UNIVERSITY OF THE
FREE STATE
UNIVERSITEIT VAN DIE
VRYSTAAT
YUNIVESITHI YA
FREISTATA

UFS·UV
NATURAL AND
AGRICULTURAL SCIENCES
NATUUR- EN
LANDBOUWETENSKAPPE

MATHEMATICAL STATISTICS
AND ACTUARIAL SCIENCE
WISKUNDIGE STATISTIEK
EN AKTUARIËLE WETENSKAP

A Robust Model for Use in Sequential Regression Multiple Imputation

M. J. von Maltitz, A. J. van der Merwe

February 9, 2015

Multiple imputation (MI) is considered the best solution to the missing data problem when it is necessary to separate the imputation task from the analysis task. One flexible form of MI is sequential regression multiple imputation (SRMI), a method in which individual variables can be modelled according to specific generalised univariate regression models. While there has been substantial research into SRMI, there remains a need for a regression model in SRMI that can handle heavy-tailed and skew regression errors. This paper introduces such a model, based on the skew Student's t-distribution, and shows that this model is robust in the presence of non-Normal errors, meaning that it can be used as the default model within SRMI for continuous data.

Keywords: Multiple imputation, sequential regression multiple imputation, robust imputation, Bayesian estimation, skew t-distribution.

1. Introduction

Large survey data sets often suffer from nonresponse, mis-measured, or lost data. Moreover, most data analysis procedures are not designed to handle these missing data, leading to invalid and inefficient inference about a population (Schafer & Graham 2002). Many analyses use either complete-case analysis or a simple method of imputing missing data, such as single imputations. Even if single imputations are accurate, however, they do not capture the uncertainty inherent in the imputation. This is one of the reasons for the development of multiple imputation. The science of multiple imputation has evolved so as to remove the onus of imputing from the analyst (Meng 1994), so that public-use datasets can be prepared by imputation experts and offered to experts in the analysis

arena without substantial loss in estimation validity and/or efficiency, provided the guidelines for the use of multiply imputed data are followed.

We assume that a random process, known as the missing data mechanism (MDM), causes data to become missing. In brief, there are three mechanisms by which data is said to be missing — 'missing at random' (MAR), 'missing completely at random' (MCAR), or 'missing not at random' (MNAR). In the MAR MDM, the distribution of positions of the missing data entries is assumed to be independent of the missing data in the analysis, or $\Pr(R|Y_{com}, \theta) = \Pr(R|Y_{obs}, \theta)$, where R is the missing data mechanism, Y_{com} is the theoretical complete data set, θ is the unknown parameter of the data, and Y_{obs} is the observed part of the data. In the case of MCAR, a special case of the MAR mechanism, the positions of the missing data entries are assumed to be independent of all of the variables in the analysis, i.e. $\Pr(R|Y_{com}, \theta) = \Pr(R|\theta)$, using the same notation as before. In MNAR MDM, the positions of the missing data entries are assumed to be at least dependent on data that is missing from the dataset, or, more basically, the distribution of missingness is not MAR. Once again using the same notation, for MNAR we have $\Pr(R|Y_{com}, \theta) \neq \Pr(R|Y_{obs}, \theta)$.

Missing data problems generally require a solution that has the following capabilities (Rubin 1987, p. 11). *Firstly*, it should be possible to use standard complete-data analysis methods on the data sets that have been filled in. *Secondly*, the imputation technique and adjustments to the subsequent analysis should yield valid inferences that produce both estimates that adjust for observed differences between respondents and non-respondents and standard errors of these estimates that reflect the reduced sample size and an adjustment for observed differences between respondents and non-respondents. *Finally*, the imputation technique should display the sensitivity of inferences to various plausible models for nonresponse. Non-imputation methods such as available-case analysis and re-weighting procedures, as well as single imputation methods where a single value is imputed for every missing datum, do not adhere fully to these three conditions. The uncertainty inherent in imputation is not included in the overarching analysis, so inferences are often biased, inefficient, or invalid (see for example Brand 1998, Schafer & Graham 2002, Ardington, Lam, Leibbrandt & Welch 2006, Saunders, Morrow-Howell, Spitznagel, Doré, Proctor & Pescarino 2006).

Multiple imputation (MI) was first proposed by Donald Rubin in the 1970's (Rubin 1978) as one solution to survey nonresponse problems. Multiple imputation covers a broad range of methods of imputation that impute several plausible values for each missing value in a data set. Rubin wished to create theoretically appropriate, yet practical methods that would split the analysis task from the imputation task. Expert imputers could then publish complete data that could be used in a statistically appropriate manner for any future analysis of that data set.

Within MI, imputed values reflect the variation within an imputation model as well as sensitivity to different imputation models, and the analysis of the resultant multiply-imputed data can be viewed as simulating predictive distributions of desired summary statistics under imputation models. The entire process behind MI and analysis is then divided into three areas, namely the modelling task, or specifying a hypothetical joint distribution, the imputation task, or deriving a predictive posterior distribution for the incomplete variable(s), and the analysis task, or estimating parameters of interest from the completed data. It is now widely accepted that MI is a viable and conservative method of handling incomplete data when the imputation and analysis tasks are to be separated (Fay 1992, Meng 1994, Nielson 2003, Schafer 2003, Rubin 2003, Zhang 2003, van Buuren 2007).

There are multiple sources of uncertainty in MI. Rubin (2003) points out that these often complement each other to make MI "self-correcting" for approximately valid statistical inference. Rubin

(2003) lists these three forms of uncertainty:

1. There is almost always uncertainty in choosing the correct imputation model and MDM (ignorable or non-ignorable)
2. Even with complete knowledge of the form of an imputation model governed by unknown parameters, there is uncertainty in the parameters' values used to create the imputations.
3. Given both the imputation model and its parameters, there is residual uncertainty to be reflected when drawing imputed values

Multiple imputation using Bayesian statistical methods can reflect all of these uncertainties: the first, by drawing imputations under different imputation models; the second, by randomly drawing parameters from their posterior distributions and thereby attempting to make the MI "proper" or "confidence proper" (see Rubin 1976, Rubin 1996); and the third, by randomly drawing imputed values from their predictive distribution, given the fixed parameters drawn previously.

One of the main problems of MI in a Bayesian context is that a multivariate model needs to be chosen for the observed data. In practice, however, survey data consists of many variables distributed in many different ways, and often displays seemingly unsystematic patterns of missing data. These properties of survey data make joint modelling approaches extremely difficult to implement, since typical multivariate distributions aren't flexible enough to accommodate such varying structure.

A more recent MI approach uses sequences of appropriate univariate multiple regression models to multiply impute missing data. Hence the name Sequential Regression Multiple Imputation (SRMI). This approach is also known as the fully conditional specification (FCS) or approach (for reasons that will be explained shortly), or MI through chained equations (MICE), as well as stochastic relaxation, regression switching, variable-by-variable imputation, partially incompatible MCMC, iterated univariate imputation, or the ordered pseudo-Gibbs sampler. This paper will refer to all of these methods with the common acronym SRMI, since they are all essentially the same procedure.[1] In the sequential procedure, each variable can be modelled individually within the imputation process. Imputers can have much more control over imputations from variables with inherent restrictions, which is not easily done when variables are jointly modelled in an imputation procedure. This method of MI was proposed by van Buuren, Boshuizen & Knook (1999), and independently by Raghunathan, Lepkowski, van Hoewyk & Solenberger (2001), although the system had been used even earlier by researchers such as Kennickell (1991).

2. The SRMI process

2.1. Overview

In essence, SRMI works in a two-dimensional process as follows (Raghunathan et al. 2001, He & Raghunathan 2009). Reviewing our standard notation, let $Y_j\,(j = 1, \ldots, p)$ denote the variables with missing values, X denote the matrix of q fully observed variables, and let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ denote the $p-1$ variables in Y excluding Y_j. In SRMI, a conditional

[1] The reason SRMI is chosen above the more common FCS acronym, is due to the more explanatory nature of the terms 'sequential regression', which adequately describe some of the steps within the procedure.

model $P\left(Y_j|Y_{-j}, X, \theta_j\right)$ is specified for each Y_j, with θ_j denoting the respective model parameters. The first dimension of the SRMI process is a single iteration, or pass, of the process, essentially 'filling in' the missing data values. The second dimension of the procedure is the repetition of this 'filling in' process, using the previously filled-in values. Thus, in each iteration of the imputation procedure, θ_j is drawn from $P\left(\theta_j|Y_j^{obs}, Y_{-j}, X\right)$ using the observed part of of the variable Y_j, Y_j^{obs}, and the completed Y_{-j} (from the previous iteration if there was one), and X, and the missing part of the variable Y_j, Y_j^{mis} is then imputed. The conditional model process is repeated, cycling through all the Y_j's. Each conditional density is modelled through the appropriate regression model, chosen for the distribution of each variable.

Note that the first round of imputations, i.e. the first iteration, is slightly different, as mentioned above in the text "... from the previous iteration *if there is one*". Raghunathan et al. (2001) breaks down the first iteration in detail. The joint conditional density of $Y_1, Y_2, \ldots, Y_p$ given X can be factored as

$$\begin{aligned} &f\left(Y_1, Y_2 \ldots, Y_p|X, \theta_1, \theta_2 \ldots, \theta_p\right) = \\ &\quad f_1\left(Y_1|X, \theta_1\right) f_2\left(Y_2|X, Y_1, \theta_2\right) \ldots f_p\left(Y_p|X, Y_1, Y_2, \ldots, Y_{p-1}, \theta_p\right) \end{aligned}$$

where $f_1, \ldots, f_p$ are the conditional density functions and θ_j is a vector of parameters in the respective conditional distribution. So the first iteration of the SRMI procedure conditions only on the data that has been filled in already in that iteration.

When the missing data have a non-monotone pattern, the target distribution is the joint conditional distribution of Y_{mis} and θ given Y_{obs}, $P(Y_{mis}, \theta|Y_{obs})$. Simulating from this distribution can be done using the MCMC method, as given by Zhang (2003), which proceeds as follows:

1. Replace the missing data Y_{mis} by some assumed values.
2. Simulate θ from the resulting completed data posterior $P(\theta|Y_{obs}, Y_{mis})$. Let $\theta^{(t)}$ be the current simulated value of θ from this complete data posterior distribution.
3. (Imputation or I-step): The next iterative sample of Y_{mis}, namely $Y_{mis}^{(t+1)}$, can be drawn from the conditional predictive distribution of Y_{mis} given Y_{obs} and $\theta^{(t)}$:
$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis}|Y_{obs}, \theta^{(t)}\right)$$
4. (Posterior or P-step): Conditioning on $Y_{mis}^{(t+1)}$, the next simulated value of θ can be drawn from its completed data posterior distribution,
$$\theta^{(t+1)} \sim P\left(\theta|Y_{obs}, Y_{mis}^{(t+1)}\right).$$
5. Repeating steps 3 and 4 from a staring value of θ, say, $\theta^{(0)}$, yields a Markov chain $\{(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2 \ldots\}$. The stationary distribution is the target distribution, $P(Y_{mis}, \theta|Y_{obs})$.

Consequently, the marginal stationary distributions of the subsequence $\{\theta^{(t)} : t = 1, 2 \ldots\}$ and $\{Y_{mis}^{(t)} : t = 1, 2 \ldots\}$ are the observed data posterior distribution $P(\theta|Y_{obs})$ and the posterior predictive distribution $P(Y_{mis}|Y_{obs})$ respectively. When t is sufficiently large, $\theta^{(t)}$ can be viewed as a single simulation from the observed data posterior distribution $P(\theta|Y_{obs})$ and $Y_{mis}^{(t)}$ can be viewed as a single simulation from the posterior predictive distribution $P(Y_{mis}|Y_{obs})$.

Traditionally, the regression models for the incomplete variables Y_j take one of the following forms (He & Raghunathan 2009):

1. a Normal linear OLS if Y_j is continuous;
2. a logistic regression if Y_j is binary;
3. a polytomous or generalised logit regression if Y_j is categorical with more than two categories;
4. a Poisson loglinear model if Y_j contains count data;
5. a two-stage model if Y_j is semi-continuous, with a logistic regression used to model the zero/non-zero status of Y_j, and a Normal OLS regression to model the value of Y_j if it is non-zero.

Imputations can be made from these models using one of a variety of software packages (He & Raghunathan 2009). For more information on the modelling and simulation procedures from these standard models, see Raghunathan et al. (2001). Brand (1998) provides a non-parametric nearest neighbour method that can be used within SRMI.

2.2. Non-Normal errors in the imputation regressions

He & Raghunathan (2009) contribute to the SRMI area by assessing several methods of Normality-based SRMI when the underlying conditional distributions of the variables are non-Normal. He & Raghunathan's (2009) study is important for the present paper which is concerned with non-Normal errors in SRMI. In a simulation study, He & Raghunathan (2009) assess the following sequential imputation methods when these methods are applied to data that is non-Normal, with missing values that are MCAR:

- Sequential Normal linear regressions
- Predictive mean matching (PMM)
- Local residual draw (LRD)
- Adjustment of Normal regression by sampling from observed residuals (or expanded residual draw, ERD)
- Adjustment by fitting Tukey's g-and-h distribution to errors

Each of these imputation methods are applied sequentially and multiply (with each dataset imputed five times) on the following simulated data, with 20% missing values generated completely at random: $Y_1 \sim U(0, 2)$, $Y_2 = 1 + Y_1 + \epsilon_2$, and $Y_3 = 1 + Y_1 + Y_2 + \epsilon_3$. The authors consider two sets (one with less variation and one with more variation) of each of the following distributions for each of ϵ_2 and ϵ_3: Lognormal, centred Student's t, and Uniform. Their study's results show that all of the methods are reasonable for estimating means, proportions, and coverage rates (although the sequential Normal method is the worst for the proportions). However, when estimating a regression coefficient for a regression on the completed data, all of the methods are left wanting when the ϵ's follow the distributions with the wider variances, although the Normal distribution often seems quite robust to the misspecification. The key conclusion from this study is that it is extremely important for a researcher to analyse the incomplete data thoroughly before applying an imputation method, since it is shown that simply applying a regular Normal method (even one

adjusting from non-Normal errors) might not be adequate for a particular estimation procedure in the presence of errors with non-Normal distributions and large variances.

3. A New Robust Sequential Regression Model

3.1. Introduction

Investigating the literature on MI and SRMI suggests that there is a need for a robust model within SRMI that can handle heavy-tailed and possibly skew data. Such a model could be chosen as a default within an SRMI routine instead of the Normal regression model, because this default model would be able to handle non-Normal errors (including heavy tails and skewness).

One model which could fill the role of a robust sequential regression model is the Student's t-distribution. With heavier tails than the Normal distribution, and the possibility of incorporating a skewness parameter, the t-distribution model could serve as a robust counterpart to the Normal OLS regression model (even with the PMM, LRD and ERD adaptations included). If the errors are indeed Normal, then this robust model will be able to reduce to the Normal case by increasing the degrees of freedom of the t-distribution. In this paper, the t model will be built using a Bayesian paradigm.

The objective of this paper is to show that the skew t-distribution in SRMI can reproduce the the error distribution under a variety of Normal and non-Normal symmetric and skew specifications. Additionally, beyond simply replicating the original distributions, we would like to show that the imputations made from the skew t-distribution have good coverage of the original data points that are made missing.

4. The Skew Student t-Distribution

We follow the setup presented in Fonseca, Ferreira & Migon (2008, p. 326). Consider a linear regression model in which an observation vector $y = (y_1, \ldots, y_n)'$ satisfies

$$y = X\beta + Z\delta + \epsilon$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are the regression coefficients, δ is a skewness parameter, Z is a diagonal matrix with diagonal elements $z_i > 0$, $i = 1, 2, \ldots, n$ as skewness coefficients, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is the error vector and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. according to the Student-t distribution with location zero, scale parameter σ and ν degrees of freedom. Here $X = [x_1, \ldots, x_n]'$ is the $n \times p$ matrix of explanatory variables and is taken to be full rank p. We denote the model parameters by $\theta = (\beta, \delta, \sigma, \nu) \in \mathbb{R}^{p+1} \times (0, \infty)^2$. The likelihood function is given by:

$$L(\beta, \sigma, \nu | y, X) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)^n \nu^{n\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)^n \pi^{n/2} \sigma^n} \prod_{i=1}^{n} \left[\nu + \left(\frac{y_i - x_i'\beta - \delta z_i}{\sigma}\right)^2\right]^{-(\nu+1)/2}. \tag{1}$$

The likelihood for the t-distribution given in Equation 1 can be restructured as follows:

$$L \propto \prod_{i=1}^{n} \left(\frac{\lambda_i \tau}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{\tau}{2}\left(y_i - x_i'\beta - \delta z_i\right)^2\right] \times \prod_{i=1}^{n} \left[\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \lambda_i^{\nu/2-1} \exp\left(\frac{-\nu\lambda_i}{2}\right)\right] \tag{2}$$

where $\tau = \sigma^{-2}$ and the λ_i are weights indicating the influence of each observation on ν. Integrating out the λ_i in Equation 2 yields Equation 1.

4.1. Fitting the skew t-distribution

When the t-distribution is used for errors on the posterior predictive distribution, generating the imputations is simply a matter of applying the posterior-drawn regression parameters to the covariates and adding an appropriate t error. The challenge is to find the degrees of freedom for this error. This involves a Gibbs sampling process for the parameters β, τ, $z_i, i = 1, \ldots, n$, δ, $\lambda_i, i = 1, \ldots, n$, and ν, while ν itself is drawn via a Metropolis-Hastings algorithm in each step of the Gibbs sampler. The Gibbs sampler requires the formulation of the conditional posterior distributions for each of the parameters of the model.

For each observation $i, i = 1, \ldots, n$, and covariate $q, q = 0, 1 \ldots, p$, $\tilde{y}_{iq} = y_i - \beta_{-q} X_{-q} - \delta z_i$, with $-q$ representing all variables in X besides variable q. In other words, for $q = 0$:

$$\tilde{y}_{i0} = y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip} - \delta z_i$$

For $q = 1$:

$$\tilde{y}_{i1} = y_i - \beta_0 - \beta_2 x_{i2} - \beta_3 x_{i3} - \ldots - \beta_p x_{ip} - \delta z_i$$

For $q = 2, \ldots, p$:

$$\tilde{y}_{iq} = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_{q-1} x_{i(q-1)} - \beta_{q+1} x_{i(q+1)} - \ldots - \beta_p x_{ip} - \delta z_i$$

Finally, for $q = p$:

$$\tilde{y}_{ip} = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_{p-1} x_{i(p-1)} - \delta z_i$$

We also define $\tilde{\tilde{y}}_i = y_i - \beta x_i - \delta z_i$ separate to $\hat{y}_i = y_i - \beta x_i$, where x_i is the i^{th} row of the data matrix, corresponding to the covariates for observation i.

With skewness a part of the $\tilde{y}_{iq}$, the same conditional distributions exist for the β_q:

$$\beta_q | y, \beta_{-q}, \tau, \Lambda \sim$$
$$N\left\{\left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1} \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq} \tilde{y}_{iq} + \frac{\mu_{\beta_q}}{\sigma_{\beta_q}^2}\right), \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1}\right\},$$

where x_{iq} is element (i, q) of the data matrix X (and when $q = 0$, $x_{i0} = 1$ for all i), and μ_{β_q} and $\sigma_{\beta_q}^2$ are the conjugate Normal prior mean and variance for β_q respectively. Note that the matrix Λ is the diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_n$. Once again, $\mu_{\beta_q} = 0$ and $\sigma_{\beta_q}^2 = 10000$.

For τ, we have that:

$$\tau | y, \beta, \Lambda \sim \Gamma\left\{\frac{n}{2} + a_\tau, \left(\frac{1}{2} \sum_{i=1}^{n} \lambda_i \tilde{\tilde{y}}_i^2 + 2b_\tau\right)^{-1}\right\},$$

where a_τ and b_τ are the conjugate Gamma prior parameters for τ.

The conditional posterior for the $z_i, i = 1, \ldots, n$ is derived to be:

$$z_i|y, \beta, \tau, \delta, \Lambda \sim N\left\{\left(\tau\lambda_i\delta^2 + 1\right)^{-1}\tau\lambda_i\delta\hat{y}_i, \left(\tau\lambda_i\delta^2 + 1\right)^{-1}\right\} I\left(Z_i > 0\right),$$

where $I\left(Z_i > 0\right)$ is an indicator function to ensure that only positive z_i exist (in order to make sense of the sign of the skewness parameter δ).

The conditional posterior distribution of the skewness parameter, δ, is given can be shown to be:

$$\begin{aligned}&\delta|y, \beta, \tau, \Lambda, z_1, \ldots, z_n \sim \\ &N\left\{\left(\tau\sum_{i=1}^{n}\lambda_i z_i^2 + \frac{1}{\sigma_\delta^2}\right)^{-1}\left(\tau\sum_{i=1}^{n}\lambda_i z_i\hat{y}_i + \frac{\mu_\delta}{\sigma_\delta^2}\right), \left(\tau\sum_{i=1}^{n}\lambda_i z_i^2 + \frac{1}{\sigma_\delta^2}\right)^{-1}\right\},\end{aligned}$$

where μ_δ and σ_δ^2 are the conjugate Normal prior parameters for δ.

For the λ_i, it can be shown that

$$\lambda_i|y, \beta, \tau, \nu, \delta, z_1, \ldots, z_n \sim \Gamma\left\{\frac{1}{2}\left(\nu + 1\right), \left[\frac{1}{2}\left(\tau\tilde{\tilde{y}}_i^2 + \nu\right)\right]^{-1}\right\},$$

with the skewness built into the distribution by replacing $\hat{y}_i$ with $\tilde{\tilde{y}}_i$.

The posterior for ν, conditional on Λ, and its priors, are given in the following equations.

$$p\left(\nu|y, \Lambda\right) \propto \frac{\nu^{\frac{1}{2}\nu n}}{2^{\frac{1}{2}\nu n}\left[\Gamma\left(\frac{\nu}{2}\right)\right]^n}\left|\Lambda\right|^{\frac{1}{2}\nu - 1}\exp\left[-\frac{1}{2}\nu\sum_{i=1}^{n}\lambda_i\right]p\left(\nu\right), \tag{3}$$

with the prior on ν taking one of four forms, namely the truncated exponential[2], the Independence Jeffrey's prior, the probability matching prior or reference priors for the orders $\left(\nu, \mu, \sigma^2\right)$, $\left(\nu, \sigma^2, \mu\right)$, and $\left(\mu, \nu, \sigma^2\right)$, and the reference priors for the orders $\left(\mu, \sigma^2, \nu\right)$, $\left(\sigma^2, \mu, \nu\right)$, and $\left(\sigma^2, \nu, \mu\right)$. For the sake of brevity, and because it is not the purpose of this paper to compare performance of priors, we will use the established truncated exponential prior:

$$p\left(\nu\right) \propto e^{-\nu\xi}, \nu > 2, \xi = 0, 1. \tag{4}$$

Working with the natural log posterior and log priors, which is easier, we have,

$$\begin{aligned}\log\left(p\left(\nu|y, \lambda\right)\right) \propto &\frac{1}{2}\nu n\log\left(\nu\right) - \frac{1}{2}\nu n\log\left(2\right) - n\log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ &- \left(\frac{1}{2}\nu - 1\right)\sum_{i=1}^{n}\log\left(\lambda_i\right) - \left(\frac{1}{2}\nu - 1\right)\sum_{i=1}^{n}\lambda_i - \nu\xi.\end{aligned} \tag{5}$$

The algorithm for the Gibbs sampler (and Metropolis sampler for ν) when we wish to incorporate skewness into the imputation model utilises the conditional distributions listed above.

4.2. Simulating from the predictive posterior distribution

The Gibbs sampler described above allows one to draw a single set of parameters that is used to generate a response prediction based on a new set of observed covariate values. By using the single

[2]This distribution is truncated so that the mean and the variance exist (Sahu, Dey & Branco 2003).

draw of each parameter in the model for the data, and then drawing with error, one effectively draws from the predictive posterior of the data.

This is the procedure that is followed within SRMI: the skew t-distribution regression model is fitted to the observed data, the Gibbs sampler (eventually, after burn-in) provides a single draw of each of the parameters from the approximate joint posterior of the parameters, and this parameter set is used to generate a prediction (with error) for the responses that are missing (but whose covariates are complete).

5. Simulation Methodology

The simulation study presented in this research will be an analysis of the robustness of a misspecified sequential imputation method based on the t-distribution (and its skew specification), as a continuation of the work presented by He & Raghunathan (2009). The objective of this study is to find a Normal-family imputation method is robust in the presence of non-Normal data.

The simulation study will also evaluate the situations where predictive mean matching (PMM), local residual draw (LRD), and expanded residual draw (ERD) can reduce bias in SRMI procedures. While these methods have already been tested for Normality-based SRMI by He & Raghunathan (2009), it would be of interest to see if they remain useful when the symmetric t-distribution is used in SRMI, and to see if these adaptations can compare in effectiveness to the skew specification of the t-distribution.

5.1. Assessment Methods

The purpose of this paper is to use the robust model in SRMI to replicate the original simulated data after it has been made incomplete. The overall analysis of multiple completed data sets is unnecessary, so this paper will refrain from running these post-imputation analyses and computing $RBIAS$ and $RRMSE$ as was done by He & Raghunathan (2009). The only results that need to be assessed are the fit of the completed data to the original data and the fit of the imputation draws to the values that were made missing.[3] This requires the construction of two quantile-quantile (QQ) plots, and a statistic to measure the deviance of these plots from the optimal solution. For an imputation method to be robust, the model should replicate the original data and predict plausible imputations.

1. *Firstly*, for one data scenario (with $n = 200$) and one MDM, a plot of the quantiles of the completed data (for each variable with missingness) is drawn against the quantiles of the original data (for each corresponding original complete variable). Since MI creates multiple completed datasets and the overall analyses after MI are averaged over these multiple completed data sets to obtain a final estimate, a 'pooling' procedure is followed when calculating the quantiles of the completed data — the five MI completed data sets for a particular MI method are pooled before the quantiles are calculated.

   For each variable with missingness, the mean squared error (MSE) of the deviation of the quantiles of the completed data from the quantiles of the original data is then computed.

[3] For an imputation method to be robust, the model should replicate the original data and predict plausible imputations.

Additionally, the MSE of the QQ plot for the incomplete (INC) data and complete-case[4] (CC) data are also calculated for comparison. Across multiple simulations within a data scenario and MDM, a distribution of QQ plot MSE calculations is then obtained. The average of these MSE calculations for an imputation method are reported for each data scenario and MDM combination.

This assessment allows one to compare post-imputation distributions with the original data distributions, as well as with the distributions under the incomplete data and the data set where incomplete observations are deleted.

2. *Secondly*, for each data scenario (with $n = 200$) and MDM combination, 200 multiply imputed data sets are created under each SRMI model. For each variable with missingness, the 1%, 2%, ..., 99% equal-tail coverage intervals of the imputed values are calculated. The proportion of the original data points that fall inside their 1%, 2%, ..., 99% imputed intervals is then determined. For an imputation method that perfectly replicates the original data, one should find that, for one variable with missingness, p% of the original data points that were made missing should fall within the p% imputation intervals for that data point. The MSE of the QQ plot of these coverage intervals from the 45° line is reported.

   This assessment allows one to make sure that the imputation model is predicting individual data points within expected intervals.

5.2. Simulated data

Complete data is generated under four different data scenarios. The data is then made incomplete using alternating MCAR and MAR mechanisms, and re-filled using various SRMI models, namely the Normal and t, with their PMM, LRD and ERD adaptations for skewness, as well as the skew t model.

5.2.1. Data Scenarios

In this study, simulated data consists of four variables, Y_1, Y_2, Y_3, and Y_4, where, $Y_1 = \epsilon_1$, $Y_2 = 1 + Y_1 + \epsilon_2$, $Y_3 = 1 + Y_1 + Y_2 + \epsilon_3$, and $Y_3 = 1 + Y_1 + Y_2 + Y_3 + \epsilon_4$.

The complete-data model errors take one of four sets of forms, namely, symmetric Normality, moderate tails and with skewness, heavy tails and with skewness, mixed Tukey's gh distributions[5] and, finally, extreme deviation from Normality (with larger error variances), a scenario replicated and expanded on from those created in He & Raghunathan (2009).

Details on the exact distributions placed on each error, $\epsilon_1, \ldots, \epsilon_4$ are given in Appendix A.

5.2.2. Missing Data Mechanisms

In this study, two MDMs are simulated, namely one MCAR mechanism one MAR mechanism. For the MCAR mechanism, every data point has a 20% chance of being deleted in one simulation. This

[4] Complete-case data or case-deleted data simply removes all incomplete observations from the data set

[5] For more information on the forms of the gh distribution that were chosen, see He & Raghunathan (2006).

does not guarantee 20% missingness, but, since the MAR mechanism does not either, this point is moot.

For the MAR MDM, a logistic regression is sequentially applied to each variable to generate a probability for each observation in the current variable to be missing, based on the values of the values of the previous variables. For more detail on the mechanism, see the Appendix B.

Across 100 simulations of these MDMs we have around 20% missingness for each variable individually under the MCAR MDM and around 49% under this MDM for CC, and around 17% missingness for each variable under individually under the MAR MDM and around 43% under this MDM for CC.[6] These values are given in Table 1 in Appendix C.

To ensure the the MDMs are indeed MCAR and MAR, we examine the difference between the mean of the original complete data and the mean of the incomplete data. The results are given in Table 2 in Appendix C.[7] In summary, the MCAR MDM is making no difference to the mean of the data, while the MAR MDM results show that the mean of the incomplete data is higher than that of the original complete data. This means the MAR MDM is successfully weeding out smaller values in the data set.

6. Simulation Analysis

6.1. Distributional coverage

In Appendix C, Tables 3 to 7 give the MSE of the QQ plots comparing the quantiles of the incomplete and completed data with the quantiles of the original data. Lower numbers are more desirable. The best result for each variable (per data scenario and MDM) is highlighted in bold, while methods with MSEs within 5% of the best method are italicised.

To summarise all of the tables we can perform the following rank analysis. If we rank the MSEs for each imputation method under a particular variable, MDM, and data scenario, we are left with 30 ranks for each method under each complete data sample size. These ranks are summarised in a boxplot, Figure 1, sorted by mean rank.

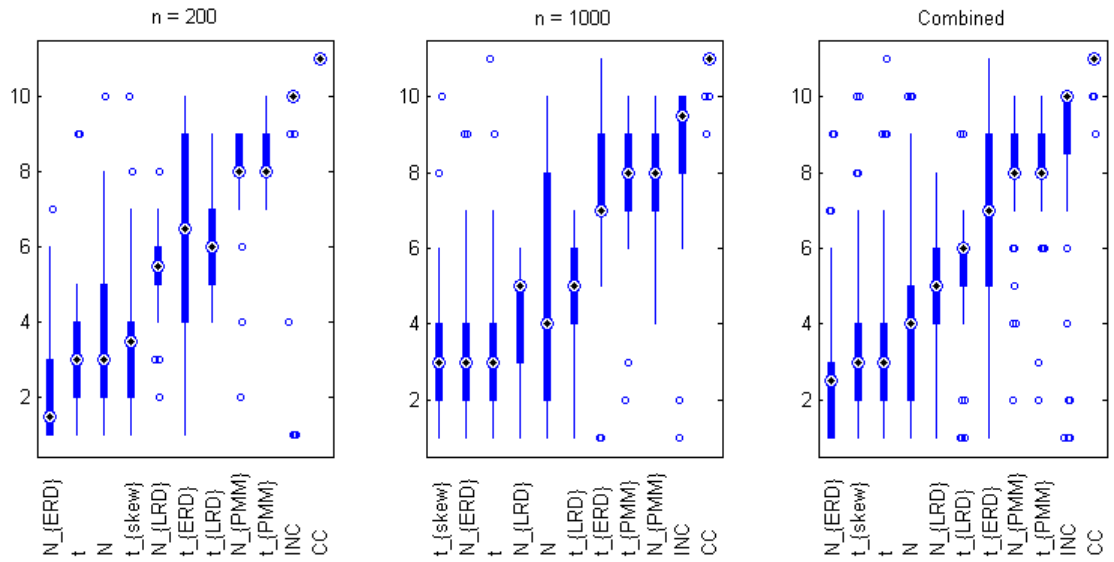
In this graphic, it is clear that the skew t model performs well under large sample sizes, and only the Normal model with ERD seems to perform better in general. However, since this is merely a crude summary of all the simulation scenarios, it may be necessary to look at the tables in more detail.

It is clear from the tables that under both MCAR and MAR MDMs, complete case data have distributions that deviate the most from the original data, although this difference is somewhat muted under the MCAR mechanism (as expected).

Under the assumption of Normal errors in Data Scenario 1, Normal, Normal with ERD, and the skew t imputation models perform the best under both the MCAR and MAR MDMs. In Data Scenario 2, with moderate t errors, the Normal model with ERD performs well, but only when $n = 200$. The Normal, t and skew t models perform well under both MDMs, both choices of N, and for all incomplete variables. Under Data Scenario 3, the Normal model and Normal model

[6] If $n = 1000$ observations are simulated, then there is no discernible difference in the missingness figures.
[7] If $n = 1000$ observations are simulated the results are very similar; no marked differences are notable.

Figure 1: Boxplots of QQ MSE ranks across variables, MDMs and data scenarios



with ERD generally perform well when $n = 200$. The t and skew t imputation models consistently perform adequately. Looking at Data Scenario 4, several imputation models perform well, including the Normal and the Normal with ERD, the t, and the skew t. The LRD adaptations of both the Normal and the t models are also adequate. These results all hold for both sample sizes. Data Scenario 5 holds mixed results. Strangely, no model is able to replicate the distribution of the original data's Y_2 better than the incomplete data. The error on this variable was Lognormal, so further investigation into this might be warranted in future research. Amongst the imputation models, the best performers for replicating the original distribution of Y_2 are the Normal and t models with the PMM and LRD adaptations. For Y_3, the t and skew t models perform well, together with the Normal and t models incorporating the LRD adjustment. For Y_4, once more the incomplete data seems close in distribution to the original data, while the Normal model with ERD, the t model, and the skew t model perform best amongst the imputation models.

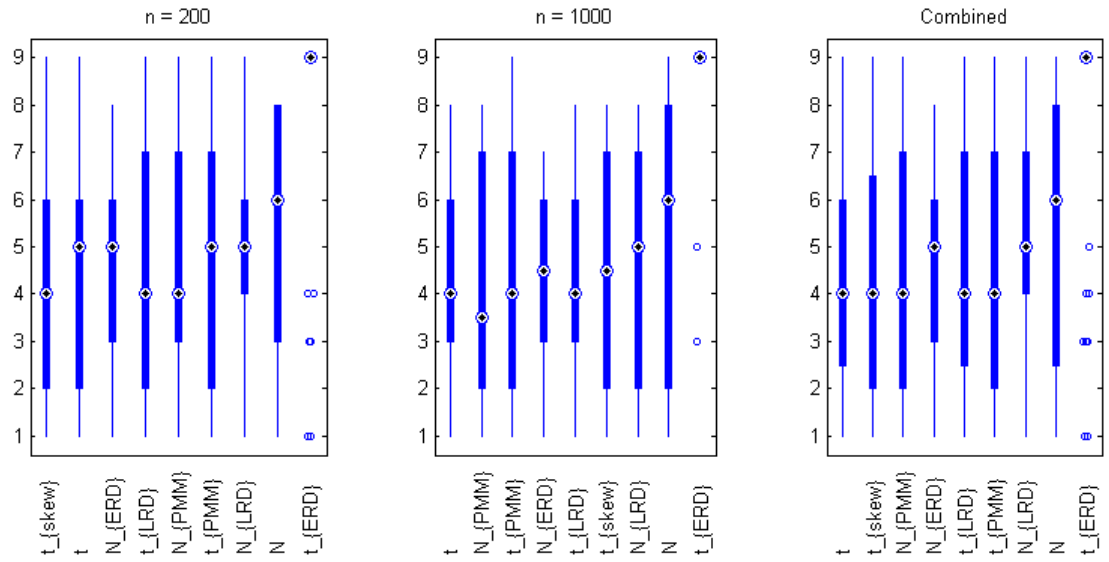
Across all models, one may argue that the skew t model is the most robust. This model shows fewer weaknesses than the other models, while always remaining relatively close to the performance of the best imputation model if it is not the best model itself. It can also be noted that when the distributional assumptions of the errors are less pronounced, *i.e.* when $n = 200$, it is more difficult to choose a better imputation model. However, with $n = 1000$, we are more able to gauge the effectiveness of some of the imputation models, for example, the skew t model.

Before continuing with the second analysis, it is important to keep in mind that the errors across the three variables are uncorrelated, and that a certain amount of 'averaging' of errors across an observation may or may not have allowed less robust models to appear better than they are. However, this should not concern us too much, since it is clear that even with this advantage, the traditionally less robust models appear are still shown to be less robust than the skew t model.

6.2. Imputation coverage

Tables 8 to 12 provide the MSEs of the QQ plots that compare the coverage of the actual distribution of the imputations over the original values of the data that was made missing. Once again, lower

Figure 2: Boxplots of imputation coverage MSE ranks across variables, MDMs and data scenarios



numbers are more desirable, and the best result for each variable (per data scenario and MDM) is highlighted in bold, while methods with MSEs within 5% of the best method are italicised. Take note that the maximum MSE for a method across the three variables is also given as a measure of the worst error the method made within the given data scenario and sample size.

Once again, the ranks of these MSEs are summarised in a boxplot, Figure 2, sorted by the mean of the ranks for each method.

While little information can be gained from this crude ranking analysis, one can at least note that the symmetric and skew t models are performing far better than the Normal model in general, and that the t model with ERD is provides terrible imputation coverage intervals.

One can look at the Tables 3-7 in Appendix C in more detail, but there still seems to be no systematic evidence of one imputation method providing more accurate coverage of the original data points before these were made missing.

The only result that is systematically clear in the tables is the result already seen in Figure 2; that the t model with the ERD adjustment is inadmissable as an imputation method. The errors that the algorithm has to 'donate' to the imputations are simply too wide, too often.

By examining averages across scenarios and MDMs, one will find that the Normal model with either PMM, LRD or ERD adjustments, along with the t model with PMM or LRD adjustments, all perform better than the unadjusted Normal, unadjusted t, and the skew t models when $n = 1000$. However, when $n = 200$, the skew t model provides the best coverages on average, followed by the t model. This is also the case if we average the MSEs across sample size. Certainly, the unadjusted t and the skew t models generally perform better than the unadjusted Normal model.

According to maximums across scenarios, sample sizes, and MDMs, the Normal with PMM, LRD, or ERD, and the t with PMM or LRD seem to provide better coverages.

The excellent performance of the models with modest adjustments for skewness, namely the PMM, LRD, and ERD on Normal errors, may suggest that in the context of imputation, it suffices to not be able to draw errors outside of the realm of observed errors in the data set. If, however, it is

important to allow for errors in the imputations that are wider than the existing observed errors (for example, if extreme proportions of the data sets are missing), then one could assume that these adaptations to the symmetric models will not suffice.

In conclusion, the second analysis shows that the most accurate of the pure distribution-based imputation methods is the skew t model, followed by the symmetric t model. If adaptations to incorporate observed skewness are deemed suitable for the data set, the Normal model with any adaptation (PMM, LRD, or ERD) will suffice, and will significantly reduce computation time.

7. Conclusion

It is clear that at the very least Normal SRMI should be used instead of incomplete data analysis or CC data analysis. The Normal model has proved to be relatively robust to misspecification within the SRMI approach to the extent of the data variations presented in this paper when compared with complete case or incomplete data analysis.

However, an imputer can make a better choice of SRMI model based on the results of this paper. Often it seems that the Normal model with a PMM, LRD or ERD adaptation on the imputed errors will suffice in order to accommodate observed skewness. The advantage of this is the quicker computation time, since the t and skew t models are much more complex in their implementation. It also seems as though the t models with PMM or LRD adaptations are rather robust. The t model with ERD is not recommended due to it's generation of errors that lead to poor coverage of imputations over the original data points before they were made missing. Unfortunately, the t model with any adaptation does not share the computational simplicity of the Normal model, with or without adaptations.

If one prefers to allow for errors in imputations that are outside of the limits of the errors that are actually observed, a more robust adaptation-free approach should be considered. If this is the case, the obvious choice is the skew t approach. In this paper this method has been shown to be favourable under many of the simulation scenarios. Moreover, it would seem that the skew t model, while not always being the best choice of imputation model, has shown no serious weaknesses in the context of this simulation study even when compared to the adjusted Normal models. The skew t model is, therefore, an acceptable choice of imputation model should the error distributions in the data not be known. The disadvantage of increased computation time (compared with the Normal model) is more than offset by the model's flexibility.

References

Ardington, C., Lam, D., Leibbrandt, M. & Welch, M. (2006), 'The sensitivity to key data imputations of recent estimates of income poverty and inequality in south africa', *Economic Modelling* **23**, 822–835.

Brand, J. P. L. (1998), Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets, PhD thesis, Erasmus University, Rotterdam.

Fay, R. E. (1992), When are inferences from multiple imputation valid?, *in* 'Proceedings of the Survey Research Methods Section', American Statistical Association, Alexandria, VA., pp. 227–232.

Fonseca, T. C. O., Ferreira, M. A. R. & Migon, H. S. (2008), 'Objective bayesian analysis for the student-t regression model', *Biometrika* **95**(2), 325–333.

He, Y. & Raghunathan, T. E. (2006), 'Tukey's *gh* distribution for multiple imputation', *The American Statistician* **60**, 251–256.

He, Y. & Raghunathan, T. E. (2009), 'On the performance of sequential regression multiple imputation methods with non normal error distributions', *Communications in Statistics—Simulation and Computation* **38**, 856–883.

Kennickell (1991), Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation, *in* 'Proceedings of the Survey Research Methods Section', American Statistical Association, pp. 112–121.

Meng, X.-L. (1994), 'Multiple-imputation inferences with uncongenial sources of input', *Statistical Science* **9**(4), 538–558.

Nielson, S. F. (2003), 'Proper and improper multiple imputation', *International Statistical Review* **71**(3), 593–607.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. & 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. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Rubin, D. B. (1978), Multiple imputation in sample surveys — a phenomenological bayesian approach to nonresponse, *in* 'Proceedings of the Survey Research Methods Section', American Statistical Association, Washington, D.C., pp. 20–34.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

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

Rubin, D. B. (2003), 'Discussion on multiple imputation', *International Statistical Review* **71**(3), 619–625.

Sahu, S. K., Dey, D. K. & Branco, M. D. (2003), 'A new class of multivariate skew distributions with applications to bayesian regression models', *The Canadian Journal of Statistics* **31**(2), 129–150.

Saunders, J. A., Morrow-Howell, N., Spitznagel, E., Doré, P., Proctor, E. K. & Pescarino, R. (2006), 'Imputing missing data: A comparison of methods for social work researchers', *Social Work Research* **30**(1), 19–35.

Schafer, J. L. (2003), 'Multiple imputation in multivariate problems when the imputation and analysis methods differ', *Statistica Nederlandica* **57**(1), 19–35.

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

van Buuren, S. (2007), 'Multiple imputation of discrete and continuous data by fully conditional specification', *Statistical Methods in Medical Research* **16**, 219–242.

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

Zhang, P. (2003), 'Multiple imputation: Theory and method', *International Statistical Review* **71**(3), 581–592.

Appendix A Data Scenarios

1. **Normality (and symmetry)**:

$$\epsilon_j \sim N(0,1) \text{ for } j = 1,2,3,4.$$

2. **Moderate tails, with skewness**:

$$\begin{aligned}
&\epsilon_1 \sim N(0,1)\\
&\epsilon_2 \sim t_6\\
&\epsilon_3 = \alpha_3 - \omega_3 \text{ where } \alpha_3 \sim t_6 \text{ and } \omega_3 \sim N(0,1)\\
&\epsilon_4 = \alpha_4 - 2\omega_4 \text{ where } \alpha_4 \sim t_6 \text{ and } \omega_4 \sim N(0,1)
\end{aligned}$$

3. **Heavy tails, with skewness**:

$$\begin{aligned}
&\epsilon_1 \sim N(0,1)\\
&\epsilon_2 \sim t_3\\
&\epsilon_3 = \alpha_3 - \omega_3 \text{ where } \alpha_3 \sim t_3 \text{ and } \omega_3 \sim N(0,1)\\
&\epsilon_4 = \alpha_4 - 2\omega_4 \text{ where } \alpha_4 \sim t_3 \text{ and } \omega_4 \sim N(0,1)
\end{aligned}$$

4. **Mixed *gh* distributions**: Again $\epsilon_1 \sim N(0,1)$. For the remaining error distributions ϵ_2, ϵ_3, and ϵ_4, various possibilities of Tukey's gh distribution are chosen.

   For all of the errors in data scenario 4, $\mu = 0$ and $\sigma = 1$. However, the g and h parameters are varied as follows:

   - For ϵ_2, $g = 1$ and $h = -0.25$. This creates a downward-sloping monotonic exponential-type distribution.
   - For ϵ_3, $g = 0.75$ and $h = 0.25$. This generates a right-skewed distribution.
   - For ϵ_4, $g = 1$ and $h = 0$. This is the well-known Lognormal distribution.

5. **Extreme deviation from Normality**: In this data scenario, extreme deviation from Normality, and larger error variances are generated. Let the vector of errors for Y_j, $j = 1, \ldots, 4$ be $\xi_j = [\epsilon_{1j}\ \epsilon_{2j}\ \epsilon_{3j}\ \ldots\ \epsilon_{nj}]'$. Also, let $U_j = [u_{1j}\ u_{2j}\ u_{3j}\ \ldots\ u_{nj}]'$. For each observation i, $i = 1, \ldots, n$, the error ϵ_{ij} for the this observation on each variable is constructed as follows:

- $\xi_1 \sim N(0,1)$
- $\epsilon_{i2} = \frac{u_{i2}-E(U_2)}{Var(U_2)} \times \sqrt{3Var(Y_1)}$, $u_2 = 1 + \exp(1+Z)$, $Z \sim N(0,1)$. So ξ_2 is a vector of centred and widely scaled Lognormal errors.
- $\epsilon_{i3} = \frac{u_{i3}-E(U_3)}{Var(U_3)} \times \sqrt{2\left[Var(Y_1) + Var(Y_2)\right]}$, $u_3 = W$, $W \sim t_3$. So ξ_3 is a vector centred and widely scaled t_3 errors.
- $\epsilon_{i4} = \frac{u_{i4}-E(U_4)}{Var(U_4)} \times \sqrt{Var(Y_1) + Var(Y_2) + +Var(Y_2)}$, $u_4 = W - 2Z$, $W \sim t_3$, $Z \sim N(0,1)$. So ξ_4 is a vector centred and widely scaled right-skewed t_3 errors.

This arrangement of error distributions is arguably the most extremely deviated from Normality of the five data scenarios.

Appendix B MAR MDM

- Y_1 is complete.
- The probability that observation i is missing in Y_2 is:

$$p_{i,2} = 0.4\left[1 + \exp(-0.3 - 0.3y_{i,1})\right]^{-1}$$

Once these probabilities are calculated, each observation with probability less that an independent draw from a $U(0,1)$ distribution is made missing.

- The probability that observation i is missing in Y_3 is:

$$p_{i,3} = 0.4\left[1 + \exp(-0.3 - 0.3y_{i,1} - 0.3y_{i,2})\right]^{-1}$$

If $y_{i,2}$ is already missing, this term is ignored, ensuring the MAR MDM does not become an MNAR MDM. Data points are made missing in the same way as for Y_2.

- Finally, the probability that observation i is missing in Y_4 is:

$$p_{i,3} = 0.4\left[1 + \exp(-0.3 - 0.3y_{i,1} - 0.3y_{i,2} - 0.3y_{i,3})\right]^{-1}$$

If $y_{i,2}$ or $y_{i,3}$ or both are already missing, the missing terms are ignored, ensuring once more that the MAR MDM does not become an MNAR MDM. Data points are made missing in the same way as for Y_2 and Y_3.

Appendix C Tables

Table 1: Missingness from the MCAR and MAR MDMs

MDM	**Variable/CC**	**Data scenario**				
		1	2	3	4	5
MCAR	Y2	20.1%	20.0%	20.4%	20.2%	19.6%
	Y3	19.7%	19.4%	20.3%	20.0%	20.0%
	Y4	20.1%	19.6%	20.4%	19.5%	19.7%
	CC	48.7%	48.1%	49.6%	48.7%	48.5%
MAR	Y2	16.9%	17.3%	17.0%	17.3%	17.5%
	Y3	16.9%	16.8%	17.0%	17.1%	17.4%
	Y4	17.4%	17.3%	17.6%	16.9%	17.3%
	CC	42.7%	43.1%	42.9%	42.8%	43.4%

Table 2: Difference after data is made missing

Data scenario 1					**Data scenario 2**				
MDM	**Y1**	**Y2**	**Y3**	**Y4**	**MDM**	**Y1**	**Y2**	**Y3**	**Y4**
MCAR		0.00	0.00	0.01	MCAR		-0.01	-0.01	0.01
MCAR CC	0.00	0.00	0.00	-0.01	MCAR CC	-0.01	0.00	-0.01	-0.04
MAR		-0.03	-0.11	-0.33	MAR		-0.04	-0.15	-0.39
MAR CC	-0.16	-0.20	-0.36	-0.73	MAR CC	-0.16	-0.26	-0.46	-0.91
Data scenario 3					**Data scenario 4**				
MDM	**Y1**	**Y2**	**Y3**	**Y4**	**MDM**	**Y1**	**Y2**	**Y3**	**Y4**
MCAR		0.00	-0.01	-0.02	MCAR		-0.01	0.00	-0.03
MCAR CC	0.00	0.01	0.00	0.00	MCAR CC	-0.02	-0.01	-0.03	-0.06
MAR		-0.04	-0.14	-0.40	MAR		-0.03	-0.13	-0.34
MAR CC	-0.16	-0.25	-0.47	-0.91	MAR CC	-0.19	-0.21	-0.40	-0.80
Data scenario 5									
MDM	**Y1**	**Y2**	**Y3**	**Y4**					
MCAR		0.01	-0.01	0.10					
MCAR CC	0.00	0.01	0.01	0.10					
MAR		-0.05	-0.16	-0.13					
MAR CC	-0.15	-0.25	-0.48	-0.49					

Table 3: QQ MSE for Data Scenario 1

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.006	0.002	0.007	0.002	0.027	0.005	0.037	0.023	0.085	0.017	0.190	0.153
CC	0.022	0.006	0.061	0.057	0.079	0.018	0.203	0.206	0.259	0.064	0.771	0.746
N	*0.003*	**0.001**	*0.003*	*0.001*	**0.009**	*0.002*	0.009	*0.002*	*0.024*	**0.004**	**0.019**	*0.004*
N_{PMM}	0.004	0.002	0.004	0.001	0.018	0.004	0.018	0.003	0.046	0.018	0.046	0.021
N_{LRD}	0.003	*0.001*	0.003	*0.001*	0.010	*0.002*	0.010	*0.002*	0.025	0.005	0.022	*0.004*
N_{ERD}	**0.003**	*0.001*	**0.003**	**0.001**	*0.009*	*0.002*	**0.008**	*0.002*	*0.024*	*0.004*	*0.019*	*0.004*
t	*0.003*	*0.001*	0.003	*0.001*	*0.009*	0.002	0.009	*0.002*	*0.025*	*0.004*	0.020	*0.004*
t_{PMM}	0.004	0.002	0.004	0.001	0.018	0.004	0.018	0.003	0.045	0.018	0.049	0.021
t_{LRD}	0.004	*0.001*	0.003	*0.001*	0.010	**0.002**	0.009	**0.002**	0.026	0.005	0.022	0.004
t_{ERD}	0.006	0.003	0.005	0.002	0.010	0.003	0.010	0.003	**0.024**	0.005	0.020	0.004
t_{skew}	*0.003*	*0.001*	*0.003*	*0.001*	*0.009*	*0.002*	*0.009*	*0.002*	*0.024*	*0.004*	0.021	**0.004**

Table 4: QQ MSE for Data Scenario 2

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.015	0.002	0.016	0.003	0.029	0.007	0.047	0.026	0.133	0.025	0.283	0.154
CC	0.074	0.007	0.146	0.061	0.113	0.023	0.330	0.206	0.474	0.083	1.284	0.756
N	0.013	**0.001**	0.010	*0.001*	*0.016*	*0.004*	0.014	0.004	*0.066*	**0.012**	*0.065*	0.011
N_{PMM}	0.013	0.002	0.013	0.002	0.022	0.005	0.017	0.005	0.097	0.022	0.105	0.019
N_{LRD}	0.011	0.002	0.010	0.001	0.017	*0.004*	0.015	0.004	0.076	0.013	0.072	0.012
N_{ERD}	0.012	*0.001*	0.011	*0.001*	**0.015**	0.004	**0.013**	0.003	**0.065**	0.013	**0.063**	0.011
t	**0.010**	0.002	**0.009**	*0.001*	0.017	*0.004*	*0.014*	0.004	0.070	*0.012*	*0.064*	0.011
t_{PMM}	0.013	0.002	0.012	0.002	0.021	0.005	0.016	0.005	0.096	0.021	0.104	0.019
t_{LRD}	0.011	0.002	0.010	0.001	0.017	0.004	0.015	0.003	0.075	0.014	0.075	0.011
t_{ERD}	0.011	0.002	0.011	0.001	*0.016*	**0.004**	*0.014*	**0.003**	0.073	0.021	0.071	0.017
t_{skew}	*0.010*	0.002	*0.009*	**0.001**	0.017	0.004	0.014	0.004	*0.067*	*0.012*	*0.065*	**0.010**

Table 5: QQ MSE for Data Scenario 3

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.015	0.004	0.014	0.004	0.037	0.008	0.060	0.024	0.154	0.030	0.307	0.170
CC	0.053	0.017	0.106	0.074	0.143	0.032	0.373	0.210	0.464	0.113	1.242	0.780
N	*0.010*	0.013	*0.009*	0.011	*0.028*	0.006	*0.021*	0.005	*0.077*	0.015	*0.063*	*0.012*
N_{PMM}	0.013	0.005	0.012	0.004	0.036	0.009	0.029	0.009	0.116	0.037	0.110	0.044
N_{LRD}	0.011	*0.003*	*0.010*	**0.002**	0.028	0.007	0.022	0.005	0.082	0.015	0.068	0.013
N_{ERD}	**0.010**	0.011	**0.009**	0.009	**0.027**	0.006	**0.020**	*0.004*	0.079	**0.014**	**0.060**	0.012
t	0.011	0.003	*0.009*	*0.002*	*0.027*	*0.005*	0.022	**0.004**	**0.075**	0.015	*0.061*	*0.012*
t_{PMM}	0.013	0.005	0.011	0.003	0.037	0.009	0.030	0.009	0.118	0.037	0.110	0.043
t_{LRD}	0.012	**0.003**	0.010	0.002	0.029	0.007	0.023	0.005	0.082	0.016	0.069	0.013
t_{ERD}	*0.011*	0.004	*0.009*	0.003	*0.028*	0.007	0.023	0.006	0.080	0.023	0.065	0.023
t_{skew}	0.011	*0.003*	*0.010*	*0.002*	*0.028*	**0.005**	0.022	*0.004*	0.083	*0.015*	*0.062*	**0.012**

Table 6: QQ MSE for Data Scenario 4

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.005	0.001	0.005	0.002	0.017	0.003	0.034	0.018	0.064	0.012	0.185	0.118
CC	0.016	0.003	0.058	0.040	0.057	0.011	0.214	0.154	0.221	0.042	0.848	0.615
N	*0.001*	**0.000**	**0.001**	*0.000*	**0.003**	*0.001*	*0.003*	*0.001*	*0.010*	*0.002*	0.009	*0.002*
N_{PMM}	0.003	0.001	0.002	0.001	0.008	0.003	0.006	0.002	0.030	0.011	0.025	0.009
N_{LRD}	0.001	0.000	0.001	0.000	0.003	0.001	0.003	0.001	0.010	*0.002*	0.009	0.002
N_{ERD}	**0.001**	*0.000*	*0.001*	*0.000*	*0.003*	**0.001**	*0.003*	*0.001*	**0.010**	*0.002*	**0.008**	*0.002*
t	0.001	0.000	0.001	**0.000**	0.003	*0.001*	*0.003*	*0.001*	*0.010*	**0.002**	*0.009*	**0.001**
t_{PMM}	0.002	0.001	0.002	0.001	0.008	0.003	0.006	0.002	0.030	0.011	0.024	0.009
t_{LRD}	0.001	0.000	0.001	0.000	0.003	0.001	0.003	0.001	0.010	0.002	0.009	0.002
t_{ERD}	0.004	0.003	0.004	0.002	0.006	0.002	0.005	0.002	0.014	0.003	0.012	0.002
t_{skew}	0.001	0.000	0.001	*0.000*	*0.003*	*0.001*	**0.003**	**0.001**	*0.010*	*0.002*	0.009	*0.002*

Table 7: QQ MSE for Data Scenario 5

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	**0.028**	**0.010**	**0.020**	**0.009**	0.080	*0.016*	0.114	0.029	**1.476**	0.423	**1.232**	0.632
CC	0.087	0.034	0.141	0.123	0.357	0.064	0.513	0.274	13.450	1.346	8.227	1.809
N	0.042	0.034	0.030	0.025	0.093	0.023	0.095	0.015	1.989	0.632	1.630	0.569
N_{PMM}	0.032	0.013	*0.020*	0.011	0.094	0.021	0.109	0.014	2.766	0.495	1.739	0.454
N_{LRD}	0.030	0.010	*0.020*	0.009	0.086	**0.015**	0.092	*0.012*	2.026	0.489	2.419	0.466
N_{ERD}	0.032	0.014	0.021	0.011	0.078	0.017	0.091	0.013	1.592	**0.422**	1.446	**0.431**
t	0.045	0.035	0.029	0.024	**0.066**	0.021	*0.083*	0.013	1.739	0.453	1.415	0.453
t_{PMM}	0.036	*0.010*	0.021	*0.009*	0.093	0.021	0.112	0.013	3.005	0.511	1.764	0.473
t_{LRD}	0.036	0.011	*0.021*	*0.009*	0.095	0.018	0.107	**0.012**	2.094	0.481	2.479	0.489
t_{ERD}	0.033	0.027	0.021	0.018	0.121	0.050	0.157	0.040	6.667	6.489	5.470	5.188
t_{skew}	0.045	0.035	0.027	0.023	0.075	0.020	**0.082**	0.013	1.666	*0.426*	1.502	*0.450*

Table 8: QQ MSE for imputations under Data Scenario 1

	MDM:	**MCAR**			**MAR**			
n	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	**7.3**	87.3	23.7	93.4	30.4	**13.5**	93.4
	N_{PMM}	67.6	43.4	37.7	**19.1**	21.0	112.5	112.5
	N_{LRD}	41.4	77.7	72.5	47.8	29.6	32.4	77.7
	N_{ERD}	9.7	77.9	67.9	73.3	27.5	17.4	77.9
	t	8.1	65.1	46.1	75.3	20.3	26.8	75.3
	t_{PMM}	66.6	**13.2**	48.8	44.5	29.8	157.8	157.8
	t_{LRD}	35.2	88.4	69.8	56.0	**17.7**	24.2	88.4
	t_{ERD}	349.3	58.2	**22.0**	484.0	123.8	20.7	484.0
	t_{skew}	10.5	66.8	41.8	71.1	23.8	19.1	**71.1**
1000	N	5.4	3.4	6.1	4.7	*5.9*	*11.8*	11.8
	N_{PMM}	5.0	2.7	8.5	8.2	8.2	24.9	24.9
	N_{LRD}	5.8	4.5	7.8	2.8	12.8	15.7	15.7
	N_{ERD}	*4.6*	3.1	6.1	2.0	**5.8**	19.3	19.3
	t	9.3	2.8	**2.1**	2.6	11.0	**11.5**	**11.5**
	t_{PMM}	7.0	**2.1**	10.2	4.2	7.5	27.7	27.7
	t_{LRD}	5.9	6.7	5.8	**1.6**	13.7	12.2	13.7
	t_{ERD}	295.5	95.0	4.4	262.7	59.5	39.6	295.5
	t_{skew}	**4.5**	2.7	2.5	3.5	8.2	13.7	13.7

Table 9: QQ MSE for imputations under Data Scenario 2

	MDM:	**MCAR**			**MAR**			
n	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	21.2	70.5	8.6	25.8	**16.6**	90.1	90.1
	N_{PMM}	16.4	33.0	22.7	36.6	97.4	**30.5**	97.4
	N_{LRD}	12.1	51.3	11.2	35.5	50.4	85.8	**85.8**
	N_{ERD}	10.6	52.5	8.2	39.3	36.8	116.0	116.0
	t	12.0	74.4	13.1	53.2	35.1	102.7	102.7
	t_{PMM}	21.2	61.0	17.8	**19.7**	100.5	51.0	100.5
	t_{LRD}	14.1	93.8	**5.0**	48.5	33.5	77.7	93.8
	t_{ERD}	56.5	**19.3**	515.4	32.1	100.9	164.7	515.4
	t_{skew}	**9.7**	75.0	8.1	46.5	29.4	116.5	116.5
1000	N	34.6	8.4	**13.5**	17.8	4.5	8.4	34.6
	N_{PMM}	30.7	5.5	25.8	10.7	4.6	7.1	30.7
	N_{LRD}	20.2	7.7	21.2	4.7	2.7	8.7	21.2
	N_{ERD}	27.0	7.6	18.6	14.1	4.4	6.1	27.0
	t	22.4	**3.2**	23.6	**3.9**	3.4	**3.8**	23.6
	t_{PMM}	21.2	11.6	15.8	7.3	5.8	5.4	**21.2**
	t_{LRD}	**18.3**	5.8	14.6	5.4	**2.0**	9.9	18.3
	t_{ERD}	90.5	69.1	758.3	31.2	105.5	612.8	758.3
	t_{skew}	25.9	5.0	14.6	9.2	6.4	4.1	25.9

Table 10: QQ MSE for imputations under Data Scenario 3

	MDM:	**MCAR**			**MAR**			
n	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	21.7	55.3	**23.1**	66.7	76.7	32.8	**76.7**
	N_{PMM}	**15.7**	201.6	27.4	279.4	57.2	55.8	279.4
	N_{LRD}	29.4	183.7	83.9	79.6	35.4	20.7	183.7
	N_{ERD}	23.5	86.7	51.3	67.6	67.5	14.3	86.7
	t	28.0	**51.4**	62.4	94.8	63.4	**10.9**	94.8
	t_{PMM}	*16.4*	178.2	*23.5*	263.9	33.5	30.2	263.9
	t_{LRD}	29.8	234.5	67.6	154.8	**22.3**	20.2	234.5
	t_{ERD}	43.8	392.1	583.6	**53.9**	44.4	558.8	583.6
	t_{skew}	23.1	69.3	63.3	62.1	92.7	17.4	92.7
1000	N	37.9	17.4	**4.6**	51.2	**3.1**	**2.4**	51.2
	N_{PMM}	**2.1**	**5.0**	12.1	**5.1**	11.2	6.9	12.1
	N_{LRD}	8.7	8.1	19.7	5.4	9.4	9.3	19.7
	N_{ERD}	5.9	7.8	14.8	13.0	8.1	5.1	14.8
	t	7.5	7.8	13.3	9.2	7.5	4.7	13.3
	t_{PMM}	2.7	5.4	8.5	10.4	11.3	5.2	**11.3**
	t_{LRD}	8.5	5.6	16.6	7.7	14.1	5.3	16.6
	t_{ERD}	7.6	200.9	835.6	6.4	221.8	711.8	835.6
	t_{skew}	10.5	7.9	14.0	13.5	5.8	3.5	14.0

Table 11: QQ MSE for imputations under Data Scenario 4

	MDM:	**MCAR**			**MAR**			
n	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	85.2	21.2	24.0	59.4	82.3	50.1	85.2
	N_{PMM}	23.3	18.5	153.2	47.2	**30.5**	19.6	153.2
	N_{LRD}	24.1	18.5	27.5	55.9	68.2	**10.9**	68.2
	N_{ERD}	85.8	**15.3**	**13.2**	61.8	45.2	24.7	85.8
	t	26.6	17.6	23.3	28.4	63.3	59.6	63.3
	t_{PMM}	17.7	33.3	106.0	47.1	34.5	24.5	106.0
	t_{LRD}	12.8	18.0	19.3	30.2	34.1	16.1	**34.1**
	t_{ERD}	717.5	711.0	381.2	803.0	559.4	444.8	803.0
	t_{skew}	**11.5**	24.0	31.0	**16.3**	41.5	93.2	93.2
	N	10.9	25.6	65.5	53.3	11.2	92.5	92.5
1000	N_{PMM}	5.8	**6.2**	*24.7*	**9.9**	16.9	20.4	24.7
	N_{LRD}	6.3	28.7	31.5	49.7	5.7	49.7	49.7
	N_{ERD}	5.5	13.9	28.1	43.6	5.9	36.3	43.6
	t	14.2	21.9	57.9	52.1	**4.7**	74.6	74.6
	t_{PMM}	6.3	8.4	**23.8**	11.1	16.3	**17.4**	**23.8**
	t_{LRD}	**4.9**	27.6	27.7	38.3	8.0	35.3	38.3
	t_{ERD}	599.4	532.5	208.1	646.4	607.8	233.6	646.4
	t_{skew}	15.8	22.8	67.6	55.7	6.8	74.1	74.1

Table 12: QQ MSE for imputations under Data Scenario 5

	MDM:	**MCAR**			**MAR**			
n	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	302.7	166.0	47.3	177.8	43.5	49.6	302.7
	N_{PMM}	**7.8**	57.1	25.3	42.6	21.6	54.7	**57.1**
	N_{LRD}	17.3	43.6	41.7	20.7	31.6	71.2	71.2
	N_{ERD}	119.2	50.8	19.6	22.6	**10.7**	55.7	119.2
	t	53.7	20.9	**16.7**	15.4	20.5	57.5	57.5
	t_{PMM}	125.2	17.8	32.2	38.0	32.0	**35.8**	125.2
	t_{LRD}	31.9	**13.7**	40.3	58.1	25.4	47.0	58.1
	t_{ERD}	379.0	824.3	2721.6	367.2	1272.0	2691.9	2721.6
	t_{skew}	31.3	*13.8*	21.5	**7.9**	20.5	60.2	60.2
1000	N	103.5	189.6	**2.5**	271.4	191.0	**2.7**	271.4
	N_{PMM}	14.1	**7.8**	13.9	10.1	3.5	11.7	**14.1**
	N_{LRD}	25.4	8.4	10.0	**7.4**	7.5	8.3	25.4
	N_{ERD}	**8.5**	19.5	12.8	10.5	4.9	11.6	19.5
	t	24.1	16.9	12.7	12.7	**1.4**	9.8	24.1
	t_{PMM}	20.7	11.3	12.6	9.4	2.2	17.4	20.7
	t_{LRD}	25.4	12.5	7.1	10.5	6.4	10.1	25.4
	t_{ERD}	532.4	750.4	2719.8	428.4	974.4	2506.6	2719.8
	t_{skew}	21.3	10.8	7.3	14.0	1.9	14.7	21.3