q + \mathbf{D}^{(q)}_\psi)$ and $\hat{\boldsymbol{\beta}}^{(q)} = (\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}'_q\mathbf{x}^*_q$, then

$$\begin{aligned}
&\exp\left\{-\frac{\tau_q}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q + \mathbf{D}^{(q)}_\psi)\boldsymbol{\beta}^{(q)} - 2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\} \\
=\ & \exp\left\{-\frac{\tau_q}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)} - 2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q(\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}'_q\mathbf{x}^*_q\right]\right\} \\
=\ & \exp\left\{-\frac{\tau_q}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)} - 2(\boldsymbol{\beta}^{(q)})'A_qA_q^{-1}\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\mathbf{x}^*_q\right]\right\} \\
\propto\ & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)} - A_q^{-1}\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\mathbf{x}^*_q)'\tau_qA_q(\boldsymbol{\beta}^{(q)} - A_q^{-1}\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\mathbf{x}^*_q)\right]\right\} \\
\Rightarrow\ & \boldsymbol{\beta}^{(q)} \sim \text{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}) \text{ with} \\
& \hat{\boldsymbol{\beta}}^{(q)} = \tau_q\mathbf{V}^{(q)}\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\mathbf{x}^*_q \text{ and } \mathbf{V}^{(q)} = \frac{1}{\tau_q}(\widetilde{\mathbf{X}}'_q\mathbf{D}^{(q)}_\phi\widetilde{\mathbf{X}}_q + \mathbf{D}^{(q)}_\psi)^{-1},
\end{aligned}$$

where $\mathbf{x}^*_q = (x^*_{1,q}, \ldots, x^*_{n,q})'$, and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where
$\widetilde{\mathbf{x}}_{i,q} = (1, I(x^*_{i,1} = 2), \ldots, I(x^*_{i,1} = L_1), \ldots, I(x^*_{i,k} = 2), \ldots, I(x^*_{i,k} = L_k), x^*_{i,k+1}, \ldots, x^*_{i,q-1})$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}^{(q)}_\phi$ is a $n \times n$ diagonal matrix with entries $\phi_{i,q}$ for $i = 1, \ldots, n$ and $\mathbf{D}^{(q)}_\psi$ is a $q \times q$ diagonal matrix with entries $\psi_{j,q}$ for $j = 0, \ldots, q-1$.

2. The full conditional distribution of τ_q is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{\sqrt{\frac{\tau_q\phi_{i,q}}{2\pi}}\exp\left[-\frac{\tau_q\phi_{i,q}}{2}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\}\\
&\prod_{j=0}^{q-1}\left\{\sqrt{\frac{\tau_q\psi_{j,q}}{2\pi}}\exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right]\right\}\frac{1}{\tau_q}\\
\propto\ & \tau_q^{\frac{n}{2}}\tau_q^{\frac{q}{2}}\tau_q^{-1}\prod_{i=1}^{n}\exp\left[-\frac{\tau_q\phi_{i,q}}{2}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\prod_{j=0}^{q-1}\exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right]\\
=\ & \tau_q^{\frac{n+q}{2}-1}\exp\left[-\frac{\tau_q}{2}\sum_{i=1}^{n}\phi_{i,q}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\exp\left[-\frac{\tau_q}{2}\sum_{j=0}^{q-1}\psi_{j,q}(\beta_j^{(q)})^2\right]\\
=\ & \tau_q^{\frac{n+q}{2}-1}\exp\left\{-\frac{\tau_q}{2}\left[(\mathbf{x}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'\mathbf{D}_\phi^{(q)}(\mathbf{x}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\psi^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
\Rightarrow\ & \tau_q\sim\text{Gamma}\left(\frac{n+q}{2},\frac{(\mathbf{x}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'\mathbf{D}_\phi^{(q)}(\mathbf{x}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\psi^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right),
\end{aligned}$$

where $\mathbf{x}^*_q=(x^*_{1,q},\ldots,x^*_{n,q})'$, and $\widetilde{\mathbf{X}}_q=(\widetilde{\mathbf{x}}_{1,q},\ldots,\widetilde{\mathbf{x}}_{n,q})'$, where
$\widetilde{\mathbf{x}}_{i,q}=(1,I(x^*_{i,1}=2),\ldots,I(x^*_{i,1}=L_1),\ldots,I(x^*_{i,k}=2),\ldots,I(x^*_{i,k}=L_k),x^*_{i,k+1},\ldots,x^*_{i,q-1})$ for $i=1,\ldots,n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n\times n$ diagonal matrix with entries $\phi_{i,q}$ for $i=1,\ldots,n$ and $\mathbf{D}_\psi^{(q)}$ is a $q\times q$ diagonal matrix with entries $\psi_{j,q}$ for $j=0,\ldots,q-1$.

3. The full conditional distribution of $\psi_{j,q}$ is proportional to

$$\begin{aligned}
&\prod_{j=0}^{q-1}\left\{\sqrt{\frac{\tau_q\psi_{j,q}}{2\pi}}\exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right]\frac{1}{\Gamma(\frac{\nu}{2})}(\frac{\nu}{2})^{(\frac{\nu}{2})}\psi_{j,q}^{\frac{\nu}{2}-1}\exp\left[-\frac{\nu\psi_{j,q}}{2}\right]\right\}\\
\propto\ & \prod_{j=0}^{q-1}\psi_{j,q}^{\frac{1}{2}}\psi_{j,q}^{\frac{\nu}{2}-1}\exp\left[-\psi_{j,q}\left(\frac{\tau_q}{2}(\beta_j^{(q)})^2+\frac{\nu}{2}\right)\right]\\
\Rightarrow\ & \psi_{j,q}\sim\text{Gamma}\left(\frac{\nu+1}{2},\frac{\tau_q}{2}(\beta_j^{(q)})^2+\frac{\nu}{2}\right),\ j=0,\ldots,q-1.
\end{aligned}$$

4. The full conditional distribution of $\phi_{i,q}$ is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{\sqrt{\frac{\tau_q\phi_{i,q}}{2\pi}}\exp\left[-\frac{\tau_q\phi_{i,q}}{2}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\frac{1}{\Gamma(\frac{\nu}{2})}(\frac{\nu}{2})^{(\frac{\nu}{2})}\phi_{i,q}^{\frac{\nu}{2}-1}\exp\left[-\frac{\nu\phi_{i,q}}{2}\right]\right\}\\
\propto\ & \prod_{i=1}^{n}\phi_{i,q}^{\frac{1}{2}}\phi_{i,q}^{\frac{\nu}{2}-1}\exp\left[-\phi_{i,q}\left(\frac{\tau_q}{2}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2+\frac{\nu}{2}\right)\right]\\
\Rightarrow\ & \phi_{i,q}\sim\text{Gamma}\left(\frac{\nu+1}{2},\frac{\tau_q}{2}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2+\frac{\nu}{2}\right),\ i=1,\ldots,n,
\end{aligned}$$

where
$\widetilde{\mathbf{x}}_{i,q}=(1,I(x^*_{i,1}=2),\ldots,I(x^*_{i,1}=L_1),\ldots,I(x^*_{i,k}=2),\ldots,I(x^*_{i,k}=L_k),x^*_{i,k+1},\ldots,x^*_{i,q-1})$

for $i = 1, \ldots, n$.

For $x^*_{i,q}$ is a latent variable ($x_{i,q}$ is binary), we use the following representation for each of the sequential regression model $p(x^*_{i,q}|x^*_{i,1}, \ldots, x^*_{i,q-1}, \boldsymbol{\theta}^{(q)})$:

$$\begin{aligned} x^*_{i,q} &= \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)} + \epsilon_{i,q}, \text{ where } x_{i,q} = I(x^*_{i,q} > 0) \\ \epsilon_{i,q} &\sim \mathrm{N}(0,1), \end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(1, I(x^*_{i,1} = 2), \ldots, I(x^*_{i,1} = L_1), \ldots, I(x^*_{i,k} = 2), \ldots, I(x^*_{i,k} = L_k), x^*_{i,k+1}, \ldots, x^*_{i,q-1})$, for $i = 1, \ldots, n$ and $\boldsymbol{\beta}^{(q)} = (\beta_0^{(q)}, \beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$. We then propose the following priors:

$$\begin{aligned} p(\beta_j^{(q)}|\tau_q, \psi_{j,q}) &\sim \mathrm{N}(0, \frac{1}{\tau_q\psi_{j,q}}), \ j = 0, \ldots, q-1, \\ p(\tau_q) &\propto \frac{1}{\tau_q}, \\ p(\psi_{j,q}) &\sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}), \ j = 0, \ldots, q-1. \end{aligned}$$

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$\begin{aligned} & \prod_{i=1}^{n} \left\{p(x^*_{i,q}|\boldsymbol{\beta}^{(q)})\right\} \prod_{j=0}^{q-1} \left\{p(\beta_j^{(q)}|\tau_q, \psi_{j,q})p(\psi_{j,q})\right\} p(\tau_q) \\ = & \prod_{i=1}^{n} \left\{ \sqrt{\frac{1}{2\pi}} \exp\left[-\frac{1}{2}(x^*_{i,q} - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right] \right\} \\ & \prod_{j=0}^{q-1} \left\{ \sqrt{\frac{\tau_q\psi_{j,q}}{2\pi}} \exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right] \frac{1}{\Gamma(\frac{\nu}{2})}(\frac{\nu}{2})^{(\frac{\nu}{2})}\psi_{j,q}^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu\psi_{j,q}}{2}\right] \right\} \frac{1}{\tau_q}. \end{aligned}$$

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \tau_q, \Psi_q$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\prod_{j=0}^{q-1}\sqrt{\frac{\tau_q\psi_{j,q}}{2\pi}}\exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right]\\
\propto\;& \exp\left\{-\frac{1}{2}\left[(\mathbf{x}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\mathbf{x}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\tau_q\mathbf{D}_\psi^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
\propto\;& \exp\left\{-\frac{1}{2}\left[(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\tau_q\mathbf{D}_\psi^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
=\;& \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})+(\boldsymbol{\beta}^{(q)})'\tau_q\mathbf{D}_\psi^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
\propto\;& \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\tau_q\mathbf{D}_\psi^{(q)})\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\},
\end{aligned}$$

Let $A_q=(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\tau_q\mathbf{D}_\psi^{(q)})$ and $\hat{\boldsymbol{\beta}}^{(q)}=(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}_q'\mathbf{x}^*_q$, then

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\tau_q\mathbf{D}_\psi^{(q)})\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\}\\
=\;& \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}_q'\mathbf{x}^*_q\right]\right\}\\
=\;& \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'A_qA_q^{-1}\widetilde{\mathbf{X}}_q'\mathbf{x}^*_q\right]\right\}\\
\propto\;& \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)}-A_q^{-1}\widetilde{\mathbf{X}}_q'\mathbf{x}^*_q)'A_q(\boldsymbol{\beta}^{(q)}-A_q^{-1}\widetilde{\mathbf{X}}_q'\mathbf{x}^*_q)\right]\right\}\\
\Rightarrow\;& \boldsymbol{\beta}^{(q)}\sim\text{MVN}(\hat{\boldsymbol{\beta}}^{(q)},\mathbf{V}^{(q)})\text{ with }\hat{\boldsymbol{\beta}}^{(q)}=\mathbf{V}^{(q)}\widetilde{\mathbf{X}}_q'\mathbf{x}^*_q\text{ and }\mathbf{V}^{(q)}=(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\tau_q\mathbf{D}_\psi^{(q)})^{-1},
\end{aligned}\tag{7.1}$$

where $\mathbf{x}^*_q=(x^*_{1,q},\ldots,x^*_{n,q})'$, and $\widetilde{\mathbf{X}}_q=(\widetilde{\mathbf{x}}_{1,q},\ldots,\widetilde{\mathbf{x}}_{n,q})'$, where
$\widetilde{\mathbf{x}}_{i,q}=(1,I(x^*_{i,1}=2),\ldots,I(x^*_{i,1}=L_1),\ldots,I(x^*_{i,k}=2),\ldots,I(x^*_{i,k}=L_k),x^*_{i,k+1},\ldots,x^*_{i,q-1})$ for $i=1,\ldots,n$. The matrix $\mathbf{D}_\psi^{(q)}$ is a $q\times q$ diagonal matrix with entries $\psi_{j,q}$ for $j=0,\ldots,q-1$.

2. The full conditional distribution of τ_q is proportional to

$$\begin{aligned}
&\prod_{j=0}^{q-1}\left\{\sqrt{\frac{\tau_q\psi_{j,q}}{2\pi}}\exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right]\right\}\frac{1}{\tau_q}\\
\propto\ & \tau_q^{\frac{q}{2}}\tau_q^{-1}\prod_{j=0}^{q-1}\exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right]\\
=\ & \tau_q^{\frac{q}{2}-1}\exp\left[-\frac{\tau_q}{2}\sum_{j=0}^{q-1}\psi_{j,q}(\beta_j^{(q)})^2\right]\\
=\ & \tau_q^{\frac{q}{2}-1}\exp\left\{-\frac{\tau_q}{2}\left[(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\psi^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
\Rightarrow\ & \tau_q\sim\text{Gamma}\left(\frac{q}{2},\frac{(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\psi^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right),
\end{aligned}\tag{7.2}$$

where $\mathbf{D}_\psi^{(q)}$ is a $q\times q$ diagonal matrix with entries $\psi_{j,q}$ for $j=0,\ldots,q-1$.

3. The full conditional distribution of $\psi_{j,q}$ is proportional to

$$\begin{aligned}
&\prod_{j=0}^{q-1}\left\{\sqrt{\frac{\tau_q\psi_{j,q}}{2\pi}}\exp\left[-\frac{\tau_q\psi_{j,q}}{2}(\beta_j^{(q)})^2\right]\frac{1}{\Gamma(\frac{\nu}{2})}(\frac{\nu}{2})^{(\frac{\nu}{2})}\psi_{j,q}^{\frac{\nu}{2}-1}\exp\left[-\frac{\nu\psi_{j,q}}{2}\right]\right\}\\
\propto\ & \prod_{j=0}^{q-1}\psi_{j,q}^{\frac{1}{2}}\psi_{j,q}^{\frac{\nu}{2}-1}\exp\left[-\psi_{j,q}\left(\frac{\tau_q}{2}(\beta_j^{(q)})^2+\frac{\nu}{2}\right)\right]\\
\Rightarrow\ & \psi_{j,q}\sim\text{Gamma}\left(\frac{\nu+1}{2},\frac{\tau_q}{2}(\beta_j^{(q)})^2+\frac{\nu}{2}\right),\ j=0,\ldots,q-1.
\end{aligned}\tag{7.3}$$

For $x_{i,q}^*$ is a latent variable ($x_{i,q}$ is ordinal), we use the following representation for each of the sequential regression model $p(x_{i,q}^*|x_{i,1}^*,\ldots,x_{i,q-1}^*,\boldsymbol{\theta}^{(q)})$:

$$\begin{aligned}
x_{i,q}^* &= \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)}+\epsilon_{i,q},\ \text{where}\ \ x_{i,q}=j^q\ \ \text{if}\ \ \gamma_{j^q-1}^{(q)}<x_{i,q}^*<\gamma_{j^q}^{(q)},\\
\epsilon_{i,q} &\sim \text{N}(0,1),
\end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q}=$
$(1,I(x_{i,1}^*=2),\ldots,I(x_{i,1}^*=L_1),\ldots,I(x_{i,k}^*=2),\ldots,I(x_{i,k}^*=L_k),x_{i,k+1}^*,\ldots,x_{i,q-1}^*)$, for $i=1,\ldots,n$ and $\boldsymbol{\beta}^{(q)}=(\beta_0^{(q)},\beta_1^{(q),2},\ldots,\beta_1^{(q),L_1},\ldots,\beta_k^{(q),2},\ldots,\beta_k^{(q),L_k},\beta_{k+1}^{(q)},\ldots,\beta_{q-1}^{(q)})'$. We then propose the following priors:

$$\begin{aligned}
p(\beta_j^{(q)}|\tau_q,\psi_{j,q}) &\sim \text{N}(0,\frac{1}{\tau_q\psi_{j,q}}),\ j=0,\ldots,q-1,\\
p(\tau_q) &\propto \frac{1}{\tau_q},\\
p(\psi_{j,q}) &\sim \text{Gamma}(\frac{\nu}{2},\frac{\nu}{2}),\ j=0,\ldots,q-1,
\end{aligned}$$

and we place an improper uniform prior on $\boldsymbol{\gamma}^{(q)}$ i.e.

$$p(\boldsymbol{\gamma}^{(q)}) \propto I(\boldsymbol{\gamma}^{(q)} \in \Omega^{(q)}),$$

where $\Omega^{(q)} = \left\{\gamma_{j^q}^{(q)} : \gamma_0^{(q)} = -\infty < \gamma_1^{(q)} = 0 < \gamma_2^{(q)} < \ldots < \gamma_{J_{q-1}}^{(q)} < \gamma_{J_q}^{(q)} = \infty\right\}$.

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$\begin{aligned}
& \prod_{i=1}^{n} \left\{p(x_{i,q}^*|\boldsymbol{\beta}^{(q)}, \boldsymbol{\gamma}^{(q)}) p(\boldsymbol{\gamma}^{(q)})\right\} \prod_{j=0}^{q-1} \left\{p(\beta_j^{(q)}|\tau_q, \psi_{j,q}) p(\psi_{j,q})\right\} p(\tau_q) \\
= & \prod_{i=1}^{n} \left\{ \sqrt{\frac{1}{2\pi}} \exp\left[-\frac{1}{2}(x_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right] \left[\sum_{j^q=1}^{J_q} I(x_{i,q} = j^q) I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)})\right]\right\} \\
& \prod_{j=0}^{q-1} \left\{ \sqrt{\frac{\tau_q \psi_{j,q}}{2\pi}} \exp\left[-\frac{\tau_q \psi_{j,q}}{2}(\beta_j^{(q)})^2\right] \frac{1}{\Gamma(\frac{\nu}{2})} (\frac{\nu}{2})^{(\frac{\nu}{2})} \psi_{j,q}^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu\psi_{j,q}}{2}\right]\right\} \frac{1}{\tau_q}.
\end{aligned}$$

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \tau_q, \psi_{j,q}, \boldsymbol{\gamma}^{(q)}$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is given by Equation 7.1.

2. The full conditional distribution of τ_q is given by Equation 7.2.

3. The full conditional distribution of $\psi_{j,q}$ is given by Equation 7.3.

4. The full conditional distribution of $\gamma_{j^q}^{(q)}$ for $j^q \in \{1, \ldots, J_q\}$ is proportional to

$$\prod_{i=1}^{n} \left[I(x_{i,q} = j^q) I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) + I(x_{i,q} = j^q + 1) I(\gamma_{j^q}^{(q)} < x_{i,q}^* < \gamma_{j^q+1}^{(q)})\right],$$

which is uniformly distributed on the interval

$$\left[\max\left\{\max\left\{x_{i,q}^* : x_{i,q} = j^q\right\}, \gamma_{j^q-1}^{(q)}\right\}, \min\left\{\min\left\{x_{i,q}^* : x_{i,q} = j^q + 1\right\}, \gamma_{j^q+1}^{(q)}\right\}\right],$$

where $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_{J_q}^{(q)} = \infty$, for $j^q \in \{1, \ldots, J_q\}$.

Appendix D - Simulation study in Section 5.1.1 of Chapter 5

In this section we present details of the simulation study in Section 5.1.1 of Chapter 5. We first describe how we simulated an incomplete data set, and we then describe the analysis we considered.

Data generation

We simulate $x_{i,1}, x_{i,2}, \ldots, x_{i,9},\ i = 1, \ldots, 1000$ in the following way:

$$
\begin{aligned}
x_{i,1} \quad & \text{simulated from a discrete distribution taking values} \in \{1,2,3\} \\
& \text{with probabilities 0.3, 0.4 and 0.3 respectively.} \\
x_{i,2} &\sim p(x_{i,2}|x_{i,1}, \boldsymbol{\theta}^{(2)}) \\
x_{i,3} &\sim \mathrm{N}(5, \frac{1}{\epsilon_{i,3}}),\ \epsilon_{i,3} \sim \mathrm{Gamma}(2,2) \\
x^*_{i,4} &\sim \mathrm{N}(1 + 2x_{i,3}\frac{2}{\epsilon_{i,4}}),\ \epsilon_{i,4} \sim \mathrm{Gamma}(1,1) \\
x_{i,4} &= I(x^*_{i,4} > 11) \\
x^*_{i,5} &\sim \mathrm{N}(2 + 5 * I(x_{i,1} = 2) + I(x_{i,1} = 3) + x_{i,3}, \frac{3}{\epsilon_{i,5}}),\ \epsilon_{i,5} \sim \mathrm{Gamma}(4,4) \\
x_{i,5} &= j^5\ I(c^{(5)}_{j^5-1} < x^*_{i,5} < c^{(5)}_{j^5}),\quad j^5 = 1, \ldots, 4 \\
x_{i,6} &\sim \mathrm{N}(8, \frac{5}{\epsilon_{i,6}}),\ \epsilon_{i,6} \sim \mathrm{Gamma}(2.5, 2.5) \\
x_{i,7} &\sim \mathrm{N}(2 + 2x_{i,6} - x_{i,3}, \frac{5}{\epsilon_{i,7}}),\ \epsilon_{i,7} \sim \mathrm{Gamma}(3,3) \\
x^*_{i,8} &\sim \mathrm{N}(3 + 2x_{i,7}, \frac{1}{\epsilon_{i,8}}),\ \epsilon_{i,8} \sim \mathrm{Gamma}(1.5, 1.5) \\
x_{i,8} &= j^8\ I(c^{(8)}_{j^8-1} < x^*_{i,8} < c^{(8)}_{j^8}),\quad j^8 = 1, \ldots, 8 \\
x_{i,9} &\sim \mathrm{N}(1 + 4 * I(x_{i,2} = 2) - 3 * I(x_{i,2} = 3) + 5 * I(x_{i,2} = 4) + 2 * I(x_{i,2} = 5) + \\
& \quad 5x_{i,7}, \frac{3}{\epsilon_{i,9}}),\ \epsilon_{i,9} \sim \mathrm{Gamma}(6,6)
\end{aligned}
$$

where

$$
p(x_{i,2} = j^2 | x_{i,1}, \boldsymbol{\theta}^{(2)}) = \frac{\exp(\theta^{(2)}_{0,j^2} + \theta^{(2)}_{1,j^2} I(x_{i,1} = 2) + \theta^{(2)}_{2,j^2} I(x_{i,1} = 3))}{\sum_{s=1}^{5} \exp(\theta^{(2)}_{0,s} + \theta^{(2)}_{1,s} I(x_{i,1} = 2) + \theta^{(2)}_{2,s} I(x_{i,1} = 3))}
$$

where $j^2 \in \{1, \ldots, 5\}$ and $\theta^{(2)}_{0,1} = \theta^{(2)}_{1,1} = \theta^{(2)}_{2,1} = 0$ for identifiability, $\theta^{(2)}_{0,2} = 1, \theta^{(2)}_{1,2} = -1, \theta^{(2)}_{2,2} = -2, \theta^{(2)}_{0,3} = -1, \theta^{(2)}_{1,3} = 3, \theta^{(2)}_{2,3} = 1, \theta^{(2)}_{0,4} = -1, \theta^{(2)}_{1,4} = 2, \theta^{(2)}_{2,4} = 2, \theta^{(2)}_{0,5} = -1, \theta^{(2)}_{1,5} = 2, \theta^{(2)}_{2,5} = 1$, and $I(\cdot)$ is the indicator function. The threshold parameters $c^{(5)}_0 = -\infty, c^{(5)}_1 = 6.962, c^{(5)}_2 =$

$9.292, c_3^{(5)} = 11.59, c_4^{(5)} = \infty$ and $c_0^{(8)} = -\infty, c_1^{(8)} = -2, c_2^{(8)} = 4, c_3^{(8)} = 9, c_4^{(8)} = 13, c_5^{(8)} = 17, c_6^{(8)} = 22, c_7^{(8)} = 28, c_8^{(8)} = \infty$.

We introduce missing values into variables $\mathbf{x}_2, \ldots, \mathbf{x}_9$ in the following way:

$$
\begin{aligned}
p(m_{i,2} = 1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,3} = 1) &= \left\{ \frac{\exp(1 - 5 * I(x_{i,2} = 2) - 4 * I(x_{i,2} = 3))}{1 + \exp(1 - 5 * I(x_{i,2} = 2) - 4 * I(x_{i,2} = 3))} \right\} (1 - m_{i,2}) \text{ (MAR)}, \\
p(m_{i,4} = 1) &= \left\{ \frac{\exp(-5.5 + 5x_{i,3})}{1 + \exp(-5.5 + 5x_{i,3})} \right\} (1 - m_{i,3}) \text{ (MAR)}, \\
p(m_{i,5} = 1) &= \left\{ \frac{\exp(-5.5 + 5x_{i,3})}{1 + \exp(-5.5 + 5x_{i,3})} \right\} (1 - m_{i,3}) \text{ (MAR)}, \\
p(m_{i,6} = 1) &= 0.3 \text{ (MCAR)} \\
p(m_{i,7} = 1) &= \left\{ \frac{\exp(60 - 8x_{i,6})}{1 + \exp(60 - 8x_{i,6})} \right\} (1 - m_{i,6}) \text{ (MAR)}, \\
p(m_{i,8} = 1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,9} = 1) &= \left\{ \frac{\exp(6 - x_{i,7})}{1 + \exp(6 - x_{i,7})} \right\} (1 - m_{i,7}) \text{ (MAR)},
\end{aligned}
$$

where $m_{i,j}$ be the missing data indicator for $x_{i,j}$, where $m_{i,j} = 1$ indicates $x_{i,j}$ is missing and $m_{i,j} = 0$ indicates $x_{i,j}$ is observed.

Analysis

We obtain the following estimates from an analysis of the imputed datasets:

- Estimates of the means of $\mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_9$.
- Estimates of the proportion of units with $\mathbf{x}_2 = 1, \ldots, 5$.
- Estimates of the regression coefficients of a linear regression from the following regression models: $p(\mathbf{x}_2|\mathbf{x}_1), p(\mathbf{x}_7|\mathbf{x}_6, \mathbf{x}_3), p(\mathbf{x}_9|\mathbf{x}_2, \mathbf{x}_7)$.

Appendix E - "P-steps" for LFRM

In this section, we present the details on deriving the full conditional distributions of the parameters in the "P-steps" mentioned in the LFRM strategy in Chapter 5. The full conditional distribution of $\boldsymbol{\theta}^{(1)}$ where $\mathbf{x}_1$ follows a multinomial distribution and Metropolis-Hastings update for the multinomial logistic regression model have been presented in the LFRM section in Chapter 5. Here, we will present the details on deriving the full conditional distributions of the remaining parameters in the "P-steps", such as $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is continuous, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)})$ for $\mathbf{x}_q^*$ when

$\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is binary) and $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is ordinal).

For $x_{i,q}^*$ is continuous, we use the following representation for each of the sequential regression model $p(x_{i,q}^*|x_{i,1}^*, \ldots, x_{i,q-1}^*, \boldsymbol{\theta}^{(q)})$:

$$\begin{aligned} x_{i,q}^* &= \beta_0^{(q)} + \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)} + \epsilon_i, \\ \epsilon_{i,q} &\sim \mathrm{N}(0, \frac{1}{\phi_q}), \end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$, for $i = 1, \ldots, n$ and $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$. We then propose the the following prior for $\boldsymbol{\beta}^{(q)}$ (Park and Casella (2008)):

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \phi_q) &\sim \mathrm{MVN}(\mathbf{0}, \frac{1}{\phi_q}\mathbf{D}_\eta^{(q)}), \text{ with } \mathbf{D}_\eta^{(q)} = \mathrm{diag}(\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2), \\ p(\eta_{j,q}^2|\lambda_q^2) &= \frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right), \ j = 1, \ldots, q-1, \end{aligned}$$

and we consider a diffuse hyperprior on λ_q^2 of the form

$$p(\lambda_q^2) = \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right), \ \lambda_q^2 > 0, \ r > 0, \ \delta > 0,$$

which is a gamma distribution with shape and rate r and δ respectively. We also specify the prior distributions for ϕ_q and $\beta_0^{(q)}$ as follows:

$$\begin{aligned} p(\phi_q) &\propto \frac{1}{\phi_q}, \\ p(\beta_0^{(q)}) &\propto 1. \end{aligned}$$

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$\begin{aligned} & \prod_{i=1}^n \left\{p(x_{i,q}^*|\beta_0^{(q)}, \boldsymbol{\beta}^{(q)}, \phi_q)\right\} \prod_{j=1}^{q-1} \left\{p(\beta_j^{(q)}|\phi_q, \eta_{j,q}^2)p(\eta_{j,q}^2|\lambda_q^2)\right\} p(\phi_q)p(\lambda_q^2)p(\beta_0^{(q)}) \\ = & \prod_{i=1}^n \left\{\sqrt{\frac{\phi_q}{2\pi}}\exp\left[-\frac{\phi_q}{2}(\widetilde{x}_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\} \\ & \prod_{j=1}^{q-1} \left\{\sqrt{\frac{\phi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\phi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\frac{\lambda_q^2}{2}\exp\left[\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right]\right\} \frac{1}{\phi_q}\frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right), \end{aligned}$$

where $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$.

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \phi_q, \beta_0^{(q)}$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{\sqrt{\frac{\phi_q}{2\pi}}\exp\left[-\frac{\phi_q}{2}(\widetilde{x}_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\}\prod_{j=1}^{q-1}\left\{\sqrt{\frac{\phi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\phi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\right\}\\
\propto\;& \exp\left\{-\frac{\phi_q}{2}\left[(\widetilde{\mathbf{x}}_q^{*}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{x}}_q^{*}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
\propto\;& \exp\left\{-\frac{\phi_q}{2}\left[(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
=\;& \exp\left\{-\frac{\phi_q}{2}\left[(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\}\\
\propto\;& \exp\left\{-\frac{\phi_q}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\mathbf{D}_\eta^{(q)})\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\},
\end{aligned}$$

Let $A_q=(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\mathbf{D}_\eta^{(q)})$ and $\hat{\boldsymbol{\beta}}^{(q)}=(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^{*}$, then

$$\begin{aligned}
&\exp\left\{-\frac{\phi_q}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\mathbf{D}_\eta^{(q)})\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\}\\
=\;& \exp\left\{-\frac{\phi_q}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^{*}\right]\right\}\\
=\;& \exp\left\{-\frac{\phi_q}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'A_qA_q^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^{*}\right]\right\}\\
\propto\;& \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)}-A_q^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^{*})'\phi_qA_q(\boldsymbol{\beta}^{(q)}-A_q^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^{*})\right]\right\}\\
\Rightarrow\;& \boldsymbol{\beta}^{(q)}\sim \mathrm{MVN}(\hat{\boldsymbol{\beta}}^{(q)},\mathbf{V}^{(q)}) \text{ with}\\
&\hat{\boldsymbol{\beta}}^{(q)}=A_q^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^{*} \text{ and } \mathbf{V}^{(q)}=\frac{1}{\phi_q}(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\mathbf{D}_\eta^{(q)})^{-1}.
\end{aligned}$$

where $\widetilde{\mathbf{x}}_q^{*}=(\widetilde{x}_{1,q}^{*},\ldots,\widetilde{x}_{n,q}^{*})'$ with $\widetilde{x}_{i,q}^{*}=x_{i,q}^{*}-\beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q=(\widetilde{\mathbf{x}}_{1,q},\ldots,\widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q}=(I(x_{i,1}^{*}=2),\ldots,I(x_{i,1}^{*}=L_1),\ldots,I(x_{i,k}^{*}=2),\ldots,I(x_{i,k}^{*}=L_k),x_{i,k+1}^{*},\ldots,x_{i,q-1}^{*})$ for $i=1,\ldots,n$. The matrix $\mathbf{D}_\eta^{(q)}$ is a $(q-1)\times(q-1)$ diagonal matrix with entries $\frac{1}{\eta_{j,q}^2}$ for $j=1,\ldots,q-1$.

2. The full conditional distribution of $\eta_{j,q}^2,\ j=1,\ldots,q-1$ is proportional to

$$\prod_{j=1}^{q-1}\left\{\sqrt{\frac{\phi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\phi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\frac{\lambda_q^2}{2}\exp\left[\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right]\right\}.$$

Let $g = \frac{1}{\eta_{j,q}^2} \Rightarrow \left|\frac{d\eta_{j,q}^2}{dg}\right| = \frac{1}{g^2}$. Hence

$$
\begin{aligned}
& \prod_{j=1}^{q-1} \left\{ \sqrt{\frac{\phi_q}{2\pi\eta_{j,q}^2}} \exp\left[-\frac{\phi_q}{2\eta_{j,q}^2} (\beta_j^{(q)})^2 \right] \frac{\lambda_q^2}{2} \exp\left[\frac{-\lambda_q^2 \eta_{j,q}^2}{2} \right] \right\} \\
\propto & \prod_{j=1}^{q-1} g^{\frac{1}{2}} g^{-2} \exp\left\{ -\frac{g\phi_q}{2} (\beta_j^{(q)})^2 - \frac{\lambda_q^2}{2g} \right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \left[g\phi_q (\beta_j^{(q)})^2 - \frac{\lambda_q^2}{g} \right] \right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \frac{\lambda_q^2 (\beta_j^{(q)})^2 \phi_q}{g\lambda_q^2} \left(g^2 - \frac{\lambda_q^2}{(\beta_j^{(q)})^2 \phi_q} \right) \right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \frac{\lambda_q^2}{(\mu_q')^2 g} \left[g^2 - (\mu_q')^2 \right] \right\}, \mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2 \phi_q}} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \frac{\lambda_q^2}{(\mu_q')^2 g} \left[(g - \mu_q')^2 + 2g\mu_q' \right] \right\} \\
\propto & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left[-\frac{1}{2} \frac{\lambda_q^2}{(\mu_q')^2 g} (g - \mu_q')^2 \right] \\
\Rightarrow & \; g \sim \sqrt{\frac{\lambda_q'}{2\pi}} g^{-\frac{3}{2}} \exp\left(-\frac{\lambda_q' (g - \mu_q')^2}{2(\mu_q')^2 g} \right), \; g > 0,
\end{aligned}
$$

which is a inverse-Gaussian distribution with parameters

$$
\mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2 \phi_q}} \quad \text{and} \quad \lambda_q' = \lambda_q^2, \quad j = 1, \ldots, q-1.
$$

3. The full conditional distribution of λ_q^2 is proportional to

$$
\begin{aligned}
& \prod_{j=1}^{q-1} \left[\frac{\lambda_q^2}{2} \exp\left(\frac{-\lambda_q^2 \eta_{j,q}^2}{2} \right) \right] \frac{\delta^r}{\Gamma(r)} (\lambda_q^2)^{r-1} \exp\left(-\delta\lambda_q^2 \right) \\
\propto & \prod_{j=1}^{q-1} \left[\frac{\lambda_q^2}{2} \exp\left(\frac{-\lambda_q^2 \eta_{j,q}^2}{2} \right) \right] (\lambda_q^2)^{r-1} \exp\left(-\delta\lambda_q^2 \right) \\
= & \left(\frac{\lambda_q^2}{2} \right)^{q-1} (\lambda_q^2)^{r-1} \exp\left[-\lambda_q^2 \left(\frac{\sum_{j=1}^{q-1} \eta_{j,q}^2}{2} \right) \right] \exp\left(-\delta\lambda_q^2 \right) \\
\propto & \left(\lambda_q^2 \right)^{(q-1)+r-1} \exp\left[-\lambda_q^2 \left(\sum_{j=1}^{q-1} \frac{\eta_{j,q}^2}{2} + \delta \right) \right] \\
\Rightarrow & \; \lambda_q^2 \sim \text{Gamma}\left((q-1) + r, \sum_{j=1}^{q-1} \frac{\eta_{j,q}^2}{2} + \delta \right).
\end{aligned}
$$

4. The full conditional distribution of ϕ_q is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{\sqrt{\frac{\phi_q}{2\pi}}\exp\left[-\frac{\phi_q}{2}(\widetilde{x}^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\}\prod_{j=1}^{q-1}\left\{\sqrt{\frac{\phi_q}{2\eta^2_{j,q}}}\exp\left[-\frac{\phi_q}{2\eta^2_{j,q}}(\beta_j^{(q)})^2\right]\right\}\frac{1}{\phi_q}\\
\propto\;& \phi_q^{-1}\phi_q^{(\frac{n}{2})}\phi_q^{(\frac{q-1}{2})}\exp\left[-\phi_q\left(\frac{(\widetilde{\mathbf{x}}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{x}}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right)\right]\\
\Rightarrow\;& \phi_q\sim\text{Gamma}\left(\frac{n+(q-1)}{2},\frac{(\widetilde{\mathbf{x}}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{x}}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right),
\end{aligned}$$

where $\widetilde{\mathbf{x}}^*_q=(\widetilde{x}^*_{1,q},\ldots,\widetilde{x}^*_{n,q})'$ with $\widetilde{x}^*_{i,q}=x^*_{i,q}-\beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q=(\widetilde{\mathbf{x}}_{1,q},\ldots,\widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q}=(I(x^*_{i,1}=2),\ldots,I(x^*_{i,1}=L_1),\ldots,I(x^*_{i,k}=2),\ldots,I(x^*_{i,k}=L_k),x^*_{i,k+1},\ldots,x^*_{i,q-1})$ for $i=1,\ldots,n$. The matrix $\mathbf{D}_\eta^{(q)}$ is a $(q-1)\times(q-1)$ diagonal matrix with entries $\frac{1}{\eta^2_{j,q}}$ for $j=1,\ldots,q-1$.

5. The full conditional distribution of $\beta_0^{(q)}$ is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{\sqrt{\frac{\phi_q}{2\pi}}\exp\left[-\frac{\phi_q}{2}(x^*_{i,q}-\beta_0^{(q)}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\}\\
=\;&\prod_{i=1}^{n}\left\{\sqrt{\frac{\phi_q}{2\pi}}\exp\left[-\frac{\phi_q}{2}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)}-\beta_0^{(q)})^2\right]\right\}\\
=\;&\prod_{i=1}^{n}\left\{\sqrt{\frac{\phi_q}{2\pi}}\exp\left[-\frac{\phi_q}{2}\left((x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2-2(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}+(\beta_0^{(q)})^2\right)\right]\right\}\\
\propto\;&\prod_{i=1}^{n}\left\{\sqrt{\frac{\phi_q}{2\pi}}\exp\left[-\frac{\phi_q}{2}\left((\beta_0^{(q)})^2-2(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right]\right\}\\
\propto\;&\exp\left[-\frac{\phi_q}{2}\sum_{i=1}^{n}\left((\beta_0^{(q)})^2-2(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right]\\
=\;&\exp\left[-\frac{\phi_q}{2}\left(\sum_{i=1}^{n}(\beta_0^{(q)})^2-2\sum_{i=1}^{n}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right]\\
=\;&\exp\left[-\frac{n\phi_q}{2}\left((\beta_0^{(q)})^2-2\beta_0^{(q)}\frac{\sum_{i=1}^{n}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}\right)\right]\\
\propto\;&\exp\left[-\frac{n\phi_q}{2}\left(\beta_0^{(q)}-\frac{\sum_{i=1}^{n}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}\right)^2\right]\\
\Rightarrow\;&\beta_0^{(q)}\sim\text{N}(\frac{\sum_{i=1}^{n}(x^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n},\frac{\phi_q^{-1}}{n}),
\end{aligned}$$

where
$\widetilde{\mathbf{x}}_{i,q}=(I(x^*_{i,1}=2),\ldots,I(x^*_{i,1}=L_1),\ldots,I(x^*_{i,k}=2),\ldots,I(x^*_{i,k}=L_k),x^*_{i,k+1},\ldots,x^*_{i,q-1})$ for $i=1,\ldots,n$.

For $x^*_{i,q}$ is a latent variable ($x_{i,q}$ is binary), we use the following representation for each of the sequential regression model $p(x^*_{i,q}|x^*_{i,1},\ldots,x^*_{i,q-1},\boldsymbol{\theta}^{(q)})$:

$$\begin{aligned} x^*_{i,q} &= \beta_0^{(q)} + \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)} + \epsilon_i, \text{ where } x_{i,q} = I(x^*_{i,q} > 0) \\ \epsilon_{i,q} &\sim \mathrm{N}(0,1), \end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(I(x^*_{i,1} = 2),\ldots,I(x^*_{i,1} = L_1),\ldots,I(x^*_{i,k} = 2),\ldots,I(x^*_{i,k} = L_k), x^*_{i,k+1},\ldots,x^*_{i,q-1})$, for $i = 1,\ldots,n$ and $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2},\ldots,\beta_1^{(q),L_1},\ldots,\beta_k^{(q),2},\ldots,\beta_k^{(q),L_k},\beta_{k+1}^{(q)},\ldots,\beta_{q-1}^{(q)})'$. We then propose the following prior for $\boldsymbol{\beta}^{(q)}$ (Park and Casella (2008)):

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q) &\sim \mathrm{MVN}(\mathbf{0}, \mathbf{D}_\eta^{(q)}), \text{ with } \mathbf{D}_\eta^{(q)} = \mathrm{diag}(\eta^2_{1,q},\ldots,\eta^2_{q-1,q}), \\ p(\eta^2_{j,q}|\lambda^2_q) &= \frac{\lambda^2_q}{2}\exp\left(\frac{-\lambda^2_q\eta^2_{j,q}}{2}\right),\ j = 1,\ldots,q-1, \end{aligned}$$

and we consider a diffuse hyperprior on λ^2_q of the form

$$p(\lambda^2_q) = \frac{\delta^r}{\Gamma(r)}(\lambda^2_q)^{r-1}\exp\left(-\delta\lambda^2_q\right),\ \lambda^2_q > 0,\ r > 0,\ \delta > 0,$$

which is a gamma distibution with shape and rate r and δ respectively. We also specify the prior distributions for $\beta_0^{(q)}$ as follows:

$$p(\beta_0^{(q)}) \propto 1.$$

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$\begin{aligned} & \prod_{i=1}^n \left\{p(x^*_{i,q}|\beta_0^{(q)},\boldsymbol{\beta}^{(q)})\right\} \prod_{j=1}^{q-1}\left\{p(\beta_j^{(q)}|\eta^2_{j,q})p(\eta^2_{j,q}|\lambda^2_q)\right\} p(\lambda^2_q)p(\beta_0^{(q)}) \\ = & \prod_{i=1}^n \left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(\widetilde{x}^*_{i,q} - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\} \\ & \prod_{j=1}^{q-1}\left\{\sqrt{\frac{1}{2\pi\eta^2_{j,q}}}\exp\left[-\frac{1}{2\eta^2_{j,q}}(\beta_j^{(q)})^2\right]\frac{\lambda^2_q}{2}\exp\left[\frac{-\lambda^2_q\eta^2_{j,q}}{2}\right]\right\}\frac{1}{\phi_q}\frac{\delta^r}{\Gamma(r)}(\lambda^2_q)^{r-1}\exp\left(-\delta\lambda^2_q\right), \end{aligned}$$

where $\widetilde{x}^*_{i,q} = x^*_{i,q} - \beta_0^{(q)}$.

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda^2_q, \beta_0^{(q)}$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(\widetilde{x}^*_{i,q}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\}\prod_{j=1}^{q-1}\left\{\sqrt{\frac{1}{2\pi\eta^2_{j,q}}}\exp\left[-\frac{1}{2\eta^2_{j,q}}(\beta^{(q)}_j)^2\right]\right\}\\
\propto\ &\exp\left\{-\frac{1}{2}\left[(\widetilde{\mathbf{x}}^*_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{x}}_q-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\eta\boldsymbol{\beta}^{(q)}\right]\right\}\\
\propto\ &\exp\left\{-\frac{1}{2}\left[(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\eta\boldsymbol{\beta}^{(q)}\right]\right\}\\
=\ &\exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})'\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}^{(q)}_\eta\boldsymbol{\beta}^{(q)}\right]\right\}\\
\propto\ &\exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q+\mathbf{D}^{(q)}_\eta)\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\},
\end{aligned}$$

Let $A_q=(\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q+\mathbf{D}^{(q)}_\eta)$ and $\hat{\boldsymbol{\beta}}^{(q)}=(\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{x}}^*_q$, then

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q+\mathbf{D}^{(q)}_\eta)\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\}\\
=\ &\exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q(\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{x}}^*_q\right]\right\}\\
=\ &\exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'A_qA_q^{-1}\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{x}}^*_q\right]\right\}\\
\propto\ &\exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)}-A_q^{-1}\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{x}}^*_q)'A_q(\boldsymbol{\beta}^{(q)}-A_q^{-1}\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{x}}^*_q)\right]\right\}\\
\Rightarrow\ &\boldsymbol{\beta}^{(q)}\sim\text{MVN}(\hat{\boldsymbol{\beta}}^{(q)},\mathbf{V}^{(q)})\text{ with }\hat{\boldsymbol{\beta}}^{(q)}=A_q^{-1}\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{x}}^*_q\text{ and }\mathbf{V}^{(q)}=(\widetilde{\mathbf{X}}'_q\widetilde{\mathbf{X}}_q+\mathbf{D}^{(q)}_\eta)^{-1},
\end{aligned}\tag{7.4}$$

where $\widetilde{\mathbf{x}}^*_q=(\widetilde{x}^*_{1,q},\ldots,\widetilde{x}^*_{n,q})'$, where $\widetilde{x}^*_{i,q}=x^*_{i,q}-\beta^{(q)}_0$ and $\widetilde{\mathbf{X}}_q=(\widetilde{\mathbf{x}}_{1,q},\ldots,\widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q}=(I(x^*_{i,1}=2),\ldots,I(x^*_{i,1}=L_1),\ldots,I(x^*_{i,k}=2),\ldots,I(x^*_{i,k}=L_k),x^*_{i,k+1},\ldots,x^*_{i,q-1})$ for $i=1,\ldots,n$. The matrix $\mathbf{D}^{(q)}_\eta$ is a $(q-1)\times(q-1)$ diagonal matrix with entries $\frac{1}{\eta^2_{j,q}}$ for $j=1,\ldots,q-1$.

2. The full conditional distribution of $\eta^2_{j,q},\ j=1,\ldots,q-1$ is proportional to

$$\prod_{j=1}^{q-1}\left\{\sqrt{\frac{1}{2\pi\eta^2_{j,q}}}\exp\left[-\frac{1}{2\eta^2_{j,q}}(\beta^{(q)}_j)^2\right]\frac{\lambda_q^2}{2}\exp\left[\frac{-\lambda_q^2\eta^2_{j,q}}{2}\right]\right\}.$$

Let $g = \frac{1}{\eta_{j,q}^2} \Rightarrow \left|\frac{d\eta_{j,q}^2}{dg}\right| = \frac{1}{g^2}$. Hence

$$\begin{aligned}
& \prod_{j=1}^{q-1} \left\{ \sqrt{\frac{1}{2\pi\eta_{j,q}^2}} \exp\left[-\frac{1}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right] \frac{\lambda_q^2}{2} \exp\left[\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right] \right\} \\
\propto & \prod_{j=1}^{q-1} g^{\frac{1}{2}} g^{-2} \exp\left\{-\frac{g}{2}(\beta_j^{(q)})^2 - \frac{\lambda_q^2}{2g}\right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\left[g(\beta_j^{(q)})^2 - \frac{\lambda_q^2}{g}\right]\right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2(\beta_j^{(q)})^2}{g\lambda_q^2}\left(g^2 - \frac{\lambda_q^2}{(\beta_j^{(q)})^2}\right)\right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2}{(\mu_q')^2 g}\left[g^2 - (\mu_q')^2\right]\right\}, \mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2}} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2}{(\mu_q')^2 g}\left[(g-\mu_q')^2 + 2g\mu_q'\right]\right\} \\
\propto & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2}{(\mu_q')^2 g}(g-\mu_q')^2\right\} \\
\Rightarrow & \; g \sim \sqrt{\frac{\lambda_q'}{2\pi}} g^{-\frac{3}{2}} \exp\left\{-\frac{\lambda_q'(g-\mu_q')^2}{2(\mu_q')^2 g}\right\}, \; g > 0, \qquad (7.5)
\end{aligned}$$

which is a inverse-Gaussian distribution with parameters

$$\mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2}} \;\text{ and }\; \lambda_q' = \lambda_q^2, \quad j = 1, \ldots, q-1.$$

3. The full conditional distribution of λ_q^2 is proportional to

$$\begin{aligned}
& \prod_{j=1}^{q-1} \left[\frac{\lambda_q^2}{2} \exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right)\right] \frac{\delta^r}{\Gamma(r)} (\lambda_q^2)^{r-1} \exp\left(-\delta\lambda_q^2\right) \\
\propto & \prod_{j=1}^{q-1} \left[\frac{\lambda_q^2}{2} \exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right)\right] (\lambda_q^2)^{r-1} \exp\left(-\delta\lambda_q^2\right) \\
= & \left(\frac{\lambda_q^2}{2}\right)^{q-1} (\lambda_q^2)^{r-1} \exp\left[-\lambda_q^2\left(\frac{\sum_{j=1}^{q-1}\eta_{j,q}^2}{2}\right)\right] \exp\left(-\delta\lambda_q^2\right) \\
\propto & \left(\lambda_q^2\right)^{(q-1)+r-1} \exp\left[-\lambda_q^2\left(\sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2} + \delta\right)\right] \\
\Rightarrow & \; \lambda_q^2 \sim \text{Gamma}\left((q-1)+r, \sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2} + \delta\right). \qquad (7.6)
\end{aligned}$$

4. The full conditional distribution of $\beta_0^{(q)}$ is proportional to

$$
\begin{aligned}
& \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(x_{i,q}^{*}-\beta_0^{(q)}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\} \\
= & \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)}-\beta_0^{(q)})^2\right]\right\} \\
= & \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}\left((x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2-2(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}+(\beta_0^{(q)})^2\right)\right]\right\} \\
\propto & \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}\left((\beta_0^{(q)})^2-2(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right]\right\} \\
\propto & \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left((\beta_0^{(q)})^2-2(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right] \\
= & \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}(\beta_0^{(q)})^2-2\sum_{i=1}^{n}(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right] \\
= & \exp\left[-\frac{n}{2}\left((\beta_0^{(q)})^2-2\beta_0^{(q)}\frac{\sum_{i=1}^{n}(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}\right)\right] \\
\propto & \exp\left[-\frac{n}{2}\left(\beta_0^{(q)}-\frac{\sum_{i=1}^{n}(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}\right)^2\right] \\
\Rightarrow & \beta_0^{(q)}\sim \mathrm{N}(\frac{\sum_{i=1}^{n}(x_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n},\frac{1}{n}),
\end{aligned}
\tag{7.7}
$$

where
$\widetilde{\mathbf{x}}_{i,q}=(I(x_{i,1}^{*}=2),\ldots,I(x_{i,1}^{*}=L_1),\ldots,I(x_{i,k}^{*}=2),\ldots,I(x_{i,k}^{*}=L_k),x_{i,k+1}^{*},\ldots,x_{i,q-1}^{*})$ for $i=1,\ldots,n$.

For $x_{i,q}^{*}$ is a latent variable ($x_{i,q}$ is ordinal), we use the following representation for each of the sequential regression model $p(x_{i,q}^{*}|x_{i,1}^{*},\ldots,x_{i,q-1}^{*},\boldsymbol{\theta}^{(q)})$:

$$
\begin{aligned}
x_{i,q}^{*} &= \beta_0^{(q)}+\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)}+\epsilon_i,\ \text{where } x_{i,q}=j^q \text{ if } \gamma_{j^q-1}^{(q)}<x_{i,q}^{*}<\gamma_{j^q}^{(q)} \\
\epsilon_{i,q} &\sim \mathrm{N}(0,1),
\end{aligned}
$$

where $\widetilde{\mathbf{x}}_{i,q}=$
$(I(x_{i,1}^{*}=2),\ldots,I(x_{i,1}^{*}=L_1),\ldots,I(x_{i,k}^{*}=2),\ldots,I(x_{i,k}^{*}=L_k),x_{i,k+1}^{*},\ldots,x_{i,q-1}^{*})$, for $i=1,\ldots,n$ and $\boldsymbol{\beta}^{(q)}=(\beta_1^{(q),2},\ldots,\beta_1^{(q),L_1},\ldots,\beta_k^{(q),2},\ldots,\beta_k^{(q),L_k},\beta_{k+1}^{(q)},\ldots,\beta_{q-1}^{(q)})'$. We then propose the the following prior for $\boldsymbol{\beta}^{(q)}$ (Park and Casella (2008)):

$$
\begin{aligned}
p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q) &\sim \mathrm{MVN}(\mathbf{0},\mathbf{D}_{\eta}^{(q)}),\ \text{with } \mathbf{D}_{\eta}^{(q)}=\mathrm{diag}(\eta_{1,q}^2,\ldots,\eta_{q-1,q}^2), \\
p(\eta_{j,q}^2|\lambda_q^2) &= \frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right),\ j=1,\ldots,q-1,
\end{aligned}
$$

and we consider a diffuse hyperprior on λ_q^2 of the form

$$p(\lambda_q^2) = \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right),\ \lambda_q^2 > 0,\ r > 0,\ \delta > 0,$$

which is a gamma distibution with shape and rate r and δ respectively. We also specify the prior distributions for $\beta_0^{(q)}$ as follows:

$$p(\beta_0^{(q)}) \ \propto \ 1.$$

and we place an improper uniform prior on $\boldsymbol{\gamma}^{(q)}$ i.e.

$$p(\boldsymbol{\gamma}^{(q)}) \propto I(\boldsymbol{\gamma}^{(q)} \in \Omega^{(q)}),$$

where $\Omega^{(q)} = \left\{\gamma_{j^q}^{(q)} : \gamma_0^{(q)} = -\infty < \gamma_1^{(q)} = 0 < \gamma_2^{(q)} < \ldots < \gamma_{J_{q-1}}^{(q)} < \gamma_{J_q}^{(q)} = \infty\right\}$.

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{p(x_{i,q}^*|\beta_0^{(q)},\boldsymbol{\beta}^{(q)},\boldsymbol{\gamma}^{(q)})p(\boldsymbol{\gamma}^{(q)})\right\}\prod_{j=1}^{q-1}\left\{p(\beta_j^{(q)}|\eta_{j,q}^2)p(\eta_{j,q}^2|\lambda_q^2)\right\}p(\lambda_q^2)p(\beta_0^{(q)}) \\
= &\prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(\widetilde{x}_{i,q}^* - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\left[\sum_{j^q=1}^{J_q} I(x_{i,q}=j^q)I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)})\right]\right\} \\
&\prod_{j=1}^{q-1}\left\{\sqrt{\frac{1}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{1}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\frac{\lambda_q^2}{2}\exp\left[\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right]\right\}\frac{1}{\phi_q}\frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right),
\end{aligned}$$

where $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$.

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is given by Equation 7.4.

2. The full conditional distribution of $\eta_{j,q}^2,\ j = 1,\ldots,q-1$ is given by Equation 7.5.

3. The full conditional distribution of λ_q^2 is given by Equation 7.6.

4. The full conditional distribution of $\beta_0^{(q)}$ is given by Equation 7.7.

5. The full conditional distribution of $\gamma_{j^q}^{(q)}$ for $j^q \in \{1,\ldots,J_q\}$ is proportional to

$$\prod_{i=1}^{n}\left[I(x_{i,q}=j^q)I(\gamma_{j^q-1}^{(q)} < x_{i,q}^* < \gamma_{j^q}^{(q)}) + I(x_{i,q}=j^q+1)I(\gamma_{j^q}^{(q)} < x_{i,q}^* < \gamma_{j^q+1}^{(q)})\right],$$

which is uniformly distributed on the interval

$$\left[\max\left\{\max\left\{x^*_{i,q} : x_{i,q} = j^q\right\}, \gamma^{(q)}_{j^q-1}\right\}, \min\left\{\min\left\{x^*_{i,q} : x_{i,q} = j^q + 1\right\}, \gamma^{(q)}_{j^q+1}\right\}\right],$$

where $\gamma^{(q)}_0 = -\infty$, $\gamma^{(q)}_1 = 0$ and $\gamma^{(q)}_{J_q} = \infty$, for $j^q \in \{1, \ldots, J_q\}$.

Appendix F - Simulation study in Section 5.2.1 of Chapter 5

In this section we present details of the simulation study in Section 5.2.1 of Chapter 5. We first describe how we simulated an incomplete data set, we then describe the analysis we considered.

Data generation

We simulate $x_{i,1}, x_{i,2}, \ldots, x_{i,25},\ i = 1, \ldots, 1000$ in the following way:

$x_{i,1}$		simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.3, 0.4 and 0.3 respectively.
$x_{i,2}$		simulated from a discrete distribution taking values $\in \{1, 2, 3, 4, 5\}$ with each probability of 0.2.
$x_{i,3}$		simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.2, 0.2 and 0.6 respectively.
$x_{i,4}$		simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.25, 0.35 and 0.4 respectively.
$x_{i,5}$	$\sim$	$p(x_{i,5}\|x_{i,1}, \boldsymbol{\theta}^{(5)})$
$x_{i,6}$	$\sim$	$\mathrm{N}(0, 1)$
$x_{i,7}$	$\sim$	$\mathrm{N}(1 + 2 * I(x_{i,5} = 2) - I(x_{i,5} = 3) + I(x_{i,5} = 4), 4)$
$x_{i,8}$	$\sim$	$\mathrm{N}(3, 1)$
$x_{i,9}$		simulated from a discrete distribution taking values $\in \{1, 2\}$ with probabilities 0.5, and 0.5 respectively.
$x_{i,10}$		simulated from a discrete distribution taking values $\in \{1, 2\}$ with probabilities 0.3, and 0.7 respectively.
$x^*_{i,11}$	$\sim$	$\mathrm{N}(3 - x_{i,7}, 4)$,
$x_{i,11}$	$=$	$j^{11}\ I(c^{(11)}_{j^{11}-1} < x^*_{i,11} < c^{(11)}_{j^{11}}),\quad j^{11} = 1, \ldots, 3$
$x_{i,12}$		simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.25, 0.5 and 0.25 respectively.

$$
\begin{aligned}
x_{i,13} &\sim \mathrm{N}(5,1)\\
x_{i,14} &\sim \mathrm{N}(1+2x_{i,13},1)\\
x^*_{i,15} &\sim \mathrm{N}(x_{i,8}+x_{i,13},4),\\
x_{i,15} &= I(x^*_{i,15}>8)\\
x^*_{i,16} &\sim \mathrm{N}(5+x_{i,14},9),\\
x_{i,16} &= I(x^*_{i,16}>16)\\
x_{i,17} &\quad \text{simulated from a discrete distribution taking values} \in \{1,2\}\\
&\quad \text{with probabilities 0.45, and 0.55 respectively.}\\
x^*_{i,18} &\sim \mathrm{N}(2+x_{i,13}-x_{i,15},9),\\
x_{i,18} &= j^{18}\ I(c^{(18)}_{j^{18}-1}<x^*_{i,18}<c^{(18)}_{j^{18}}),\quad j^{18}=1,\ldots,3\\
x^*_{i,19} &\sim \mathrm{N}(3+2x_{i,14}-x_{i,16},1),\\
x_{i,19} &= j^{19}\ I(c^{(19)}_{j^{19}-1}<x^*_{i,19}<c^{(19)}_{j^{19}}),\quad j^{19}=1,\ldots,5\\
x_{i,20} &\quad \text{simulated from a discrete distribution taking values} \in \{1,2,3,4\}\\
&\quad \text{with probabilities 0.25 each.}\\
x_{i,21} &\sim \mathrm{N}(0,1)\\
x_{i,22} &\sim \mathrm{N}(7,2)\\
x_{i,23} &\sim \mathrm{N}(4,1)\\
x_{i,24} &\sim \mathrm{N}(13,3)\\
x_{i,25} &\sim \mathrm{N}(5+I(x_{i,1}=2)-2*I(x_{i,1}=3)+x_{i,7}+2x_{i,13},4)
\end{aligned}
$$

where

$$
p(x_{i,5}=j^5|x_{i,1},\boldsymbol{\theta}^{(5)})=\frac{\exp(\theta^{(5)}_{0,j^5}+\theta^{(5)}_{1,j^5}I(x_{i,1}=2)+\theta^{(5)}_{2,j^5}I(x_{i,1}=3))}{\sum_{s=1}^{5}\exp(\theta^{(5)}_{0,s}+\theta^{(5)}_{1,s}I(x_{i,1}=2)+\theta^{(5)}_{2,s}I(x_{i,1}=3))}
$$

where $j^5\in\{1,\ldots,4\}$ and $\theta^{(5)}_{0,1}=\theta^{(5)}_{1,1}=\theta^{(5)}_{2,1}=0$ for identifiability, $\theta^{(5)}_{0,2}=1,\theta^{(5)}_{1,2}=-3,\theta^{(5)}_{2,2}=-1.5,\theta^{(5)}_{0,3}=-1,\theta^{(5)}_{1,3}=2,\theta^{(5)}_{2,3}=1,\theta^{(5)}_{0,4}=-0.5,\theta^{(5)}_{1,4}=2,\theta^{(5)}_{2,4}=-2$, and $I(\cdot)$ is the indicator function. The threshold parameters $c^{(11)}_0=-\infty,c^{(11)}_1=0.2017,c^{(11)}_2=2.225,c^{(11)}_3=\infty$, $c^{(18)}_0=-\infty,c^{(18)}_1=4.366,c^{(18)}_2=6.511,c^{(18)}_3=\infty$ and $c^{(19)}_0=-\infty,c^{(19)}_1=19.14,c^{(19)}_2=22.52,c^{(19)}_3=24.53,c^{(19)}_4=27.60,c^{(19)}_5=\infty$.

We then introduce missing values into variables $\mathbf{x}_5,\mathbf{x}_9,\mathbf{x}_{11},\mathbf{x}_{13},\mathbf{x}_{15},\mathbf{x}_{16},\mathbf{x}_{17},\mathbf{x}_{18},\mathbf{x}_{25}$ in the

following way:

$$
\begin{aligned}
p(m_{i,5}=1) &= \left\{\frac{\exp(f(x_{i,2}))}{1+\exp(f(x_{i,2}))}\right\}(1-m_{i,2}) \text{ (MAR), where} \\
f(x_{i,2}) &= -1+I(x_{i,2}=2)-4*I(x_{i,2}=3)-5*I(x_{i,2}=4)+3*I(x_{i,2}=5), \\
p(m_{i,9}=1) &= \left\{\frac{\exp(-1-5x_{i,7})}{1+\exp(-1-5x_{i,7})}\right\}(1-m_{i,7}) \text{ (MAR)}, \\
p(m_{i,11}=1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,13}=1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,15}=1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,16}=1) &= 0.3 \text{ (MCAR)}, \\
p(m_{i,17}=1) &= \left\{\frac{\exp(-1+2x_{i,6})}{1+\exp(-1+2x_{i,6})}\right\}(1-m_{i,6}) \text{ (MAR)}, \\
p(m_{i,18}=1) &= \left\{\frac{\exp(2-8x_{i,8})}{1+\exp(2-8x_{i,8})}\right\}(1-m_{i,8}) \text{ (MAR)}, \\
p(m_{i,25}=1) &= \left\{\frac{\exp(1-5*I(x_{i,1}=2)-4*I(x_{i,1}=3))}{1+\exp(1-5*I(x_{i,1}=2)-4*I(x_{i,1}=3))}\right\}(1-m_{i,1}) \text{ (MAR)},
\end{aligned}
$$

where $m_{i,j}$ be the missing data indicator for $x_{i,j}$, where $m_{i,j}=1$ indicates $x_{i,j}$ is missing and $m_{i,j}=0$ indicates $x_{i,j}$ is observed.

Analysis

We obtain the following estimates from an analysis of the imputed datasets:

- Estimates of the means of $\mathbf{x}_6, \mathbf{x}_{13}$.
- Estimates of the proportion of units with $\mathbf{x}_{11}=1,2,3$ and $\mathbf{x}_{15}=1$
- Estimates of the regression coefficients of the linear regression model of $\mathbf{x}_{25}$ on all other variables.

Appendix G - "P-steps" for MFRM

In this section, we present the details on deriving the full conditional distributions of the parameters in the "P-steps" mentioned in the MFRM strategy in Chapter 5. The full conditional distribution of $\boldsymbol{\theta}^{(1)}$ where $\mathbf{x}_1$ follows a multinomial distribution and Metropolis-Hastings update for the multinomial logistic regression model have been presented in the MFRM section in Chapter 5. Here, we will present the details on deriving the full conditional distributions of the remaining parameters in the "P-steps", such as $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q)$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is continuous, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)})$ for

$\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is binary) and $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)})$ for $\mathbf{x}_q^*$ when $\mathbf{x}_q^*$ is a latent variable ($\mathbf{x}_q$ is ordinal).

For $x_{i,q}^*$ is continuous, we use the following representation for each of the sequential regression model $p(x_{i,q}^*|x_{i,1}^*, \ldots, x_{i,q-1}^*, \boldsymbol{\theta}^{(q)})$:

$$\begin{aligned} x_{i,q}^* &= \beta_0^{(q)} + \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)} + \epsilon_{i,q}, \\ \epsilon_{i,q} &\sim \mathrm{N}(0, \frac{1}{\tau_q \phi_{i,q}}), \end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$, for $i = 1, \ldots, n$ and $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$. We then propose the following prior for $\boldsymbol{\beta}^{(q)}$ (Park and Casella (2008)):

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \psi_q) &\sim \mathrm{MVN}(\mathbf{0}, \frac{1}{\psi_q}\mathbf{D}_\eta^{(q)}), \text{ with } \mathbf{D}_\eta^{(q)} = \mathrm{diag}(\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2), \\ p(\eta_{j,q}^2|\lambda_q) &= \frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right), \ j = 1, \ldots, q-1, \end{aligned}$$

and we consider a diffuse hyperprior on λ_q^2 of the form

$$p(\lambda_q^2) = \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right), \ \lambda_q^2 > 0, \ r > 0, \ \delta > 0,$$

which is a gamma distribution with shape and rate r and δ respectively. We also specify the prior distributions for $\psi_q, \beta_0^{(q)}, \tau_q, \Phi_q$ as follows:

$$\begin{aligned} p(\psi_q) &\propto \frac{1}{\psi_q}, \\ p(\beta_0^{(q)}) &\propto 1, \\ p(\tau_q) &\propto \frac{1}{\tau_q}, \\ p(\phi_{i,q}) &\sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}), \ i = 1, \ldots, n. \end{aligned}$$

The joint posterior distribution of the parameters in this model is given up to a normalising

constant by:

$$
\begin{aligned}
& \prod_{i=1}^{n} \left\{ p(x_{i,q}^{*} | \beta_0^{(q)}, \boldsymbol{\beta}^{(q)}, \tau_q, \phi_{i,q}) p(\phi_{i,q}) \right\} \\
& \prod_{j=1}^{q-1} \left\{ p(\beta_j^{(q)} | \psi_q, \eta_{j,q}^2) p(\eta_{j,q}^2 | \lambda_q^2) \right\} p(\tau_q) p(\psi_q) p(\lambda_q^2) p(\beta_0^{(q)}) \\
= & \prod_{i=1}^{n} \left\{ \sqrt{\frac{\tau_q \phi_{i,q}}{2\pi}} \exp\left[-\frac{\tau_q \phi_{i,q}}{2} (\widetilde{x}_{i,q}^{*} - \widetilde{\mathbf{x}}_{i,q} \boldsymbol{\beta}^{(q)})^2 \right] \frac{1}{\Gamma(\frac{\nu}{2})} (\frac{\nu}{2})^{(\frac{\nu}{2})} \phi_{i,q}^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu \phi_{i,q}}{2} \right] \right\} \\
& \prod_{j=1}^{q-1} \left\{ \sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}} \exp\left[-\frac{\psi_q}{2\eta_{j,q}^2} (\beta_j^{(q)})^2 \right] \frac{\lambda_q^2}{2} \exp\left[\frac{-\lambda_q^2 \eta_{j,q}^2}{2} \right] \right\} \frac{1}{\psi_q} \frac{1}{\tau_q} \frac{\delta^r}{\Gamma(r)} (\lambda_q^2)^{r-1} \exp\left(-\delta \lambda_q^2 \right),
\end{aligned}
$$

where $\widetilde{x}_{i,q}^{*} = x_{i,q}^{*} - \beta_0^{(q)}$.

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \tau_q, \Phi_q$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is proportional to

$$
\begin{aligned}
& \prod_{i=1}^{n} \sqrt{\frac{\tau_q \phi_{i,q}}{2\pi}} \exp\left[-\frac{\tau_q \phi_{i,q}}{2} (\widetilde{x}_{i,q}^{*} - \widetilde{\mathbf{x}}_{i,q} \boldsymbol{\beta}^{(q)})^2 \right] \prod_{j=1}^{q-1} \sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}} \exp\left[-\frac{\psi_q}{2\eta_{j,q}^2} (\beta_j^{(q)})^2 \right] \\
\propto & \exp\left\{ -\frac{1}{2} \left[(\widetilde{\mathbf{x}}_q^{*} - \widetilde{\mathbf{X}}_q \boldsymbol{\beta}^{(q)})' \tau_q \mathbf{D}_{\phi}^{(q)} (\widetilde{\mathbf{x}}_q^{*} - \widetilde{\mathbf{X}}_q \boldsymbol{\beta}^{(q)}) + (\boldsymbol{\beta}^{(q)})' \mathbf{D}_{\psi,\eta}^{(q)} \boldsymbol{\beta}^{(q)} \right] \right\} \\
\propto & \exp\left\{ -\frac{1}{2} \left[(\widetilde{\mathbf{X}}_q \hat{\boldsymbol{\beta}}^{(q)} - \widetilde{\mathbf{X}}_q \boldsymbol{\beta}^{(q)})' \tau_q \mathbf{D}_{\phi}^{(q)} (\widetilde{\mathbf{X}}_q \hat{\boldsymbol{\beta}}^{(q)} - \widetilde{\mathbf{X}}_q \boldsymbol{\beta}^{(q)}) + (\boldsymbol{\beta}^{(q)})' \mathbf{D}_{\psi,\eta}^{(q)} \boldsymbol{\beta}^{(q)} \right] \right\} \\
= & \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^{(q)})' \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q (\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^{(q)}) + (\boldsymbol{\beta}^{(q)})' \mathbf{D}_{\psi,\eta}^{(q)} \boldsymbol{\beta}^{(q)} \right] \right\} \\
\propto & \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}^{(q)})' (\widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)}) \boldsymbol{\beta}^{(q)} - 2 (\boldsymbol{\beta}^{(q)})' \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q \hat{\boldsymbol{\beta}}^{(q)} \right] \right\}
\end{aligned}
$$

Let $A_q = (\widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)})$ and $\hat{\boldsymbol{\beta}}^{(q)} = (\widetilde{\mathbf{X}}_q' \widetilde{\mathbf{X}}_q)^{-1} \widetilde{\mathbf{X}}_q' \widetilde{\mathbf{x}}_q^{*}$, then

$$
\begin{aligned}
& \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}^{(q)})' (\widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)}) \boldsymbol{\beta}^{(q)} - 2 (\boldsymbol{\beta}^{(q)})' \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q \hat{\boldsymbol{\beta}}^{(q)} \right] \right\} \\
= & \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}^{(q)})' A_q \boldsymbol{\beta}^{(q)} - 2 (\boldsymbol{\beta}^{(q)})' \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q (\widetilde{\mathbf{X}}_q' \widetilde{\mathbf{X}}_q)^{-1} \widetilde{\mathbf{X}}_q' \widetilde{\mathbf{x}}_q^{*} \right] \right\} \\
= & \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}^{(q)})' A_q \boldsymbol{\beta}^{(q)} - 2 (\boldsymbol{\beta}^{(q)})' A_q A_q^{-1} \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{x}}_q^{*} \right] \right\} \\
\propto & \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}^{(q)} - A_q^{-1} \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{x}}_q^{*})' A_q (\boldsymbol{\beta}^{(q)} - A_q^{-1} \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{x}}_q^{*}) \right] \right\} \\
\Rightarrow & \ \boldsymbol{\beta}^{(q)} \sim \text{MVN}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}) \text{ with} \\
& \ \hat{\boldsymbol{\beta}}^{(q)} = \mathbf{V}^{(q)} \widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{x}}_q^{*} \text{ and } \mathbf{V}^{(q)} = (\widetilde{\mathbf{X}}_q' \tau_q \mathbf{D}_{\phi}^{(q)} \widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)})^{-1}.
\end{aligned}
$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*, \ldots, \widetilde{x}_{n,q}^*)'$ where $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n \times n$ diagonal matrix with entries $\phi_{i,q}$ for $i = 1, \ldots, n$ and $\mathbf{D}_{\psi,\eta}^{(q)}$ is a $(q-1) \times (q-1)$ diagonal matrix with entries $\frac{\psi_q}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

2. The full conditional distribution of $\eta_{j,q}^2,\ j = 1, \ldots, q-1$ is proportional to

$$\prod_{j=1}^{q-1} \left\{ \sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}} \exp\left[-\frac{\psi_q}{2\eta_{j,q}^2} (\beta_j^{(q)})^2 \right] \frac{\lambda_q^2}{2} \exp\left[\frac{-\lambda_q^2 \eta_{j,q}^2}{2} \right] \right\}.$$

Let $g = \frac{1}{\eta_{j,q}^2} \Rightarrow \left| \frac{d\eta_{j,q}^2}{dg} \right| = \frac{1}{g^2}$. Hence

$$\begin{aligned}
& \prod_{j=1}^{q-1} \left\{ \sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}} \exp\left[-\frac{\psi_q}{2\eta_{j,q}^2} (\beta_j^{(q)})^2 \right] \frac{\lambda_q^2}{2} \exp\left[\frac{-\lambda_q^2 \eta_{j,q}^2}{2} \right] \right\} \\
\propto & \prod_{j=1}^{q-1} g^{\frac{1}{2}} g^{-2} \exp\left\{ -\frac{g\psi_q}{2} (\beta_j^{(q)})^2 - \frac{\lambda_q^2}{2g} \right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \left[g\psi_q (\beta_j^{(q)})^2 - \frac{\lambda_q^2}{g} \right] \right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \frac{\lambda_q^2 (\beta_j^{(q)})^2 \psi_q}{g\lambda_q^2} \left(g^2 - \frac{\lambda_q^2}{(\beta_j^{(q)})^2 \psi_q} \right) \right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \frac{\lambda_q^2}{(\mu_q')^2 g} \left[g^2 - (\mu_q')^2 \right] \right\}, \mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2 \psi_q}} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \frac{\lambda_q^2}{(\mu_q')^2 g} \left[(g - \mu_q')^2 + 2g\mu_q' \right] \right\} \\
\propto & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{ -\frac{1}{2} \frac{\lambda_q^2}{(\mu_q')^2 g} (g - \mu_q')^2 \right\} \\
\Rightarrow & \ g \sim \sqrt{\frac{\lambda_q'}{2\pi}} g^{-\frac{3}{2}} \exp\left\{ -\frac{\lambda_q' (g - \mu_q')^2}{2(\mu_q')^2 g} \right\},\ g > 0,
\end{aligned}$$

which is a inverse-Gaussian distribution with parameters

$$\mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2 \psi_q}} \ \text{ and } \ \lambda_q' = \lambda_q^2, \quad j = 1, \ldots, q-1.$$

3. The full conditional distribution of λ_q^2 is proportional to

$$\begin{aligned}
& \prod_{j=1}^{q-1}\left[\frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right)\right]\frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right) \\
\propto\ & \prod_{j=1}^{q-1}\left[\frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right)\right](\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right) \\
=\ & \left(\frac{\lambda_q^2}{2}\right)^{q-1}(\lambda_q^2)^{r-1}\exp\left[-\lambda_q^2\left(\frac{\sum_{j=1}^{q-1}\eta_{j,q}^2}{2}\right)\right]\exp\left(-\delta\lambda_q^2\right) \\
\propto\ & \left(\lambda_q^2\right)^{(q-1)+r-1}\exp\left[-\lambda_q^2\left(\sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2}+\delta\right)\right] \\
\Rightarrow\ & \lambda_q^2 \sim \text{Gamma}\left((q-1)+r, \sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2}+\delta\right).
\end{aligned}$$

4. The full conditional distribution of ψ_q is proportional to

$$\begin{aligned}
& \prod_{j=1}^{q-1}\left\{\sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\psi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\right\}\frac{1}{\psi_q} \\
\propto\ & \psi_q^{\frac{q-1}{2}}\psi_q^{-1}\exp\left[-\psi_q\left(\frac{(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right)\right] \\
\Rightarrow\ & \psi_q \sim \text{Gamma}\left(\frac{q-1}{2}, \frac{(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right).
\end{aligned}$$

where $\mathbf{D}_\eta^{(q)}$ is a $(q-1)\times(q-1)$ diagonal matrix with entries $\frac{1}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

5. The full conditional distribution of $\beta_0^{(q)}$ is proportional to

$$
\begin{aligned}
& \prod_{i=1}^{n}\left\{\sqrt{\frac{\tau_q\phi_{i,q}}{2\pi}}\exp\left[-\frac{\tau_q\phi_{i,q}}{2}(x_{i,q}^*-\beta_0^{(q)}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\} \\
\propto\ & \prod_{i=1}^{n}\left\{\exp\left[-\frac{\tau_q\phi_{i,q}}{2}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)}-\beta_0^{(q)})^2\right]\right\} \\
=\ & \prod_{i=1}^{n}\left\{\exp\left[-\frac{\tau_q\phi_{i,q}}{2}\left((x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2-2(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}+(\beta_0^{(q)})^2\right)\right]\right\} \\
\propto\ & \prod_{i=1}^{n}\left\{\exp\left[-\frac{\tau_q\phi_{i,q}}{2}\left((\beta_0^{(q)})^2-2(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right]\right\} \\
=\ & \exp\left[-\frac{\tau_q}{2}\sum_{i=1}^{n}\left(\phi_{i,q}(\beta_0^{(q)})^2-2\phi_{i,q}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right] \\
=\ & \exp\left[-\frac{\tau_q}{2}\left(\sum_{i=1}^{n}\phi_{i,q}(\beta_0^{(q)})^2-2\sum_{i=1}^{n}\phi_{i,q}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right] \\
=\ & \exp\left[-\frac{\tau_q}{2}\left((\beta_0^{(q)})^2\sum_{i=1}^{n}\phi_{i,q}-2\beta_0^{(q)}\sum_{i=1}^{n}\phi_{i,q}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\right)\right] \\
=\ & \exp\left[-\frac{\tau_q\sum_{i=1}^{n}\phi_{i,q}}{2}\left((\beta_0^{(q)})^2-2\beta_0^{(q)}\frac{\sum_{i=1}^{n}\phi_{i,q}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{\sum_{i=1}^{n}\phi_{i,q}}\right)\right] \\
\propto\ & \exp\left[-\frac{\tau_q\sum_{i=1}^{n}\phi_{i,q}}{2}\left(\beta_0^{(q)}-\frac{\sum_{i=1}^{n}\phi_{i,q}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{\sum_{i=1}^{n}\phi_{i,q}}\right)^2\right] \\
\Rightarrow\ & \beta_0^{(q)}\sim \mathrm{N}(\frac{\sum_{i=1}^{n}\phi_{i,q}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{\sum_{i=1}^{n}\phi_{i,q}},(\tau_q\sum_{i=1}^{n}\phi_{i,q})^{-1}),
\end{aligned}
$$

where
$\widetilde{\mathbf{x}}_{i,q}=(I(x_{i,1}^*=2),\ldots,I(x_{i,1}^*=L_1),\ldots,I(x_{i,k}^*=2),\ldots,I(x_{i,k}^*=L_k),x_{i,k+1}^*,\ldots,x_{i,q-1}^*)$ for $i=1,\ldots,n$.

6. The full conditional distribution of τ_q is proportional to

$$
\begin{aligned}
& \prod_{i=1}^{n}\left\{\sqrt{\frac{\tau_q\phi_{i,q}}{2\pi}}\exp\left[-\frac{\tau_q\phi_{i,q}}{2}(\widetilde{x}_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\}\frac{1}{\tau_q} \\
\propto\ & \tau_q^{\frac{n}{2}}\tau_q^{-1}\prod_{i=1}^{n}\exp\left[-\frac{\tau_q\phi_{i,q}}{2}(\widetilde{x}_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right] \\
=\ & \tau_q^{\frac{n}{2}-1}\exp\left[-\frac{\tau_q}{2}\sum_{i=1}^{n}\phi_{i,q}(\widetilde{x}_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right] \\
=\ & \tau_q^{\frac{n}{2}-1}\exp\left\{-\frac{\tau_q}{2}\left[(\widetilde{\mathbf{x}}_q^*-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'\mathbf{D}_\phi^{(q)}(\widetilde{\mathbf{x}}_q^*-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})\right]\right\} \\
\Rightarrow\ & \tau_q\sim \mathrm{Gamma}\left(\frac{n}{2},\frac{(\widetilde{\mathbf{x}}_q^*-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'\mathbf{D}_\phi^{(q)}(\widetilde{\mathbf{x}}_q^*-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})}{2}\right),
\end{aligned}
$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*, \ldots, \widetilde{x}_{n,q}^*)'$ with $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n \times n$ diagonal matrix with entries $\phi_{i,q}$ for $i = 1, \ldots, n$.

7. The full conditional distribution of $\phi_{i,q}$ is proportional to

$$\begin{aligned}
& \prod_{i=1}^{n} \left\{ \sqrt{\frac{\tau_q \phi_{i,q}}{2\pi}} \exp\left[-\frac{\tau_q \phi_{i,q}}{2} (\widetilde{x}_{i,q}^* - \widetilde{\mathbf{x}}_{i,q} \boldsymbol{\beta}^{(q)})^2 \right] \frac{1}{\Gamma(\frac{\nu}{2})} (\frac{\nu}{2})^{(\frac{\nu}{2})} \phi_{i,q}^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu \phi_{i,q}}{2} \right] \right\} \\
\propto & \prod_{i=1}^{n} \phi_{i,q}^{\frac{1}{2}} \phi_{i,q}^{\frac{\nu}{2}-1} \exp\left[-\phi_{i,q} \left(\frac{\tau_q}{2} (\widetilde{x}_{i,q}^* - \widetilde{\mathbf{x}}_{i,q} \boldsymbol{\beta}^{(q)})^2 + \frac{\nu}{2} \right) \right] \\
\Rightarrow & \ \phi_{i,q} \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\tau_q}{2} (\widetilde{x}_{i,q}^* - \widetilde{\mathbf{x}}_{i,q} \boldsymbol{\beta}^{(q)})^2 + \frac{\nu}{2} \right),
\end{aligned}$$

where $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and
$\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$.

For $x_{i,q}^*$ is a latent variable ($x_{i,q}$ is binary), we use the following representation for each of the sequential regression model $p(x_{i,q}^* | x_{i,1}^*, \ldots, x_{i,q-1}^*, \boldsymbol{\theta}^{(q)})$:

$$\begin{aligned}
x_{i,q}^* &= \beta_0^{(q)} + \widetilde{\mathbf{x}}_{i,q} \boldsymbol{\beta}^{(q)} + \epsilon_{i,q}, \text{ where } \ x_{i,q} = I(x_{i,q}^* > 0) \\
\epsilon_{i,q} &\sim \text{N}(0, 1),
\end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$, for $i = 1, \ldots, n$ and $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$. We then propose the following priors for $\boldsymbol{\beta}^{(q)}$ (Park and Casella (2008)):

$$\begin{aligned}
p(\boldsymbol{\beta}^{(q)} | \boldsymbol{\eta}_q, \psi_q) &\sim \text{MVN}(\mathbf{0}, \frac{1}{\psi_q} \mathbf{D}_\eta^{(q)}), \text{ with } \mathbf{D}_\eta^{(q)} = \text{diag}(\eta_{1,q}^2, \ldots, \eta_{q-1,q}^2), \\
p(\eta_{j,q}^2 | \lambda_q) &= \frac{\lambda_q^2}{2} \exp\left(\frac{-\lambda_q^2 \eta_{j,q}^2}{2} \right), \ j = 1, \ldots, q-1,
\end{aligned}$$

and we consider a diffuse hyperprior on λ_q^2 of the form

$$p(\lambda_q^2) = \frac{\delta^r}{\Gamma(r)} (\lambda_q^2)^{r-1} \exp\left(-\delta \lambda_q^2\right), \ \lambda_q^2 > 0, \ r > 0, \ \delta > 0,$$

which is a gamma distribution with shape and rate r and δ respectively. We also specify the prior distributions for ψ_q and $\beta_0^{(q)}$ as follows:

$$\begin{aligned}
p(\psi_q) &\propto \frac{1}{\psi_q}, \\
p(\beta_0^{(q)}) &\propto 1.
\end{aligned}$$

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$
\begin{aligned}
& \prod_{i=1}^{n}\left\{p(x_{i,q}^{*}|\beta_0^{(q)},\boldsymbol{\beta}^{(q)})\right\}\prod_{j=1}^{q-1}\left\{p(\beta_j^{(q)}|\psi_q,\eta_{j,q}^2)p(\eta_{j,q}^2|\lambda_q^2)\right\}p(\psi_q)p(\lambda_q^2)p(\beta_0^{(q)}) \\
= & \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(\widetilde{x}_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\} \\
& \prod_{j=1}^{q-1}\left\{\sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\psi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\frac{\lambda_q^2}{2}\exp\left[\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right]\right\}\frac{1}{\psi_q}\frac{1}{\tau_q} \\
& \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right),
\end{aligned}
$$

where $\widetilde{x}_{i,q}^{*}=x_{i,q}^{*}-\beta_0^{(q)}$.

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)},\boldsymbol{\eta}_q,\lambda_q^2,\psi_q,\beta_0^{(q)}$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is proportional to

$$
\begin{aligned}
& \prod_{i=1}^{n}\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(\widetilde{x}_{i,q}^{*}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\prod_{j=1}^{q-1}\sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\psi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right] \\
\propto & \exp\left\{-\frac{1}{2}\left[(\widetilde{\mathbf{x}}_q^{*}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{x}}_q^{*}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_{\psi,\eta}^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\} \\
\propto & \exp\left\{-\frac{1}{2}\left[(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}-\widetilde{\mathbf{X}}_q\boldsymbol{\beta}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_{\psi,\eta}^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\} \\
= & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q(\boldsymbol{\beta}^{(q)}-\hat{\boldsymbol{\beta}}^{(q)})+(\boldsymbol{\beta}^{(q)})'\mathbf{D}_{\psi,\eta}^{(q)}\boldsymbol{\beta}^{(q)}\right]\right\} \\
\propto & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q+\mathbf{D}_{\psi,\eta}^{(q)})\boldsymbol{\beta}^{(q)}-2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\},
\end{aligned}
$$

Let $A_q = (\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)})$ and $\hat{\boldsymbol{\beta}}^{(q)} = (\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^*$, then

$$\begin{aligned}
& \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)})\boldsymbol{\beta}^{(q)} - 2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q\hat{\boldsymbol{\beta}}^{(q)}\right]\right\} \\
= \quad & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)} - 2(\boldsymbol{\beta}^{(q)})'\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q(\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q)^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^*\right]\right\} \\
= \quad & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)})'A_q\boldsymbol{\beta}^{(q)} - 2(\boldsymbol{\beta}^{(q)})'A_qA_q^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^*\right]\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}^{(q)} - A_q^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^*)'A_q(\boldsymbol{\beta}^{(q)} - A_q^{-1}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^*)\right]\right\} \\
\Rightarrow \quad & \boldsymbol{\beta}^{(q)} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}^{(q)}, \mathbf{V}^{(q)}) \text{ with} \\
& \hat{\boldsymbol{\beta}}^{(q)} = \mathbf{V}^{(q)}\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{x}}_q^* \text{ and } \mathbf{V}^{(q)} = (\widetilde{\mathbf{X}}_q'\widetilde{\mathbf{X}}_q + \mathbf{D}_{\psi,\eta}^{(q)})^{-1}, \qquad (7.8)
\end{aligned}$$

where $\widetilde{\mathbf{x}}_q^* = (\widetilde{x}_{1,q}^*, \ldots, \widetilde{x}_{n,q}^*)'$ with $\widetilde{x}_{i,q}^* = x_{i,q}^* - \beta_0^{(q)}$ and $\widetilde{\mathbf{X}}_q = (\widetilde{\mathbf{x}}_{1,q}, \ldots, \widetilde{\mathbf{x}}_{n,q})'$, where $\widetilde{\mathbf{x}}_{i,q} = (I(x_{i,1}^* = 2), \ldots, I(x_{i,1}^* = L_1), \ldots, I(x_{i,k}^* = 2), \ldots, I(x_{i,k}^* = L_k), x_{i,k+1}^*, \ldots, x_{i,q-1}^*)$ for $i = 1, \ldots, n$. The matrix $\mathbf{D}_\phi^{(q)}$ is a $n \times n$ diagonal matrix with entries $\phi_{i,q}$ for $i = 1, \ldots, n$ and $\mathbf{D}_{\psi,\eta}^{(q)}$ is a $(q-1) \times (q-1)$ diagonal matrix with entries $\frac{\psi_q}{\eta_{j,q}^2}$ for $j = 1, \ldots, q-1$.

2. The full conditional distribution of $\eta_{j,q}^2,\ j = 1, \ldots, q-1$ is proportional to

$$\prod_{j=1}^{q-1}\left\{\sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\psi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\frac{\lambda_q^2}{2}\exp\left[\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right]\right\}.$$

Let $g = \frac{1}{\eta_{j,q}^2} \Rightarrow \left|\frac{d\eta_{j,q}^2}{dg}\right| = \frac{1}{g^2}$. Hence

$$\begin{aligned}
& \prod_{j=1}^{q-1} \left\{ \sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}} \exp\left[-\frac{\psi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right] \frac{\lambda_q^2}{2} \exp\left[\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right] \right\} \\
\propto & \prod_{j=1}^{q-1} g^{\frac{1}{2}} g^{-2} \exp\left\{-\frac{g\psi_q}{2}(\beta_j^{(q)})^2 - \frac{\lambda_q^2}{2g}\right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\left[g\psi_q(\beta_j^{(q)})^2 - \frac{\lambda_q^2}{g}\right]\right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2(\beta_j^{(q)})^2\psi_q}{g\lambda_q^2}\left(g^2 - \frac{\lambda_q^2}{(\beta_j^{(q)})^2\psi_q}\right)\right\} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2}{(\mu_q')^2 g}\left[g^2 - (\mu_q')^2\right]\right\}, \mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2\psi_q}} \\
= & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2}{(\mu_q')^2 g}\left[(g-\mu_q')^2 + 2g\mu_q'\right]\right\} \\
\propto & \prod_{j=1}^{q-1} g^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda_q^2}{(\mu_q')^2 g}(g-\mu_q')^2\right\} \\
\Rightarrow & \; g \sim \sqrt{\frac{\lambda_q'}{2\pi}} g^{-\frac{3}{2}} \exp\left\{-\frac{\lambda_q'(g-\mu_q')^2}{2(\mu_q')^2 g}\right\}, \; g > 0, \qquad (7.9)
\end{aligned}$$

which is a inverse-Gaussian distribution with parameters

$$\mu_q' = \sqrt{\frac{\lambda_q^2}{(\beta_j^{(q)})^2\psi_q}} \;\text{ and }\; \lambda_q' = \lambda_q^2, \;\; j = 1, \ldots, q-1.$$

3. The full conditional distribution of λ_q^2 is proportional to

$$\begin{aligned}
& \prod_{j=1}^{q-1} \left[\frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right)\right] \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right) \\
\propto & \prod_{j=1}^{q-1} \left[\frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta_{j,q}^2}{2}\right)\right] (\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right) \\
= & \left(\frac{\lambda_q^2}{2}\right)^{q-1} (\lambda_q^2)^{r-1}\exp\left[-\lambda_q^2\left(\frac{\sum_{j=1}^{q-1}\eta_{j,q}^2}{2}\right)\right]\exp\left(-\delta\lambda_q^2\right) \\
\propto & \left(\lambda_q^2\right)^{(q-1)+r-1}\exp\left[-\lambda_q^2\left(\sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2} + \delta\right)\right] \\
\Rightarrow & \; \lambda_q^2 \sim \text{Gamma}\left((q-1)+r, \sum_{j=1}^{q-1}\frac{\eta_{j,q}^2}{2} + \delta\right). \qquad (7.10)
\end{aligned}$$

4. The full conditional distribution of ψ_q is proportional to

$$\begin{aligned}
& \prod_{j=1}^{q-1}\left\{\sqrt{\frac{\psi_q}{2\pi\eta_{j,q}^2}}\exp\left[-\frac{\psi_q}{2\eta_{j,q}^2}(\beta_j^{(q)})^2\right]\right\}\frac{1}{\psi_q} \\
\propto\ & \psi_q^{\frac{q-1}{2}}\psi_q^{-1}\exp\left[-\psi_q\left(\frac{(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right)\right] \\
\Rightarrow\ & \psi_q \sim \text{Gamma}\left(\frac{q-1}{2}, \frac{(\boldsymbol{\beta}^{(q)})'\mathbf{D}_\eta^{(q)}\boldsymbol{\beta}^{(q)}}{2}\right). \qquad (7.11)
\end{aligned}$$

where $\mathbf{D}_\eta^{(q)}$ is a $(q-1)\times(q-1)$ diagonal matrix with entries $\frac{1}{\eta_{j,q}^2}$ for $j=1,\ldots,q-1$.

5. The full conditional distribution of $\beta_0^{(q)}$ is proportional to

$$\begin{aligned}
& \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(x_{i,q}^*-\beta_0^{(q)}-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\right\} \\
=\ & \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)}-\beta_0^{(q)})^2\right]\right\} \\
=\ & \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}\left((x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2-2(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}+(\beta_0^{(q)})^2\right)\right]\right\} \\
\propto\ & \prod_{i=1}^{n}\left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}\left((\beta_0^{(q)})^2-2(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right]\right\} \\
\propto\ & \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left((\beta_0^{(q)})^2-2(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right] \\
=\ & \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}(\beta_0^{(q)})^2-2\sum_{i=1}^{n}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})\beta_0^{(q)}\right)\right] \\
=\ & \exp\left[-\frac{n}{2}\left((\beta_0^{(q)})^2-2\beta_0^{(q)}\frac{\sum_{i=1}^{n}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}\right)\right] \\
\propto\ & \exp\left[-\frac{n}{2}\left(\beta_0^{(q)}-\frac{\sum_{i=1}^{n}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}\right)^2\right] \\
\Rightarrow\ & \beta_0^{(q)} \sim \text{N}(\frac{\sum_{i=1}^{n}(x_{i,q}^*-\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})}{n}, \frac{1}{n}). \qquad (7.12)
\end{aligned}$$

For $x_{i,q}^*$ is a latent variable ($x_{i,q}$ is ordinal), we use the following representation for each of the sequential regression model $p(x_{i,q}^*|x_{i,1}^*,\ldots,x_{i,q-1}^*,\boldsymbol{\theta}^{(q)})$:

$$\begin{aligned}
x_{i,q}^* &= \beta_0^{(q)}+\widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)}+\epsilon_{i,q},\ \text{where } x_{i,q}=j^q \text{ if } \gamma_{j^q-1}^{(q)}<x_{i,q}^*<\gamma_{j^q}^{(q)}, \\
\epsilon_{i,q} &\sim \text{N}(0,1),
\end{aligned}$$

where $\widetilde{\mathbf{x}}_{i,q} =$
$(I(x^*_{i,1} = 2), \ldots, I(x^*_{i,1} = L_1), \ldots, I(x^*_{i,k} = 2), \ldots, I(x^*_{i,k} = L_k), x^*_{i,k+1}, \ldots, x^*_{i,q-1})$, for $i = 1, \ldots, n$ and $\boldsymbol{\beta}^{(q)} = (\beta_1^{(q),2}, \ldots, \beta_1^{(q),L_1}, \ldots, \beta_k^{(q),2}, \ldots, \beta_k^{(q),L_k}, \beta_{k+1}^{(q)}, \ldots, \beta_{q-1}^{(q)})'$. We then propose the following prior for $\boldsymbol{\beta}^{(q)}$ (Park and Casella (2008)):

$$\begin{aligned} p(\boldsymbol{\beta}^{(q)}|\boldsymbol{\eta}_q, \psi_q) &\sim \text{MVN}(\mathbf{0}, \frac{1}{\psi_q}\mathbf{D}_\eta^{(q)}), \text{ with } \mathbf{D}_\eta^{(q)} = \text{diag}(\eta^2_{1,q}, \ldots, \eta^2_{q-1,q}), \\ p(\eta^2_{j,q}|\lambda_q) &= \frac{\lambda_q^2}{2}\exp\left(\frac{-\lambda_q^2\eta^2_{j,q}}{2}\right), \ j = 1, \ldots, q-1, \end{aligned}$$

and we consider a diffuse hyperprior on λ_q^2 of the form

$$p(\lambda_q^2) = \frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left(-\delta\lambda_q^2\right), \ \lambda_q^2 > 0, \ r > 0, \ \delta > 0,$$

which is a gamma distribution with shape and rate r and δ respectively. We also specify the prior distributions for $\psi_q, \beta_0^{(q)}$ as follows:

$$\begin{aligned} p(\psi_q) &\propto \frac{1}{\psi_q}, \\ p(\beta_0^{(q)}) &\propto 1, \end{aligned}$$

and we place an improper uniform prior on $\boldsymbol{\gamma}^{(q)} = \left\{\gamma_{j^q}^{(q)} : j^q \in \{1, \ldots, J_q\}\right\}$ i.e.

$$p(\boldsymbol{\gamma}^{(q)}) \propto I(\boldsymbol{\gamma}^{(q)} \in \Omega^{(q)}),$$

where $\Omega^{(q)} = \left\{\gamma_{j^q}^{(q)} : \gamma_0^{(q)} = -\infty < \gamma_1^{(q)} = 0 < \gamma_2^{(q)} < \ldots < \gamma_{J_{q-1}}^{(q)} < \gamma_{J_q}^{(q)} = \infty\right\}$.

The joint posterior distribution of the parameters in this model is given up to a normalising constant by:

$$\begin{aligned} &\prod_{i=1}^n \left\{p(x^*_{i,q}|\beta_0^{(q)}, \boldsymbol{\beta}^{(q)}, \boldsymbol{\gamma}^{(q)})p(\boldsymbol{\gamma}^{(q)})\right\} \prod_{j=1}^{q-1}\left\{p(\beta_j^{(q)}|\psi_q, \eta^2_{j,q})p(\eta^2_{j,q}|\lambda_q^2)\right\} p(\psi_q)p(\lambda_q^2)p(\beta_0^{(q)}) \\ = &\prod_{i=1}^n \left\{\sqrt{\frac{1}{2\pi}}\exp\left[-\frac{1}{2}(\widetilde{x}^*_{i,q} - \widetilde{\mathbf{x}}_{i,q}\boldsymbol{\beta}^{(q)})^2\right]\left[\sum_{j^q=1}^{J_q} I(x_{i,q} = j^q)I(\gamma^{(q)}_{j^q-1} < x^*_{i,q} < \gamma^{(q)}_{j^q})\right]\right\} \\ &\prod_{j=1}^{q-1}\left\{\sqrt{\frac{\psi_q}{2\pi\eta^2_{j,q}}}\exp\left[-\frac{\psi_q}{2\eta^2_{j,q}}(\beta_j^{(q)})^2\right]\frac{\lambda_q^2}{2}\exp\left[\frac{-\lambda_q^2\eta^2_{j,q}}{2}\right]\right\}\frac{1}{\psi_q}\frac{1}{\tau_q}\frac{\delta^r}{\Gamma(r)}(\lambda_q^2)^{r-1}\exp\left[-\delta\lambda_q^2\right], \end{aligned}$$

where $\widetilde{x}^*_{i,q} = x^*_{i,q} - \beta_0^{(q)}$.

Hence, full conditional distributions of $\boldsymbol{\beta}^{(q)}, \boldsymbol{\eta}_q, \lambda_q^2, \psi_q, \beta_0^{(q)}, \boldsymbol{\gamma}^{(q)}$ can be derived in the following way:

1. The full conditional distribution of $\boldsymbol{\beta}^{(q)}$ is given by Equation 7.8.

2. The full conditional distribution of $\eta^2_{j,q},\ j = 1, \ldots, q-1$ is given by Equation 7.9.

3. The full conditional distribution of λ^2_q is given by Equation 7.10.

4. The full conditional distribution of ψ_q is given by Equation 7.11.

5. The full conditional distribution of $\beta_0^{(q)}$ is given by Equation 7.12.

6. The full conditional distribution of $\gamma_{j^q}^{(q)}$ for $j^q \in \{1, \ldots, J_q\}$ is proportional to

$$\prod_{i=1}^{n} \left[I(x_{i,q} = j^q) I(\gamma^{(q)}_{j^q-1} < x^*_{i,q} < \gamma^{(q)}_{j^q}) + I(x_{i,q} = j^q + 1) I(\gamma^{(q)}_{j^q} < x^*_{i,q} < \gamma^{(q)}_{j^q+1}) \right],$$

which is uniformly distributed on the interval

$$\left[\max \left\{ \max \left\{ x^*_{i,q} : x_{i,q} = j^q \right\}, \gamma^{(q)}_{j^q-1} \right\}, \min \left\{ \min \left\{ x^*_{i,q} : x_{i,q} = j^q + 1 \right\}, \gamma^{(q)}_{j^q+1} \right\} \right],$$

where $\gamma_0^{(q)} = -\infty$, $\gamma_1^{(q)} = 0$ and $\gamma_{J_q}^{(q)} = \infty$, for $j^q \in \{1, \ldots, J_q\}$.

Appendix H - Simulation study in Section 5.3.1 of Chapter 5

In this section we present details of the simulation study in Section 5.3.1 of Chapter 5. We first describe how we simulate an incomplete data set, we then describe the analyses we considered.

Data generation

We simulate $x_{i,1}, x_{i,2}, \ldots, x_{i,25},\ i = 1, \ldots, 1000$ in the following way:

$x_{i,1}$ simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.3, 0.4 and 0.3 respectively.

$x_{i,2}$ simulated from a discrete distribution taking values $\in \{1, 2, 3, 4, 5\}$ with each probability of 0.2.

$x_{i,3}$ simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.2, 0.2 and 0.6 respectively.

$x_{i,4}$ simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.25, 0.35 and 0.4 respectively.

$$x_{i,5} \sim p(x_{i,5}|x_{i,1}, \boldsymbol{\theta}^{(5)}),$$

$$x_{i,6} \sim \mathrm{N}(0, \frac{1}{\epsilon_{i,6}}),\ \epsilon_{i,6} \sim \mathrm{Gamma}(2, 2)$$

$$x_{i,7} \sim \mathrm{N}(1 + 2 * I(x_{i,5} = 2) - I(x_{i,5} = 3) + I(x_{i,5} = 4), \frac{2}{\epsilon_{i,7}}),\ \epsilon_{i,7} \sim \mathrm{Gamma}(4, 4)$$

$$x_{i,8} \sim \mathrm{N}(3, 1)$$

$x_{i,9}$ simulated from a discrete distribution taking values $\in \{1, 2\}$ with probabilities 0.5, and 0.5 respectively.

$x_{i,10}$ simulated from a discrete distribution taking values $\in \{1, 2\}$ with probabilities 0.3, and 0.7 respectively.

$$x^*_{i,11} \sim \mathrm{N}(3 - x_{i,7}, \frac{2}{\epsilon_{i,11}}),\ \epsilon_{i,11} \sim \mathrm{Gamma}(1, 1)$$

$$x_{i,11} = j^{11}\ I(c^{(11)}_{j^{11}-1} < x^*_{i,11} < c^{(11)}_{j^{11}}),\quad j^{11} = 1, \ldots, 3$$

$x_{i,12}$ simulated from a discrete distribution taking values $\in \{1, 2, 3\}$ with probabilities 0.25, 0.5 and 0.25 respectively.

$$
\begin{aligned}
x_{i,13} &\sim \mathrm{N}(5, \frac{1}{\epsilon_{i,13}}),\ \epsilon_{i,13} \sim \mathrm{Gamma}(3,3) \\
x_{i,14} &\sim \mathrm{N}(1 + 2x_{i,13}, 1) \\
x^*_{i,15} &\sim \mathrm{N}(x_{i,8} + x_{i,13}, \frac{2}{\epsilon_{i,15}}),\ \epsilon_{i,15} \sim \mathrm{Gamma}(1.5, 1.5) \\
x_{i,15} &= I(x^*_{i,15} > 8) \\
x^*_{i,16} &\sim \mathrm{N}(5 + x_{i,14}, 9), \\
x_{i,16} &= I(x^*_{i,16} > 16) \\
x_{i,17} & \quad \text{simulated from a discrete distribution taking values} \in \{1,2\} \\
& \quad \text{with probabilities 0.45, and 0.55 respectively.} \\
x^*_{i,18} &\sim \mathrm{N}(2 + x_{i,13} - x_{i,15}, \frac{3}{\epsilon_{i,18}}),\ \epsilon_{i,18} \sim \mathrm{Gamma}(5,5) \\
x_{i,18} &= j^{18}\ I(c^{(18)}_{j^{18}-1} < x^*_{i,18} < c^{(18)}_{j^{18}}),\quad j^{18} = 1, \ldots, 3 \\
x^*_{i,19} &\sim \mathrm{N}(3 + 2x_{i,14} - x_{i,16}, 1), \\
x_{i,19} &= j^{19}\ I(c^{(19)}_{j^{19}-1} < x^*_{i,19} < c^{(19)}_{j^{19}}),\quad j^{19} = 1, \ldots, 5 \\
x_{i,20} & \quad \text{simulated from a discrete distribution taking values} \in \{1,2,3,4\} \\
& \quad \text{with probabilities 0.25 each.} \\
x_{i,21} &\sim \mathrm{N}(0,1) \\
x_{i,22} &\sim \mathrm{N}(7,2) \\
x_{i,23} &\sim \mathrm{N}(4,1) \\
x_{i,24} &\sim \mathrm{N}(13,3) \\
x_{i,25} &\sim \mathrm{N}(5 + I(x_{i,1} = 2) - 2 * I(x_{i,1} = 3) + x_{i,7} + 2x_{i,13}, \frac{2}{\epsilon_{i,25}}) \\
& \quad \epsilon_{i,25} \sim \mathrm{Gamma}(3.5, 3.5)
\end{aligned}
$$

where

$$
p(x_{i,5} = j^5 | x_{i,1}, \boldsymbol{\theta}^{(5)}) = \frac{\exp(\theta^{(5)}_{0,j^5} + \theta^{(5)}_{1,j^5} I(x_{i,1} = 2) + \theta^{(5)}_{2,j^5} I(x_{i,1} = 3))}{\sum_{s=1}^{5} \exp(\theta^{(5)}_{0,s} + \theta^{(5)}_{1,s} I(x_{i,1} = 2) + \theta^{(5)}_{2,s} I(x_{i,1} = 3))}
$$

where $j^2 \in \{1, \ldots, 4\}$ and $\theta^{(5)}_{0,1} = \theta^{(5)}_{1,1} = \theta^{(5)}_{2,1} = 0$ for identifiability, $\theta^{(5)}_{0,2} = 1, \theta^{(5)}_{1,2} = -3, \theta^{(5)}_{2,2} = -1.5, \theta^{(5)}_{0,3} = -1, \theta^{(5)}_{1,3} = 2, \theta^{(5)}_{2,3} = 1, \theta^{(5)}_{0,4} = -0.5, \theta^{(5)}_{1,4} = 2, \theta^{(5)}_{2,4} = -2$, and $I(\cdot)$ is the indicator function. The threshold parameters $c^{(11)}_0 = -\infty, c^{(11)}_1 = 0.2017, c^{(11)}_2 = 2.225, c^{(11)}_3 = \infty$, $c^{(18)}_0 = -\infty, c^{(18)}_1 = 4.366, c^{(18)}_2 = 6.511, c^{(18)}_3 = \infty$ and $c^{(19)}_0 = -\infty, c^{(19)}_1 = 19.14, c^{(19)}_2 = 22.52, c^{(19)}_3 = 24.53, c^{(19)}_4 = 27.60, c^{(19)}_5 = \infty$.

We then introduce missing values into variables $\mathbf{x}_5, \mathbf{x}_9, \mathbf{x}_{11}, \mathbf{x}_{13}, \mathbf{x}_{15}, \mathbf{x}_{16}, \mathbf{x}_{17}, \mathbf{x}_{18}, \mathbf{x}_{25}$ in the

following way:

$$
\begin{aligned}
p(m_{i,5}=1) &= \left\{\frac{\exp(f(x_{i,2}))}{1+\exp(f(x_{i,2}))}\right\}(1-m_{i,2})\ \text{(MAR), where}\\
f(x_{i,2}) &= -1+I(x_{i,2}=2)-4*I(x_{i,2}=3)-5*I(x_{i,2}=4)+3*I(x_{i,2}=5),\\
p(m_{i,9}=1) &= \left\{\frac{\exp(-1-5x_{i,7})}{1+\exp(-1-5x_{i,7})}\right\}(1-m_{i,7})\ \text{(MAR)},\\
p(m_{i,11}=1) &= 0.3\ \text{(MCAR)},\\
p(m_{i,13}=1) &= 0.3\ \text{(MCAR)},\\
p(m_{i,15}=1) &= 0.3\ \text{(MCAR)},\\
p(m_{i,16}=1) &= 0.3\ \text{(MCAR)},\\
p(m_{i,17}=1) &= \left\{\frac{\exp(-1+2x_{i,6})}{1+\exp(-1+2x_{i,6})}\right\}(1-m_{i,6})\ \text{(MAR)},\\
p(m_{i,18}=1) &= \left\{\frac{\exp(2-8x_{i,8})}{1+\exp(2-8x_{i,8})}\right\}(1-m_{i,8})\ \text{(MAR)},\\
p(m_{i,25}=1) &= \left\{\frac{\exp(1-5*I(x_{i,1}=2)-4*I(x_{i,1}=3))}{1+\exp(1-5*I(x_{i,1}=2)-4*I(x_{i,1}=3))}\right\}(1-m_{i,1})\ \text{(MAR)},
\end{aligned}
$$

where $m_{i,j}$ be the missing data indicator for $x_{i,j}$, where $m_{i,j}=1$ indicates $x_{i,j}$ is missing and $m_{i,j}=0$ indicates $x_{i,j}$ is observed.

Analysis

We obtain the following estimates from an analysis of the imputed datasets:

- Estimates of the means of $\mathbf{x}_6, \mathbf{x}_{13}$.
- Estimates of the proportion of units with $\mathbf{x}_{11}=1,\ldots,3,\ \mathbf{x}_{15}=1$
- Estimates of the regression coefficients of the linear regression model of $\mathbf{x}_{25}$ on all other variables.

Appendix I - Density plots of social security, tax and child support payment from the CPS data

In this section, we present the density plots of the continuous variables (social security, tax and child support payment) from the CPS data in Chapter 6 which show spikes at zero. Figure 7.9, Figure 7.10 and Figure 7.11 show the density plots of social security, tax and child support payment from the CPS data respectively.

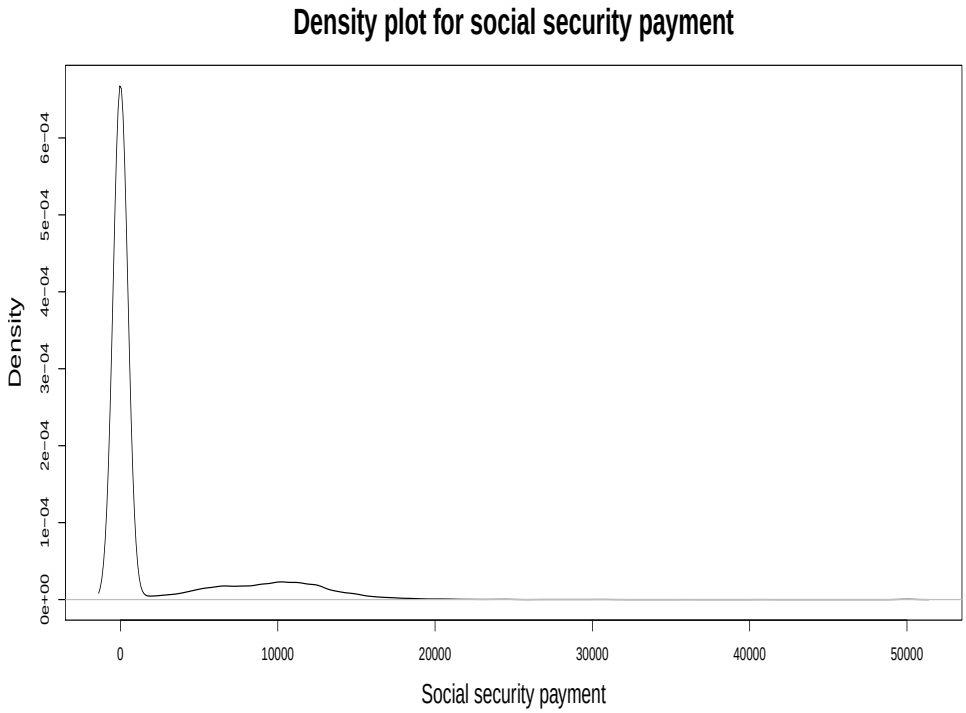


Figure 7.9: Density plot of social security payment for households in the CPS data.

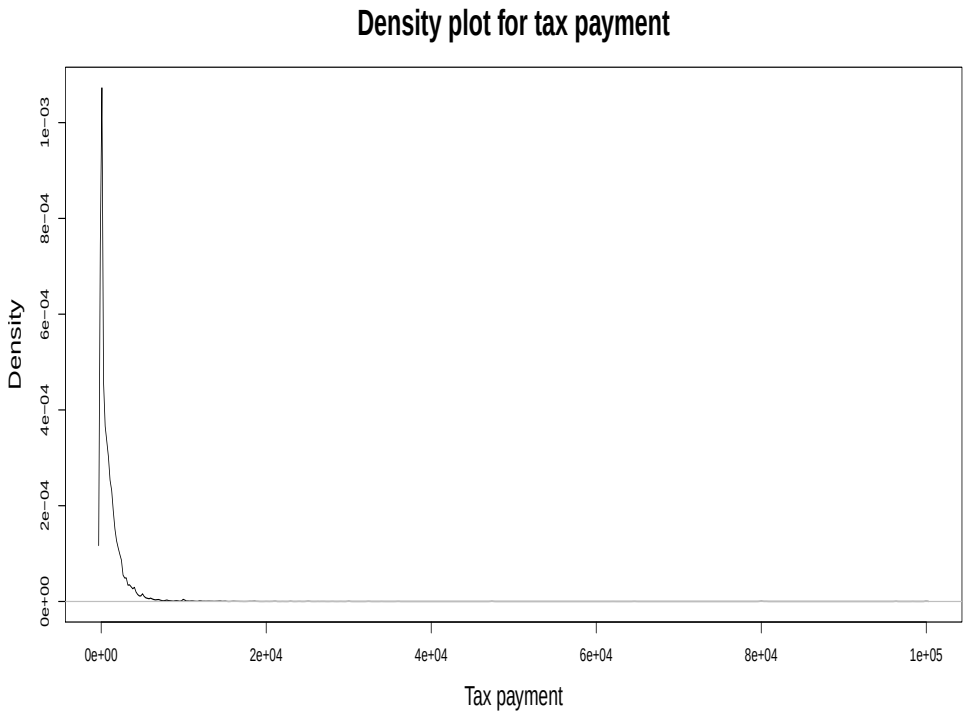


Figure 7.10: Density plot of tax payment for households in the CPS data.

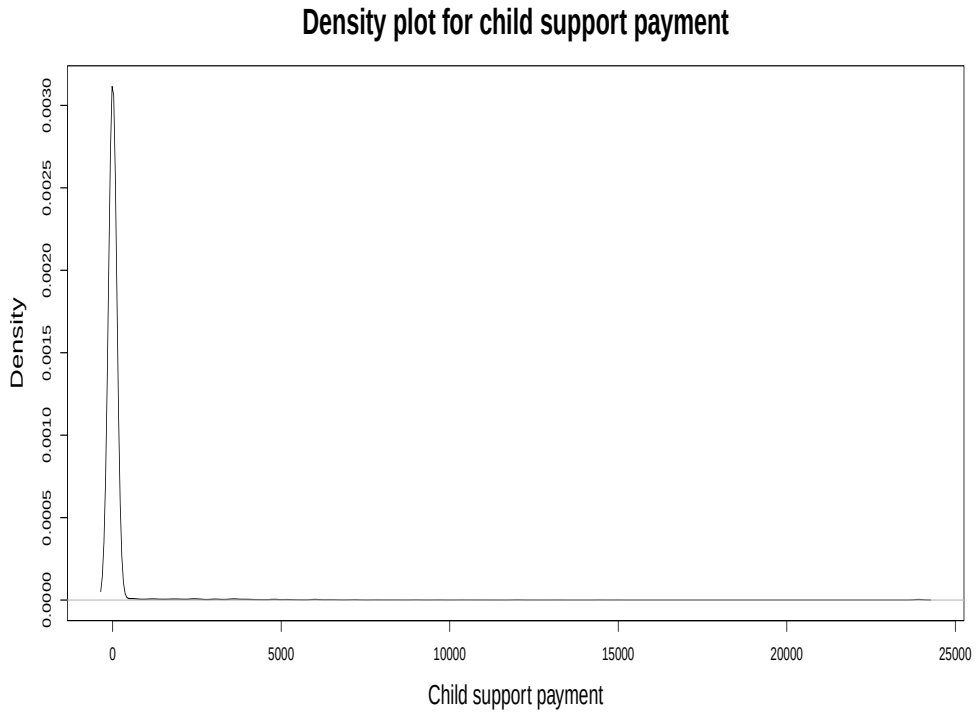


Figure 7.11: Density plot of child support payment in the CPS data.

Appendix J - Estimands considered for simulation study analysis in Chapter 6

In the section we provide details on the estimands considered in the simulation study in Chapter 6. We have selected the following estimands:

- Estimates of the means of $\mathbf{x}_5, \mathbf{x}_5 * I(\mathbf{x}_5 > 0)$.
- Estimates of the proportion of units with $\mathbf{x}_2 = 1, \ldots, 5$
- Estimates of the regression coefficients of a linear regression from the following regression models: $p(\mathbf{x}_2|\mathbf{x}_1), p(\mathbf{x}_5 * I(\mathbf{x}_5 > 0)|\mathbf{x}_1 * I(\mathbf{x}_5 > 0), \mathbf{x}_3 * I(\mathbf{x}_5 > 0))$.

Bibliography

James H. Albert and Siddhartha Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple Imputation by Chained Equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

Vic Barnett and Toby Lewis. *Outliers in Statistical Data (2nd ediition)*. John Wiley & Sons, 1994.

Todd E. Bodner. What Improves with Increased Missing Data Imputations? *Structural Equation Modeling*, 15(4):651–675, 2008.

G.E.P. Box and D.R. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

J. R. Carpenter and M. G. Kenward. Missing data in randomised controlled trials a practical guide. *London School of Hygiene*, 2007.

J.R. Carpenter and M.G. Kenward. Missing data in clinical trials a practical guide. *National Institute for Health Research*, 2008.

Caroline J. Chantry, Cynthia R. Howard, and Peggy Auinger. Full Breastfeeding Duration and Associated Decrease in Respiratory Tract Infection in US Children. *Pediatrics*, 117 (2):425–432, 2006.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

Ralph B. D'Agostino, Jr. and Donald B. Rubin. Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association*, 95(451): 749–759, 2000.

Alexander R. De LEON and K.C. Carrièger. General mixed data model: Extension of general location and grouped continuous models. *The Canadian Journal of Statistics*, 35(4):533–548, 2007.

Jörg Drechsler and Jerome P. Reiter. Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data. *Privacy in Statistical Databases*, 5262:227–238, 2008.

G.T. Duncan, S.A. Keller-McNulty, and S.L. Stokes. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. 2002.

Stephen E. Fienberg and Julie McIntyre. Data Swapping: Variations on a Theme by Dalenius and Reiss. *Privacy in Statistical Databases*, 3050:14–29, 2004.

Wayne A. Fuller. Masking procedures for Microdata Disclosure Limitation. *Journal of Official Statistics*, 9(2):383–406, 1993.

Alan E. Gelfand and Adrian F.M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

Andrew Gelman and T.P. Speed. Characterizing a Joint Probability Distribution by Conditionals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1): 185–188, 1993.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis (2nd edition)*. Chapman & Hall/CRC, 2004.

Harvey Goldstein, James Carpenter, Michael G. Kenward, and Kate A Levin. Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197, 2009.

John. W. Graham, Allison. E. Olchowski, and Tamika. D. Gilreath. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Society for Prevention Research*, 8:206213, 2007.

W.K. Hastings. Monte Carlo sampling methods using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *University of Massachusetts, Amherst, MA, Tech. Rep*, 2007.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 1970.

J.W. Hogan and T. Lancaster. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13:17–48, 2004.

Joseph G. Ibrahim, Stuart R. Lipsitz, and Ming-Hui Chen. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61:173–190, 1999.

Joseph G. Ibrahim, Ming-Hui Chen, Stuart R. Lipsitz, and Amy H. Herring. Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

Geetha Jagannathan and Rebecca N. Wright. Privacy-preserving imputation of missing data. *Data & Knowledge Engineering*, 65:40–56, 2008.

Shahjahan Khan. Predictive Distribution of Regression Vector and Residual Sum of Squares for Normal Multiple Regression Model. *Communications in Statistics - Theory and Methods*, 33(10):2423–2441, 2005.

Satkartar K. Kinney and David B. Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63(3):690698, 2007.

Katherine J. Lee and John B. Carlin. Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology*, 171(5), 2010.

Fan Li, Yaming Yu, and Donald B. Rubin. Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines. *Department of Statistical Science*, 2012.

Stuart R. Lipsitz and Joseph G. Ibrahim. A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83(4):916–922, 1996.

R.J.A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993a.

Roderick J.A. Little. Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9(2):407–426, 1993b.

Roderick J.A. Little. Regression with missing X's: A Review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

Roderick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing data (2nd edition)*. Wiley-Interscience, 2002.

L. Liu, J. Wang, J. Liu, and J. Zhang. Privacy Preservation in Social Networks with Sensitive Edge Weights. *SIAM Data Mining Conference*, 2009.

Xiao-Li Meng. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4):538573, 1994.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

Robin Mitra and Jerome P. Reiter. Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*, 30:627–641, 2011.

Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27(1):85–95, 2001.

O. J. Reichman, Matthew B. Jones, and Mark P. Schildhauer. Challenges and Opportunities of Open Data in Ecology. *Science*, 331(703), 2011.

Jerome P. Reiter. Using CART to Generate Partially Synthetic Public Use Microdata. *Journal of Official Statistics*, 21(3):441–462, 2005a.

Jerome P. Reiter. Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology*, 2003.

Jerome P. Reiter. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168:185–205, 2005b.

Jerome P. Reiter. Simultaneous use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology*, 30:235–242, 2004.

Jerome P. Reiter and Robin Mitra. Estimating Risks of Identification Disclosure in Partially Synthetic Data. *The Journal of Privacy and Confidentiality*, 1(1):99–110, 2009.

J.P. Reiter, T.E. Raghunathan, and D. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16, 2003.

D.B. Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience, 1987.

Donald B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976.

Donald B. Rubin. Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1):318, 2003.

Donald B. Rubin and John Barnard. Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, 86(4):948955, 1999.

Joseph L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, 1997.

Joseph L. Schafer and John W. Graham. Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2):147–177, 2002.

Russell J. Steele, Naisyin Wang, and Adrian E. Raftery. Inference from Multiple Imputation for Missing Data Using Mixtures of Normals. *Statistical Methodology*, 7(3):351364, 2010.

Lynn A. Streeter, Robert E. Kraut, Henry C. Lucas Jr, and Laurence Caby. How Open Data Networks Influence Business Performance and Market Structure. *Communications of the ACM*, 99(7), 1996.

Martin A. Tanner and Wing Hung Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Luke Tierney. Markov Chains for Exploring Posterior Distributions. *Annals of Statistics*, 22(4):1701–1728, 1994.

S. Van Buuren. mice: Multivariate Imputation by Chained Equations. *Journal of Statistical Software*, 45(3), 2011.

S. Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242, 2007.

S. Van Buuren, H.C. Boshuizen, and D.L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694, 1999.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377399, 2011.

Leon C.R.J. Willenborg and Ton de Waal. Elements of statistical disclosure control. 2001.

Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *The Journal of Privacy and Confidentiality*, 1(1):111–124, 2009.

Nengjun Yi and Shizhong Xu. Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics Society of America*, 179(2):1045–1055, 2008.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67, 2006.

Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. *in Proceedings of the 24th International Conference on Data Engineering (ICDE08), Cancun, Mexico*, pages 506–515, 2008.

Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.