# Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models

Min Lee

Robin Mitra*

*School of Mathematics*

*University of Southampton, Southampton, SO17 1BJ, UK*

*E-mail: R.Mitra@soton.ac.uk.

**Abstract**

Multiple imputation is a commonly used approach to deal with missing values. In this approach, an imputer repeatedly imputes the missing values by taking draws from the posterior predictive distribution for the missing values conditional on the observed values, and releases these completed data sets to analysts. With each completed data set the analyst performs the analysis of interest, treating the data as if it were fully observed. These analyses are then combined with standard combining rules, allowing the analyst to make appropriate inferences which take into account the uncertainty present due to the missing data. In order to preserve the statistical properties present in the data, the imputer must use a plausible distribution to generate the imputed values. In data sets containing variables with different measurement scales, e.g. some categorical and some continuous variables, Multivariate Imputation by Chained Equations (MICE) 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 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, and use data augmentation methods to draw imputations from a proper posterior distribution using Markov Chain Monte Carlo (MCMC). We compare the performance of our method with MICE using simulation studies and on a genuine data set taken from a breast feeding study.

*Keywords:*Data Augmentation; Latent variable; Missing data; Multiple imputation

# 1 INTRODUCTION

Missing data are a common problem in many fields and this can complicate statistical analyses of interest. In addition, failure to handle the missing data appropriately can result in biased inferences being made from the data. It is thus important to treat the missing data problem with care. Methods such as complete case analysis or ad-hoc single imputation strategies, while straight-forward to implement, might not always be appropriate to use (White & Carlin (2010)).

Multiple imputation is a widely used and appealing approach to deal with missing values (Rubin (1996), Gelman et al. (2005), Wallace et al. (2010), Moreno-Betancur & Latouche (2013), Lee et al. (2012) and Wang et al. (2011)). The approach models the joint distribution of the variables in the data set and, using this model, imputes missing values from their posterior predictive distribution conditional on the observed data. This is done $m$ times to yield $m$ multiply imputed data sets. Analysts can perform the same standard analysis that would have been performed on fully observed data on each of these completed data sets, and combine the analyses across data sets using simple combining rules, thereby incorporating the additional uncertainty due to the missing values.

It is crucial to draw imputations from an appropriate distribution to preserve relationships present in the data set. There are often complications in drawing imputations from their posterior predictive distribution: (1) the pattern of missing values in the data set may not be monotone, and (2) the variables in the data set may be continuous or categorical. One way to simplify this problem is to model the joint distribution of variables in the data set through a sequence of generalised linear models (GLMs) (Lipsitz & Ibrahim (1996), Ibrahim et al. (2005) and Mitra & D.D. (2010)) where the the link function of the GLM is determined by the measurement scale of the variable being modelled; missing values can then be imputed on a variable by variable basis. This approach has some advantages over multivariate imputation models (Schafer (1997)), as it is easier to detect model inadequacies in each GLM than having to assess the fit over a joint model for the whole dataset. Other benefits of this modelling approach are noted in Chen & Ibrahim (2001). Still, to draw imputations with a non-monotone pattern of missing values from their joint posterior predictive distribution requires Markov Chain Monte Carlo (MCMC) methods (Little & Rubin (2002)).

In this article we use the decomposition above and data augmentation methods (Albert & Chib (1993)) to draw missing values from their posterior predictive distribution. In data sets with continuous, binary and ordinal variables such an approach allows imputations to be drawn within a Gibbs sampler. When data sets include nominal variables, we develop a Metropolis within Gibbs sampler to impute missing values with an innovative Metropolis-Hastings proposal distribution. We illustrate the performance of this imputation strategy on simulated data as well as on a genuine breast feeding study. We also compare the performance of this strategy with the approach of multiple imputation via chained equations (MICE) in these scenarios. MICE is a commonly used method for imputing missing values in these types of situations (Raghunathan et al. (2001), Van Buuren (2011)) and so it would be interesting to see if there were any advantages gained by using the more formal modelling strategy proposed.

The remainder of the article is structured as follows; Section 2 briefly reviews the approach of multiple imputation to deal with the problem of missing values.

Section 3 describes the modelling strategy we propose for imputing missing values, and briefly describes the approach of MICE to impute missing values, Section 4 illustrates the performance of both approaches in a simulation study, Section 5 illustrates the performances of both approaches in a genuine breast feeding study, finally Section 6 presents some concluding remarks.

## 2 Missing data and multiple imputation

In this section we briefly describe the missing data framework and how the multiple imputation procedure is used for inference. We suppose that we have a $n \times p$ 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 $\mathbf{X}$. Also denote an $n \times p$ missing data indictor matrix $M = (m_{1,j}, \ldots, m_{n,j})$, 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.

If the conditional distribution for $p(M|\mathbf{X}, \boldsymbol{\phi}) = p(M|\mathbf{X}_{obs}, \boldsymbol{\phi})$ where $\boldsymbol{\phi}$ are the parameters of the distribution, then the data are said to be missing at random (MAR). If in fact $p(M|\mathbf{X}, \boldsymbol{\phi}) = p(M|\boldsymbol{\phi})$ so that the missing values are a random sample in $\mathbf{X}$, then the data are said to be missing completely at random (MCAR). If data are not MAR or MCAR then the data are said to be not missing at random (NMAR). In this article we focus on MAR situations.

We 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 $u$. With the presence of missing data, analysts may no longer be able to obtain these estimates. Common ad-hoc approaches to dealing with the missing values, while easy to apply, can have undesirable properties. Complete case analysis can result in biased estimates when the mechanism is not MCAR, while single imputation methods tend to underestimate the uncertainty due to the presence of missing values (Schafer & Graham (2002), Little & Rubin (2002)), leading to biased confidence intervals.

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,

$$\bar{q}_m = \frac{\sum_{k=1}^{m} q_k}{m} \text{ and } T_m = (1 + 1/m)\frac{\sum_{k=1}^{m}(q_k - \bar{q})^2}{m-1} + \frac{\sum_{k=1}^{m} u_k}{m}. \tag{1}$$

Analysts can use $\bar{q}_m$ as a point estimate for $Q$, and $T_m$ as an estimate of the variance in $\bar{q}_m$, which incorporates the additional uncertainty due to the presence of missing values. Confidence intervals can be constructed using a t-distribution with $\nu$ degrees of freedom where

$$\nu = (m-1)\left(1 + \frac{1}{m+1}\frac{\frac{\sum_{k=1}^{m} u_k}{m}}{\frac{\sum_{k=1}^{m}(q_k - \bar{q})^2}{m-1}}\right)^2.$$

(Rubin (1987) and Rubin & Schenker (1986)). Modified degrees of freedom have been proposed by Rubin & Barnard (1999) when $n$ is small, while Steele et al. (2010) explore alternative strategies to making inferences when $m$ is small.

A key challenge in the multiple imputation approach is to be able to draw imputations from a plausible distribution. 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 $\Theta$ 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 $\mathbf{X}^{(t)}_{com}$) and $p(\Theta|(\mathbf{X}^{(t)}_{mis}, \mathbf{X}_{obs})$ at each iteration $t$. Once samples $(\mathbf{X}^{(t)}_{mis}, \Theta^{(t)})$ 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, $\mathbf{X}^{(t)}_{com}$, 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 & Wong (1987)). We now describe a modelling strategy to impute missing values through data augmentation in data sets containing variables with different measurement scales.

# 3 MODELLING STRATEGY

In this section we describe our modelling strategy to multiply impute missing values. First we describe the model when the data set only contains continuous, binary and ordinal variables; we then extend the model to accommodate nominal variables in the data set. Lastly we briefly describe an alternative commonly used multiple imputation strategy based on chained equations.

## 3.1 Model for continuous, binary and ordinal variables

We model the joint distribution $p(\mathbf{X}|\Theta)$ using a sequence of conditional regression models:

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

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, $\phi_k$ 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 & 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

$$
\begin{aligned}
&x_{i,k} = j^k \; I(\gamma^{(k)}_{j^k-1} < x^*_{i,k} < \gamma^{(k)}_{j^k}), \text{ where} \\
&p(x^*_{i,k}|x_{i,1},\ldots,x_{i,k-1},\boldsymbol{\theta}^{(k)}) = \mathrm{N}(\beta^{(k)}_0 + \beta^{(k)}_1 x_{i,1} + \cdots + \beta^{(k)}_{k-1} x_{i,k-1}, 1),
\end{aligned}
$$

with $j^k \in \{1,\ldots,J_k\}$, $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)},\boldsymbol{\gamma}^{(k)})$, where $\boldsymbol{\beta}^{(k)} = (\beta^{(k)}_0,\ldots,\beta^{(k)}_{k-1})$ as before, and $\boldsymbol{\gamma}^{(k)} = \left\{\gamma^{(k)}_{j^k} : j^k \in \{1,\ldots,J_k\}\right\}$ are threshold parameters, with $\gamma^{(k)}_0 = -\infty$, $\gamma^{(k)}_1 = 0$ and $\gamma^{(k)}_{J_k} = \infty$.

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 & Chib (1993). Benefits of latent variable in modelling data sets with these types of variables have been noted by De LEON & Carriègre (2007).

To complete a Bayesian specification we place prior distributions on all parameters. We place independent Jeffreys priors on all regression coefficients and residual precisions, i.e.

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

and

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

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

With this modelling approach and choice of prior distributions, the full conditional distribution 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. The full conditional distributions for all unknowns are presented in Appendix A; here we only present the full conditional distribution that imputes values for $x^*_{i,k}$ with missing $x_{i,k}$ (for notational convenience we assume that if $x_{i,k}$ is continuous then $x_{i,k} = x^*_{i,k}$). The full conditional distribution for $x^*_{i,k}$ (given $m_{i,k} = 1$) follows 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

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

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 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 2, 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 unordered categorical 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

$$\begin{aligned} p(x_{i,k} = j^k | x_{i,1}^*, \ldots, x_{i,k-1}^*, \boldsymbol{\theta}_{j^k}^{(k)}) &= \pi_{i,j^k}^{(k)} \\ &= \frac{\exp(\beta_{0,j^k}^{(k)} + \beta_{1,j^k}^{(k)} x_{i,1}^* + \cdots + \beta_{k-1,j^k}^{(k)} x_{i,k-1}^*)}{\sum_{s=1}^{L_k} \exp(\beta_{0,s}^{(k)} + \beta_{1,s}^{(k)} x_{i,1}^* + \cdots + \beta_{k-1,s}^{(k)} x_{i,k-1}^*)}, \end{aligned} \tag{3}$$

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 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 $\beta_{q,j^k}^{(k)} x_{i,q}^*$ with $\sum_{j^q=2}^{L_q} \beta_{q,j^k}^{(k),j^q} I(x_{i,q}^* = j^q)$ in Equation 3. 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.

$$p(\boldsymbol{\theta}^{(k)}) = \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{4}$$

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. 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

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

where

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

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 anymore.

Of course we must also sample the parameters, $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\theta}^{(q)}_1, \ldots, \boldsymbol{\theta}^{(q)}_{L_q})$ 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 $j^q = 1, \ldots, L_q$, where we set $\boldsymbol{\theta}^{(q)}_1 = \mathbf{0}$ 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, see Appendix A for the form of this distribution. 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

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

where $I(\cdot)$ is the indicator function and $\pi(\boldsymbol{\theta}^{(q)})$ is the diffuse prior given by Equation 4.

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

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

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 $t$, $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}^{(t)}_{com})$. We generate proposals for $\boldsymbol{\theta}^{(q)}$ from a $\mathrm{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ distribution where $\boldsymbol{\mu}^{(t)}$ is the value that maximises $l(\boldsymbol{\theta}^{(q)}; \mathbf{X}^{(t)}_{com})$ and $\boldsymbol{\Sigma}^{(t)}$ is given by $-E[\frac{\delta^2}{\delta\beta^{(q)}_{j,w}\delta\beta^{(q)}_{\tilde{j},\tilde{w}}} l(\boldsymbol{\theta}^{(q)}; \mathbf{X}^{(t)}_{com})]^{-1}_{\boldsymbol{\theta}^{(q)}=\boldsymbol{\mu}^{(t)}}$.

In the simulation studies and applications to the breast-feeding study in Section 4 and Section 5, 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.

## 3.3 Multivariate imputation via chained equations

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 $D$ 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 $p_1(\mathbf{x}_1|D)$, denote the imputed variable $\mathbf{x}_1^{(t)}$. Then missing values in $\mathbf{x}_2$ are imputed using $p_2(\mathbf{x}_2|\mathbf{x}_1^{(t)}, D)$ to create an imputed variable $\mathbf{x}_2^{(t)}$, this continues sequentially until an imputed variable $\mathbf{x}_p^{(t)}$ is created from $p_k(\mathbf{x}_p|\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{p-1}^{(t)}, D)$. 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 $g_k(\mathbf{x}_k|D, \mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{k-1}^{(t)}, \mathbf{x}_{k+1}^{(t-1)}, \ldots, \mathbf{x}_p^{(t-1)})$ 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 $g_k(\mathbf{x}_k|D, \mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{k-1}^{(t)}, \mathbf{x}_{k+1}^{(t-1)}, \ldots, \mathbf{x}_p^{(t-1)})$ are not dervied from any joint posterior distribution as was the case with FRM. Problems with this modelling approach have been noted by Gelman & 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 section we illustrate the performance of both modelling approaches in a simulation study.

# 4 SIMULATION STUDY

We now illustrate the performance of our imputation modelling strategy through a simulation. 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 a 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 B. We replicate this data generation 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 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. See Appendix B for a full list of the estimands considered. Using the $m$ imputed data sets, we apply the MI combining rules described in (1) to construct point and interval estimates for these estimands. We also construct esimates 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 generation process and so decomposes the joint distribution of the imputation model as described in (2). 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 B.

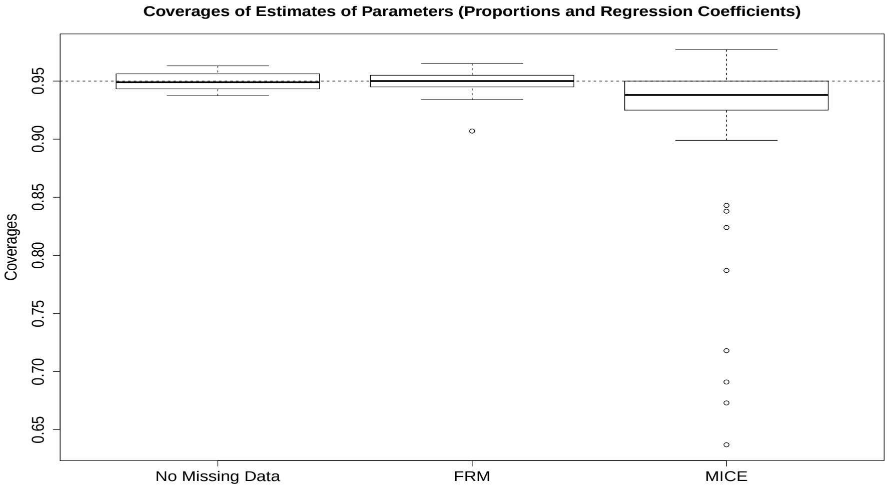

Figure 1: The coverages of estimands in Scenario 1.

Figure 1 presents boxplots 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.

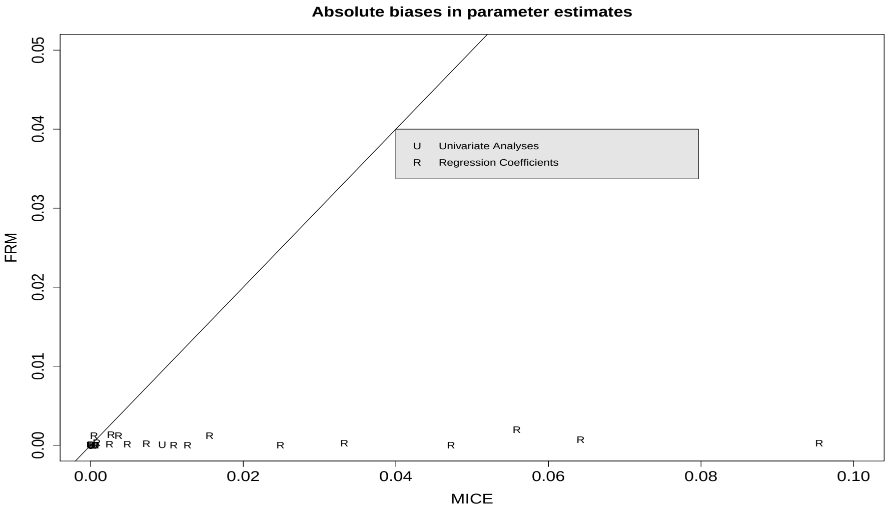

Figure 2: Absolute biases in estimates in Scenario 1.

To determine whether these low coverages seen from using MICE are due to a result in biases in the estimates, in Figure 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 3 and 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.

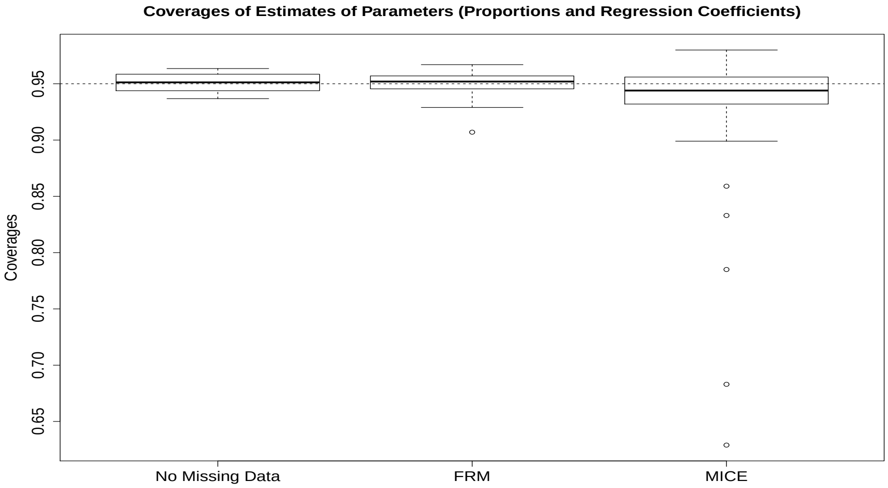

Figure 3: The coverages of estimands in Scenario 2.

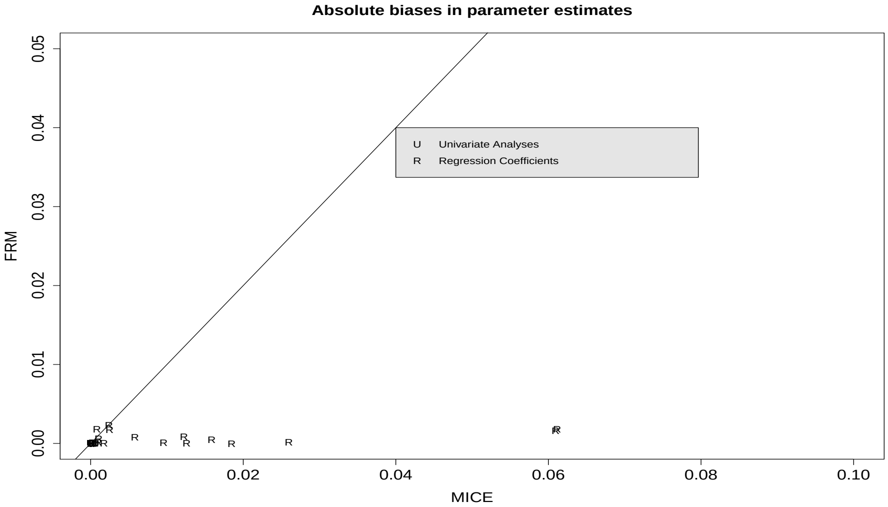

Figure 4: Absolute biases in estimates in Scenario 2.

# 5 APPLICATION TO A BREASTFEEDING 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.

## 5.1 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 breastfeeding on children's cognitive development.

The study recorded the Peabody individual assessment math score (PIATM) administered at five or six years of age, which we use as the measure of child's cognitive development. Following Mitra & Reiter (2011) we dichotomize 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 breastfed for less than 24 weeks, while the second group, denoted the treatment group, comprises those units who were breastfed 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 PIATM 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 & 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 & Reiter (2011) we also categorised the last two variables due to their highly non-normal distributions (see Appendix C 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, see Appendix C for more details.

Analysts can fit a regression model using PIATM 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 breastfeeding on cognitive development; we are simply using this analysis to illustrate the FRM approach to imputing 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 C for more details). There were 1313 complete cases 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 Section 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 C. 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 C. 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 & Rubin (2000)). We could have repeated the analysis without including the response in the imputations.

## 5.2 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 they 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 can 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

Table 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) | 85.611 (4.897) |
| Treatment effect. | 1.609 (0.948) | 1.275 (1.586) | 2.218 (1.230) |
| The number of years between 1979 and when the mother gave birth. | -0.025 (0.099) | 0.001 (0.182) | -0.050 (0.135) |
| The child's race-Black | -1.244 (1.117) | -0.933 (2.078) | -0.610(1.495) |
| The child's race-Other | 2.346 (0.971) | 4.159 (1.736) | 4.153 (1.384) |
| Spouse present at birth. | -1.595 (1.438) | -1.860 (2.879) | -1.492 (1.832) |
| Partner present at birth. | -0.159 (1.104) | 0.177 (1.818) | -0.118(1.603) |
| 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) | 0.062(0.929) |
| Grandparents were present at birth. | -1.078 (1.151) | -1.175 (1.867) | -0.335 (1.581) |
| Mother's intelligence. | 0.081 (0.018) | 0.057 (0.032) | 0.054(0.026) |
| Mother's highest educational. | 0.606 (0.212) | 0.628 (0.355) | 0.556(0.313) |
| Child's birth weight. | 0.023 (0.021) | 0.026 (0.032) | 0.027(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) | -0.388 (0.172) |
| Weeks that the mother worked-level 2 | 1.190 (0.984) | 2.194 (1.556) | 2.658 (1.666) |
| Weeks that the mother worked-level 3 | 0.038 (1.261) | 1.653 (2.083) | 1.790 (2.191) |
| Weeks that the mother worked-level 4 | 2.156 (1.102) | 3.392 (2.003) | 5.142 (1.778) |
| Child was born premature-level 2 | 1.265 (0.919) | -0.233 (1.531) | -0.137 (1.180) |
| Child was born premature-level 3 | 2.009 (2.248) | -0.226 (3.718) | -0.104(3.068) |

given in Appendix C.

Table 1 presents estimates of the regression coefficients and variances obtained from fitting the regression model for PIATM 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 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 MICE is about 25.3% 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.

In Figure 5 we also plot the absolute biases in the regression coefficient estimates

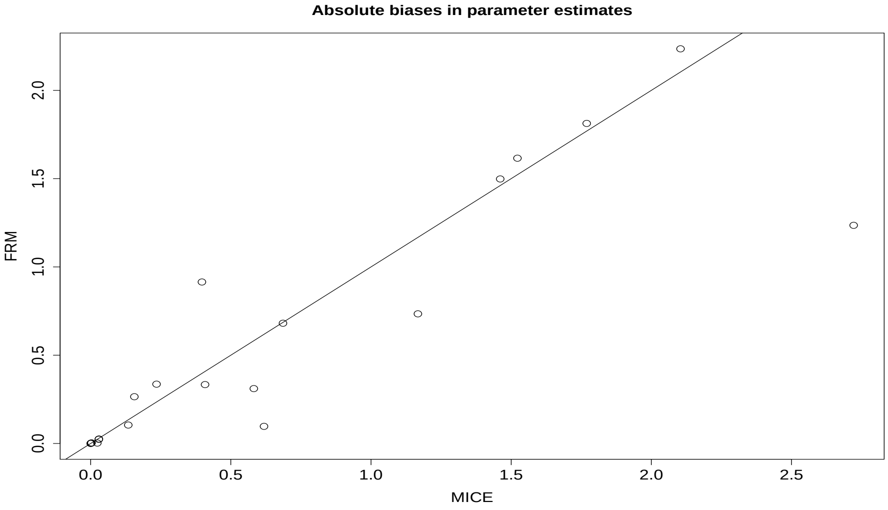

Figure 5: Absolute biases in parameter estimates using FRM against MICE

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.

## 5.3 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 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 0.1 points of the Peabody score (standard errors from using FRM and MICE are 0.727 and 0.736 respectively).

While we cannot be certain which approach most closely reflects the estimates that

Table 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) |

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.

# 6 CONCLUDING REMARKS

We have proposed a modelling strategy to impute missing values in data sets that contain both categorical and continuous variables. 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 propose a Metropolis within Gibbs sampler with a novel Metropolis-Hastings proposal that can efficiently impute the missing values. This can be seen to reduce the bias in estimates over using MICE.

It would 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, variables are also sometimes part continuous and part discrete, e.g. those that arise from skip patterns; incorporating these variables would require a more complex modelling strategy. It would also be interesting to consider broadening the class of models used, for example using logistic regression models for binary responses, potentially using computational techniques proposed in Kinney & Dunson (2007).

## 7 Acknowledgements

The authors would like to thank Professor Jennifer Hill for providing the data used in Section 5, and Professor Jon Forster for helpful suggestions in the preparation of this manuscript.

## References

Albert, J. H. & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**(422), 669–679.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple Imputation by Chained Equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, **20**(1), 40–49.

Box, G. & Cox, D. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**(2), 211–252.

Chantry, C. J., Howard, C. R., & Auinger, P. (2006). Full Breastfeeding Duration and Associated Decrease in Respiratory Tract Infection in US Children. *Pediatrics*, **117**(2), 425–432.

Chen, M.-H. & Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, **57**, 43–52.

D'Agostino, Jr., R. B. & Rubin, D. B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association*, **95**(451), 749–759.

De LEON, A. R. & Carriègre, K. (2007). General mixed data model: Extension of general location and grouped continuous models. *The Canadian Journal of Statistics*, **35**(4), 533–548.

Gelman, A. & Speed, T. (1993). Characterizing a Joint Probability Distribution by Conditionals. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**(1), 185–188.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis (2nd edition)*. Chapman & Hall/CRC.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D., & Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, **61**(1), 74–85.

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, **9**(3), 173–197.

Ibrahim, J. G., Lipsitz, S. R., & Chen, M.-H. (1999). 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.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association*, **100**(469), 332–346.

Kinney, S. K. & Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**(3), 690698.

Lee, K. J., Galati, J. C., Simpson, J. A., & Carlin, J. B. (2012). Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. *Statistics in Medicine*, **31**(30), 4164–4174.

Lipsitz, S. R. & Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, **83**(4), 916–922.

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

Little, R. J. & Rubin, D. B. (2002). *Statistical Analysis with Missing data (2nd edition)*. Wiley-Interscience.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, **9**(4), 538573.

Mitra, R. & D.D., D. (2010). Two level stochastic search variable selection in glms with missing predictors.

Mitra, R. & Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*, **30**, 627–641.

Moreno-Betancur, M. & Latouche, A. (2013). Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in Medicine*.

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, **27**(1), 85–95.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**(434), 473–489.

Rubin, D. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**, 366–374.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley-Interscience.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, **57**(1), 318.

Rubin, D. B. & Barnard, J. (1999). Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, **86**(4), 948955.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC.

Schafer, J. L. & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**(2), 147–177.

Steele, R. J., Wang, N., & Raftery, A. E. (2010). Inference from Multiple Imputation for Missing Data Using Mixtures of Normals. *Statistical Methodology*, **7**(3), 351364.

Tanner, M. A. & Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**(398), 528–540.

Van Buuren, S. (2011). mice: Multivariate Imputation by Chained Equations. *Journal of Statistical Software*, **45**(3).

Van Buuren, S., Boshuizen, H., & Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**(6), 681–694.

Wallace, M. L., Anderson, S. J., & Mazumdar, S. (2010). A stochastic multiple imputation algorithm for missing covariate data in tree-structured survival analysis. *Statistics in Medicine*, **29**(29), 3004–3016.

Wang, C.-N., Little, R., Nan, B., & Harlow, S. D. (2011). A hot-deck multiple imputation procedure for gaps in longitudinal recurrent event histories. *Biometrics*, **67**, 1573–1582.

White, I. R. & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, **29**, 2920–2931.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, **30**(4), 377399.

# 8 Appendices

In this section we provide details that complement the material presented in the main text. Appendix A presents the full conditional distributions needed to implement FRM for imputing missing values. Appendix B presents details of how the simulation studies in Section 4 were implemented. Appendix C presents some more details of the data analysis in Section 5.

## 8.1 Appendix A - FRM strategy to impute missing values

In this section we present the full conditional distributions in the Metropolis within Gibbs sampler required to generate imputed datasets using FRM mentioned in Section 3. We first express the joint posterior distribution of all unknowns; these comprise model parameters $\Theta$, 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 following Section 3, we order the variables so that $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are nominal, each $\mathbf{x}_q$, for $q = 1, \ldots, k$ 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 $\pi_{j^1}^{(1)} = p(x_{i,1} = j^1|\boldsymbol{\theta}^{(1)}) = \frac{\exp(\beta_{0,j^1}^{(1)})}{\sum_{s=1}^{L_1} \exp(\beta_{0,s}^{(1)})}$, $j^1 = 1, \ldots, L_1$. 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 $q = 2, \ldots, k$,

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

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

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

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),

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

We specify the standard non-informative prior for $\boldsymbol{\theta}^{(q)}$,

$$p(\boldsymbol{\theta}^{(q)}) = p(\boldsymbol{\beta}^{(q)}, \phi_q) \propto \phi_q^{-1}.$$

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,

$$p(\boldsymbol{\theta}^{(q)}) = p(\boldsymbol{\beta}^{(q)}) \propto 1.$$

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),

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

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 $\boldsymbol{\theta}^{(q)}$ by,

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

where

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

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

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 $x_{i,q} = x^*_{i,q}$, for $q \in \{1, \ldots, p\}$.

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

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

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,

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

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

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

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

which is a normal distribution $\mathrm{N}(\widetilde{\mu}_{i,q},\phi_q^{-1})$ truncated to $(\gamma^{(q)}_{j^q-1},\gamma^{(q)}_{j^q})$ where the $\gamma^{(q)}_{j^q}$ are 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 normal distribution $\mathrm{N}(\boldsymbol{\mu}^{(t)},\boldsymbol{\Sigma}^{(t)})$ , and $\boldsymbol{\mu}^{(t)}$ is the value that maximises $l(\boldsymbol{\theta}^{(q)};\mathbf{X}^{(t)}_{com})$ and $\boldsymbol{\Sigma}^{(t)}$ is given by $-E[\frac{\delta^2}{\delta\beta^{(q)}_{j,w}\delta\beta^{(q)}_{\tilde{j},\tilde{w}}}l(\boldsymbol{\theta}^{(q)};\mathbf{X}^{(t)}_{com})]^{-1}_{\boldsymbol{\theta}^{(q)}=\boldsymbol{\mu}^{(t)}}$.

Next sample values for $\boldsymbol{\theta}^{(q)} = (\boldsymbol{\beta}^{(q)},\phi_q),\ q\in\{k+1,\ldots,p\}$ when $\mathbf{x}^*_q$ is continuous from the joint posterior distribution of $\boldsymbol{\beta}^{(q)} = (\beta^{(q)}_0,\beta^{(q),2}_1,\ldots,\beta^{(q),L_1}_1,\ldots,\beta^{(q),2}_k,\ldots,\beta^{(q),L_k}_k,\beta^{(q)}_{k+1},\ldots,\beta^{(q)}_{q-1})$ and $\phi_q$ 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

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

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

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

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

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

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 $j^q \in \{1, \ldots, J_q\}$.

## 8.2 Appendix B - Simulation study in Section 4

In this section we present details of the simulation study in Section 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:

$$
\begin{aligned}
x_{i,1} & \quad \text{simulated from a discrete distribution taking values} \in \{1,2,3\} \\
& \quad \text{with probabilities 0.3, 0.4 and 0.3 respectively.} \\
x_{i,2} & \sim p(x_{i,2}|x_{i,1}, \boldsymbol{\theta}^{(2)}), \\
x_{i,3} & \sim \mathrm{N}(5,1), \\
x^*_{i,4} & \sim \mathrm{N}(1+2x_{i,3},1), \\
x_{i,4} & = I(x^*_{i,4} > 11) \\
x^*_{i,5} & \sim \mathrm{N}(2+5*I(x_{i,1}=2)+I(x_{i,1}=3)+x_{i,3},2), \\
x_{i,5} & = j^5\ I(c^{(5)}_{j-1} < x^*_{i,5} < c^{(5)}_j), \quad j^5 = 1,\ldots,4 \\
x_{i,6} & \sim \mathrm{N}(8,3), \\
x_{i,7} & \sim \mathrm{N}(2+2x_{i,6}-x_{i,3},5), \\
x^*_{i,8} & \sim \mathrm{N}(3+2x_{i,7},5), \\
x_{i,8} & = j^8\ I(c^{(8)}_{j-1} < x^*_{i,8} < c^{(8)}_j), \quad j^8 = 1,\ldots,8 \\
x_{i,9} & \sim \mathrm{N}(1+4*I(x_{i,2}=2)-3*I(x_{i,2}=3)+5*I(x_{i,2}=4)+2*I(x_{i,2}=5)+5x_7,3),
\end{aligned}
$$

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 $j^2 \in \{1,\dots,5\}$ 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 $\mathbf{x}_2,\dots,\mathbf{x}_9$ 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 proportion of units with $\mathbf{x}_2 = 1,\dots,5,\ \mathbf{x}_4 = 1,\ \mathbf{x}_5 = 1,\dots,4,\ \mathbf{x}_8 = 1,2,\dots,8$

- 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.

## 8.3 Appendix C - Breast-feeding data analysis

In this section we present the details of the breast-feeding data analysis mentioned in Section 5. We first describe the ordering of the variables we used in the imputation. As noted in Section 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 6.

From Figure 7 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

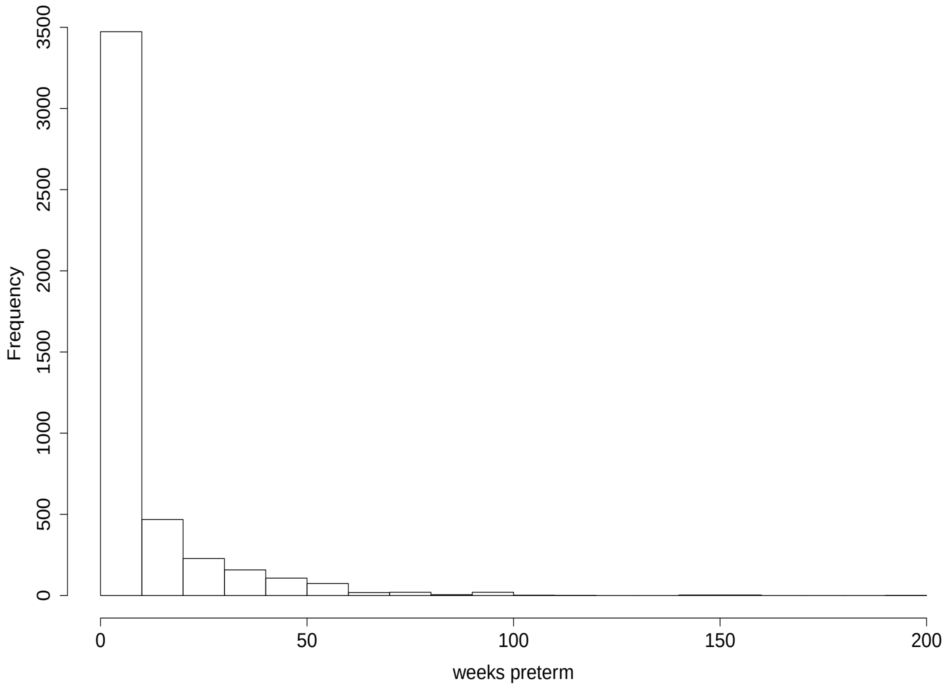

Figure 6: Histogram of weeks mother worked in the year before giving birth for subjects in the breast-feeding study.

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 & 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 transformation, of the residuals in regression models where a transformation of

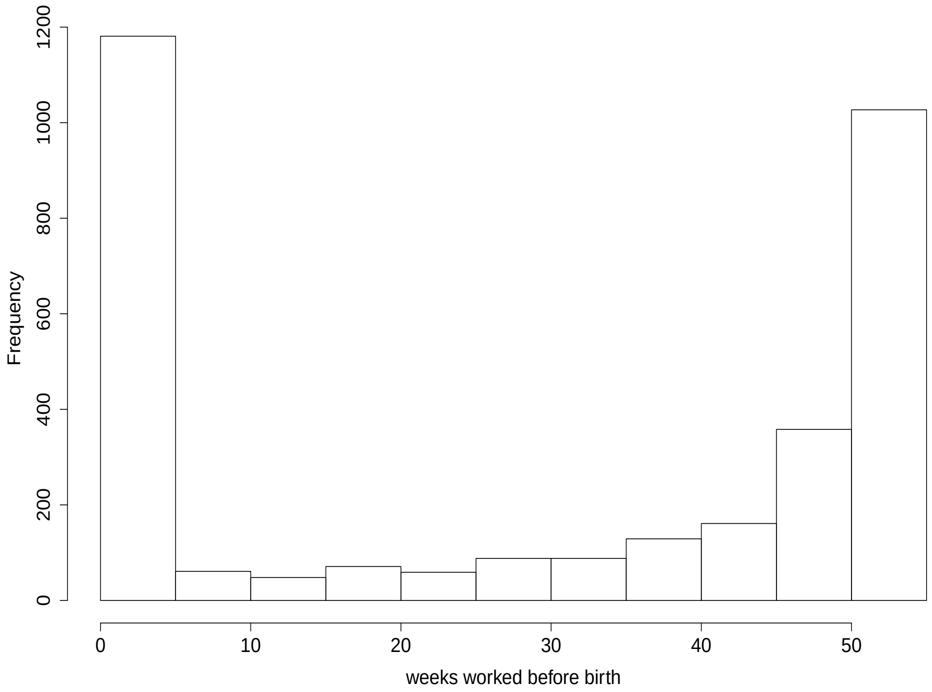

Figure 7: Histogram of weeks preterm for subjects in the breast feeding study.

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.

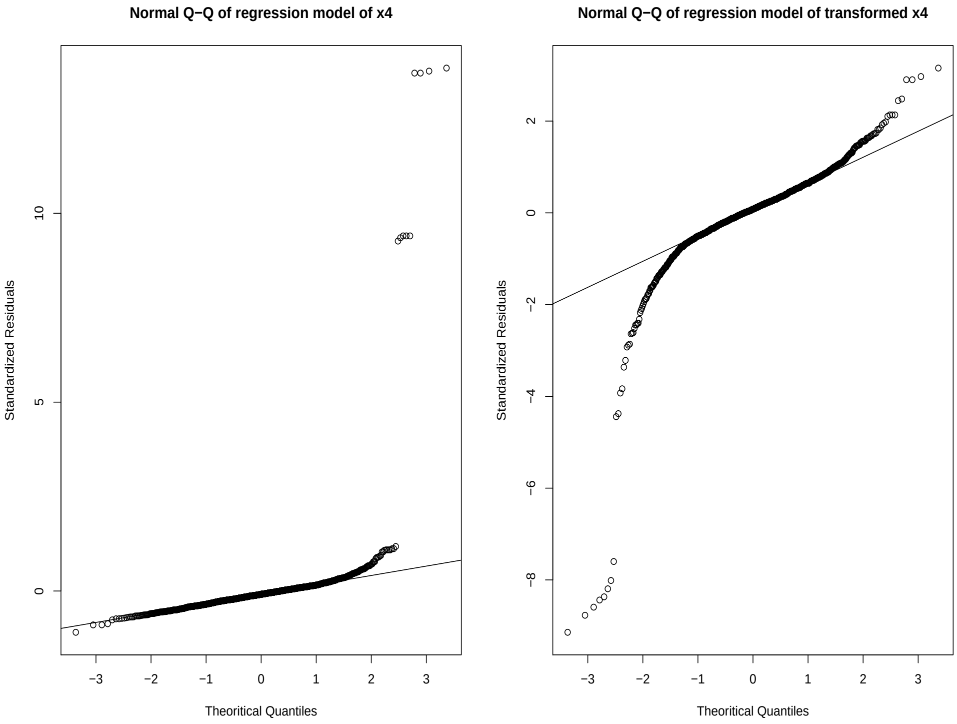

Figure 8: Q-Q plot of the residuals in the linear regression model of family income, $\mathbf{x}_4$ before(left) and after(right) transformation (natural log).

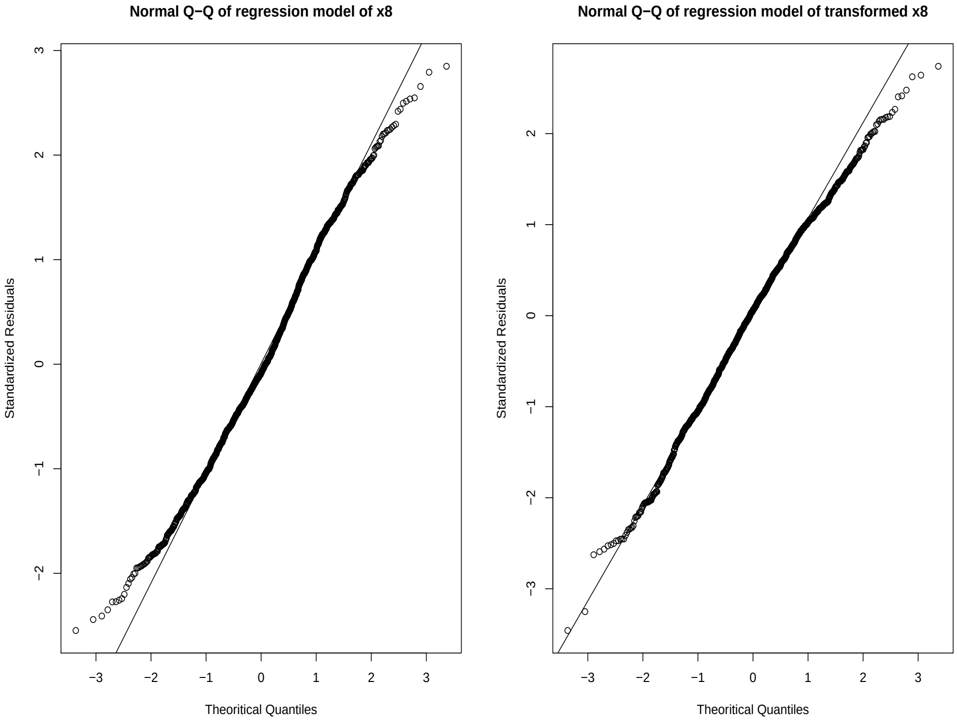

Figure 9: Q-Q plot of the residuals in the linear regression model of mother's intelligence as measured by an armed forces qualification test, $\mathbf{x}_8$ before(left) and after(right) transformation (square root).

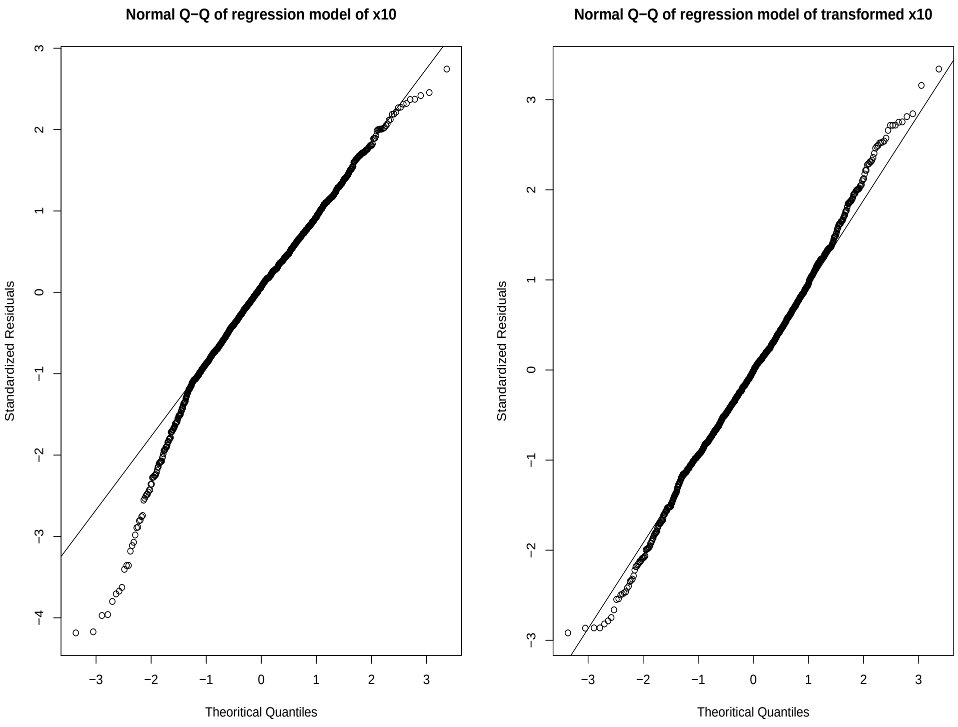

Figure 10: Q-Q plot of the residuals in the linear regression model of child's birth weight, $\mathbf{x}_{10}$ before(left) and after(right) transformation (square).

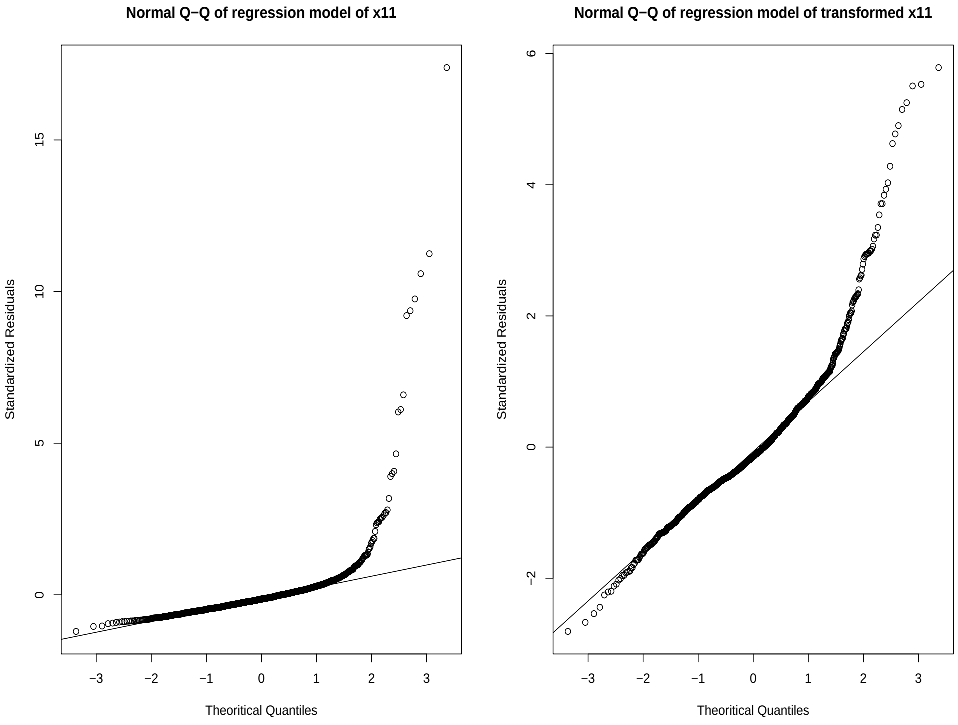

Figure 11: Q-Q plot of the residuals in the linear regression model of number of days that the child spent in hospital, $\mathbf{x}_{11}$ before(left) and after(right) transformation (natural log).

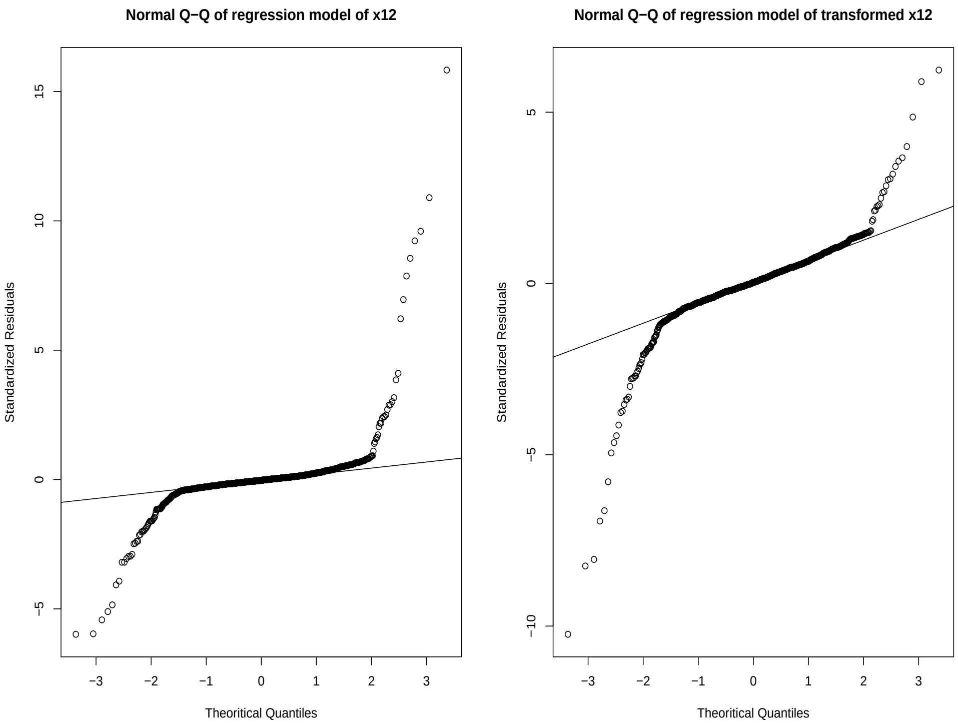

Figure 12: Q-Q plot of the residuals in the linear regression model of number of days that the mother spent in hospital, $\mathbf{x}_{12}$ before(left) and after(right) transformation (natural log).

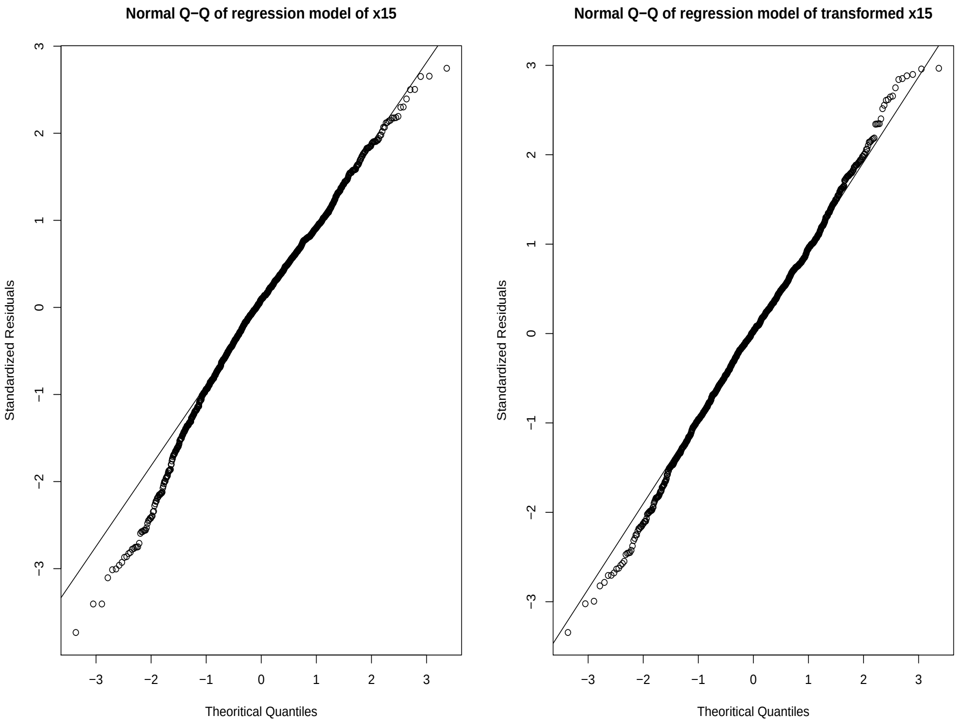

Figure 13: Q-Q plot of the residuals in the linear regression model of Peabody individual assessment test math score (PI-ATM), $\mathbf{x}_{15}$ before(left) and after(right) transformation (square).