Extending the Reach of Sequential Regression Multiple Imputation

M. J. von Maltitz*

August 16, 2013

Abstract

It has been shown in numerous studies that missing data should not be ignored when survey data is being analysed. Dropping incomplete cases can lead to biases and inefficient analyses. Other schemes commonly used to handle incomplete data, such as single imputation, can lead to falsely accurate analyses and incorrect conclusions. While expert statisticians can model missing data within an overall analysis, it is necessary to separate the handling of missing data and the analysis tasks so that non-experts can obtain valid and reliable inferences from basic survey data analysis.

The missing data handling needs to incorporate the uncertainty associated with the way in which the data became missing in the first place, the uncertainty of the model with which missing data points are predicted, and the uncertainty associated with each survey observation. These uncertainties are incorporated by the multiple imputation (MI) methods.

The paper presented will review the use of an easily applied form of MI known as sequential regression multiple imputation (SRMI), and will discuss the way in which the analysis of the multiple datasets arising from MI are combined for a single set of results. The purpose of this paper is to familiarise researchers with these important (and easily implemented) methods of handling missing data, essentially extending the reach of SRMI.

*Department of Mathematical Statistics and Actuarial Science, University of the Free State, South Africa

1 Introduction

This research paper is written by a South African, for South Africans, and with a very specific goal: to encourage South African researchers to make use of multiply imputed datasets to obtain better inferences from originally incomplete data.

While proper imputation of missing data might seem quite complicated and technical, it needn't be. There are multiple imputation methods that are robust and easy to implement, with a variety of software packages available to handle multiple imputation.

However, this paper also elaborates on that idea that a researcher need not be an expert imputer, but at least needs to be able to use multiply imputed datasets within their research. If a researcher still does not want to fill in missing data themselves, this is not a problem. Multiple imputation neatly splits the imputation task from the analysis task. Researchers need only be willing to use multiply imputed datasets that expert imputers have prepared for the end-users – the researchers themselves.

Even the South African Statistical authorities, Statistics South Africa, are using imputation methods that have been shown to be, time and again, inferior to the multiple imputation methods used in numerous first world countries. In their most recent publications on the 2011 Census and the latest Quarterly Labour Survey they publish the fact that they use re-weighting and single hotdeck imputation to fill in the missing data (StatsSA 2012, StatsSA 2013).

This paper elaborates on the methods used to handle incomplete data that are deemed outdated, and summarises the correct methods that are being used today. The research also lists and briefly elaborates on published evidence that the methods used by many research authorities in South Africa today are not the right methods to use, except under extremely rare circumstances.

1.1 Incomplete Data

Incomplete data refers to any data set that contains missing values within one or more of the variables within that data set. If the missing data points are only found within a single variable, then the problem is univariate, but, most often it is the case that missing values are spread across several variables within a data set. These missing values are often the result of subjects not responding to certain questions in a survey questionnaire in the cross-sectional case, or subjects dropping out of a study in the longitudinal case. However, missing values can also be the result of several other factors, for example, anomalous responses that are deleted.

In the statistical world, it is assumed that a random process causes data to become missing. This random process is known as the missing data mechanism (MDM), or the 'missingness mechanism'. 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 mechanism, 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 MDM, 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 version 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 notations as before. Of course, this implies that the missing entries are entirely independently randomly missing. In the last case, the 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 notations, this means that for MNAR, $\Pr(R|Y_{com}, \theta) \neq \Pr(R|Y_{obs}, \theta)$.

1.2 Archaic Methods of Handling Incomplete Data

In order understand the evolution towards an acceptable solution to the missing data problem, we need to review the historical methods used to handle incomplete data. These methods include non-imputation procedures and single-imputation procedures.

1.2.1 Non-imputation procedures

The non-imputation methods of handling incomplete data include complete-case analysis, available-case analysis, weighting (or re-weighting) procedures, indicator methods, and model-based procedures.

Complete-case analysis on incomplete data In complete-case analysis, only cases containing values for each of the variables in the analysis are retained in the analysis procedures. This can raise the problem of serious bias in the analysis if the data is originally incomplete (Little & Rubin 2002), including problems relating to invalid and/or inefficient estimates. One must note, however, that these possible biases may not always exist in complete-case analysis, but rather that the extent of bias will depend on the mechanism by which data is deemed to be missing. Particularly, if the data is missing completely at random (MCAR), then there will be no bias in complete-case analysis of multivariate data with missing entries (Schafer 2003). This is logical, since if data is missing completely at random, any incomplete cases dropped from the complete-case analysis can be thought of as sampled units dropped in a second random stage of sampling. However, even if this is the case, the resulting inferences from this list-wise deletion may be inefficient, since the sample size is reduced, essentially unnecessarily.

To overcome the possible biases in complete-case analysis, many methods of dealing with incomplete data have been suggested. These methods are split into two camps — the

non-imputation procedures, and the imputation procedures. The former methods are introduced below, while the latter procedures are reviewed in Subsection 1.2.2.

Available-case analysis Available-case analysis estimates different parameters of interest using different subsets of the dataset, basically creating estimates of interest according to the data that is available. While use of all of the available data is sensible, analytical procedures are difficult to perform under these circumstances (Schafer & Graham 2002).

Re-weighting The re-weighting procedure drops incomplete cases and assigns weights to the remaining observations (determined by additional or auxiliary variables) so that the remaining cases more accurately reflect the distribution of the complete data. In this way generalised estimating equations can be modified to provide valid inferences when the MDM is MCAR or MAR (Kenward & Carpenter 2007). However, "[w]eighting can eliminate bias due to differential response related to the variables used to model the response probabilities, but it cannot correct for biases related to variables that are unused or unmeasured" (Schafer & Graham 2002, p.157). In other words, if the probability of response is determined by unmeasured variables, which is entirely possible, then this method becomes less attractive.

Additionally, re-weighting complete observations so that they are representative of the population sampled implies calculating weights which are generated from estimated probabilities of non-response. These estimated probabilities are inferred from the data or from auxiliary variables, but the overall weighted analysis often does not include an uncertainty component due to the estimation of these probabilities from the data.

Indicator method In the indicator method, summarised by Brand (1998), for each incomplete independent variable x_j, the regression term $\beta_j x_j$ can simply be replaced by $\beta_{0j} x_j(1 - R_j) + \beta_j R_j x_j$, where R_j is the response indicator of x_j. This simply adjusts the intercept when the value is missing, and as such, can lead to biased estimates under a number of conditions. A better replacement would be $\beta_{0j} x_j(1 - R_j) + R_j \beta_j x_j + \sum_{k \in mis; k \neq j} R_j(1 - R_k)\beta_{jk} x_j$, but this method increases the number of parameters by a great deal, and, therefore, may not be more efficient than list-wise deletion (Brand 1998).

Maximum likelihood Rubin (1976) introduces this maximum likelihood estimation procedure that integrates out missing data. Schafer & Graham (2002, p.162) mention that, "[u]nder MAR, the marginal distribution of the observed data... provides the correct likelihood for the unknown parameters [of the data,] θ, provided that the model for the complete data is realistic". Thus, even the maximum likelihood (ML) method of imputation can suffer from serious drawbacks — a lack of robustness in estimates when the model deviates from the fully parametric model assumed for the complete data, and

the fact that the ML method needs a large sample for ML estimates to be approximately unbiased and Normally distributed (Schafer & Graham 2002).

Additional drawbacks mentioned by Brand (1998) include the following: the fact that the convergence of the Expectation Maximisation (EM) algorithm that is used in this method can be very slow in cases where there is a large proportion of missing data; that convergence to a global maximum is not guaranteed; that standard errors and correlation matrices of point estimates are not directly available from EM and their calculation can be complicated; that ML is designed for large samples and has limitations for small samples; and that EM requires statistical expertise.

Extending the ML method by including priors to formulate posteriors alleviates the inconvenience of the large samples being required. However, these posteriors may be extremely complex and may require numerical integration or Monte Carlo techniques in order to solve them, similar to the final non-imputation method presented below.

Integration One could also attempt to integrate out the missing data within an incomplete data set, in a way that is summarised by Carpenter & Kenward (2007). Suppose that we split out incomplete data set into outcome variables, Y, and covariates, Z. Let R be the MDM. Our initial model is $f(Y, Z, R) = f(Y, Z) \Pr(R|Y, Z)$. If the MDM is MCAR or MAR, then it is ignorable, and if the overall analysis model and the model for missingness share no parameter space, we can integrate over the missing observation outcomes and covariates, Y_{mis} and Z_{mis}, as follows:

$$\begin{aligned} & f\left(Y_{obs}, Z_{obs}\right) \\ &= f\left(Y_{obs}|Z_{obs}\right) f\left(Z_{obs}\right) \\ &= \int\int f\left(Y_{obs}, Y_{mis}|Z_{obs}, Z_{mis}\right) f\left(Z_{mis}|Z_{obs}\right) f\left(Z_{obs}\right) dZ_{mis} dY_{mis}, \end{aligned}$$

where Y_{obs} and Z_{obs} are the observed parts of the outcome and covariate matrices, respectively.

This integration is often analytically intractable, and so many methods have been developed and applied to tackle the problem, including the expectation-maximisation (EM) algorithm, Monte Carlo Newton Raphson and Monte Carlo likelihood, mean score methods, and fully Bayesian methods based on Markov Chain Monte Carlo (MCMC) modelling (Kenward & Carpenter 2007).

1.2.2 Single imputation before complete-case analysis

Alternatively, if complete-case analysis methods are to be used on an dataset that is originally incomplete, data might be filled in by several single imputation procedures, including substitution, cold-deck imputation, unconditional and conditional mean (or

mean/mode) substitution, imputation from unconditional distributions or (single) hot-deck imputation, and imputation from conditional distributions. The term 'single' in the concept, 'single imputation methods' implies imputing only one value for each missing datum.

Substitution Substitution, occurring at the fieldwork stage of a survey, substitutes non-respondents with respondents not originally selected for interview. Again, possible bias may exist in parameters drawn from analysis if the non-respondents differ systematically from the respondents.

Cold-decking Cold-deck imputation substitutes missing values with values from outside the current dataset, such as a previous wave of the current survey. As with substitution, possible bias may exist due to a systematic difference between non-respondents and the respondents from which the imputed values are taken.

Unconditional mean substitution Unconditional mean substitution simply replaces a missing value in a variable with the mean of the available data for that variable. While the means of the variables will be preserved by this process, the standard errors will be reduced, leading to parameter estimates that seem more significant than they actually are. A variation on this mean substitution is mean/mode substitution. The difference between these two methods lies in their handling of categorical variables. For mean substitution the mean of the corresponding indicator variables created from a categorical variable is used, whereas in mean/mode substitution the mode of the categorical variables are used for imputations.

Conditional mean substitution Conditional mean substitution regresses the complete part of a variable on other variables and predicts values for the incomplete part of that variable. The missing values are imputed using the fitted values from the regression model. This method is not recommended for analysis of covariances or correlations, as the strengths of the relationships between the imputation-filled variables and the rest of the dataset are overstated.

Hot-decking Imputation from unconditional distributions (hot-decking) chooses a value for the missing entries in a variable from the observed values of that variable. In this case bias in analysis on the completed data is still possible, but it is more likely to occur in regressions equations based on the completed data than in measures of central tendency (Saunders, Morrow-Howell, Spitznagel, Doré, Proctor & Pescarino 2006). Ardington, Lam, Leibbrandt & Welch (2006) also correctly point out that observed outliers or anomalies can affect the analyses more than they should be allowed to do so since any outlier or anomaly has a chance of being drawn to replace a missing value.

Imputation from conditional distributions Imputation from conditional distributions implies simulating a draw from the distribution $\Pr(Y_{mis}|Y_{obs},\theta) = \Pr(Y_{obs},Y_{mis},\theta) / \Pr(Y_{obs}|\theta)$, where θ is again the unknown parameter of the data. Since θ is unknown, an estimate of θ, $\hat{\theta}$, must be made from Y_{obs}, after which a draw can be made from $\Pr(Y_{mis}|Y_{obs},\hat{\theta})$. This method requires a correctly specified model for $\Pr(Y_{mis}|Y_{obs},\theta)$, but if this is the case it will produce "nearly unbiased estimates for many population quantities under [the] MAR [mechanism]" (Schafer & Graham 2002, p. 159). For more on these procedures, both imputation and non-imputation, see Little & Rubin (2002) and Schafer & Graham (2002).

1.2.3 Requirements for good imputations

Incomplete data problems generally require a solution that has the following capabilities, according to Rubin (1987, p. 11). *Firstly*, it should be possible to utilise standard complete-data analysis methods on the data sets that have been filled in. *Secondly*, the imputation technique and adjustments to the follow-up 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 multiple imputation technique should display the sensitivity of inferences to various plausible models for nonresponse. As Meng (1994, p. 538) puts it, "[f]rom an inferential point of view, perhaps the most fundamental reason for imputation is that a data collector's assessment and information about the data, both observed and unobserved, can be incorporated into the imputations."

The single imputation methods mentioned in this subsection have the advantage of allowing existing complete-case analysis methods to be used on the filled-in data set. Additionally, the imputer's knowledge can be incorporated into the imputation procedure. The drawbacks, however, are that the complete-data methods that will be used assume the imputed values are known. This means that inferences based on the data are systematically sharper than they should be, and quantities based on variability (e.g. correlations) can be biased. Moreover, if the nonresponse mechanisms are not understood, no accommodation is being made for the uncertainty of not knowing which nonresponse models for imputation are appropriate. In essence, the single imputation procedures only sufficiently adhere to one of the three properties needed from a solution to incomplete data problems (Rubin 1987).

2 Multiple Imputation

Multiple imputation (MI) was first proposed by Donald Rubin in the 1970's as one solution to survey nonresponse problems. Rubin (1978) suggested that guidelines be established for imputers to be able to follow, rather than have to create *ad hoc* measures

to solve the nonresponse problem every time it arose. Rubin (1978) also mentions that a goal of, or the plan for MI, was that the imputed values reflect the variation within an imputation model as well as sensitivity to different imputation models, and that the analysis of the resultant multiply-imputed data be viewed as simulating predictive distributions of desired summary statistics under imputation models.

Multiple imputation essentially covers a broad category of methods of imputation that impute several plausible values for each missing value in a data set. Rubin (1978) mentions that the interest in multiple imputation may have grown for three reasons:

1. Surveys seemed to be suffering more and more from nonresponse;
2. There was a growing awareness that the existing standard methods of handling nonresponse were unsatisfactory; and,
3. Both mathematically and computationally, this topic was proving to be a rich statistical research area.

Multiple imputation was shown to be all about transferring the researcher's uncertainty inherent in the guesswork involved in filling in missing data points to the final analysis results. Bayesian statistics provided a way for these uncertainties to filter through to the final product – data sets analysable by complete-case analysis methods. The goal was to simplify a complicated joint model of the data and the MDM so that a posterior distribution of data parameters could be created, from which predictions could be made for the missing data points. In this way, three uncertainties could be incorporated into the final product. *Firstly*, the uncertainty concerning the MDM is incorporated into the joint model for the data and the MDM. *Secondly*, the uncertainty concerning the actual model that should be used for imputations is incorporated into the process by obtaining a joint posterior of the data and the MDM and drawing parameter values (with error) from this distribution. *Thirdly*, the uncertainty of the values that should be drawn, or thinking in another way, the uncertainty about the observations whose data was missing, is incorporated by the prediction draws that are drawn (with error) based on the drawn model parameter values. The main confounding aspect within this process is the joint modelling of the MDM and the data, a task which could be extremely complicated.

The entire process behind MI and analysis is then divided into three areas, namely the modelling task (specifying a hypothetical joint distribution), the imputation task (deriving a predictive posterior distribution for the incomplete variable(s)), and the analysis task (estimating parameters of interest from the completed data).

From the start of the development of his methods (which, in time, have been proved to be the fundamental groundwork for this entire research area), Rubin's (1978, p. 20) objective was to build "statistically sound tools for handling nonresponse in general purpose surveys", and to "be concerned with both *theoretical appropriateness* and *practical utility*" [emphasis added]. From these statements we can see that the aim of MI was two-fold from the beginning: statistical appropriateness *and* tractability, in particular for application in survey nonresponse. Rubin continues his explanation of the guidelines

under which he developed his methods, writing that "handling nonresponse must mean displaying how different the answers from the surveys might have been if the nonrespondents had responded" (Rubin 1978, p. 20). At this early stage in the development of MI, Rubin already knew that key to this research area would be to separate the analysis task from the imputation task. This is evident when he writes, "in multipurpose surveys, some form of imputation is just about the only practically possible method for handling nonresponse, because the data set will be used to address many questions now and in the future. Remodelling the missing data process each time a new question is to be asked of the data base seems to be impractical, while creating an imputed data set is quite practical" (Rubin 1978, p. 21).

Multiple imputation is viewed as a flexible alternative to likelihood methods for a range of incomplete data problems (Schafer & Graham 2002), as well as a range of nonresponse models. As in single imputation, the knowledge of the imputer can be incorporated into the imputation procedure, and, once the imputations have been completed, complete-case analysis procedures can be used to analyse the data. However, the primary advantage of multiple imputation is the inflation of uncertainty in the analysis estimates. As was mentioned before, MI covers a class of methods that impute several plausible values for a single missing data entry. Once the missing values have been imputed, several completed datasets are left to be analysed by complete-case methods. A simple set of rules is then used to combine the estimates from the separate analyses of the several datasets, and the uncertainty of these estimates is then formed from the sample variation as well as variation in the imputed values themselves. So the estimates derived from MI adjust for observed differences between respondents and nonrespondents and the standard errors of these estimates reflect the reduced sample size and an adjustment for observed differences between respondents and nonrespondents. Hence MI methods technically adhere to all three of the guidelines set out by Rubin (1987).

Bayesian statistical theory goes hand-in-hand with MI (Meng 1994), as the imputed values for a single missing data entry can be draws from the predictive posterior distribution for the non-missing data — the unknown values can be modelled directly given the known and explicit model parameters.

The drawbacks of MI, according to Rubin (1987), are as follows: that more work is required to produce multiple imputations rather than single imputations; that more space is needed to store multiply-imputed data sets; and that more work is needed to analyse multiply-imputed data sets. However, with the advance in computing power in modern times, these three disadvantages pale in comparison to the advantages of applying multiple imputations in solving incomplete-data problems, even if we choose to create a large number of multiply-imputed data sets for each incomplete-data problem (though this will not often be necessary).

2.1 Early Multiple Imputation Breakthroughs

In Rubin's (1976) paper, he shows that if the missing data is MAR (or MCAR) and the parameter spaces for the model parameters governing the data, θ, and those governing the MDM, ϕ, the posterior distribution of the model parameters for the data distribution, θ, ignoring the process that causes missing data equals the correct posterior distribution of θ, and the posterior distributions of θ and ϕ are independent.

To summarise the Bayesian aspect of Rubin's (1976) research, if the data are MAR and the parameters of the MDM and the overall data are distinct, the joint posterior distribution of the parameters of the data and the MDM is the same as the correct posterior of the parameter of the data. These distributions are the correct posterior distributions for the parameter of the data if and only if the conditional expectation of the MDM, given the pattern of missing data, the observed data and the underlying parameter for the data, is a constant positive value. Note that the missing data need not be MAR *and* the observed data be observed at random (i.e. we need not have missing data that is MCAR), but rather only that the data are MAR and the parameters of the missing data process and the overall data are distinct.

This work has boiled down to a very useful fact: "When response is unrelated to values of missing variables within subgroups defined by observed covariates, the non-response is called ignorable" (Glynn, Laird & Rubin 1993, p. 984). Rubin (1978, p. 21) simplifies this idea, stating that "when mechanisms used to sample units and record data are known (possibly probabilistic) functions of recorded values, the mechanisms are said to be ignorable." It's more important to note, however, that if the MDM is ignorable and is ignored, the inferences based on the observed data are valid inferences concerning the original population parameter of interest.

This meant that imputers could make the weak assumption that the MDM is MAR, and the joint posterior of the data and the MDM would simply be the distribution of the data. The most difficult aspect of MI had been made substantially less complex. All that is left was the joint modelling assumption of the multivariate data.

Further research into the use of the MAR MDM made MI even more attractive. Schafer (2003) notes that assuming the MAR mechanism when the data may have a more complex MDM is at least a step in the right direction, and should be done rather than not impute at all — more effort can thus be placed on modelling the data correctly, which may have stronger consequences than mis-modelling the MDM. In any case, a more general imputation model than analysis model (*i.e.* using all data for imputation when only a portion will be used for analysis) often means larger standard errors on the MI estimates (Schafer 2003), which shows the MAR mechanism is a conservative step in the right direction.

2.2 The Combining Rules

Once multiple datasets have been imputed from the same starting point, inferences on the datasets can be combined using a simple set of rules as originally defined by Rubin (1987), and explained below.

Suppose that Q is a scalar population quantity to be estimated from the sample data taken in a survey, and that an estimate of this quantity, $\hat{Q}$ and standard error $\sqrt{U}$ could be easily calculated if Y_{mis} were available. In MI, Y_{mis} is replaced by $m > 1$ simulated versions, $Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots, Y_{mis}^{(m)}$, leading to m estimates and their respective standard errors, $\left(\hat{Q}_j, \sqrt{U_j}\right), j = 1, \ldots, m$. An overall estimate for Q is

$$\bar{Q} = \frac{1}{m} \sum_{j=1}^{m} \hat{Q}_j \tag{1}$$

with a standard error of $\sqrt{T}$, where

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \tag{2}$$

$$\bar{U} = \frac{1}{m} \sum_{j=1}^{m} U_j \tag{3}$$

and

$$B = \frac{1}{m-1} \sum_{j=1}^{m} \left(\hat{Q}_j - \bar{Q}\right)^2. \tag{4}$$

T is the total variance of the estimator, while $\bar{U}$ is the regular within-imputation variance, and B is the between-imputation variance.

Note that if all the imputations are made under the same MDM, the variability measured by B stems from the inability of the observed data to predict the missing data under the given MDM. In contrast, if the MDM is varied across the m imputed data sets, B would describe a measure of the sensitivity of the MI process to the choice of the MDM (Rubin 1978). This idea is one that has not received much attention over the years. It is most common to fix the MDM and impute several data sets, rather than to impute under differing MDMs and incorporate the variation from these changing MDMs into B. This procedure, however, could be used to investigate sensitivity to the MDM of the estimates from MI (rather than being used to investigate the appropriateness of a single posited MDM).

If $\frac{(\hat{Q}-Q)}{\sqrt{\bar{U}}}$ is approximately $N(0,1)$ with complete data, as is assumed to be the case in many regression contexts for example, then, in the imputed case, according to Rubin & Schenker (1986):

$$\frac{\left(\hat{Q}-Q\right)}{\sqrt{T}} \sim t_v \tag{5}$$

where:

$$v = (m-1)\left[1+\left(\frac{m}{m+1}\right)\frac{\bar{U}}{B}\right]^2, \tag{6}$$

or

$$v = (m-1)\left(1+\frac{1}{r}\right)^2 \tag{7}$$

and

$$r = \left(1+\frac{1}{m}\right)\frac{B}{\bar{U}} \tag{8}$$

It is worth repeating that the Equations (5)-(8) are for a complete data analysis that is based on the Normal distribution.

The latter, r, is the relative increase in variance due to nonresponse (Schafer & Graham 2002). The degrees of freedom vary from $(m-1)$ to ∞ according to the rate of missing information in the dataset. According to Rubin (1987), the rate of missing information is given by:

$$\gamma = \frac{r+\frac{2}{v+3}}{r+1}, \tag{9}$$

where r is as above. The estimated rate of missing information, $\hat{\gamma}$, is approximately $r/(r+1)$. Through simple rearranging, this can be written as $(1+1/m)B/T$, the form for the rate of missing information given by Little & Rubin (2002).

Schafer & Graham (2002) also note that with large degrees of freedom (or alternatively when the variation in the estimates between imputations is small compared to the overall variation), then there is not much that can be gained from increasing m, the number of imputed datasets.

Additionally, Schenker, Raghunathan, Chiu, Makuc, Zhang & Cohen (2006) show that when the rate of missing information is low, point estimates from MI vary little from those obtained through single imputation. This is naturally due to a small value of B

that arises when there is little missing information. These authors also mention that the rate of missing information is regularly less than the proportion of nonresponse, due to the predictive power of other variables within the incomplete data set.

Barnard & Rubin (1999) provide a further refinement to the expression for the degrees of freedom when the completed data sets are based on limited degrees of freedom, say, v_{com} (when there are no missing values). In this case, v is replaced by v^*, given by

$$v^* = \left(v^{-1} + \hat{v}_{obs}^{-1}\right)^{-1}, \tag{10}$$

where

$$\hat{v}_{obs} = (1 - \hat{\gamma}) \left(\frac{v_{com} + 1}{v_{com} + 3}\right) v_{com}. \tag{11}$$

This modified degrees of freedom increases monotonically in v_{com}, is always less than or equal to v_{com}, and is equal to the original degrees of freedom, v, when v_{com} is infinite.

In order to determine the actual number of imputed datasets that should be created, Rubin (1987) also provides a measure of efficiency, measured in standard errors, and based on the rate of missing information, γ, or at least $\hat{\gamma}$. It is given by:

$$\lambda = \left(1 + \frac{\gamma}{m}\right)^{-\frac{1}{2}} \tag{12}$$

This measure essentially compares the size of the standard error after m imputations with the size of the standard error after an infinite number of imputations.

Although the number of datasets that should be completed is often debated, a small number of completed datasets, say, around 10, often suffices in order to obtain precise estimates (assuming that the fraction of missing information is not extreme). According to Little & Rubin (2002, p. 209),

> "In those cases where inference from the complete-data posterior distribution is based on multivariate [N]ormality (or the multivariate t), posterior moments of θ can be reliably estimated from a surprisingly small number, D, of draws of the missing data Y_{mis} (e.g., D = 2–10), if the fraction of missing information is not too large."

2.3 Sequential Regeression Multiple Imputation

With the MI task reduced to choosing a multivariate model for the data (given the MDM is at most MAR), and the combining rules well-defined, imputation experts started to seek easier ways of obtaining the joint posterior of the data. In practice, 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 multivariate modelling approaches extremely difficult to implement, since typical multivariate distributions aren't flexible enough to accommodate such varying structure. This led to research into the chaining of univariate models into an approximate posterior distribution.

This MI approach uses sequences of appropriate 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) approach, 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.

In this sequential procedure, each variable can be modelled individually within the imputation process. In this way, imputers can have a great deal more control over imputations from variables with inherent restrictions than they do when the variables are jointly modelled in an imputation procedure.

This method of MI was proposed by van Buuren, Boshuizen & Knook (1999) and elaborated on by Raghunathan, Lepkowski, van Hoewyk & Solenberger (2001), although the system had been used even earlier by researchers such as Kennickell (1991).

2.3.1 The SRMI process

Overview In essence, SRMI works in a three-phase process, as explained by Raghunathan et al. (2001) and He & Raghunathan (2009), and summarised below. Reviewing our standard notation, let $Y_j\,(j = 1, \ldots, p)$ denote the variables with missing values, X denote our 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 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 phase of the SRMI process is a single of the sequential models over an incomplete data set, essentially 'filling in' the missing data values. The second phase of the procedure is the repetition of this 'filling in' process, using the previously filled-in values. Thus, in each pass of the imputation procedure within these first two phases, θ_j is drawn from $P\left(Y_j|Y_j^{obs}, Y_{-j}, X\right)$ 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 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

$$f\left(Y_1, Y_2 \ldots, Y_p | X, \theta_1, \theta_2 \ldots, \theta_p\right) = \\ f_1\left(Y_1 | X, \theta_1\right) f_2\left(Y_2 | X, Y_1 \theta_2\right) \ldots f_p\left(Y_p | X, Y_1, Y_2, \ldots, Y_{p-1}, \theta_p\right) \quad (13)$$

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 pass of the SRMI procedure conditions only on the data that has been filled in already in that iteration.

Step-by-step SRMI and Gibbs sampling Given an incomplete dataset, the the dataset's incomplete variables are sorted from the variable with the least missing entries to the variable with the most missing entries. Let the variable with the most missing values be the vector Y_1, the variable with the next fewest missing be Y_2, *etc.*, until Y_p. Let X again be the part of the dataset that is complete. Finally, let θ_j once more be the vector of unknown regression and dispersion parameters in the conditional model for Y_j. The sorting of the dataset follows as an extension to the fact that in model-based imputations the joint conditional density of $Y_1, Y_2, \ldots, Y_p$ given X can be factored as in Equation (13).

The first round of imputations then begins; the variable with the least amount of missing data entries (apart from the complete variables) is selected. This variable is regressed on the complete data according to a regression model that is assumed to fit the distribution of the variable, as mentioned above. The model first processed is illustrated in Equation (13) by f_1. The regression is Bayesian by nature, but utilises a diffuse or non-informative prior. If $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$ then the prior for each model is $\pi(\Theta) \propto 1$. A set of regression parameters is then drawn from the regression model and a single draw from the predictive posterior of the model (the predictive distribution of the missing values given the observed values) is made for every missing data entry in that variable. These draws are the imputed values for that variable.

The SRMI process then selects the variable with the next fewest missing values, and the procedures in the second step are repeated. A new regression model, illustrated by f_2 in Equation (13), is chosen according to the assumed distribution of Y_2, the variable now being regressed. This new variable is regressed on the complete data and the newly completed variable from the previous step (i.e. the variable with the least missing values, all of which have now been imputed with a single imputation). Again a set of regression parameters is drawn from the new regression model and a single draw from the predictive posterior of this model is made for every missing data entry in the variable. This step is repeated until all of the variables in the dataset are filled in by appropriate regression predictions. By the nature of this process, the term 'sequential regression imputation' is justified.

The reason that the data is sorted according to missingness is explained by the fact that the starting distribution in the Gibbs sampling procedure should be as close as

possible to the target distribution $P(Y_{mis}|Y_{obs})$, since the Gibbs sampling procedure can be heavily influenced by the initial distribution (Brand 1998, p. 53). By filling in the data set variable by variable, from least missing to most missing, we obtain the best possible starting distribution.

Once an entire dataset has been filled in with imputed values for the original missing entries, this completed dataset is subjected to an update round, round two, starting essentially at the second step above. The process involved in the updating rounds differs slightly to that of steps two and three above.

For missingness without a particular pattern, a Gibbs sampling algorithm must be developed to improve upon the previous round's estimates. Raghunathan et al. (2001) suggest that the missing values in Y_j at round $(w+1)$ need to be drawn from the conditional density:

$$f_j^* \left(Y_j | X, \theta_1^{(w+1)}, Y_1^{(w+1)}, \ldots, \theta_{j-1}^{(w+1)}, Y_{j-1}^{(w+1)}, \theta_{j+1}^{(w)}, Y_{j+1}^{(w)}, \ldots, \theta_p^{(w)}, Y_p^{(w)} \right) \tag{14}$$

where $Y_i^{(w)}$ is the vector Y_i that was filled in with imputed values in round w. Equation 14 is computed based on the joint distribution specified in Equation (13). This draw process would be extremely difficult to complete, since the density in Equation (14) is difficult to compute in most practical situations without restrictions (Raghunathan et al. 2001, He & Raghunathan 2009). However, Raghunathan et al. (2001) propose that, instead, the draw in round $(w+1)$ for Y_j is taken from the predictive distribution corresponding to the conditional density:

$$g_j \left(Y_j | X, Y_1^{(w+1)}, Y_2^{(w+1)}, \ldots, Y_{j-1}^{(w+1)}, Y_{j+1}^{(w)}, \ldots, Y_p^{(w)}, \phi \right) \tag{15}$$

where ϕ is a vector of regression parameters with diffuse prior.

In other words, in imputation rounds after the first the values that were originally missing in each variable are now predicted from regression models regressing those variables on all of the other variables in the dataset, implying that the variables with values imputed from the first round are used as regressors in the second round in addition to the newly updated variables from the current round. This process can be viewed as an approximation to the Gibbs sampling procedure in Equation (14). In some particular cases this approximation is equivalent to drawing values from a posterior predictive distribution under a fully parametric model. For example, if all of the variables are continuous and Normally distributed with constant variance, then the algorithm governing Equation (15) converges to a joint predictive distribution under a multivariate Normal distribution with an improper prior for the mean and covariance matrix (Raghunathan et al. 2001).

This fourth step is then repeated as many times as the researcher deems fit (usually to a point where the inferences made on the data during subsequent rounds converge).

2.3.2 Regression Models Commonly Used in SRMI

From Equation (15) it is evident that a particular regression model needs to be utilised according to the assumed distribution of the variable in question, in order to obtain predictions for the missing data in that variable. Regular regression models considered are the Ordinary Least Squares (OLS) regression model for a variable that is Normally distributed, the logistic Generalised Linear Model (GLM) for a variable that is dichotomous or binary, the Poisson GLM for a variable that displays count data, and a polytomous regression model for variables with three or more categories.

In this section, it is important to note that the univariate outcome variable is denoted by Y, while the covariates used in the regression of Y are denoted by X. This is as opposed to the regular notation in this thesis that regards Y as an incomplete data matrix, and X as a complete one. Similarly, Y_{mis} and X_{mis} are the outcome and covariate rows respectively where the outcome is missing, and Y_{obs} and X_{obs} are the outcome and covariate rows where the outcome is observed.

Normal data When the variable in question is distributed Normally, i.e. $Y \sim N\left(\mu, \sigma^2 I\right)$, then the OLS regression model is applicable, where $E\left[Y\right] = X\beta$. This method is very similar to that reviewed by Zhang (2003) in the predictive model (PM) method for monotone missingness. As noted in Subsection 2.3.1 random draw from the posterior of the parameters and σ^2 is needed, and from there a random draw can be made from the predictive posterior of the variable.

The parameter estimates for OLS are easily shown to be $\hat{\beta} = \left(X'X\right)^{-1} X'Y$. In order to generate a random draw from the posterior of σ^2 we note that:

$$U = \frac{\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right)}{\sigma^2} \sim \chi^2_{n-k} \tag{16}$$

where n is the number of observations in the regression and k is the number of parameters.

Generating a random draw, u, from the χ^2_{n-k} distributions, and using the parameter estimates, $\hat{\beta}$, one can generate an estimate for σ^2, namely, σ^2_*, using the following equation:

$$\sigma^2_* = \frac{\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right)}{u} \tag{17}$$

Using this estimate one can draw a set of parameters, β^*, from the posterior distribution of the parameters, using:

$$\beta^* = \hat{\beta} + Tz_1, \tag{18}$$

where T is the symmetric square root of $(X'X)^{-1}$, the covariance matrix of $\hat{\beta}$, and z_1 is a random draw from the Standard Normal distribution.

Using β^* and σ_*^2, one can impute missing values using the following equation:

$$Y_{mis}^* = X_{mis}\beta^* + \sigma_* z_2, \tag{19}$$

where z_2 is another random draw from the Standard Normal distribution.

Binary data When the variable in question is binary, one should implement a special case of the Binomial model, in which $Y \sim Bin(n, \pi)$. With dichotomous data the elements of n are ones. Parameters are estimated from the general logistic regression model, with the logit link function, namely:

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = X\beta \tag{20}$$

Maximum likelihood estimates of the parameters β, and therefore also of the vector of probabilities $\pi = \frac{\exp(X_{mis}\beta)}{1+\exp(X_{mis}\beta)}$, are obtained by maximizing the following log-likelihood function:

$$l(\pi; Y) = \sum_{i=1}^{n} [Y_i \ln \pi_i + (1 - Y_i) \ln(1 - \pi_i)] \tag{21}$$

From Equation (20) we have that

$$\pi_i = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)} \tag{22}$$

and therefore

$$\ln(\pi_i) = X_i\beta - \ln[1 + \exp(X_i\beta)] \tag{23}$$

and

$$\ln(1 - \pi_i) = -\ln[1 + \exp(X_i\beta)] \tag{24}$$

Using these results in Equation (21) yields the following log-likelihood function to be maximised:

$$\begin{aligned} l(\pi; Y) &= \sum_{i=1}^{n} \{Y_i [X_i\beta - \ln[1 + \exp(X_i\beta)]] - (1 - Y_i) \ln[1 + \exp(X_i\beta)]\} \\ &= \sum_{i=1}^{n} \{Y_i X_i\beta - \ln[1 + \exp(X_i\beta)]\} \end{aligned} \tag{25}$$

For maximum likelihood estimation, the scores with respect to the $(p+1)$ elements of β are required, $U_0, U_1, \ldots, U_p$, or in other words, the derivatives of the log-likelihood function with respect to the elements of β, as well as the information matrix, F. The estimates are then obtained by solving the iterative equation $F^{(m-1)}\hat{\beta}^{(m)} = F^{(m-1)}\hat{\beta}^{(m-1)} + U^{(m-1)}$, where the superscripts denote the number of the iteration. The initial settings for the elements of $\hat{\beta}$ are zeros. Estimates are taken once convergence has been achieved, and at that stage the covariance matrix is taken as the inverse of the information matrix. For more details on the process, see Dobson (2002).

To impute missing values from this distribution, a random draw, β^*, is drawn from the posterior of the parameters as before in Equation (18), although this time a different MLE estimate $\hat{\beta}$ is used. Then a vector of probabilities is generated:

$$\pi_* = \frac{\exp\left(X_{mis}\beta^*\right)}{1 + \exp\left(X_{mis}\beta^*\right)} \tag{26}$$

Finally, a vector of Uniform random variables is generated that has the same length as π_*, and this vector is compared with π_*. If an element of the vector of Uniforms is less than or equal to the corresponding element of π_* then a '1' is imputed for the missing value associated with that element of π_*. Alternatively, if an element of the vector of Uniforms is greater than the corresponding element of π_* then a '0' is imputed for the missing value associated with that element of π_*. This process details approximate draws from the predictive posterior of the missing values (Raghunathan et al. 2001).

Count data For count data, where $Y \sim Pois\left(\lambda\right)$, the Poisson regression model is used. The mean of Y is λ, and is modelled as follows:

$$\lambda = \exp\left(X\beta\right) \tag{27}$$

The linear predictor is $g\left(\lambda\right) = X\beta$, where $g\left(.\right)$ is the log link function.

Once more a random draw, β^*, is taken from the posterior of the parameters of the regression model, as before in Equation (18). A parameter set, λ^*_{mis}, is then generated as follows:

$$\lambda^*_{mis} = \exp\left(X_{mis}\beta^*\right) \tag{28}$$

A missing datum is then imputed by drawing a random number from a Poisson distribution with the element of λ^*_{mis} corresponding to that missing datum as the distribution's parameter.

Categorical and ordinal data For Y that can take one of k values, $j = 1, 2, \ldots, k$, let $\pi_j = \Pr(Y = j|X)$. A polytomous regression model is fitted, relating Y to X as follows:

$$\log\left(\frac{\pi_j}{\pi_k}\right) = X\beta_j, \;\; j = 1, 2, \ldots, k-1. \tag{29}$$

With the restriction that $\sum_{j=1}^{k} \pi_j = 1$, then $\pi_k = [1 + \sum_{j=1}^{k-1} \exp(X\beta_j)]^{-1}$. Let $\hat{\beta}$ be the MLE estimate for a polytomous regression with regression coefficients $(\beta_1, \beta_2, \ldots, \beta_{k-1})$, and let V be the asymptotic covariance matrix with Cholesky decomposition T.

Again a random draw, β^*, is taken from the posterior of the parameters of the regression model, as before in Equation (18). Now let

$$P_i^* = \frac{\exp\left(X_{mis}\beta_i^*\right)}{1 + \sum_{i=1}^{k-1} \exp\left(X_{mis}\beta_i^*\right)}, \tag{30}$$

where β_i^* is the appropriate elements of β^*, $i = 1, 2, \ldots, k-1$, and $P_k^* = 1 - \sum_{i=1}^{k-1} P_i^*$.

Then let $C_0 = 0$, $C_j = \sum_{i=1}^{j} P_i^*$ and $C_k = 1$, the cumulative sums of the probabilities. To impute values, generate a random Uniform number u and take j as the imputed category if $C_{j-1} \leq u \leq C_j$. As with count data, the imputations are from approximate predicitve posterior distributions, since the corresponding parameter draws are from asymptotic Normal approximate posterior distributions.

Other sequential procedures The models detailed above are by no means the only possibilities. Many other Bayesian (and non-Bayesian) models can be incorporated into the SRMI process, including models to adjust for heavy-tailed data and skew distributions. Since these topics are complicated in their own right, they will not be discussed here. It need only mentioned that the list of models above is by no means an exhaustive one.

2.4 The Beauty of Multiple Imputation

There are multiple sources of uncertainty inherent in incomplete data that can be adjusted for within MI. Rubin (2003) points out that these uncertainties often complement each other within MI, 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

MI can reflect all of these uncertainties: the first, by drawing 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.

Zhang (2003, pp. 581, 584) lists the three uncertainties in a slightly different way:

1. Uncertainty due to modelling the joint distribution of the response variables and the missingness indicators, i.e. the uncertainty from $P(Y, R|\theta, \phi)$.
2. Uncertainty due to the sampling from a given imputation model assuming that the observed data and the values of the model parameters are known, i.e. the uncertainty from $P(Y_{mis}|Y_{obs}, \theta)$
3. Uncertainty about the values of the model parameters; the uncertainty for selecting the imputation model; i.e. uncertainty from $P(\theta|Y_{obs})$.

According to Rubin (2003) one can also use MI to investigate changes in the completed-data inference resulting from changes in the assumed process for creating missing data, when there is a desire for such sensitivity analysis, for example, testing whether the missing data mechanism is MCAR or MAR, since, in the former case, imputation will not change the complete-case analysis results, while in the latter case, results may change significantly.

Rubin (2003) believes that the combining rules for multiply imputed data from a sensible but imperfect model will lead to slightly conservative inferences (coverage slightly larger than nominal). In other words, the MI and combining process is self-correcting with the result that imperfect MI tends to be confidence proper.

Nielson (2003) also emphasises the fact that MI is self-correcting. This is because the between-to-within variance ratio is upwardly biased, leading to small degrees of freedom attached to the resulting inference, and thus valid MI confidence intervals with a realistic number of imputations even when the MI variance estimate is downwardly biased. So, the parameter estimates have distributions that are heavy-tailed, leading to more uncertainty even if the estimate itself is too low.

Another advantage of MI is concerned with the ignorability of the MDM. Rubin (2003) suggests that researchers should not be held up by the belief that their MDM is nonignorable. The primary concern should be to build a MI procedure around an ignorable MDM that builds the relationships between variables, and to then test the sensitivity to a nonignorable model. This is because one cannot determine whether the mechanism is, in fact, nonignorable, by the very definition of nonignorability.

It can also be shown that the MI analysis and maximum likelihood estimation techniques often produce similar results (in large samples and with diffuse priors) if their distributional assumptions are equivalent (Schafer 2003). This would seem to make MI an unnecessary evolution. However, the critical point to remember is that MI allows the separation of the data collection and analysis tasks, which remains a particularly high priority in the large-sample survey data sets that are typically made available today. One alternative to MI, the design-based approach (Fay 1992), places significant burden

on the data analyser, rather than allow a specialised imputer handle the missing data problem. So, in essence the MI procedure can be entirely modular (van Buuren 2007), splitting the imputation and analysis tasks between the two parties.

Additionally there are researchers who believe that one need only have an imputation model that is more general than the analysis model (*i.e.* incorporating more of the survey, for example), for good, valid MI results, regardless of the MDM (see, for example Rubin 1978, Rubin 2003, Schafer 2003, Zhang 2003).

So, in essence, MI has been derived for the specific Normal case, with specific rules for both imputation and analysis procedures in order to obtain perfectly unbiased, valid, efficient estimates in the analysis. However, these rules can be relaxed in numerous ways without affecting unbiasedness, validity, and efficiency in too great a manner. If the restrictions are relaxed in any one of several ways, approximately unbiased and generally valid estimates are still obtained, these estimates being more efficient than those based on the incomplete data. The process becomes entirely modular, allowing the imputer and analyser to be separate entities. The precise science that would yield perfect results can be reduced to an art form that, when incorporated into scientific analyses, makes the answers from those analyses more scientifically agreeable than naïve results.

3 Evidence Supporting Multiple Imputation

In this section, the review of papers is restricted to those using SRMI in their analysis, and specifically those that compare SRMI with other methods.

3.1 Literature using SRMI

Raghunathan et al. (2001) Raghunathan et al. (2001) investigate the differences between complete-case estimates and post-SRMI estimates in two illustrations, and then extend their analysis to a simulation study in order to compare post-SRMI results with results from an originally complete data set.

In their first example, the authors analysed a case-control study examining the relationship between cigarette smoking and primary cardiac arrest. The data were of such a nature that explicit joint modelling would have been extremely difficult; this made SRMI a natural choice for MI. See Raghunathan et al. (2001, pp. 88–89) for details.

The primary model for the data was a logistic regression regressing the log-odds ratio of cardiac arrest versus no cardiac arrest on binary variables indicating former smoker status and current smoker status, the years smoked under each of these categories, Body Mass Index (BMI), and binary indicators for gender and high school graduate status. A total of 103 (11.5%) of the cases had missing values.

The authors then used an SRMI model including only the covariates of interest in the final analysis, and found that the logistic model estimates were similar to those found for complete-case analysis. However, even with only modest changes in the estimates, the MI standard errors were smaller due to the increased number of respondents used in the analysis.

The authors then implemented an SRMI model including 50 additional variables, and found again that even though estimates were only modestly different from the complete-case analysis and their first SRMI model, the standard errors were once again smaller. This is due to the fact that many of the additional variables used were highly predictive of the BMI and smoking-related variables.

It may seem counter-intuitive that the standard errors of the estimates were smaller after accounting for the uncertainty due to missingness. However, this concept is not uncommon in the MI arena; this is the case when the imputation model is imputing correct information and the superefficiency concept discussed by Meng (1994) and Rubin (2003) is occurring — essentially, so much more information is gained from the respondents that would have been dropped that the analysis estimates are more precise than before.

In the authors' second example, they examine three models that investigate the impact of parental psychological disorders on childhood development, the impact of these disorders on reading scores, and the impact of these disorders on verbal scores. Around 60% of the data set is fully observed. Raghunathan et al. (2001) compare post-SRMI results with results after a fully Bayesian MI approach, and with complete-case results.

While the fully post-Bayesian MI and post-SRMI analyses do not differ greatly, the authors find once more that the confidence intervals after MI are substantially smaller than those under complete-case analysis in the first model. They also find that "the complete-case estimates of the effects of parental psychological disorders on the child's reading and verbal scores are quite different than those obtained by the analysis of the multiply imputed data. This is not surprising because the data on reading and verbal scores are not missing completely at random and are related to the risk group as well as the number of symptoms of the child" (Raghunathan et al. 2001, pp. 91–92). In fact, the complete-case analyses paint a picture far more severe than is truly the case when it comes to these models — controlling for disorders, verbal and reading scores for higher-risks groups are not as far below those of normal groups as complete-case analysis would suggest.

Finally, in a simulation study, Raghunathan et al. (2001) show that confidence intervals from complete-data regression analysis are indeed narrower than those from post-SRMI analyses, although the point estimates are similar. This shows that the standard errors are well-calibrated by the SRMI procedure.

Faris et al. (2002) Faris, Ghali, Brant, Norris, Galbraith & Knudtson (2002) compare Multivariate Normal (MN) MI, SRMI, Cublic Spline Regression MI, and data en-

hancement (through merging of additional data) in a study of 6 065 cardiac care patients in Alberta, Canada, in 1995.

The outcome of interest is the binary variable measuring 1-year mortality. Categorical variables in the predictor set are split into binary indicators. The original data set of consists of 6 276 individuals, but is reduced in the data enhancement, in which 6 026 individuals are matched with hospital discharge administrative data in the 'enhanced' data sets — the combined clinical and administrative data sets. Several of the variables overlap in these two data sets. For the two categorical variables in the study that did not overlap (left ejection fraction and Duke Index of coronary artery disease severity), missing values were recoded into an additional observed category (but only in the enhanced data set). On comparison between the individuals with and without administrative data, the authors note that one may assume that the individuals with administrative data are simply a random sample of the original individuals. Sources discussing the merits are drawbacks of this method are given by Faris et al. (2002, p. 186).

The authors then compare logistic regression results from mortality regressed on the clinical data after MN MI on the original data, SRMI on the original data, complete-case analysis of the enhanced data set, as well as enhancement and SRMI combined, and finally, enhancement and SRMI with the administrative variables included in the imputation model. The imputations from the SRMI procedure are taken after only five rounds in this study, as the Gibbs sampler seems to converge at that point. Ten imputed data sets are used in MI analyses.

In order to assess the procedures, *firstly* the logistic model fit was assessed using (1) the C statistic, which is the area under the receiver operating characteristic or ROC curves[1] with bootstrapped confidence intervals, and (2) the changes in deviance from the null model, i.e. $-2 \times \log L$ of the model versus $-2 \times \log L$ of the null. *Secondly*, the ability of the coefficients to predict the outcomes for the complete cases in a 1996 follow up survey was assessed.

The authors find the following: for the 1995 data methods, the C statistic is greatest for SRMI, followed by CSR, MN, and lastly, enhancement. Validating using the 1996 data shows the greatest C statistic for SRMI again, followed by CSR and enhancement, followed by MN. In both cases, the SRMI C statistic rank first in more than 90% of the bootstrapped estimates. The SRMI model also has the largest decrease in deviance from the null model for 1995, followed by CSR, then enhancement, then MN; SRMI is best in more than 99% of the bootstrapped values. Validating using the 1996 data shows SRMI above CSR, followed by MN, followed by enhancement; SRMI is again first in 98% of the bootstrapped estimates. The enhanced SRMI combination performs better than the original enhanced fit, while the administrative SRMI model shows no improvement over the original SRMI model. All of these results show the superiority of the SRMI method in this case, although the performance of all the methods are relatively satisfactory.

[1] ROC curves will have a maximum value of one if when those dying all have larger fitted values than those surviving.

Ardington et al. (2006) Ardington et al. (2006) make use of SRMI to test the sensitivity of poverty and inequality measures in South Africa to the imputations of certain covariates. The authors summarise a vast literature (based on studies of incomplete data) that show how poverty and inequality within racial groups both increased in South Africa over the period between 1996 and 2001. Their aim was to use SRMI to validate these findings.

Moreover, the authors also analyse the effect of a high proportion of household incomes totalling zero (around 25%). They adjust their imputation procedure to take account of this drastic proportion, rather than ignoring the zeros or arbitrarily assigning small amounts to these households, both practices having been previously thought of as being acceptable. Besides this analysis, the authors also check the sensitivity of the results to the assumptions regarding point estimates for income that were generated from the income bands recorded in the surveys in several ways. For more details, see Ardington et al. (2006).

The authors followed the imputation processes and recommendations made by Raghunathan et al. (2001), and estimated mean household income, a poverty head count index, and the Gini coefficient inequality measure from the resulting multiply imputed data sets. The results were combined using the regular combining rules presented in Subsection 2.2. The authors use province of residence, urban/rural location, and race as complete predictors, while the incomplete variables, ordered from least missing to most missing include age (a count variable), gender, emploment status (unemployed vs employed), occupation (four categories), years of education (a count variable), and income (an ordered categorical variable of 12 income bands).

Amongst other things, the authors find that the previous literature was indeed correct, although Ardington et al. (2006) are able to give better confidence intervals via the MI procedure than were offered by the default hot-deck single imputation confidence intervals that were given with the public release of the data. They find that there are small increases in poverty for the poorest of the poor, and increases in inequality across the board.

Barnes et al. (2006) Barnes, Gutierrez-Romero & Noble (2006) use SRMI on data taken from the South African census in 2001. In this census, over 50% of individuals above the age of 18 report zero income, and additionally, 16% of the income values are missing (Barnes et al. 2006). The authors of this study compare the SRMI method with the imputation method used by Statistics South Africa (StatsSA) when they published the data, namely single hot-deck imputation (S-HD). The S-HD process is the same as the HD process, except that only one value is imputed for every missing value.

As mentioned by Ardington et al. (2006), income is recorded in income bands, so the variable is categorical in essence. Other variables used in the study include age, gender, population group, employment status, occupation, education, income, province and

location, the latter two being the only complete variables. Also as done in Ardington et al. (2006), implausible values of income are recoded as follows:

- If household income is zero, income is set to missing for household members aged 15 and older, and to zero for those younger than 15.
- For those younger than 15 with recorded income greater than R6 400 per month, income is set to missing.
- For those recorded as being employed but with zero income, income is set to missing.

The results reported by this paper are minimal. The only statistics that are compared are the proportions lying in each of the income bands before and after imputation. Barnes et al.'s (2006) results are similar to those obtained by Ardington et al. (2006), as well as those obtained StatsSA's S-HD imputation procedure, although the latter seems to favour slightly higher income band imputations for all but the lowest income band (the zero band). The similarity between the results of this study and that of Ardington et al. (2006) occur in spite of the fact that age and education are modelled as Poisson variables in Ardington et al.'s (2006) paper, while they are modelled as Normal variables by Barnes et al. (2006) — since the latter authors prefer not to assume a constant failure rate for the inter-arrival times within the variable — as well as the fact that only a logit regression is used for the income bands, instead of the ordered logit used by Ardington et al. (2006). The most noteworthy conclusion is that it is obvious that no major outlier problem exists, since this would make the S-HD method's results differ substantially to those of SRMI.

A word of caution should be given after reviewing the paper by Barnes et al. (2006). It is important to remember that it is not just the point estimates that are of importance in assessing imputations, but also the confidence intervals generated by the procedure. As has been mentioned before, no single imputation procedure will produce valid confidence intervals, since the uncertainties associated with choosing an imputation model are not incorporated into the analysis estimates as they are in MI.

Schenker et al. (2006) Schenker et al. (2006) use SRMI to impute missing income data in the National Health Interview Survey (NHIS). The setting for the imputations in this study is well-suited to SRMI — some of the variables are hierarchical in nature, with family-level and individual-level measures; some variables have structural dependencies, with values depending on other variables; some variables should be imputed within bounds; incomplete variables have different non-Normal distributional forms. Each of these complications could easily be incorporated into the sequential regression models, within the following steps (Schenker et al. 2006, p. 926):

1. Impute missing values of person-level covariates and employment status for adults.
2. Create family-level covariates.

3. Impute missing values of family income and family earnings, as well as any missing values of family-level covariates(due primarily to missing person-level covariates for children).

4. Impute the proportion of family earnings to be allocated to each employed adult with missing personal earnings, and calculate the resulting personal earnings.

The authors' procedure then followed the standard SRMI sequence with four update rounds after the initial round. As usual, in the initial round, step one did not include income and employment as covariates, while and steps two to five did (once all the gaps in the data set had been filled in). However, in the update rounds employment status was not re-imputed, to avoid incompatibilities with imputed values of personal earnings. The entire process was completed five times ($m = 5$) for five imputed data sets. According to the authors, other variables were also imputed and created, but were not retained in the final public-use data. Note also that Schenker et al. (2006) maintained the hypothesised structural dependencies and truncated imputations, as mentioned before, and they also incorporated design variables as imputation covariates. Some variables were transformed via Box-Cox analyses to Normality during imputation and were transformed back afterwards (see Box & Cox 1964). Schenker et al. (2006) use about 60 predictors in the SRMI procedure, including variables related to sample design.

The results obtained are as expected; in the context of assessing MI, the MI standard errors for poverty ratios are lower than those from complete-case analysis (if both procedures produce unbiased estimates), but more than those of single imputation (which, of course, are not incorporating the true uncertainties within the imputation procedure).

Finally, the authors compare the MI results with results from poststratification reweighting, the latter producing results similar to those from complete-case analysis. It is clear that the MI makes use of more additional information than the poststratification adjustment does.

Van Buuren et al. (2006) Van Buuren et al. (2006) evaluate SRMI in three simulation studies by looking at univariate and multivariate missingness and three types of models. The univariate studies use Irish wind speed data (for linear and logistic imputation methods), and data from women on knowledge of, and attitude and behaviour towards mammography, i.e. mammographic experience (for a polytomous imputation method).

For the univariate analysis the outcome variable is replaced in each study by predicted values of the outcome given the other variables, in a sample taken from the original data. This is done 1 000 times for each data set. Then 50% missingness is induced in these simulated outcomes, according to an MCAR mechanism, and MAR mechanisms creating more missing data in larger values (MARRIGHT), more missing in tail values (MARTAIL) and more missing in the centre of the distribution (MARMID).

In the multivariate missingness study on continuous data, the wind speed data is used

and two samples of 400 observations are taken; one to approximate the mean and covariance of the original data (the simulated set) and one random sample. Missing values were then created in the data according to a specific non-monotone structure, as generated by Brand (1998, p. 110–113), and summarised in van Buuren, Brand, Groothuis-Oudshoorn & Rubin (2006, Appendix B).

The authors find that the linear regression SRMI procedure restores correlations and eliminates biases in the data set, correlation losses and biases that appear in the available case analyses.

For their third and final simulation study, in each of 500 replication, 1000 draws are made from a bivariate Normal distribution with means equal to 5, variance equal to 1, and correlation equal to 0.6. All values are positive. Missing values are generated in one of three ways:

- MARRIGHT: $\text{logit}(\Pr(Y_1 = \text{missing})) = -1+Y_2/5$, while $\text{logit}(\Pr(Y_2 = \text{missing})) = -1 + Y_1/5$.
- MARTAIL: $\text{logit}(\Pr(Y_1 = \text{missing})) = -1+0.4|Y_2|$, while $\text{logit}(\Pr(Y_2 = \text{missing})) = -1 + 0.4|Y_1|$.
- MARMID: $1 - \Pr(\text{MARTAIL})$.

For MARRIGHT there are about 50% missing entries, 75% incomplete cases, and about 25% completely missing (Y_1, Y_2) pairs, with proportionally more missing data for the higher values of Y_1 and Y_2. The average missing information generated is around 0.63, which is rather extreme. Van Buuren et al. (2006, p. 1059) mention that, "The multivariate missing data were not entirely MAR because the cases where Y_1 and Y_2 (or both) is (are) missing were more frequent for the higher values. The regression lines are, however, not affected because the nonresponse is generated symmetrically around the regression lines." Compatibility was ensured in the Gibbs sampler for the bivariate draws by chaining the multiple imputations Y_1^* and Y_2^* from the conditional models $Y_2^*|Y_1 \sim N(\mu_1^* + \beta_1^* Y_1, \sigma_1^{2*})$ and $Y_1^*|Y_2 \sim N(\mu_2^* + \beta_2^* Y_2, \sigma_2^{2*})$, where $\mu_1^*, \beta_1^*, \sigma_1^{2*}, \mu_2^*, \beta_2^*, \sigma_2^{2*}$ are draws from the appropriate posterior distributions. Incompatibility was generated by replacing the imputation step for Y_2 by $Y_2^*|Y_1 \sim N(\mu_1^* + \beta_1^* Y_1^2, \sigma_1^{2*})$, and, separately, $Y_2^*|Y_1 \sim N(\mu_1^* + \beta_1^* \log(Y_1), \sigma_1^{2*})$. The authors generate $m = 5$ completed data sets for each model, while in each data set, the Gibbs sampler is only iterated 5 times as well. The complete data model is the linear model $Y_1 = \alpha + \beta Y_2 + \varepsilon$, with analysis interest focussed on β.

The authors find that for the incompatible models, serious bias and undercoverage of the true estimate is eliminated using SRMI, meaning that incompatibility is a relatively minor issue in their SRMI applications; i.e. the SRMI procedure is robust against incompatibility.

Van der Heijden et al. (2006) In their paper, van der Heijden, Donders, Stijnen & Moons (2006) compare SRMI with SI, complete-case analysis, and the missing-indicator approach; a form of the latter is used by Faris et al. (2002) in their non-imputation data enhancement technique that was compared with the MN method and SRMI, as discussed earlier in this subsection. The data set consists of 398 consecutive patients 18 years or older who were referred to a Dutch hospital because acute pulmonary embolism (PE) was suspected. Numerous tests were completed on these patients, which found that 43% of the patients did have PE. The predictors chosen for the analysis of PE are based on those recommended by previous studies.

None of the outcome values are missing, although the covariate data is incomplete. In 38% of the patients, one or more predictors are missing, while the data is certainly not MCAR, since, based on the results of certain tests, doctors may have skipped subsequent tests, considering them to be uninformative given the prior test result. Van der Heihden et al. (2006) then use complete-case analysis, the missing-indicator method[2], SI (conditional and unconditional mean imputation), and SRMI, with convergence within the imputed data sets occurring after 5 rounds, and creating 10 imputed data sets.

The authors then use a backward selection process for the overall model on each of these five methods, and compare regression coefficients, standard errors of the coefficients, and areas under the ROC curves — the so-called C statistic of Faris et al. (2002).

Van der Heijden et al. (2006) find that the model selected from complete-case analysis is different to those selected after the imputation procedures. This is natural, since the data is not MCAR. The standard errors are smallest for the conditional mean imputation (as expected). The indicator variables for the missing-indicator approach all achieve significance, while the coefficients in this method are larger than for the other models (since there are simply more coefficients). All of the areas under the ROC curves are above 0.75, but since there are more significant (but clinically meaningless) predictors in the missing-indicator approach, this method produces the highest area, albeit surely overestimated, while conditional mean imputation and SRMI produce the lowest values. The most significant outcome of this study is that the complete-case analysis is different to that of the post-imputation analysis, warning researchers of the hazards of simply relying on complete-case analysis methods. Interestingly, the missing-indicator, SI and MI methods' results do not differ greatly in post-imputation analysis coefficient direction, magnitude and precision. This may be due to the relatively small amount of missing data. However, the authors do warn against the use of the missing-indicator approach, for several valid reasons (see van der Heijden et al. 2006, p. 1108).

Amber et al. (2007) Ambler, Omar & Royston (2007) compare the SRMI proce-

[2]In this method, a missing value in a variable is recoded into a separate indicator variable attached to the incomplete variable itself, while the missing values in the incomplete variable are recoded into zeros, for example; complete case analysis will, then, not drop the observations for which these variables are 'missing'.

dure with several others in their study of risk modelling. Complete case analysis, single imputation procedures and MI procedures are compared to SRMI in an analysis model with a binary outcome, i.e. a logistic analysis model. The single imputation procedures include mean, mean/mode and conditional mean imputation, while the MI procedures include hot-deck (HD), hot-deck by covariate pattern (HD_{CP}), hot-deck by observation (HD_{obs}) and hot-deck including outcome (HD_{Y}) procedures combined with the Approximate Bayesian Bootstrap (ABB) procedure to produced valid MI, all explained in the bullets below:

- **Hot-deck imputation (HD)**. In this MI method, the imputed values for Y_{mis} are drawn with replacement from Y_{obs}, where each element of Y_{obs} has equal probability of being drawn. Unfortunately, since the parameter of the data, θ, is not drawn from its own posterior distribution, and draws from a predictive posterior distribution conditional on this θ are not made, the HD method underestimates uncertainty. Moreover, observed outliers will have a greater influence on the post-imputation analyses, since these outliers will form part of the donor pool for missing values (Ardington et al. 2006).

- **Hot-deck imputation by covariate pattern (HD_{CP})**. This modification of the HD procedure matches fully observed categorical variables (or categorised continuous variables) for observations with missing values. These matches form the donor pool from which imputations are drawn. Note that only predictors are imputed this way.

- **Hot-deck imputation by observation (HD_{obs})**. In this modification of the HD_{CP} procedure, the entire observation is replaced by a fully observed match, with the matching made once more through fully observed categorical variables (or categorised continuous variables). Once more, only predictors are replaced this way.

- **Hot-deck imputation including outcome (HD_{Y})**. Two variations of this modification of the HD_{CP} procedure exist: one where the outcome variable is the only variable used to find matches from which imputed values are drawn, and the other, where both the outcome and predictors are used to find matches from which imputed values are drawn.

- **Bayesian Bootstrap (BB)**. This method improves on the HD method by incorporating uncertainty in θ. Suppose that each element of the population takes one of the values $d_1, \ldots, d_K$ with probabilities $\theta_1, \ldots, \theta_K$, respectively. If the improper Dirichlet prior with density $\propto \prod_{k=1}^{K} \theta_k^{-1}$ is placed on the vector $\theta = (\theta_1, \ldots, \theta_K)$,, then the posterior distribution of θ is the Dirichlet distribution with density $\propto \prod_{k=1}^{K} \theta_k^{q_k - 1}$ and K-dimensional mean vector $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$ having components given by $\hat{\theta}_k = q_k / n_{obs}$, where $q_k =$ (number of times d_k appears in Y_{obs}). The BB MI method first draws θ^* from this posterior for θ. Then the components of Y_{mis} are independently drawn from among $d_1, \ldots, d_K$ using the probabilities in θ^*.

- **Approximate Bayesian Bootstrap (ABB)**. This method is more computationally direct than the BB method. Drawing n_{obs} observations with replacement from Y_{obs}. Then from this sample, draw n_{mis} observations with replacement. In this way, rather than drawing θ from a Dirichlet posterior distribution, this method draws θ from a scaled multinomial distribution. The distributions for θ in this method have the same means and correlations as in the BB method, but have $(n_{obs} + 1)/n_{obs}$ times the variances. This method is approximately equivalent to choosing the values of θ for the conditional posterior predictive distribution $P(Y_{mis}|Y_{obs}, \theta)$ from the observed data posterior distribution $P(\theta|Y_{obs})$. Note that the former HD procedure modifications can also be combined with the ABB procedure, as is done in Ambler et al. (2007).

The data used were the medical characteristics of 20 738 aortic and/or mitral valve surgery patients in Great Britain and Ireland between 1995 and 2003, and this included in-hospital mortality as the outcome variable.

The authors find the most prevalent missingness patterns (among observations), impute the data using SRMI, create binary responses from the fitted logistic regression coefficients based on the observed data (hereafter known as the true coefficients), and then find the (rescaled) fitted probabilities of each observation belonging to each of the most prevalent patterns, also using logistic regressions. Simulated data sets are created by sampling without replacement from the completed data. For each of these data sets binary responses were created from the true coefficients, and data was made missing according to both MCAR and MAR rules separately. For MCAR, the observations were randomly assigned to a missing data pattern according to that pattern's prevalence. For MAR, the same rule is applied, although, then the covariates for an individual are only set to missing with the probability equal to that of that individual having been in that particular pattern originally (the fitted probabilities calculated before). The different MI techniques are then applied to five imputed versions of each of these simulated data sets and the overall logistic regression is again run; this allows comparison between the true coefficients and the multiply imputed coefficients. The measures used to assess the imputations over 1000 simulated datasets are as follows:

- **Measure of agreement**: This is the proportion of observations correctly classified into the correct risk group.
- **Rank correlation**: This is a Spearman rank correlation comparison between the true ranking of disease severity and the ranking after the imputation procedures on the simulated data sets.
- **Root mean squared error (RMSE)**: This is the RMSE between the fitted and true probabilities for patients.
- **Regression based calibration measure**: To assess the calibration of the fitted model, the fitted log-odds are regressed against the true log-odds for each simulation data set using a standard linear regression. The coefficients of this regression

provide information about the calibration of the fitted model. Slope coefficients close to one are wanted.

- **Regression coefficients and confidence interval coverage**: The biases between the true regression coefficients and those obtained from the completed simulated data sets is then also assessed, averaging over all the simulated data sets for a particular method.

In essence the authors show that mean/mode imputation provides an improvement over complete-case analysis. Mean imputation goes one step further, being less biased. Conditional mean imputation, however, outperforms both of these methods in this study. However, all of these methods suffer from the deficiency in not accounting for imputation model uncertainty, which is provided for in MI.

Of the MI measure, HD_{obs} performs the worst, even worse than some of the single imputation procedures, but the authors admit that this may have been due to the setup of the simulated data sets (i.e. high proportions of missing data). SRMI, HD, HD_{CP}, HD_{Y} generally perform well and very similarly with respect to the agreement, rank correlation and RMSE. Additionally, SRMI and HD_{Y} methods exhibit good calibration and provide better classification of low and high risk patients. This may be due to the fact that these methods include the outcome in the imputation procedure. The SRMI procedure also produces the lowest biases in the regression coefficients and the confidence intervals provide coverage values close to the nominal level.

Penn (2007) Penn (2007) applies SRMI to 1 747 observations (of individuals over the age of 25) from the U.S. General Social Survey (GSS) of 1998. In these observations, 11.7% are missing the income measure, primarily due to those individuals refusing to answer that specific survey question. By looking at the distributions of the income-complete and incomplete observations, Penn (2007) shows that complete-case analysis would definitely produce biased results. The purpose of the study is to verify results from the study by McBride (2001), which, after using only complete-case analyses, showed that a person's self-reported happiness is reliant on the person's standard of living compared with that of their parents. The variables used in the imputation include happiness, parents' standard of living, educational level, age, marital status, gender, health status, race, family income, working status, occupation (9 categories), and educational level. All but the last three variables are used for the completed-data analysis, so the imputation model is more general than the analysis model. Of the 1 747 observations, income is missing for 205, parents' standard of living has 26 missing values, happiness has 21 missing values, and fewer for health status and age. At least one item is missing for 13.9% of the sample.

The author set $m = 6$ in his analysis, and performs the ordered probit regression of happiness on both the completed-data and the incomplete data, for comparison. Penn (2007) finds that not only do the coefficients of the relative standard of living between parents and children grow in magnitude (meaning that if they have a better standard

of living than their parents did at their age, they tend to be happier), but the standard errors for this categorical variable decrease after imputation, moving the two statistically insignificant categories into significance. The fact that these standard errors decrease may be due to the increase of the analysed sample size, from 1 503 to 1 747, or, more likely, the imputation model is imputing correct information and the superefficiency concept discussed by Meng (1994) and Rubin (2003) is occurring again.

Van Buuren (2007) Van Buuren (2007) compares joint modelling through the MN specification to an SRMI procedure on the Fourth Dutch Growth Study data set of measures of pubertal development in 3801 Dutch girls (these were the observations with complete age, height and weight measures). About 34% of the so called Tanner stage development data is missing. The data includes menarche stage (2 categories), breast development (5 categories) and pubic hair (6 categories).

The imputations made for these incomplete variables under the MN scheme are rounded to fit into the categorical nature of the data. The two-category menarche variable is imputed using a logistic regression model, while the two incomplete categorical variables, breast development and pubic hair, are imputed using polytomous logistic regression, which in itself can raise modelling issues (see van Buuren 2007, p. 233). For each the MI procedures, five imputed data sets were created.

The author uses correspondence analysis on the two categorical variables to determine whether the inherent structure of the data is preserved between complete-case analysis and MI via the MN and SRMI methods, and finds that the SRMI based correspondence analysis preserves the canonical correlations of the complete-case analysis better than the MN method does.

Van Buuren (2007) then regresses log weight on the incomplete and completed data, and finds all three procedures (complete-case analysis, MN, and SRMI) produce the same significant model fit. Standard error from the imputed data are much narrower due to the increased sample size (from 2200 in the complete-case analysis to 3801 in the completed data sets). Suspicious of these similar results, van Buuren (2007) creates reference curves for the complete-case method versus the imputation methods. For each stage transition of breast development, a reference curve was fitted conditional on age by a series of four logistic additive models. From these curves it is clear that the MN method imputes data for breast development that does not fit the complete-case distribution across age, while the SRMI approach does succeed in doing so. For this reason, van Buuren (2007) recommends that the MN approach is not chosen above the SRMI approach when the incomplete variables are categorical in nature. It is possible that the rounding of the MN imputations adds to biases seen in the results.

He and Raghunathan (2009) He & Raghunathan (2009) contribute to this research area by assessing several methods of Normality-based SRMI when the underlying conditional distributions of the variables are non-Normal. In a simulation study, they assess

the following sequential imputation methods when these methods are (incorrectly) applied to data that is non-Normal, with missing values that are MCAR:

- **Sequential Normal linear regressions**. This is equivalent to imputation under the multivariate Normal model using Gibbs sampling (Schafer 1997, van Buuren 2007).[3]

- **Predictive mean matching**. This method by, described by Schenker & Taylor (1996) has hot-deck imputation origins. Missing Y values are imputed from nearby complete cases. The predictive mean of an observation is given as $\hat{Y}_i = X_i\beta^*$ where β^* is drawn as in the Normal method given above. For each incomplete case this method draws an observation randomly from from a set of "possible donors" which are observations with predictive mean close to that of the incomplete case. The value of Y for the chosen case is then donated to the incomplete case.

- **Local residual draw**. This method, also described by Schenker & Taylor (1996), imputes the value $Y_i^* = X_i\beta^* + r^*$, where r^* is drawn at random and with replacement from the set of complete "donor cases" as defined above. This method can adjust for the lack of fit of the Normal regression model by fitting local residuals, rather than by drawing local observed values, as in the previous method.

  There is a discussion in Schenker & Taylor (1996) on the possible bias inherent in having too many donor cases for the previous two methods, and on the possible overstated correlation inherent in having too few donor cases. For this reason Schenker & Taylor (1996) develop an adaptive technique for choosing the number of donor cases. These authors find, however, that there is little difference in results between their adaptive technique and a non-adaptive fixed number of donor cases. Additionally, He & Raghunathan (2009) also show that there isn't much change in the overall analysis if different (reasonable) fixed numbers are used for the donor cases.

- **Adjustment of Normal regression by sampling from observed residuals (or expanded residual draw)**. This method, as given by Rubin (1987), is a modification of the sequential Normal method. This method first obtains the standardised residuals:

$$\frac{Y_i - X_i\beta^*}{\sqrt{\frac{SSE}{n_{obs}-p}}}$$

  From these residuals, n_{mis} values are sampled with replacement (i.e. as many as there are missing observations), then these are multiplied by σ_*, and, finally, they are added to $X_i\beta^*$. These standardised residuals then have the correct conditional moments, but a distribution whose shape is adjusted to reflect that of the actual

[3]Another special case relating joint modelling to the SRMI approach is when three variables are modelled using logistic regressions. The joint model for this is effectively a multivariate log-linear model with no three-way interaction term (see van Buuren 2007).

error terms. In fact, this method's "donor set" for residual has just been expanded to include all complete cases, and the residuals donated are standardised, so the overall adjustment is partially parametric.

- **Adjustment by fitting Tukey's g-and-h distribution to errors**. The final method these authors analyse, is the Normal method adjusted to fit Tukey's (1977) g-and-h distribution to the error terms. Tukey (1977) proposed the gh family based on a transformation of the standard Normal Z,

$$T_{gh}(Z) = \mu + \tau \frac{e^{gZ} - 1}{g} e^{hZ^2/2}, \tag{31}$$

where μ is the location parameter, $\tau(> 0)$ is the scale parameter, and g and h are scalars that govern the skewness and kurtosis or elongation of the data, respectively. He & Raghunathan (2009) use a linear regression model for their method, with their error terms modelled using the centered gh distribution,

$$\begin{aligned} Y_i &= X_i\beta + \epsilon_i, \\ \epsilon_i &= \tau \left(\frac{e^{gZ} - 1}{g} e^{hZ^2/2} - E_{gh} \right), \end{aligned} \tag{32}$$

where $E_{gh} = \frac{1}{g\sqrt{1-h}} \left(e^{g^2/[2(1-h)]} - 1 \right)$ is the mean of the standardised gh distribution with $\mu = 0$ and $\tau = 1$ in Equation 31. First, β^* is drawn as if the error distribution is Normal, and then parameters τ, g, and h are estimated from a bootstrap sample of the observed residuals $Y_i - X_i\beta^*$ using a quantile-based method (He & Raghunathan 2009, Appendix). Then, for a missing case i, independent standard Normal Z_i's are simulated and the missing value of Y_i is estimated as $Y_i = X_i\beta^* + \tau(\frac{e^{gZ}-1}{g} e^{hZ^2/2} - E_{gh})$.

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 then values generated completely at random:

$$\begin{aligned} Y_1 &\sim U(0, 2), \\ Y_2 &= 1 + Y_1 + \epsilon_2 \\ Y_3 &= 1 + Y_1 + Y_2 + \epsilon_3 \end{aligned}$$

The authors then 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 simulation study consists of 1000 replicates, and each replicate includes 1000 cases. On each replicate, once missing values have been generated, the given SRMI procedures are applied, and in each case 5 imputed data sets are created (and within each data set

the SRMI procedure is iterated 5 times). If "donor cases" are chosen in any method, their number is restricted to 20. The quantities of interest in their study are the marginal mean of Y_3, the proportions of Y_3 that are less than its different population quantiles (5%, 25%, 50%, 75%, and 95%), and the coefficients of regressing Y_3 on Y_1 and Y_2. Inferences taken before deleting missing values are taken as a benchmark, while the results from applying complete-case analysis on the incomplete data are also used. The performance of these methods are evaluated using relative bias, $RBIAS$, and the root of the relative mean squared error, $RRMSE$.

$$RBIAS = \left| \frac{Bias}{True} \right| \times 100\% \tag{33}$$

$$RRMSE = \sqrt{\frac{MSE(Method)}{MSE(Before\ deletion)}} \tag{34}$$

Additionally, coverage rates of the 95% confidence intervals across the 1000 replicates are examined, which should be close to nominal if the imputation method is working well. A reasonable upper limit for $RBIAS$ is 5%. $RRMSE$ measures the increase of $RMSE$ relative to that of the analysis before deletion of the missing observations. In general, it is expected to be more than 1, reflecting loss of efficiency when analysing incomplete data. However, it can be less than 1 as well, implying "superefficiency" of the MI models, as discussed in Rubin (1996). To recap, this happens when the analysis used more (correct) data than would be used in the complete-case analysis of the incomplete data, i.e. it means that the imputation methods are imputing correct values.

In essence, the study shows that all of the methods are reasonable for estimating means and proportions, and for the 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. 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.

Lee and Carlin (2010) Lee & Carlin (2010) compare the SRMI method with the standard MN method in their study of estimation of regression coefficients from simulated data after MI. Their analysis is similar to that of van Buuren (2007), except the rounding of the MN method's imputations is adapted to the true distribution of the categorical variable. The simulated data sets are obtained by sampling from a synthetic

population of 971 327 girls, grades 7–10, created to resemble the sample from the US National Longitudinal Study of Adolescent Health.

The variables are synthesised sequentially, starting by drawing 1 million observations from the 3×5 race–grade table, and then adding one variable at a time using predictive simulation from regression models based on the original data. At each step, the model conditions on the previously generated values, incorporating them into complex regressions that included nonlinear relations and numerous interactions, to create sufficient population complexity. Since the outcome variable, emotional distress at wave II, is a continuous measure between 0 and 3 that is strongly positively skewed, the 0 score observations are dropped so as to not complicate a logarithmic transform. Data sets for the study each draw 1 000 individuals from this synthetic population. The analysis of the regression of the log of distress on other covariates (diet, log of distress at wave I, Black race indicator, Hispanic race indicator, grade, health and physical fitness) is primarily concerned with the main coefficient, that of diet. The so-called true values of the coefficients are obtained from the same OLS model applied to the full synthetic population. The authors use the original non-significant diet effect and an artificially inflated significant diet effect as comparisons (but both produce a similar set of results, given below).

The data are set to missing according to one of three models set out below, each model utilising a logistic regression of the following form (and where $ldistW2$ is the outcome variable):

$$\text{logit}\,\Pr(\text{missing}) = \alpha + \beta_1 diet + \beta_2 Black + \beta_3 Hisp + \beta_4 grade + \beta_5 ldistW2$$

1. Missing data on emotional distress at wave I
2. Model 1, plus independent missing data pairs on health and physical fitness
3. Model 2, plus independent missing data on diet.

For Models 1 and 2, the coefficients are fixed to create a substantial association between variables and missingness, as follows: $\alpha = 3, \beta_1 = \beta_2 = \beta_3 = 1, \beta_4 = 0.2, \beta_5 = 0.3$. The MDM for Models 1 and 2 is automatically MAR, but to make Model 3 MAR, β_1 is set to 0.

Imputations from the SRMI or MN methods are rounded to fit into the given scales. Adaptive rounding is additionally used as an option for the binary diet variable (where rounding is based on a Normal approximation to the Binomial, making use of the marginal distribution in the observed data). For the SRMI approach, diet is imputed using a logistic regression, while health and physical fitness used ordinal (proportional odds) logistic regressions. The distress variable is either transformed to a Normal distribution (via log transformation and log transformation with an offset to make 0 skewness in the observed values) or is left as is. Imputations of 0 are then replaced with the smallest value in the sample, while observations above 3 are truncated at 3. For the SRMI approach, predictive matching is also used as an option for the Normal variables,

imputing the observed value with the predictive mean closest to that of the imputation for a missing value. The regular combining rules are used on 20 imputed data sets for each method.

The authors find that the best results (under all missingness models) for the diet coefficient are obtained using the MN method that uses the zero-log-skewness adjustment or the prediction matching method in SRMI. All the methods alleviate the biases and poor coverages existing for the complete-case analysis. For the other coefficients (not associated with the MDM), all methods provided adequate results, and it is shown that precision is improved by imputing rather than using complete-case analysis. The zero-log-skewness adjusted MN method performs even better than the SRMI approach with predicative matching in the context of coverage, when the adaptive rounding is used.

This studies results are important to note, since Lee & Carlin (2010) show that the MN method can be adjusted to impute properly even for a binary variable. It may seem that the added complexity inherent in an SRMI model may not always be justified. However, sensitivity to non-Normality may make the SRMI approach seem to be the more robust option. Of course, the inherent ease of dealing with ordinal and categorical variables in the SRMI model may be enough to sway a researcher towards that option.

Von Maltitz and van der Merwe (2012) In the paper by von Maltitz & van der Merwe (2012), an overall complete-case analysis of household welfare in KwaZulu-Natal in South Africa is shown to be influenced by the application of SRMI. In the paper, missing data is multiply imputed through a general SRMI imputation model using all data from the three waves of a panel survey, and the overall analysis, a regression of household welfare on various covariates, is a panel regression model. The amount of missing data across the three waves of the survey increased drastically across the years. Although the missing data seem to be within reasonable bounds in the first wave of the survey, more data were missing in the subsequent waves, *i.e.* it changed in most cases from around 1% to 16% to 40% over the waves. In all three waves of the survey, the variables with the most missing entries are those measured for the head of the household, namely years of education of the head of household, gender of the head of household and age of the head of household, a missing proportion that moves from 15% to 26% to 55% over the three waves.

The covariates utilised in explaining household welfare include total household education, household size, household race, the household head's years of education, gender, and age, the location of the household (rural *versus* urban), and various social capital indicators, measured by household memberships in financial, production, cultural, service, political and other social groups.

The authors find that complete-case analysis finds every variable a significant indicator of household welfare, but once SRMI is applied and the analysis is re-run, household education and household size become insignificant. This means that confidence intervals have been inflated by the MI procedure, showing that incorporating the uncertainty inherent

in the missing data has indeed changed the overall inferences. In essence, social capital access is shown to be a more important determinant of welfare than household education in KwaZulu-Natal, as long as the education of the head-of-household is controlled for.

3.2 Literature Summary

It is clear from the literature that complete-case analysis is unacceptable if the incomplete data is not entirely MCAR, which is a strong assumption indeed. Single imputation methods, including single hot-deck imputations, are shown to lead to incorrectly precise estimations when compared with multiple imputation methods. It is also clear that the MN method is more valid and robust than single imputation methods, as well as MI hot-decking methods. The most parsimonious method, however, is evidently the SRMI method.

It is also important to note that SRMI has produced confidence intervals both wider and narrower than intervals from complete case analysis. If the intervals are wider, we know the uncertainty inherent in imputation has been provided for. If they are narrower, then the imputation model has either allowed so many more respondents to be included into the survey that the standard errors become smaller (than the extra uncertainty increases them by), or the concept of superefficiency is in play — where the imputation model is indeed the correct model for the originally complete data.

4 Conclusion

This paper has elaborated on the missing data problem, and has reviewed the historical (and now inadvisable) methods of dealing with this missing data problem. The paper has presented the Sequential Regression Multiple Imputation (SRMI) method in detail, and has elaborated on the rules used in handling multiple datasets arising from Multiple Imputation (MI). An argument is then made for the inherent applicability of MI methods in general. Finally, a brief overview of some of the literature providing evidence for SRMI is given.

From this paper, it is very clear that South African researchers cannot afford to be naïve about the missing data problem, and that they certainly cannot afford to remain in the rut of applying single imputation methods. Even if researchers prefer not to do multiple imputations themselves, it is important that imputation experts apply MI to handle missing data in public data sets, and that researchers familiarise themselves with the methods used to combine results from these multiply imputed data sets, so that valid, conservative inferences can be made from the incomplete data that we so often struggle with.

References

Ambler, G., Omar, R. Z. & Royston, P. (2007), 'A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome', *Statistical Methods in Medical Research* **16**, 277–298.

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.

Barnard & Rubin, D. B. (1999), 'Small-sample degrees of freedom with multiple imputation', *Biometrika* **86**, 949–955.

Barnes, H., Gutierrez-Romero, R. & Noble, M. (2006), Multiple imputation of missing data in the 2001 south african census, Working Paper 4, Centre for the Analysis of South African Social Policy, University of Oxford.

Box, G. E. P. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society* **Ser. B**(26), 211–252. (With discussion).

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.

Carpenter, J. R. & Kenward, M. G. (2007), 'Sensitivity analysis after multiple imputation under missing at random: a weighting approach', *Statistical Methods in Medical Research* **16**, 259–275.

Dobson, A. J. (2002), *An Introduction to Generalised Linear Models*, 2 edn, Chapman & Hall/CRC, Boca Raton.

Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D. & Knudtson, M. L. (2002), 'Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses', *Journal of Clinical Epidemiology* **55**, 184–191.

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.

Glynn, R., Laird, N. & Rubin, D. B. (1993), 'The performance of mixture models for nonignorable nonresponse with follow ups', *Journal of the American Statistical Association* **88**(423), 984–993.

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.

Kenward, M. G. & Carpenter, J. (2007), *Statistical Methods in Medical Research* **16**, 199–218.

Lee, K. J. & Carlin, J. B. (2010), 'Multiple imputation for missing data: Fully conditional specification vesrsus multivariate normal imputation', *American Journal of Epidemiology* **171**(5), 624–632.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2 edn, John Wiley & Sons.

McBride, M. (2001), 'Relative-income effects on subjective well-being in the cross-section', *Journal of Economic Behavior & Organization* **45**, 251–278.

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.

Penn, D. A. (2007), 'Estimating missing values from the general social survey: An application of multiple imputation', *Social Science Quarterly* **88**(2), 573–584.

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.

Rubin, D. B. & Schenker, N. (1986), 'Multiple imputation for interval estimation from samples with ignorable nonresponse', *Journal of the American Statistical Association* **81**(394), 366–374.

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. (1997), *Analysis of Incomplete Multivariate Data*, CRC Press, New York.

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.

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G. & Cohen, A. J. (2006), 'Multiple imputation of missing income data in the national health interview survey', *Journal of the American Statistical Association* **101**(475), 924–933.

Schenker, N. & Taylor, J. M. G. (1996), 'Partially parametric techniques for multiple imputation', *Computational Statistics and Data Analysis* **22**(425–446).

StatsSA (2012), Census 2011 — statistical release (revised), Technical Report P0301.4, Statistics South Africa.

StatsSA (2013), Quarterly labour force survey, quarter 2, 2013 — statistical release, Technical Report P0211, Statistics South Africa.

Tukey, J. W. (1977), Modern techniques in data analysis, MSF-Sponsored Regional Research Conference at Southeastern Massachusetts University, North Dartmouth, MA.

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.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2006), 'Fully conditional specification in multivariate imputation', *Journal of Statistical Computation and Simulation* **76**(12), 1049–1064.

van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T. & Moons, K. G. M. (2006), 'Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example', *Journal of Clinical Epidemiology* **59**, 1102–1109.

von Maltitz, M. J. & van der Merwe, A. J. (2012), 'An application of sequential regression multiple imputation on panel data', *South African Journal of Economics* **80**(1), 77–90.

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