DataFirst Technical Papers

Sequential Regression Multiple Imputation for Incomplete Multivariate Data Using Markov Chain Monte Carlo

by

Miguel Lacerda, Cally Ardington & Murray Leibbrandt

Technical Paper Series

Number 5

About the Author(s) and Acknowledgments

Support for this research was provided by the U.S. National Institute of Child Health and Human Development (Grant R01HD045581)

Recommended citation

Lacerda, M., Ardington, C., Leibbrandt., M. (2007). Sequential Regression Multiple Imputation for Incomplete Multivariate Data Using Markov Chain Monte Carlo. A DataFirst Technical Paper Number 5. Cape Town: DataFirst, University of Cape Town

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

Sequential Regression Multiple Imputation for Incomplete Multivariate Data Using Markov Chain Monte Carlo

Miguel Lacerda, Cally Ardington and Murray Leibbrandt

Abstract

This paper discusses the theoretical background to handling missing data in a multivariate context. Earlier methods for dealing with item non-response are reviewed, followed by an examination of some of the more modern methods and, in particular, multiple imputation. One such technique, known as sequential regression multivariate imputation, which employs a Markov chain Monte Carlo algorithm is described and implemented. It is demonstrated that distributional convergence is rapid and only a few imputations are necessary in order to produce accurate point estimates and preserve multivariate relationships, whilst adequately accounting for the uncertainty introduced by the imputation procedure. It is further shown that lower fractions of missing data and the inclusion of relevant covariates in the imputation model are desirable in terms of bias reduction.

KEY WORDS: Missing data; Item non-response; Missingness mechanism; Imputation; Regression; Markov chain Monte Carlo.

Support for this research was provided by the U.S. National Institute of Child Health and Human Development (Grant R01HD045581)

1 Introduction

1.1 Problem Description

The collection and analysis of survey data is as much an art as it is a science. In well-designed surveys, the choice of the sample follows careful deliberation in order to ensure that inferences concerning the population of interest are both reliable and straightforward to obtain [12]. Consideration must be given to the design of the questionnaire, sampling types and procedures, the cost of data collection, the subsequent statistical analyses of interest and the affect of the complex sample design on these analyses to name but a few of the pertinent issues. However, despite all the thought and effort that may inform the selection of a sample, survey analysts will inevitably encounter the problem of survey non-response, which threatens to distort sample statistics and introduce biases that may render the observed sample unrepresentative of the target population.

In many censuses and sample surveys, some of the units selected into the sample may not respond to one or more of the items being asked of them. Such datasets arise frequently in practice where the population of interest consists of observational units such as people, households or businesses. The problem created by non-response is, of course, that data values that were intended to be observed by the sample design are in fact missing. These missing values do not only imply a loss of efficiency in estimates due to the reduced sample size, but also that standard complete-data methods cannot be immediately used to analyse the data. Moreover, biases may arise due to systematic differences between the respondents and non-respondents. Such biases may be difficult to resolve in practice since the specific reasons for non-response cannot usually be precisely known [12].

In practical terms, missingness is regarded as a nuisance, inhibiting the analyses of real interest, rather than forming the main focus of inquiry. Handling the problem of non-response in a principled manner, however, raises conceptual difficulties and computational challenges [15]. In an attempt to side step such complications, practitioners frequently resort to ad hoc edits to force the incomplete dataset to have an appearance of completeness such that standard statistical methods may be utilised [13]. However, such methods fail to reflect the inherent uncertainty surrounding the missing values. A principled, yet computationally efficient, method for handling missing data would therefore seem desirable.

1.2 Background to the Investigation

The literature on the statistical analysis of incomplete data has flourished since the early 1970s, spurred on by advances in computer technology that made previously laborious numerical calculations a simple matter [6]. Prior to this period, missing values were handled primarily be means of naïve, editing procedures such as complete-case or available-case analysis. In 1976, Donald B. Rubin developed the first framework of inference for incomplete data that remains in use to date. The formulation of the Expectation-Maximisation, or simply EM, algorithm by Dempster, Laird and Rubin in 1977 provided a method to compute maximum likelihood estimates in many missing data problems [1]. Rather than deleting or filling in the missing values, maximum likelihood treats the missing data as random variables to be removed from the likelihood function as if they were never sampled [15].

Another approach to handling missing data that has developed considerably over the last few decades is imputation; that is, filling in the missing values with plausible substitutes such that complete-data based methods may be employed to conduct the relevant statistical analyses [14]. However, if such procedures are unprincipled, they may in fact cause more harm than good. Consequently, Little and Rubin (2002) argue in favour of explicit imputation models, rather than informal procedures such as the substitution of means.

In 1987, Rubin introduced the notion of multiple imputation in which each missing value is replaced with $m > 1$ simulated values prior to analysis [15]. Such an approach is favoured over single imputation procedures, since the uncertainty in the missing values is accounted for by the additional variation between imputations. The creation of such multiple imputations was facilitated by the developments in computer technology and the new methods for Bayesian iterative simulation, such as Markov chain Monte Carlo and the data augmentation algorithm, discovered in the late 1980s [13]. The maximum likelihood and multiple imputation methods for incomplete multivariate datasets have now become standard in many reputable statistical software packages [15].

The 1990s saw many new developments in the field of incomplete data analysis. New lines of research focus on how to handle missing values while avoiding the specification of a full parametric model for the population. New methods for non-ignorable modelling, in which the probabilities of non-response are allowed to depend upon the missing values themselves, are also proliferating. Researchers are now also beginning to assess the sensitivity of results to alternative hypotheses about the distribution of missingness [15].

1.3 Purpose of the Research

This paper will describe and evaluate the technique for multiple imputation proposed by Raghunathan *et al* (2001) and known as sequential regression multivariate imputation. This method is an application of Markov chain Monte Carlo which seeks to multiply impute missing entries with pseudo-random draws from the posterior distribution of the missing data. Such Markov chain Monte Carlo techniques have become increasingly popular in recent years and notably so in the area of Bayesian inference, where random draws from mathematically intractable posterior distributions are often desired [14].

As with all iterative simulation methods, the number of iterations necessary in order to ensure convergence is a question of both theoretical and practical interest. This is a topic that will be examined extensively in this paper in terms of the sequential regression multivariate imputation technique introduced above. Given that this algorithm is stochastic, it will converge to a probability distribution, rather than a point in the parameter space [14]. Consequently, this study will be concerned with assessing distributional convergence. The empirical findings from this investigation will be compared to the stance taken in the literature which suggests that convergence is rapid in this context.

Another subject of interest is exactly how many imputations are necessary in order for the multiple imputation model to produce unbiased and efficient estimates that accurately reflect the uncertainty of the missing data and preserve the multivariate relationships. The literature on this topic would appear to be somewhat divided. Accordingly, this matter will also be addressed in this paper in order to shed further light on the debate.

1.4 Layout of the Paper

This paper proceeds as follows. The next section provides an overview of the available literature on handling missing data. The types and patterns of non-response are explored and the various missingness mechanisms discussed. The section also considers some of the earlier methods used to address the missing data problem as well as the more modern techniques available to the analyst. Section 3 provides a guide as to the methods employed in this paper. The sequential regression multivariate imputation technique is discussed in depth, along with the methods employed to assess convergence and the optimal number of imputations. The creation of the hypothetical datasets utilised for these assessments is also described. Section 4 presents the empirical findings of this investigation. It begins by employing a hypothetical dataset to assess distributional convergence and the optimal number of imputations. The robustness of the best model is then evaluated by varying the amount of missing data and the number of covariates included in the imputation model. Finally, conclusions and recommendations for further research are presented in Section 5.

2 Literature Review

Missing or incomplete data is a pervasive problem faced by most applied researchers [9]. Although most practitioners often resort to data editing in an attempt to create an artificial appearance of completeness, ad hoc edits such as casewise deletion may actually do more harm than good. With or without missing data, the goal of the analyst is to produce valid and efficient estimates of the population of interest [15]. Simple data editing procedures without further consideration for the missingness mechanism may result in estimates which are not only biased, but do not adequately reflect the uncertainty induced by the unobserved data. Put simply, the validity and efficiency of complete-data based methods cannot be guaranteed when data are incomplete [11]. Nonetheless, the data analyst's primary focus does not rest in the estimation, prediction or recovering of missing observations. Indeed, dealing with missing data should rather be viewed as an aid in the analyst's quest for accurate inferences concerning the population of interest [15].

This section will provide an overview of the missing data problem and the methods that have been presented in the literature for handling non-response. The next subsection will consider the types and patterns of non-response that arise in different practical settings. This is followed by a discussion of the three missingness mechanisms and their theoretical consequences for data analysis. The early methods for dealing with missing data are then assessed. Thereafter, a theoretical model for imputing missing data is presented, followed by an evaluation of the methods for creating single and multiple imputations. The section is then concluded with a discussion of Markov chain Monte Carlo, an iterative procedure for multiple imputation which has become increasingly popular in recent years.

2.1 Types and Patterns of Non-Response

Survey analysts have historically distinguished between *unit non-response*, which occurs when the entire data collection procedure fails (possibly because the sampled person or household was unavailable or refused to participate at all), and *item non-response*, which implies that only partial data are available where, for example, an individual declines to respond to a subset of survey items. In practice, unit non-response is typically accounted for by reweighting the sample, whilst item non-response is either ignored (casewise deleted) or "resolved" by means of a single imputation [15]. The focus of this paper is on the latter form of non-response where the modern literature typically favours multiple imputation as the superior approach to the missing data problem.

Most multivariate datasets can be arranged in a rectangular or matrix form with rows corresponding to observational units and columns corresponding to variables. With rectangular data, there are three important classes of overall missing data patterns as illustrated graphically in Figure 1. In each panel, there are p variables denoted Y_i for $i = 1, 2, \ldots, p$. A *univariate pattern* of non-response is presented in the first panel where missing values occur on Y_p only, whilst $Y_1, Y_2, \ldots, Y_{p-1}$ are completely observed.

In the middle panel, the variables have been ordered such that if Y_j is missing for an observational unit, then $Y_{j+1}, \ldots, Y_p$ are also missing. Such a missingness pattern is referred to as a *monotone pattern* and typically arises in longitudinal studies with attrition, where Y_j represents a set of variables collected on the jth wave. Finally, the third panel displays an *arbitrary pattern* of missingness in which any set of variables may be missing for any observational unit [15]. Such a missingness pattern is characteristic of most household survey data, including that which will be analysed in this paper.

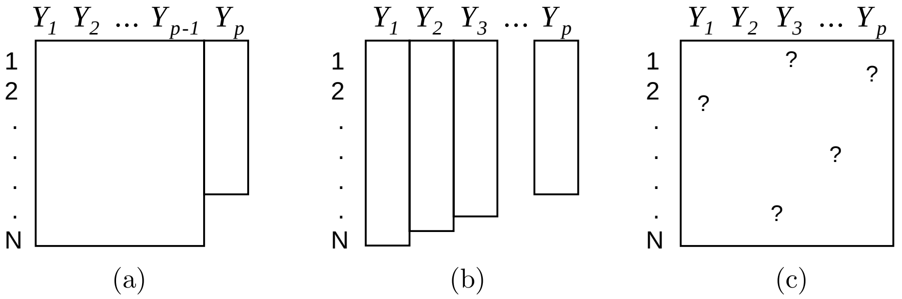
Y_1 Y_2 ... Y_{p-1} Y_p
1
2
.
.
.
.
N
(a)
Y_1 Y_2 Y_3 ... Y_p
1
2
.
.
.
.
N
(b)
Y_1 Y_2 Y_3 ... Y_p
1
2
.
.
.
.
N
?
?
?
?
?
(c)

Figure 1: Patterns of non-response in rectangular datasets: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern (Schafer and Graham, 2002, p. 150).

2.2 Missingness Mechanisms

To remain consistent with the previous nomenclature, let $\mathbf{Y}$ denote an $n \times p$ data matrix with n observational units and p variables and let $P(\mathbf{Y}|\boldsymbol{\theta})$ denote the multivariate probability distribution of the p variables governed by the parameter set of interest $\boldsymbol{\theta}$. One can then define a missingness indicator matrix $\mathbf{R}$ to identify what is observed and what is missing. The form of this latter matrix will clearly depend upon the missingness pattern inherent in $\mathbf{Y}$. For example, if the missingness pattern is univariate, $\mathbf{R}$ could represent an n-dimensional column vector of binary variables indicating whether Y_p is observed or missing for each of the n observational units. On the other hand, if a monotone pattern is observed, $\mathbf{R}$ might be an n-dimensional column vector of integer variables indicating the highest $j = \{1, 2, \ldots, p\}$ for which Y_j is observed [15]. More generally, however, $\mathbf{R} = \{r_{ij}\}$ can be regarded as an $n \times p$ matrix of binary indicators such that

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is missing} \\ 0 & \text{if } y_{ij} \text{ is observed.} \end{cases}$$

This final model of missingness allows for the arbitrary pattern in Figure 1.

In modern missing data procedures, the missingness represented by $\mathbf{R}$ is regarded as a probabilistic phenomenon, where $\mathbf{R}$ is treated as a set of random variables with a joint probability distribution [11]. In the statistical literature, this distribution is referred to as the missingness mechanism. For the discussion to follow, let $P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y})$ denote the joint probability distribution of the missingness indicator variables conditional on the response variables $\mathbf{Y}$ and governed by the nuisance parameters $\boldsymbol{\xi}$. By conditioning on $\mathbf{Y}$, this distributional form allows for the fact that the missingness $\mathbf{R}$ may be related to the data. The missingness mechanism may then be classified into three categories based on the nature of this relationship [15].

2.2.1 Missing Completely at Random

Using the notation $\mathbf{Y}_{\text{OBS}}$ and $\mathbf{Y}_{\text{MISS}}$ to represent the observed and missing portions of the dataset Y respectively, the missingness mechanism is deemed to be missing completely at random (MCAR) if

$$P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y}) = P(\mathbf{R}|\boldsymbol{\xi}). \tag{2.1}$$

In this case, the distribution of missingness or equivalently, the propensity to respond, is independent of both the observed and missing data. The missing values for a variable are therefore akin to a simple random sample of the data for that variable, such that the distribution of missing values is the same as the distribution of observed values [19].

In terms of the univariate pattern of non-response mentioned earlier, the MCAR mechanism implies that the probability that Y_p is missing for a participant does not depend upon that respondent's own observed values on $Y_1, Y_2, \ldots, Y_{p-1}$ nor does it depend upon his or her missing value on Y_p. Similarly, for the monotone pattern, MCAR means that Y_j is missing with probability unrelated to any variables in the system; that is, attrition is independent of the responses at every occasion. Finally, where an arbitrary pattern of missingness persists, the MCAR assumption requires independence between missingness and the p variables as is formalised in Equation 2.1 [15].

2.2.2 Missing at Random

The missingness mechanism is defined as missing at random (MAR) if

$$P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y}) = P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y}_{\text{OBS}}). \tag{2.2}$$

The above expression implies that the missingness $\mathbf{R}$ is independent of the missing responses $\mathbf{Y}_{\text{MISS}}$, but is dependent upon some or all of the observed variables in $\mathbf{Y}_{\text{OBS}}$ for each observational unit. Therefore, despite its name, MAR does not suggest that the missing data values are merely a random draw from the dataset $\mathbf{Y}$ as is the case for MCAR [13]. In contrast to MCAR, MAR is a less restrictive assumption in that the missing values can depend upon the response variables through the observed data. In this case, the missing values for a variable are like a simple random sample of the data for that variable within subgroups defined by the categories of the observed variables which are related to the missingness $\mathbf{R}$. Consequently, the distribution of the missing values is assumed to be the same as the distribution of the observed values within each subgroup defined by the observed variables related to missingness [19]. The data are therefore purported to be missing completely at random within each of these subgroups.

Returning to the missingness patterns discussed earlier, an MAR mechanism in the presence of a univariate pattern would imply that the probability of a response on variable Y_p may be dependent upon one or more of the observed variables $Y_1, Y_2, \ldots, Y_{p-1}$, but not upon Y_p itself. In terms of a monotone pattern, MAR means that the probability of response on the variables $Y_j, Y_{j+1}, \ldots, Y_p$ may be related only to $Y_1, Y_2, \ldots, Y_{j-1}$. Intuitively, this implies that attrition may depend upon any or all of the responses prior to the point at which the participant drops out. Consequently, MAR is often referred to as *non-informative attrition* in a longitudinal context, since the participant's propensity to respond is not "informed" by his or her missing values in the waves subsequent to dropout. Finally, with respect to an arbitrary missingness pattern, the MAR assumption implies that a participant's probabilities of response may be related only to his or her set of observed values, a set that may change from one participant to another [15].

2.2.3 Missing Not at Random

In the event that missingness is dependent not only upon the observed data, but also upon that which is missing, the missingness mechanism cannot be simplified further as was the case in Equations 2.1 and 2.2 [19]. Formally, this implies that

$$P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y}) \neq P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y}_{\mathrm{OBS}}). \tag{2.3}$$

Such a missingness mechanism is termed missing not at random (MNAR) since some residual dependence between the missingness $\mathbf{R}$ and $\mathbf{Y}_{\mathrm{MISS}}$ remains after accounting for $\mathbf{Y}_{\mathrm{OBS}}$. In terms of the univariate missingness pattern above, MNAR would imply that the propensity to respond is dependent upon Y_p itself, even after accounting for the possible relationship between missingness and $Y_1, Y_2, \ldots, Y_{p-1}$. In contrast to MAR, MNAR in the presence of a monotone pattern implies that Y_j is missing with probability related to the unobserved responses on $Y_j, Y_{j+1}, \ldots, Y_p$ subsequent to the participant's dropout. Unsurprisingly, this form of missingness mechanism is often referred to as *informative attrition* in longitudinal studies. Finally and more generally, in the presence of an arbitrary pattern of missingness, MNAR would imply that a participant's probability of response may depend upon his or her set of observed values as well as his or her unobserved values, where the set of observed and missing data is likely to differ between observational units [15].

2.3 Implications of the Missingness Mechanisms

When dealing with complete data, most traditional statistical methods assume that the data is a randomly drawn sample from the population distribution $P(\mathbf{Y}|\boldsymbol{\theta})$ governed by the parameter set $\boldsymbol{\theta}$. To the statistician, $P(\mathbf{Y}|\boldsymbol{\theta})$ has two very different interpretations. Firstly, from a frequentist or Fisherian standpoint, $P(\mathbf{Y}|\boldsymbol{\theta})$ is regarded as the repeated sampling distribution of $\mathbf{Y}$. In this context, the distribution describes the probability of observing any specific dataset among all possible datasets that could arise over hypothetical repetitions of the sampling procedure. The second interpretation treats $P(\mathbf{Y}|\boldsymbol{\theta})$ as a likelihood function for $\boldsymbol{\theta}$ conditional on the observed data, often denoted as $L(\boldsymbol{\theta}|\mathbf{Y})$ to distinguish it from the first interpretation. By substituting the realised values of $\mathbf{Y}$ into $L(\boldsymbol{\theta}|\mathbf{Y})$, the likelihood function summarises the data's evidence about the parameters $\boldsymbol{\theta}$ [15].

When data are incomplete, the full probability model to describe the data becomes the joint probability function $P(\mathbf{Y}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\xi})$. From the definitions of MCAR and MAR presented earlier, it follows immediately that the joint probability distribution of the observed data may be simplified as

$$P(\mathbf{Y}_{\mathrm{OBS}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\xi}) = \begin{cases} P(\mathbf{R}|\boldsymbol{\xi})P(\mathbf{Y}_{\mathrm{OBS}}|\boldsymbol{\theta}) & \text{if MCAR} \\ P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y}_{\mathrm{OBS}})P(\mathbf{Y}_{\mathrm{OBS}}|\boldsymbol{\theta}) & \text{if MAR.} \end{cases} \tag{2.4}$$

In the case of MCAR, the separation of the joint distribution into the product of its marginal distributions is allowed, because the missingness $\mathbf{R}$ is assumed to be independent of the observed data $\mathbf{Y}_{\mathrm{OBS}}$. Similarly, for MAR, the joint distribution may be separated into the product of the conditional distribution of $\mathbf{R}$ given $\mathbf{Y}_{\mathrm{OBS}}$ and the marginal distribution of $\mathbf{Y}_{\mathrm{OBS}}$, since missingness only depends on the observed values.

For what follows, it is further assumed that the parameter set of interest $\boldsymbol{\theta}$ and the nuisance parameter set $\boldsymbol{\xi}$ are distinct; that is, knowledge of $\boldsymbol{\theta}$ provides no information about $\boldsymbol{\xi}$ and vice versa. Such an assumption would appear to be intuitively appealing. The implication, however, is far more profound. If $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are distinct, then it follows that the joint observed-data distribution $P(\mathbf{Y}_{\mathrm{OBS}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\xi})$ may simply be replaced by the marginal observed-data distribution $P(\mathbf{Y}_{\mathrm{OBS}}|\boldsymbol{\theta})$ for the purpose of likelihood-based inferences on $\boldsymbol{\theta}$, since clearly knowledge of $\boldsymbol{\xi}$ would have no influence on such inferences. Consequently, the observed-data likelihood function for $\boldsymbol{\theta}$ may be obtained by integrating over the range of the marginal distribution of $\mathbf{Y}$ with respect to the missing data as follows

$$L(\boldsymbol{\theta}|\mathbf{Y}_{\mathrm{OBS}}) = P(\mathbf{Y}_{\mathrm{OBS}}|\boldsymbol{\theta}) = \int P(\mathbf{Y}|\boldsymbol{\theta}) \ \mathrm{d}\mathbf{Y}_{\mathrm{MISS}}. \tag{2.5}$$

Note, however, that it is not always true that this expression is the correct sampling distribution for the observed data or the correct likelihood function for $\boldsymbol{\theta}$. Donald B. Rubin (1976) was the first to identify the conditions under which Equation 2.5 is a proper sampling distribution and a proper likelihood function. Interestingly, these conditions are not identical. For Equation 2.5 to be a correct sampling distribution, the missingness mechanism should be MCAR, whilst only the MAR assumption is necessary for the expression to yield a proper likelihood function. The weaker condition for a proper likelihood function as oppose to a proper sampling distribution suggests that missing-data procedures based on likelihood principles are generally more useful than those derived solely on the basis of repeated sampling arguments. Many of the older data editing procedures such as complete-case analysis bear no relationship to the likelihood function and may be valid only under MCAR. Even in the unlikely event that MCAR does hold, such methods may still be inefficient. Hence, methods which treat Equation

2.5 as a likelihood tend to be more powerful and better suited to real world applications in which MCAR is often violated [15].

From Equation 2.5, it follows that inferences on $\boldsymbol{\theta}$ under MCAR or MAR may be based solely on the observed-data likelihood function $L(\boldsymbol{\theta}|\mathbf{Y}_{\mathrm{OBS}})$ without concern for the missingness mechanisms $P(\mathbf{R}|\boldsymbol{\xi})$ and $P(\mathbf{R}|\boldsymbol{\xi}, \mathbf{Y}_{\mathrm{OBS}})$ respectively. Consequently, the MCAR and MAR missingness mechanisms are said to be *ignorable* [19].

In contrast, an MNAR mechanism implies that Equation 2.5 is neither a proper sampling distribution nor a proper likelihood function for $\boldsymbol{\theta}$. In order to obtain either of these, one would instead need to evaluate the following integral with respect to the missing data,

$$P(\mathbf{Y}_{\mathrm{OBS}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\xi}) = \int P(\mathbf{R}|\mathbf{Y}, \boldsymbol{\xi}) P(\mathbf{Y}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{Y}_{\mathrm{MISS}}, \tag{2.6}$$

which clearly requires an explicit model for $P(\mathbf{R}|\mathbf{Y}_{\mathrm{OBS}}, \mathbf{Y}_{\mathrm{MISS}}, \boldsymbol{\xi})$. Consequently, this form of missingness is often termed *non-ignorable* missingness since the missingness mechanism cannot be ignored when drawing inferences on $\boldsymbol{\theta}$. In most instances, the missingness model $P(\mathbf{R}|\mathbf{Y}, \boldsymbol{\xi})$ is a nuisance since questions of substantive interest usually pertain to the distribution of $\mathbf{Y}$ rather than that of $\mathbf{R}$. Nonetheless, the distinctness assumed between $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ under MCAR and MAR does not apply under MNAR, since by definition the model for $\mathbf{R}$ does contain information concerning $\boldsymbol{\theta}$. Consequently, the evidence about $\boldsymbol{\theta}$ from Equation 2.6 may in fact suggest a very different story to that of Equation 2.5 [15].

2.4 Early Methods for Dealing with Missing Data

Prior to the mid 1970s, missing data was handled mostly by means of ad hoc procedures which often express little concern if any for the underlying missingness mechanism. Casewise deletion, also known commonly as listwise deletion and complete-case analysis, was the most popular of such procedures primarily due to its simplicity. This method quite simply discards any observational unit whose information is incomplete. Available-case analysis is an extension of casewise deletion, but differs in that it uses different sets of observational units to estimate different parameters. Whilst casewise deletion eliminates all cases that have any missing values on any variables regardless of the parameters being estimated, available-case analysis will only exclude those cases for which data is missing on the variables necessary to estimate the parameters of interest. Available-case analysis therefore makes more efficient use of the information contained in the sample relative to casewise deletion [15]. Indeed, available-case analysis is still the most common way in which analysts deal with missing data today and is often the default in many statistical software packages.

Both complete-case and available-case analysis implicitly assumes that non-response is missing completely at random and non-respondents are assumed to be no different from respondents. If the missingness mechanism is in fact MCAR, it follows that complete-data based methods applied to only the observed data will produce unbiased estimates of the parameter set $\boldsymbol{\theta}$, albeit subject to a loss in precision due to the smaller sample size. In practice, however, an MCAR mechanism is unlikely and the observed cases could therefore be unrepresentative of the target population. If the departures from MCAR are not serious, the impact of the bias may not be important, although it may be difficult to assess the magnitude and direction of such biases. On the other hand, where such departures are substantial, parameter estimates may be severely biased [15].

In some non-MCAR situations, it is possible to reduce biases from casewise deletion by reweighting the sample. After the removal of incomplete cases, the remaining complete cases are weighted so that their distribution more closely resembles that of the full sample with respect to the weighting variables. Weights are derived by estimating the propensity to respond from the data using, for example, logistic or probit regression models. Weighting can reduce biases due to differential response related to the variables used to model the probability of response, but it cannot account for biases that arise from variables that are unused or unmeasured [15].

In fully parametric models, maximum likelihood estimates can often be calculated directly from the incomplete data by specialised numerical methods. An example of such a method is the Expectation Maximisation, or EM, algorithm, which is a popular approach to incomplete data analysis. The key idea of EM is to solve a difficult incomplete-data estimation problem by iteratively solving an easier complete-data problem. Intuitively, this involves "filling in the missing data" with the best guess of what it might be under the current estimate of the unknown parameters. The parameters are then re-estimated from the observed and "filled-in" data. In this context, "filling in the missing data" is not a literal process as is the case for the imputation methods described next. Instead, it refers to filling in the complete data sufficient statistics in the likelihood function. This method is therefore more computationally efficient than the simulation-based imputation methods described below. However, when missing data are a nuisance rather than the focus of enquiry, a simple approximate solution with good properties may be preferable to one that is more efficient, but problem specific and complicated to implement [13].

2.5 The Imputation Model

Imputation, the practice of filling in missing data with plausible values, has emerged in recent years as an attractive alternative to analysing incomplete datasets [14]. Whilst both the complete-case and available-case analyses blindly ignore the available information for each observational unit, imputation procedures attempt to harness this information to provide reasonable estimates of the missing data. These methods can be applied to impute one value for each missing item or to impute several values to allow for the inherent uncertainty in the imputation procedure. The former is referred to as *single imputation*, whilst the latter is termed *multiple imputation* [6]. Operationally, imputation resolves the missing data problem at the outset and allows the analyst to proceed relying on familiar complete-data based statistical methods. However, a naïve or unprincipled imputation method may actually create more problems than it solves, distorting estimates, standard errors and multivariate relationships [14].

Imputations are means or draws from the predictive distribution of the missing values given the observed data and therefore require a method for creating this predictive distribution for the imputation [6]. Once the missing values have been imputed, the analyst may then proceed to draw inferences concerning $\boldsymbol{\theta}$. Theoretically, this idea may be presented in two ways. From a frequentist perspective, imputation implies simulating a random draw or mean from the conditional distribution,

$$P(\mathbf{Y}_{\mathrm{MISS}}|\mathbf{Y}_{\mathrm{OBS}}, \boldsymbol{\theta}) = \frac{P(\mathbf{Y}|\boldsymbol{\theta})}{P(\mathbf{Y}_{\mathrm{OBS}}|\boldsymbol{\theta})} \tag{2.7}$$

In practice, however, the parameter set $\boldsymbol{\theta}$ is unknown and must be estimated from the observed data. A draw or mean may then be taken from $P(\mathbf{Y}_{\mathrm{MISS}}|\mathbf{Y}_{\mathrm{OBS}}, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is an observed-sample estimate of $\boldsymbol{\theta}$ [15].

The Bayesian perspective, on the other hand, allows for prior information concerning $\boldsymbol{\theta}$ to enter explicitly into the model such that the observed-data posterior distribution on $\boldsymbol{\theta}$ is proportional to the likelihood of $\boldsymbol{\theta}$ given the data multiplied by the prior information,

$$P(\boldsymbol{\theta}|\mathbf{Y}_{\mathrm{OBS}}) \propto L(\boldsymbol{\theta}|\mathbf{Y}_{\mathrm{OBS}})\pi(\boldsymbol{\theta}) \tag{2.8}$$

where $\pi(\boldsymbol{\theta})$ represents the prior distribution of $\boldsymbol{\theta}$. Note that as the sample size increases, the sample estimate of $\boldsymbol{\theta}$ will be weighted more heavily relative to this prior information such that the prior distribution will exert less influence in inferences on $\boldsymbol{\theta}$ [15]. Where no prior information is assumed, a non-informative prior distribution may be used for $\boldsymbol{\theta}$ such that the observed-data posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ is proportional to the likelihood function. The conditional distribution of $\mathbf{Y}_{\text{MISS}}$ given $\mathbf{Y}_{\text{OBS}}$ may be obtained by averaging over the observed-data posterior distribution of $\boldsymbol{\theta}$ as follows

$$P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}) = \int P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})\ \mathrm{d}\boldsymbol{\theta} \tag{2.9}$$

where $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \boldsymbol{\theta})$ is the conditional predictive distribution of $\mathbf{Y}_{\text{MISS}}$ given $\mathbf{Y}_{\text{OBS}}$ and $\boldsymbol{\theta}$. It therefore follows that if the values for the parameters $\boldsymbol{\theta}$ can be drawn from their posterior distribution, then the corresponding draws from the conditional predictive distribution given $\mathbf{Y}_{\text{OBS}}$ and $\boldsymbol{\theta}$ are the draws from the posterior predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$. Random draws from or the posterior mean of this distribution may therefore be used to impute missing values [19].

Irrespective of the perspective adopted, imputing under either Equation 2.7 or Equation 2.9 assumes that the missingness mechanism is ignorable; that is, inferences concerning $\boldsymbol{\theta}$ are assumed to be unrelated to the distribution of missingness. Consequently, all methods which impute under this model assume either MCAR or MAR or both [15].

Three types of uncertainties are involved in the imputation process, which are best illustrated by recourse to the posterior predictive distribution discussed above. The first uncertainty is that which arises in the modelling of the joint distribution of the response variables and the missingness indicators $P(\mathbf{Y}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\xi})$. This uncertainty includes any assumptions made concerning the missingness mechanism itself. The second uncertainty pertains to the random sampling from the conditional predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \boldsymbol{\theta})$ when the values of $\mathbf{Y}_{\text{OBS}}$ and $\boldsymbol{\theta}$ are known. The third and final uncertainty arises because the value of $\boldsymbol{\theta}$ in the conditional predictive distribution is in fact unknown and therefore requires a random draw from its posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ [19]. In light of the first uncertainty, the analyst has little option but to try different imputation models and make his or her assumptions explicit. With respect to the latter two uncertainties, however, the choice of imputation procedure has a large part to play in ensuring that the uncertainties are adequately reflected at the analysis stage.

2.6 Single Imputation

In light of the above theoretical foundation, some of the more popular single imputation methods available to the analyst will now be reviewed.

2.6.1 Imputing Means from Predictive Distributions

Unconditional mean substitution is a popular approach to imputation in which missing values are replaced by the average of the observed values for that variable. Whilst the average for the variable will clearly be preserved, other properties of the distribution such as the variance and quantiles will be altered with potentially serious ramifications. In particular, the precision of estimates of $\boldsymbol{\theta}$ will be overstated since the sample variance will be downwardly biased and the sample size will be artificially inflated. Moreover, Schafer and Graham (2002) show that with 25% missing data, the coverage of a 95% confidence interval on $\boldsymbol{\theta}$ drops to only 86%, corresponding to a Type 1 error rate that is nearly three times its nominal value. Furthermore, unconditional mean imputation disregards multivariate relationships and therefore also corrupts the covariances and intercorrelations between variables [15].

Cell mean substitution, by contrast, does attempt to take account of the multivariate relationships in the dataset. Instead of imputing the observed-data mean for each variable, respondents are divided into cells or classes on the basis of several known variables and the mean values within these cells are used for imputation [6]. For example, under the univariate missingness pattern of Figure 1, cell mean imputation may be executed by estimating a regression model based on the observed data and then using this model to predict values for non-respondents on Y_p, conditional on their observed values for $Y_1, Y_2, \ldots, Y_{p-1}$. This form of cell mean imputation is often referred to as *regression imputation* and may be extended to other missingness patterns. The technique allows for an individual's response probability to depend upon the observed data, but assumes that the missingness mechanism is MCAR within cells. As with unconditional mean imputation, however, this technique continues to understate the standard errors of estimates. Furthermore, the method is not recommended for the analysis of covariances or correlations as it overstates the strength of the relationship between the dependent and predictor variables. Moreover, where no such relationship exists, this method simply reduces to ordinary mean substitution [15].

2.6.2 Imputing Draws from Predictive Distributions

The conceptual basis of the mean substitution methods presented above, i.e. to predict missing values, is somewhat misguided. It is generally more desirable to preserve a variable's distribution, or perhaps more accurately, the multivariate distribution of the dataset. Consequently, a wide array of single imputation methods has been developed to more effectively preserve distributional shape [7]. One such class of procedures known as *hot deck* *imputation* involves substituting missing values with observed values drawn from similar responding units. This non-parametric technique is common in survey practice and may involve very elaborate schemes for selecting similar observational units for imputation [6]. One such scheme involves dividing observational units into cells and then replacing each missing value within the cell with a random draw from the observed values. In some instances, an entire observational unit may be drawn at random from within a cell to impute for all the missing values on another observational unit within the same cell. This method partially resolves the issue of understating uncertainty, because the variance of each variable is not distorted to the same degree as is the case with unconditional mean substitution. However, hot deck imputation still corrupts correlations and other measures of association [15].

Several single imputation methods have also evolved to exploit information which may be available outside the realised sample. *Substitution*, a method for dealing with unit non-response at the fieldwork stage of the survey, replaces non-responding units with alternative units not selected into the sample. Erroneously, such datasets are often treated as if they are complete. Although it is true that these data do not contain missing values, this is only because non-respondents have been substituted or, equivalently, imputed with the values of respondents. Hence, if the non-response mechanism is not MCAR such that respondents do differ systematically from non-respondents, estimates of the population parameters may be seriously biased [6].

A similar method to substitution is *cold deck imputation* where missing values are substituted by a constant value from an external source, such as a value from a previous realisation of the same survey [6]. As with substitution, the imputed values of cold deck imputation are commonly regarded as part of the complete sample, thereby subjecting inferences to the biases that may arise from differences between the current sample and the external source from which the imputations are extracted.

It was noted earlier that whilst unconditional mean substitution completely disregards the possible relationships between the imputed variable and its covariates, conditional or cell mean substitution will overstate these relationships to the extent that the R^2 measure among the imputed values will be unitary [15]. In an attempt to reach a compromise between these two extremes, *stochastic mean substitution* might be employed whereby imputed values are randomly generated from a specified theoretical distribution (usually Gaussian) with mean equivalent to the cell mean and variance equal to the cell variance. By imputing random draws from the predictive distribution rather than means, this method will not dilute or exaggerate the relationship between the imputed variable and the covariates selected for the imputation. This method does not, however, resolve the understated standard error of estimate and will continue to corrupt the multivariate relationships within cells.

An immediate extension of this method is *stochastic regression imputation.* In this case, missing values are replaced by a value predicted by regression imputation plus a residual drawn to represent the uncertainty in the predicted value [6]. With an ordinary least squares model, this residual will be a random draw from the Gaussian distribution with zero mean and variance estimated by the mean squared error of the model. Stochastic regression imputation is, however, not limited to the normal linear regression model and may be implemented with generalised linear models such as Poisson and generalised logistic regression. The technique allows for an MAR missingness mechanism in that the propensity to respond may depend upon any of the variables included in the appropriate regression model.

Stochastic regression imputation is essentially an attempt to address the inherent uncertainty in sampling from the predictive distribution, an issue that is disregarded when deterministic means are imputed. However, once the missing values have been imputed, the dataset is treated as if it were completely observed, ignoring the fact that the imputed values are not much more than calculated guesses. Consequently, the standard errors of estimates will not reflect the variability of each imputed value that would arise from hypothetical repetitions of the imputation process. Recall that this variability arises due to both the uncertainty surrounding the random draw from the conditional predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \boldsymbol{\theta})$ and the uncertainty concerning the random draw of $\boldsymbol{\theta}$ from its posterior distribution. Failure to fully account for these uncertainties is the fundamental flaw with most single imputation methods, a problem which has spurred on the development of techniques for multiple imputation.

2.7 Multiple Imputation

Multiple imputation was first proposed by Rubin in the 1970s as a possible solution to the problem of survey non-response and has since emerged as a flexible alternative to likelihood methods for a wide variety of missing data problems. This class of techniques retains the attractiveness of stochastic single imputation from the predictive distribution of $\mathbf{Y}_{\text{MISS}}$, whilst simultaneously addressing the problem of understating uncertainty [15]. Multiple imputation is a principled method for handling missing data and consists of three steps. The first step is to create $m > 1$ plausible versions of the complete data by imputing each missing value m times using m independent draws from an appropriate imputation model conditional on the observed data. This procedure is presented schematically in Figure 2.

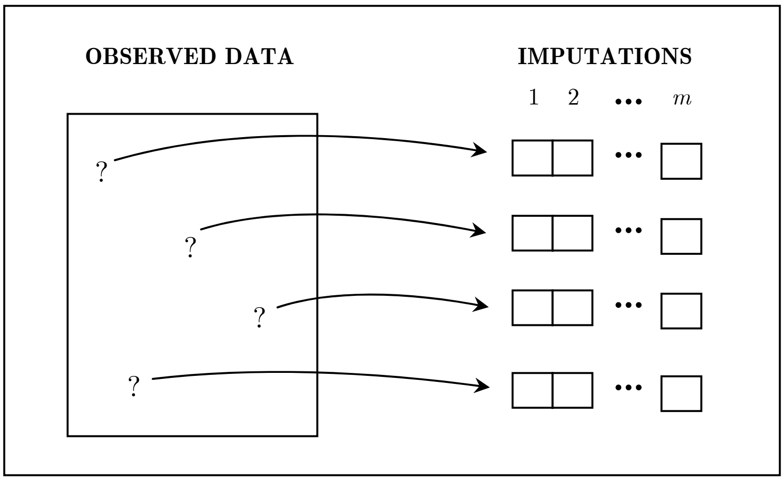

Figure 2: Schematic representation of multiple imputation (Schafer and Graham, 2002, p. 165)

Initially, m independent values of the parameters $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(m)}$ will need to be simulated for use in the model for $\mathbf{Y}_{\text{MISS}}$ given $\mathbf{Y}_{\text{OBS}}$. Treating parameters as random, rather than fixed, is an essential part of multiple imputation and the technique therefore falls more naturally into the Bayesian context [15]. In the second step, the m imputed datasets are treated as if they were entirely observed and analysed individually by standard complete-data methods [19]. Note that for $m = 1$, this procedure is akin to single imputation. In the third and final step, the results from the m analyses are combined in a simple and appropriate manner to obtain overall estimates and standard errors that reflect not only sampling variation, but also the uncertainty associated with the imputed values [15]. These estimates provide the basis for what Rubin termed *repeated imputation inference* concerning the parameter set $\boldsymbol{\theta}$ [12].

In order to understand how multiple imputation accounts for the uncertainty introduced by the imputation procedure, it is useful to consider the rules for combining the results from the m imputed datasets. This set of rules has been dubbed "Rubin's rules" in the literature, after its devisor Donald B. Rubin. After imputing $\mathbf{Y}_{\text{MISS}}$ with m sets of conditionally independent draws from the posterior predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$, or equivalently from $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ based on the observed data, the repeated imputation inference is achieved as follows. Suppose that $Q \in \boldsymbol{\theta}$ represents a scalar population quantity to be estimated and let $\hat{Q}^{(t)} = \hat{Q}(\mathbf{Y}_{\text{OBS}}, \mathbf{Y}_{\text{MISS}}^{(t)})$ be the repeated estimates along with their estimated squared standard errors $\hat{U}^{(t)} = \hat{U}(\mathbf{Y}_{\text{OBS}}, \mathbf{Y}_{\text{MISS}}^{(t)})$ from the imputed datasets $\{\mathbf{Y}_{\text{OBS}}, \mathbf{Y}_{\text{MISS}}^{(t)} : t = 1, \ldots, m\}$. Then the overall estimate of

Q is simply taken to be average of the repeated estimates,

$$\bar{Q} = \frac{1}{m}\sum_{t=1}^{m} \hat{Q}^{(t)}. \tag{2.10}$$

In a multiple imputation context, the variance of the estimate $\bar{Q}$ has two components, namely the average *within-imputation* variance,

$$\bar{U} = \frac{1}{m}\sum_{t=1}^{m} \hat{U}^{(t)} \tag{2.11}$$

and *between-imputation* variance,

$$B = \frac{1}{m-1}\sum_{t=1}^{m} (\hat{Q}^{(t)} - \bar{Q})^2. \tag{2.12}$$

The total variance is then the adjusted sum of these two components given by

$$T = \bar{U} + (1 + m^{-1})B. \tag{2.13}$$

By accounting for both within- and between-imputation variation, the uncertainties in the imputed data are generally correctly incorporated into the final inference. This overcomes the major drawback of single imputation, which underestimates total variation because it has zero between-imputation variance [19]. The uncertainties surrounding the random draws from the conditional predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \boldsymbol{\theta})$ and the posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ are reflected in the estimated variation between imputations. Furthermore, uncertainty in the choice of imputation model may also be addressed by utilising two or more models for non-response. Differences between imputation models will again be reflected in the between-imputation variation, allowing the analyst to assess the sensitivity of inferences across models. This may prove critical where the missingness mechanism is non-ignorable [6].

Naturally, confidence intervals constructed with this larger standard error of estimate will be wider, thereby increasing coverage in repeated sampling. With respect to confidence intervals and hypothesis testing, Rubin (1987) suggested the use of the Student-t approximation,

$$\sqrt{T}(\bar{Q} - Q) \sim t_v \tag{2.14}$$

with degrees of freedom

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

The degrees of freedom may vary from $m-1$ to ∞, depending on the rate of missing information. The fraction of information missing due to non-response is given by

$$\gamma = \frac{r + 2/(v+3)}{r+1}, \tag{2.15}$$

where

$$r = \frac{(1+m^{-1})B}{\bar{U}}$$

is the relative increase in total variance due to non-response [12]. Note that the fraction of missing information γ is not the same of the fraction of missing data. A high rate of missing observations on a variable does not automatically translate into high rates of missing information, since the variable may be highly correlated with other variables that are more fully observed such that little information is lost through missingness [13].

When the degrees of freedom are large, the Student-t distribution is approximately normal, the total variance is well estimated and there is little to be gained from increasing the number of imputations [15]. Rubin (1987) computed a measure of efficiency based on the rate of missing information and the number of imputations relative to the hypothetical case where $m = \infty$,

$$\lambda = \left(1 + \frac{\gamma}{m}\right)^{-1/2} . \tag{2.16}$$

Note that λ is measured in units of standard errors. Table A1 in Appendix A presents Rubin's efficiency estimates for various values of γ and m, showing that in cases with low levels of missing information, as few as two or three imputations is "nearly fully efficient" (Rubin, 1987, p. 114). With 50% missing information, only five imputations are necessary to obtain an efficiency of 95% or equivalently a standard error that is only $\sqrt{1+0.5/5} = 1.049$ times as large as an imputation with $m = \infty$ [13]. Indeed, $m = 5$ would appear to be the favourite choice in the literature and Schafer (1999) claims that there is little to no practical benefit from using more than five to ten imputations. However, some authors have expressed concern over this rule-of-thumb. For example, Royston (2004) argues that the coefficient of variation of the confidence coefficient $t_v\sqrt{T}$ is particularly high for such low values of m, resulting in unreliable confidence intervals for Q. Consequently, Royston proposes that the value of m should be chosen such that the coefficient of variation for the confidence coefficient of the worst-case parameter is less than 5%. In his empirical studies, this would require m to be at least 20 and possibly more [10]. Thus, there would appear to be no hard and fast rule for the choice of m.

The validity of multiple imputation relies on the manner in which the imputations are created and how that procedure relates to the subsequent analyses of the data [15]. If the imputation model does not preserve the distributional relationships between the missing values and the observed values, it follows that inferences on these relationships from imputed complete data will generally be biased. For example, if the imputation model does not include variables to be used in the inferences from the imputed complete data, correlations between these omitted variables and the imputed variables will be attenuated to zero. Furthermore, if the multiple imputations are not based on conditionally independent samples from the model for $\mathbf{Y}_{\text{MISS}}$ given $\mathbf{Y}_{\text{OBS}}$, the between-imputation variance will typically be understated [19]. Construction of an appropriate imputation model is therefore non-trivial and will require careful consideration of the subsequent complete-data analyses to be conducted.

2.8 Markov Chain Monte Carlo

In many missing data problems, the observed-data posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ is intractable and cannot be easily summarised or simulated. The first step in multiple imputation, however, specifies that m parameter sets be randomly drawn from this distribution, thereby allowing the analyst to further simulate the draws of $\mathbf{Y}_{\text{MISS}}$ from its predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$. Consequently, it is critical that this apparent problem be resolved. It turns out that if the observed data is augmented by an assumed set of values for $\mathbf{Y}_{\text{MISS}}$, the resulting complete-data posterior $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}}, \mathbf{Y}_{\text{MISS}})$ becomes much easier to handle. This led to the development of the data augmentation algorithm by Tanner and Wong in 1987 [17]. In recent years, this method has become increasingly popular for the purpose of multiple imputation within a Markov chain Monte Carlo framework.

Markov chain Monte Carlo is an iterative sampling scheme and has been applied successfully to a broad range of statistical problems [19]. In contrast to standard Monte Carlo methods that produce a set of independent simulated values from a desired probability distribution, Markov chain Monte Carlo produces chains in which each of the simulated values is mildly dependent on the preceding value. Formally, a Markov chain is a stochastic process with the property that any specified state in the series $\boldsymbol{\theta}^{(t)}$ is dependent only on the previous value in the chain $\boldsymbol{\theta}^{(t-1)}$ and is therefore conditionally independent of all other previous states $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(t-2)}$. Hence, a Markov chain wanders around the state space remembering only where it has been in the last period. This property turns out to be enormously useful and is exploited in the Markov chain Monte Carlo methods. The basic principle behind Markov chain Monte Carlo is that once this Markov chain has run through a sufficient number of iterations, it will find its way to the desired posterior distribution of interest. By letting the chain wander around, it will then produce a sample from this distribution that is only mildly non-independent. These sample values may then be used to describe the limiting distribution [3]. In a multiple imputation context, the target distribution of interest is the joint conditional distribution of $\mathbf{Y}_{\text{MISS}}$ and $\boldsymbol{\theta}$ given $\mathbf{Y}_{\text{OBS}}$, that is, $P(\mathbf{Y}_{\text{MISS}}, \boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$.

The Markov chain Monte Carlo method suggests the following iterative sampling scheme for the imputation of missing values. Given a current estimate $\boldsymbol{\theta}^{(t)}$ of the parameter set, impute the missing data from the conditional predictive distribution of $\mathbf{Y}_{\text{MISS}}$,

$$\mathbf{Y}_{\text{MISS}}^{(t+1)} \sim P\left(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \boldsymbol{\theta}^{(t)}\right). \tag{2.17}$$

Next, conditioning on $\mathbf{Y}_{\text{MISS}}^{(t+1)}$, a new value of $\boldsymbol{\theta}$ can be drawn from its augmented complete-data posterior distribution,

$$\boldsymbol{\theta}^{(t+1)} \sim P\left(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}}, \mathbf{Y}_{\text{MISS}}^{(t+1)}\right). \tag{2.18}$$

Recall that it is generally easier to simulate this distribution, rather than the observed-data posterior $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$. Repeating this process from a starting value $\boldsymbol{\theta}^{(0)}$ yields the stochastic sequence $\{\boldsymbol{\theta}^{(t)}, \mathbf{Y}_{\text{MISS}}^{(t)} : t = 1, 2, \ldots\}$ which converges in distribution to $P(\mathbf{Y}_{\text{MISS}}, \boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ as required. Note further that the subsequences $\{\boldsymbol{\theta}^{(t)} : t = 1, 2, \ldots\}$ and $\{\mathbf{Y}_{\text{MISS}}^{(t)} : t = 1, 2, \ldots\}$ have the posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ and the predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$ as their marginal stationary distributions respectively. For sufficiently large t, one can regard $\boldsymbol{\theta}^{(t)}$ as an approximate draw from $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ and $\mathbf{Y}_{\text{MISS}}^{(t)}$ as an approximate draw from $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$. Consequently, Tanner and Wong (1987) refer to Equation 2.17 as the Imputation or I-step and Equation 2.18 as the Posterior or P-step. With respect to the starting point $\boldsymbol{\theta}^{(0)}$, the maximum likelihood estimate based on the observed data is typically regarded as a good choice [13].

Multiple imputations of $\mathbf{Y}_{\text{MISS}}$ should ideally be independent given $\mathbf{Y}_{\text{OBS}}$. However, even after a long run of the Markov chain, such multiple imputations cannot be acquired by successive iterations, because the successive iterations of a single chain tend to be correlated. Instead, one may use the values produced after every kth iteration of a single chain, where k is large enough such that the dependence between the imputed values is negligible. Alternatively, one can generate m independent chains of length k and take the final values of each chain as the m imputations for $\mathbf{Y}_{\text{MISS}}$. Again, the choice of k should be such so as to ensure that successive imputations are statistically independent and distributional convergence in the Markov chain $\{\boldsymbol{\theta}^{(t)}, \mathbf{Y}_{\text{MISS}}^{(t)}\}$ has been reached. Monitoring distributional convergence, however, is a far more complicated task than, for example, monitoring pointwise convergence. This issue will be examined empirically in this paper. Nonetheless, the value of Markov chain Monte Carlo lies in the fact that it allows one to avoid the complicated analytical calculations of the observed-data posterior distribution for the unknown parameters $\boldsymbol{\theta}$ and the posterior predictive distribution of $\mathbf{Y}_{\text{MISS}}$ given $\mathbf{Y}_{\text{OBS}}$ [19]. Indeed, the Markov chain Monte Carlo approach to multiple imputation would appear to be the most favourable means for handling missing data in the literature at present.

3 Methodology

Survey datasets such as Statistics South Africa's *Labour Force Survey* consist of large numbers of variables of a wide variety of distributional forms. These variables may be continuous, counts, dichotomous, polytomous and even semi-continuous. Moreover, missing values are likely to be dispersed across the dataset, often resulting in the arbitrary missingness pattern presented earlier. Postulating a full imputation model of the form $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$ may be extremely difficult in such instances and is often unnecessary.

3.1 Imputation Technique

This paper will employ the multiple imputation technique proposed by Raghunathan *et al* (2001) for dealing with complex data structures where explicit full multivariate imputation models cannot be easily formulated. The technique, known as sequential regression multivariate imputation, involves imputing missing values on a variable-by-variable basis, conditioning on all the observed and imputed variables. At the outset, all variables are ordered with respect to the amount of missing data they contain. Beginning with the variable with the least number of missing values, imputations are then generated through random draws from the predictive distribution of a generalised linear model with the observed variables as covariates and parameters drawn randomly from their joint posterior distribution. After imputing for the first variable, one then proceeds to impute the variable with the second least number of missing values, conditioning on the previously imputed variable in addition to the fully observed variables. This process is repeated for each variable in the order of missing data, varying the type of regression model according to the type of dependent variable being imputed. Independent variables include all other complete variables, either observed or imputed, for each individual. The technique therefore assumes an ignorable missingness mechanism. The imputations are defined as draws from the posterior predictive distribution specified by the regression model with a non-informative prior distribution for the regression parameters [9].

The aforementioned sequence of imputations is repeated in a cyclical manner, each time replacing the previously drawn values with the latest updates. In this sense, the technique may be viewed as an application of Markov chain Monte Carlo. The imputations for $\mathbf{Y}_{\text{MISS}}$ on round $t + 1$ are taken as the random draws from the predictive distribution $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}}, \boldsymbol{\theta}^{(t)})$, conditioning on the parameter values obtained in the previous round. The regression parameters $\boldsymbol{\theta}$ are then updated by drawing from the complete-data posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}}, \mathbf{Y}_{\text{MISS}}^{(t+1)})$, augmented by the latest imputed values for $\mathbf{Y}_{\text{MISS}}$. The values for $\mathbf{Y}_{\text{MISS}}$ are then replaced by random draws from the predictive distribution, conditioning on the updated parameter set $\boldsymbol{\theta}$ [13]. This process is repeated until distributional convergence in the parameters $\boldsymbol{\theta}$ and $\mathbf{Y}_{\text{MISS}}$ is achieved. Although it is theoretically possible that the Markov chain may not converge to a stationary distribution, Raghunathan *et al* (2001) have not encountered this problem in their empirical work. Another issue of practical and theoretical interest is the optimal length of this chain for imputation purposes. Surprisingly, this topic has not received much attention in the literature and will be investigated further in this paper. Once the chain has converged, the results are stored and the entire procedure is repeated m times to produce m imputed datasets. The results from these datasets may then be combined using Rubin's rules to provide repeated imputation inferences for $\boldsymbol{\theta}$.

3.2 Variables Included in the Imputation Model

The choice of variables to be included in an imputation model is informed by three considerations. Firstly, the analyst must consider the nature of the statistical analyses to be performed on the imputed dataset. As discussed previously, failure to include variables in the imputation model that are to be used for subsequent analyses will lead to biases in inferences concerning the relationships between these omitted variables and the imputed variables [19]. Secondly, the independent variables included in the imputation model should be able to explain a reasonable proportion of the variation in the target variable. Finally, all variables that are known to have influence on the occurrence of missing data should appear in the model, such that the missingness mechanism may be assumed to be MCAR within cells composed of these variables [18].

For the current investigation, a hypothetical dataset was created based on Statistics South Africa's *Labour Force Survey* of September 2003. The original dataset was first restricted to include only the variables displayed in Table 1, which are all potential predictors for monthly earnings. Note that the prevalence of missing data on all the variables except monthly earnings is negligible, often of the order of less than 1% of the sample. By contrast, the proportion of missing data on the monthly earnings variable is non-trivial, with over one third of the sample failing to provide a point earnings value. An imputation procedure would thus seem appropriate for this variable.

In order to address the questions raised earlier concerning the optimal length of a Markov chain and the number of parallel chains required under each of the three missingness mechanisms, three artificial datasets were created by dropping all observational units with missing values on any of the variables in Table 1 and then simulating missingness under each of the three mechanisms. On each of age, hours worked per week, skills training, years of

Variables	**% Missing**	**Data Type**	**Model**
Province	–	Polytomous	–
Gender	–	Dichotomous	–
Racial Group	–	Polytomous	–
Age	0.15%	Count	Poisson
Hours Worked per Week	0.26%	Count	Poisson
Skills Training	0.33%	Dichotomous	Logistic
Years of Education	0.67%	Count	Poisson
Occupation and Sector	0.99%	Polytomous	Multinomial Logit
Employment Type	1.37%	Polytomous	Multinomial Logit
Monthly Earnings - Bands	8.75%	Polytomous	Ordered Logit
Monthly Earnings - Point	38.64%	Continuous	Lognormal

Table 1: Variables included in Imputation Model

education, occupation and sector and employment type, an MCAR mechanism was simulated by setting a randomly selected sample to missing for each variable in proportion to the missingness in the original dataset. Missing values on monthly earnings were simulated under each of the three missingness mechanisms in turn, holding the missing values constant on all other variables.

An MCAR mechanism was simulated on monthly earnings by setting a simple random sample of earnings values to missing. For the MAR mechanism, missingness was simulated so as to be dependent upon only a subset of covariates, namely province, racial group, gender, age and years of education. Disproportionate random samples were drawn from each cell created by these variables. Since some groups are clearly larger than others (for example, Gauteng province has many more observations than Limpopo province), sampling the same number of observational units from each group will induce a relationship between the propensity to respond on earnings and the covariates. The significance of these relationships was confirmed by means of Pearson's χ^2 tests. Finally, an MNAR mechanism was simulated by drawing random samples of similar size from each of the earnings bands provided in the *Labour Force Survey.* This disproportionate sampling forces a dependency between the propensity to respond and the earnings variable itself, since monthly earnings are heavily skewed with very many individuals in the lower earnings bands and much fewer numbers in the higher brackets. Consequently, proportionately more wealthier individuals had their earnings set missing relative to poorer individuals, effectively inducing a downward bias on mean monthly earnings.

With reference to the prevalence of missing data in Table 1, the sequential regression multivariate imputation technique was implemented in this study as follows. First, age was imputed using the completely observed variables, province, gender and racial group, as predictors. After imputing age, hours worked per week was imputed with province, gender, racial group and age as covariates. The procedure continues in this fashion, proceeding down the list and adding the newly imputed variables to the covariates until eventually point monthly earnings is imputed, conditional on all the complete variables that precede it. This process would comprise one run of the Markov chain and should be looped until convergence is achieved. To multiply impute the missing data, m such chains should be formed by choosing different starting points for each chain.

3.3 Creating Imputations from Generalised Linear Models

As mentioned earlier, the type of regression model employed at each step in the Markov chain will depend upon the data type of the dependent variable. Table 1 indicates the data type of each variable and the generalised linear model that is appropriate when imputing that variable. Note that since province, gender and racial group do not require imputation, they will not enter into the imputation model as dependent variables and consequently no regression model is specified. It should be further noted that the choice of a lognormal model for earnings is not obligatory. Indeed, there do exist more sophisticated models for dealing with distributions that are restricted to non-negative values, such as truncated and Tobit regression. Long (1997) provides a detailed discussion of such models. A lognormal model will be utilised here primarily for its simplicity and in light of the fact that earnings are conventionally regarded as following a lognormal distribution.

More generally, let $\boldsymbol{y}$ denote the variable to be imputed, with observed values $\boldsymbol{y}_{\text{obs}}$ and missing values $\boldsymbol{y}_{\text{miss}}$. Further, let $\mathbf{X}$ denote the most recently updated predictor matrix containing both the completely observed variables and the previously imputed variables. In addition, let that portion of $\mathbf{X}$ for which $\mathbf{Y}$ is missing be represented as $\mathbf{X}_{\text{MISS}}$. Let $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimate of the k-dimensional set of regression parameters $\boldsymbol{\beta}$. Let $\mathbf{V}$ denote the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and let $\mathbf{T}$ be the Cholesky decomposition of $\mathbf{V}$. The Cholesky decomposition produces the square root of a symmetric positive definite matrix such that $\mathbf{TT}' = \mathbf{V}$, where $\mathbf{T}$ is a lower triangular matrix [9]. Finally, let $\boldsymbol{z}$ represent a column vector of the same dimension as $\hat{\boldsymbol{\beta}}$ containing realisations from the standard normal distribution. Imputations for $\boldsymbol{y}_{\text{miss}}$ may then be generated given each of the following distributional assumptions for $\boldsymbol{y}$ [9].

3.3.1 Gaussian Distribution

If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, an ordinary least squares model of the form $\mathrm{E}[\boldsymbol{y}] = \mathbf{X}\boldsymbol{\beta}$ is appropriate for $\boldsymbol{\mu}$. Imputations for $\boldsymbol{y}_{\text{miss}}$ may thus be generated as follows [9].

1. Generate a random draw from the posterior distribution of σ^2. In order to achieve this, note that

$$U = \frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{n-k}. \tag{3.1}$$

   A random draw from the posterior distribution of σ^2 may thus be achieved by generating a realisation of U, say u, from a χ^2_{n-k} distribution and putting

$$\sigma^2_\star = \frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{u}. \tag{3.2}$$

2. Next, it is necessary to generate a random draw from the posterior distribution of $\boldsymbol{\beta}$ defined as

$$\boldsymbol{\beta}^\star = \hat{\boldsymbol{\beta}} + \mathbf{T}\boldsymbol{z}. \tag{3.3}$$

   Note that

$$\begin{aligned} \mathrm{E}[\boldsymbol{\beta}^\star] &= \mathrm{E}[\hat{\boldsymbol{\beta}} + \mathbf{T}\boldsymbol{z}] \\ &= \hat{\boldsymbol{\beta}} + \mathbf{T}\ \mathrm{E}[\boldsymbol{z}] \\ &= \hat{\boldsymbol{\beta}} & \text{since } \mathrm{E}[\boldsymbol{z}] = 0 \end{aligned}$$

   and

$$\begin{aligned} \mathrm{Var}[\boldsymbol{\beta}^\star] &= \mathrm{Var}[\hat{\boldsymbol{\beta}} + \mathbf{T}\boldsymbol{z}] \\ &= \mathrm{Var}[\mathbf{T}\boldsymbol{z}] \\ &= \mathbf{T}\mathbf{T}' & \text{since } \mathrm{Var}[\boldsymbol{z}] = \mathbf{I} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

   From the normality of $\boldsymbol{z}$, it therefore follows that $\boldsymbol{\beta}^\star$ has a multivariate normal distribution $\boldsymbol{\beta}^\star \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

3. Missing values may now be imputed as random draws from the posterior predictive distribution defined as

$$\boldsymbol{y}^{\star}_{\text{miss}} = \mathbf{X}_{\text{MISS}}\boldsymbol{\beta}^{\star} + \sigma_{\star}\boldsymbol{\upsilon}, \tag{3.4}$$

where $\boldsymbol{\upsilon}$ represents an independent column vector of the same dimension as $\boldsymbol{y}_{\text{miss}}$ containing random deviates from the standard normal distribution. The predictive distribution of $\boldsymbol{y}_{\text{miss}}$ defined in this manner is therefore multivariate normal with mean $\mathbf{X}_{\text{MISS}}\boldsymbol{\beta}^{\star}$ and variance $\sigma^2_{\star}\mathbf{I}$.

3.3.2 Poisson Distribution

If $\boldsymbol{y} \sim \text{Pois}(\boldsymbol{\lambda})$, a generalised linear model of the form $\lambda = \exp(\mathbf{X}\boldsymbol{\beta})$ is appropriate with linear predictor $g(\boldsymbol{\lambda}) = \mathbf{X}\boldsymbol{\beta}$ and log link function $g(\cdot)$ [2]. The imputations for $\boldsymbol{y}_{\text{miss}}$ may thus be obtained as follows.

1. Generate a random draw of $\boldsymbol{\beta}^{\star}$ as per Equation 3.3 in the Gaussian case.

2. Generate values of $\boldsymbol{\lambda}$ for the distribution of $\boldsymbol{y}_{\text{miss}}$ as

$$\boldsymbol{\lambda}^{\star}_{\text{miss}} = \exp(\mathbf{X}_{\text{MISS}}\boldsymbol{\beta}^{\star}). \tag{3.5}$$

3. Impute the missing values as independent draws from the Poisson distribution with parameters $\boldsymbol{\lambda}^{\star}_{\text{miss}}$. Since the inverse cumulative distribution function of a Poisson random deviate does not exist in closed analytical form, an acceptance-rejection algorithm must be employed in this regard [8].

3.3.3 Binomial Distribution

For $\boldsymbol{y} \sim \text{Binom}(\boldsymbol{n}, \boldsymbol{\pi})$, a logistic regression model is appropriate with

$$\text{logit}(\boldsymbol{\pi}) = \ln\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \mathbf{X}\boldsymbol{\beta}. \tag{3.6}$$

Noting this, imputations for $\boldsymbol{y}_{\text{miss}}$ may be simulated as follows.

1. Generate a random draw of $\boldsymbol{\beta}^{\star}$ as per Equation 3.3.

2. Predict the probability of a success for each element in $\boldsymbol{y}_{\text{miss}}$ as [5]

$$\boldsymbol{\pi}^\star_{\text{miss}} = \Pr[\boldsymbol{y}_{\text{miss}} = 1|\mathbf{X}_{\text{MISS}}] = \frac{\exp(\mathbf{X}_{\text{MISS}}\boldsymbol{\beta}^\star)}{1+\exp(\mathbf{X}_{\text{MISS}}\boldsymbol{\beta}^\star)}. \tag{3.7}$$

3. Generate a random vector $\boldsymbol{u}$ of the same dimension as $\boldsymbol{\pi}^\star_{\text{miss}}$ from the uniform distribution on $[0,1]$. Impute a one if an element of $\boldsymbol{u}$ is less than or equal to the corresponding element in $\boldsymbol{\pi}^\star_{\text{miss}}$ and impute zero otherwise [9].

3.3.4 Multinomial Distribution

Although the multinomial distribution is not a member of the exponential family, data which follow such a distribution may be modelled by a generalisation of the logistic regression model presented above. More specifically, if there are k categories to which a random variable may be classified, then one could fit $k-1$ logistic models to estimate the probability of being assigned to category $j = 1,\ldots,k-1$ relative to the base category k. Formally, the multinomial logistic model is defined as

$$\text{logit}(\boldsymbol{\pi}_{j|k}) = \ln\left(\frac{\boldsymbol{\pi}_j}{\boldsymbol{\pi}_k}\right) = \mathbf{X}\boldsymbol{\beta}_j, \tag{3.8}$$

where $\boldsymbol{\beta}_j$ represents the coefficients corresponding to category $j = 1,\ldots,k-1$ relative to the omitted category. The ordered logit model, applicable to polytomous random variables for which there exist a natural ordering of the categories, is a special case of the multinomial logit and the same principles apply for imputation purposes [5]. Imputations for polytomous variables may be generated as follows.

1. Generate a random draw of $\boldsymbol{\beta}^\star$ as per Equation 3.3 where, in this context, $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}^{\star\prime}_1, \boldsymbol{\beta}^{\star\prime}_2, \ldots, \boldsymbol{\beta}^{\star\prime}_{k-1})'$ and similarly for $\hat{\boldsymbol{\beta}}$.
2. Predict the probability that each observational unit belongs to each of the k categories using the random draw of $\boldsymbol{\beta}^\star$ [5],

$$\boldsymbol{\pi}^\star_{\text{miss},j} = \Pr[\boldsymbol{y}_{\text{miss}} = j|\mathbf{X}_{\text{MISS}}] = \frac{\exp(\mathbf{X}_{\text{MISS}}\boldsymbol{\beta}^\star_j)}{\sum_{j=1}^{k}\exp(\mathbf{X}_{\text{MISS}}\boldsymbol{\beta}^\star_j)} \quad \text{where } \boldsymbol{\beta}^\star_k = 0. \tag{3.9}$$

3. Define $\boldsymbol{\phi}_p = \sum_{j=1}^{p}\boldsymbol{\pi}^\star_{\text{miss},j}$ as the cumulative sum of the probabilities with $\boldsymbol{\phi}_0 = 0$ and $\boldsymbol{\phi}_k = 1$. To impute $\boldsymbol{y}_{\text{miss}}$, generate a random vector $\boldsymbol{u} \sim U[0,1]$ of the same dimension as $\boldsymbol{y}_{\text{miss}}$ and impute category j if $\boldsymbol{\phi}_{j-1} \leq \boldsymbol{u} \leq \boldsymbol{\phi}_j$ [9].

3.4 Assessing Convergence

After imputing all variables with a chain of generalised linear models, it is necessary to repeat the procedure k times until the sequence converges to the limiting distribution $P(\mathbf{Y}_{\text{MISS}}, \boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$. In practice, it is useful to know roughly how large a value of k is necessary for $\boldsymbol{\theta}^{(t+k)}$ to be independent of $\boldsymbol{\theta}^{(t)}$ for any $\boldsymbol{\theta}^{(t)}$ within a reasonable range of the posterior density. If such a value were known, then a burn-in period of length k would be sufficient to achieve stationarity, provided that the starting point of the algorithm was not highly unusual with respect to $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$. Moreover, after the burn-in period, every kth iterate of $\boldsymbol{\theta}$ could be taken as an independent draw from $P(\boldsymbol{\theta}|\mathbf{Y}_{\text{OBS}})$ and every kth iterate of $\mathbf{Y}_{\text{MISS}}$ could be used for imputation purposes [13].

The sequential regression multivariate imputation procedure was applied to the aforementioned hypothetical datasets, varying the number of iterations within a single chain and the number of chains themselves. Convergence was assessed primarily by comparing the results obtained after each iteration with the known results in the contrived dataset. Since the "true" distribution of earnings is known, one can compare the distribution of the imputed values to that of the true values in order to assess at what point in the chain the former provides a satisfactory approximation of the latter. The methods utilised to assess convergence in this study are described next.

3.4.1 Time Series Plots

A quick and dirty approach for assessing convergence is a time series plot of a scalar component of $\boldsymbol{\theta}$ over various iterations. For the present investigation, the pointwise convergence of mean earnings will be monitored in this manner. Convergence is assumed where values of mean earnings fluctuate within a relatively narrow horizontal band for successive iterations.

3.4.2 Autocorrelation Functions

In order to examine the relationships among successive iterates of mean earnings, it is helpful to consider the autocorrelation functions for each of the three missingness mechanisms. The lag-k autocorrelation for the stationary series $\{\mu^{(t)} : t = 1, 2, \ldots, n\}$ is defined as

$$\rho_k = \frac{\text{Cov}[\mu^{(t)}, \mu^{(t+k)}]}{\text{Var}[\mu^{(t)}]}. \tag{3.10}$$

Note that stationarity implies $\text{Var}[\mu^{(t)}] = \text{Var}[\mu^{(t+k)}]$. A sample estimate of ρ_k is given by

$$r_k = \frac{\sum_{t=1}^{n-k}(\mu^{(t)} - \bar{\mu})(\mu^{(t+k)} - \bar{\mu})}{\sum_{t=1}^{n}(\mu^{(t)} - \bar{\mu})^2}, \tag{3.11}$$

where $\bar{\mu}$ is the mean earnings over the entire series in this case [13]. A plot of r_k versus the lags within a single chain is known as a sample autocorrelation function or correlogram.

In testing the hypotheses

$$\begin{aligned}&\text{H}_0 : \rho_k = \rho_{k+1} = \rho_{k+2} = \ldots = 0\\&\text{H}_1 : \rho_k \neq 0,\end{aligned}$$

the following condition would lead to the rejection of the null hypothesis

$$|r_k| \geq z_{1-\alpha/2}\left[\frac{1}{n}\left(1 + 2\sum_{t=1}^{k-1} r_t^2\right)\right]^{1/2}, \tag{3.12}$$

where α denotes the Type I error rate [13]. Consequently, one can infer that the series has converged when r_k does not exceed the right-hand-side expression in Equation 3.12 for subsequent values of k.

3.4.3 Kernel Density Estimation

Although the methods discussed above are easy to understand and implement, they are not foolproof. Indeed, the primary concern of this investigation is distributional convergence, whilst the methods mentioned thus far have been concerned with diagnosing stationarity pertaining to a scalar element of $\boldsymbol{\theta}$, in particular mean monthly earnings. Moreover, whilst monitoring the behaviour of individual components of $\boldsymbol{\theta}$ is easier than assessing its full multidimensional distribution, convergence in the marginal distributions of the components does not necessarily imply convergence in the joint posterior. Furthermore, the marginal distributions will often converge at different rates and there is thus always the possibility that some unknown function of $\boldsymbol{\theta}$ has not yet converged [13].

In addition to the above, it may be further argued that zero correlation does not imply independence and that non-linear relationships may still exist where autocorrelations are insignificant [13]. Consequently, it would seem appropriate to complement this analysis by evaluating the full distribution of earnings after each iteration, rather than merely the first moment.

This was achieved by means of kernel density estimation, computed by means of the Epanechnikov kernel function, at various points within the chain for the three missingness mechanisms. The imputed densities were superimposed over the true densities to determine at what point in the chain the former distribution provides an adequate approximation of the latter.

3.4.4 Empirical Cumulative Distribution Functions

Distributional convergence was also assessed by plotting the empirical cumulative distribution function of the imputed and true earnings values for each of the three missingness mechanisms at various points within the Markov chain. This method merely provides an alternative graphic display to the comparison of probability density functions.

3.4.5 Gelman and Rubin's Monitoring Statistic

An explicit monitoring statistic for distributional convergence was devised by Gelman and Rubin in 1992. In order to compute this monitoring statistic, it is necessary to simulate $m > 1$ sequences with various starting points. The basic idea is then that convergence may be monitored by comparing the variation between and within chains until within-variation roughly equals between-variation. When this occurs, the distribution of each simulated sequence will be close to the distribution of all the sequences mixed together and therefore all chains will approximate the target distribution [6]. Following this principle, Gelman and Rubin's monitoring statistic is computed as follows.

Let $\psi \in \boldsymbol{\theta}$ be a scalar quantity of interest and label the draws on this estimand as $\psi_{m,k}$, where $m = 1, 2, \ldots, M$ is the number of sequences and $k = 1, 2, \ldots, K$ is the number of iterations within chain m. The between-sequence variation is computed as follows

$$B = \frac{K}{M-1} \sum_{m=1}^{M} \left(\bar{\psi}_{m\cdot} - \bar{\psi}_{\cdot\cdot}\right)^2, \tag{3.13}$$

where

$$\bar{\psi}_{m\cdot} = \frac{1}{K} \sum_{k=1}^{K} \psi_{m,k} \tag{3.14}$$

is the mean of ψ within chain m and

$$\psi_{\cdot\cdot} = \frac{1}{M} \sum_{m=1}^{M} \bar{\psi}_{m\cdot} \tag{3.15}$$

is the grand mean across all chains.

The variation within a sequence is computed as

$$\bar{V} = \frac{1}{M} \sum_{m=1}^{M} S_m^2, \tag{3.16}$$

where

$$S_m^2 = \frac{1}{K-1} \sum_{k=1}^{K} \left(\psi_{m,k} - \bar{\psi}_{m\cdot}\right)^2 \tag{3.17}$$

is the variation within chain m.

The estimate of the marginal posterior variance of ψ is then given by the weighted average of B and $\bar{V}$, namely

$$\hat{T} = \frac{K-1}{K}\bar{V} + \frac{1}{K}B. \tag{3.18}$$

Hence, when the length of the sequence K is finite, $\bar{V}$ will be an underestimate of the true posterior variance $\hat{T}$, because the individual chains will not have had adequate opportunity to wander through the entire range of the target distribution. Consequently, variation within a sequence will be less than variation between sequences. In the limit as $K \to \infty$, within-variation will tend to the true posterior variance of ψ, since the chain will be producing values from the posterior distribution of ψ [6].

The above observations provide the basis from Gelman and Rubin's monitoring statistic which is defined as

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{T}}{\bar{V}}}. \tag{3.19}$$

The ratio $\hat{T}/\bar{V}$ represents the factor by which variation within a chain is under-representative of the posterior variance in ψ and declines to one as the length of the chain tends to infinity. A value of the monitoring statistic close to one is therefore indicative of convergence in the sequence. Little and Rubin (2002) suggest that values below 1.2 are acceptable for most practical problems.

3.5 Assessing the Optimal Number of Imputations

For those who are unfamiliar with multiple imputation, the claim that three to five imputations is often sufficient may be rather surprising; in other applications of Monte Carlo, hundreds or thousands of draws are often needed to achieve an acceptable level of accuracy. In the literature, the small value of m for multiple imputation purposes is justified for two fundamental reasons [13].

Firstly, multiple imputation relies on simulation to solve only the missing data aspect of the problem. As with any iterative simulation method, one could effectively eliminate Monte Carlo error by choosing a sufficiently large value of m. However, within a multiple imputation context, the resulting gain in efficiency is typically regarded as unimportant, since the Monte Carlo error is a relatively small portion of the overall inferential uncertainty. Consequently, it might be argued that the opportunity costs of the additional resources that would be required to create and store more than a few imputations are too high [13].

The second reason why one can often obtain valid inferences for a small value of m is that Rubin's rules for combining the m complete datasets explicitly account for Monte Carlo error. Confidence interval estimates based on these rules take into account the fact that both the point and variance estimates contain a predictable amount of simulation error due to the finiteness of m. The width of the interval is therefore adjusted accordingly to maintain the appropriate probability of coverage [13].

In order to establish whether or not the empirical evidence supports these arguments, the bias, standard error of estimate and root mean squared error of mean earnings was computed for various values of m. The coverage in repeated sampling of 95% confidence intervals on mean earnings is also assessed for alternative values of m. In addition to the accuracy of scalar estimates, multiple imputation is also concerned with the preservation of multivariate relationships. In this respect, the relationship between earnings and years of education will also be investigated in terms of bias, efficiency and root mean squared error across the three missingness mechanisms. In order to achieve this, the natural logarithm of monthly earnings was regressed against province, gender, racial group, age, hours worked per week, skills training, years of education, occupation and sector and employment type. The coefficient estimate on the years of education variable will provide the relevant proxy for multivariate relationships and has a natural (approximate) interpretation as the rate of return to education.

Rubin's measure of relative efficiency presented as Equation 2.16 on page 20 will also provide a basis for assessing the optimal number of imputations under the three missingness mechanisms.

3.6 Assessing the Robustness of the Imputation Model

The robustness of the imputation model was assessed along two dimensions, namely the fraction of missing data and the number of covariates included in the model. In both instances, the bias, efficiency and root mean squared error of mean earnings and returns to education were monitored for various levels of these two factors. More specifically, the fraction of missing data was varied between 10% and 70% in intervals of 10%. Three imputation models were constructed varying the number of covariates. Recall that the MAR missingness mechanism was simulated on the earnings variable with respect to province, racial group, gender, age and years of education. The first imputation model included all variables in Table 1 on page 26; that is, the model included more variables than that which was actually responsible for missingness under MAR. This model was further employed to assess distributional convergence and the optimal number of imputations. A second imputation model was built including only the variables that induced missingness. Finally, a third model containing only province, age and racial group was also constructed; that is, the model contained less covariates than actually induced missingness. Note that years of education was deliberately omitted so as to assess the impact of the omission of key covariates on multivariate relationships. Together, these three models were utilised to assess robustness with respect to the number of covariates included the imputation model.

3.7 Statistical Software

All imputations, simulations and subsequent analyses were executed in STATA version 8. Specific programs were written to implement the imputation techniques and various simulations. The complex sample design was taken into account in all the analyses by means of STATA's suite of `svy` commands [16]. These commands use the Taylor series linearisation method to calculate variance taking weighting, stratification and clustering into account.

4 Empirical Findings

This section will attempt to address the theoretical questions raised earlier concerning the optimal number of iterations and imputations that are necessary to produce accurate results. The section will commence with an application of the various methods for monitoring convergence to the hypothetical complete datasets created from the *Labour Force Survey* of September 2003. This complete dataset will also provide the basis for monitoring bias and efficiency, whilst varying the number of imputations. The section will conclude by proposing an optimal imputation model and assessing the robustness of this model with respect to the fraction of missing data and the number of covariates included in the model.

4.1 Assessing Convergence

Mean earnings obtained after each iteration is illustrated in the time series plots presented as Figures 3 to 5 for each of the three missingness mechanisms. Note that the true mean earnings for the complete dataset (with no missing values) is R2 166 per month. After simulating missingness under each of the three missingness mechanisms, the mean monthly earnings ignoring missing data became R2 103, R1 722 and R1 318 for the MCAR, MAR and MNAR mechanisms respectively.

None of the three time series plots appears to reveal any immediately discern-

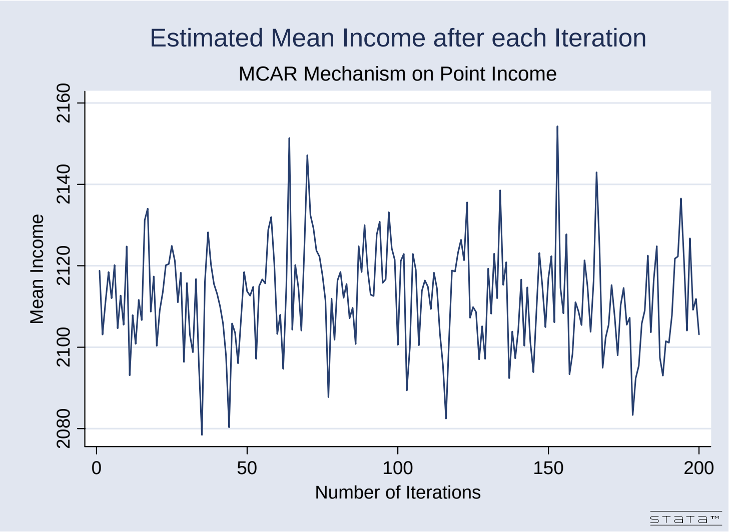

Figure 3: Estimated mean earnings after each iteration under the MCAR mechanism. Red horizontal line indicates the true mean earnings.

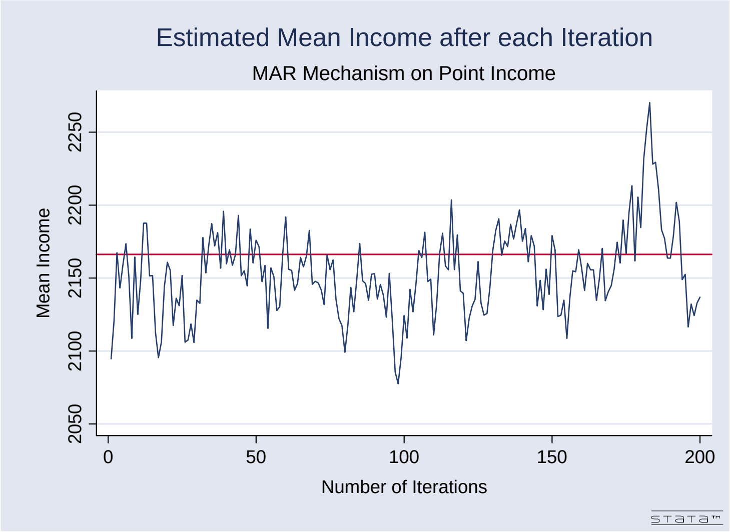

Figure 4: Estimated mean earnings after each iteration under the MAR mechanism. Red horizontal line indicates the true mean earnings.

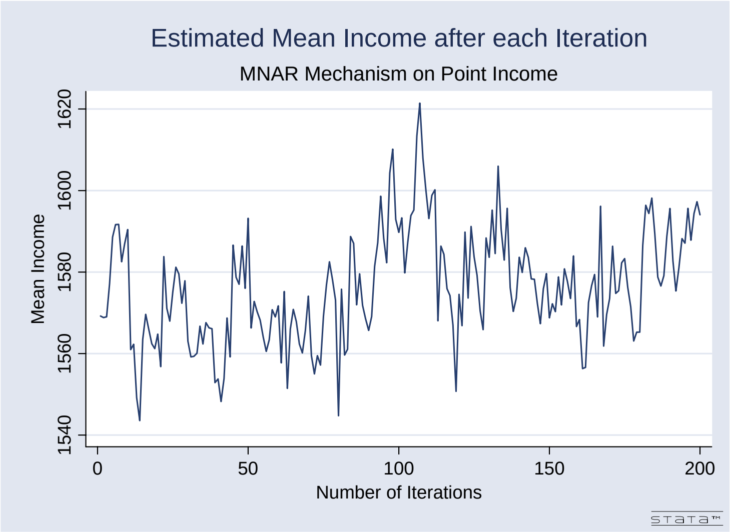

Figure 5: Estimated mean earnings after each iteration under the MNAR mechanism.

able trends and fluctuations tend to be confined to fairly narrow horizontal bands. Of the three, the MNAR missingness mechanism would appear to produce the least stable estimates of mean earnings, although one could argue that the seemingly large spike in mean earnings around the 100th iteration is merely a random phenomenon. The evidence would therefore suggest that pointwise convergence in mean earnings is achieved almost instantaneously, irrespective of the missingness mechanism.

Note, however, that although mean earnings does converge quickly, it only appears to converge to its true value under the MAR mechanism. For the MAR plot, it is noted that estimated mean earnings tends to fluctuate around the horizontal line, which represents its true value. In this respect, it would appear to take as few as five iterations before the estimated mean earnings provides a reasonable approximation of the true value.

In terms of the MCAR mechanism, however, mean earnings appears to have converged to just over R2 100 per month, which is of the same magnitude as the estimate ignoring missing data. This result is not particularly surprising given that the missing values are simply a random sample which is assumed to be no different from the observed values. The propensity to respond does not depend upon the observed variables and hence one would not expect the imputed values to differ systematically from the observed values. Indeed, it was noted earlier in this paper that MCAR is the most restrictive of the missingness mechanisms, as is clearly demonstrated here. Moreover, it must be noted that the bias observed under the MCAR mechanism is purely a result of the random sampling procedure used to set the data to missing. If the mean earnings ignoring missing data was equivalent to the true mean for the dataset without missing values, it follows that estimated mean earnings would indeed have converged to its true value.

By contrast, a systematic bias in estimated mean earnings does arise when imputing under the MNAR mechanism, even after a full 200 iterations. The severity of this bias is quite substantial and is of the order of R580 in the downward direction. This result is also not unexpected since the imputation model does in fact assume ignorable missingness and cannot account for the dependence of response probabilities on the earnings variable itself. It should, however, be noted that the imputation procedure will produce a less biased estimate of mean monthly earnings relative to the estimate obtained by performing casewise deletion. Recall that the estimate of mean monthly earnings ignoring missing data is R1 318, relative to the imputed estimates which fluctuate around approximately R1 580 per month. Clearly, both estimates are well below the parameter value of R2 166, although the results after imputation would appear to be the lesser of two evils. In practice, the missingness mechanism is likely to a combination of both MAR and MNAR, with the imputation model accounting for that portion of the non-response bias that arises due to the former mechanism.

In addition to the time series plots, it is also useful to consider the autocorrelation functions for the three missingness mechanisms. These are presented as Figures 6 to 8 for the first 100 lags of a single chain. In each plot, the dashed line corresponds to the 0.05-level critical values for testing the null hypothesis of no correlation at lag k or beyond against the alternative hypothesis of significant serial dependence.

For the MCAR mechanism, it is observed that serial dependence is only significant at a lag of one. This is to be expected since Markov chains, by definition, produce estimates which are mildly dependent upon the previous state [3]. Beyond this point, however, autocorrelation is insignificant at the 5% level, suggesting that convergence in mean earnings is reached rapidly under this mechanism.

Under the MAR mechanism, successive iterates remain significantly correlated until around the eighth lag, becoming insignificant for all lags thereafter. This would imply that a burn-in period of at least eight iterations is necessary to achieve stationarity in the mean earnings μ, where an MAR mechanism is assumed to be present. Hence, every eighth iterate of μ in a single chain might be regarded as an independent draw from the posterior distribution of this parameter.

Finally, with respect to the MNAR missingness mechanism, serial depen-

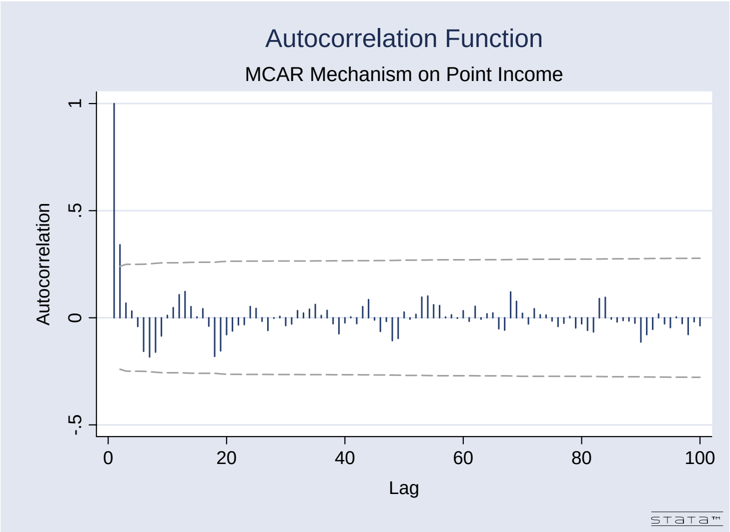

Figure 6: Autocorrelation function for mean earnings in a single chain under the MCAR mechanism. Dashed line corresponds to the 5% critical values for the test of significant serial dependence.

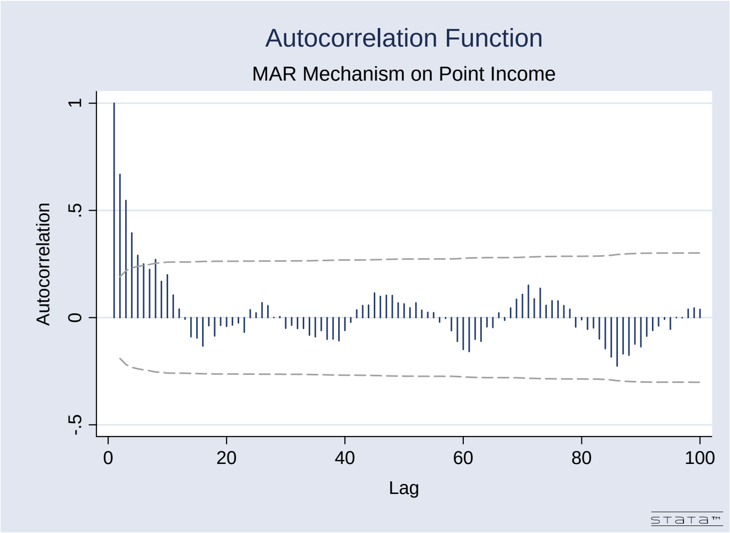

Figure 7: Autocorrelation function for mean earnings in a single chain under the MAR mechanism. Dashed line corresponds to the 5% critical values for the test of significant serial dependence.

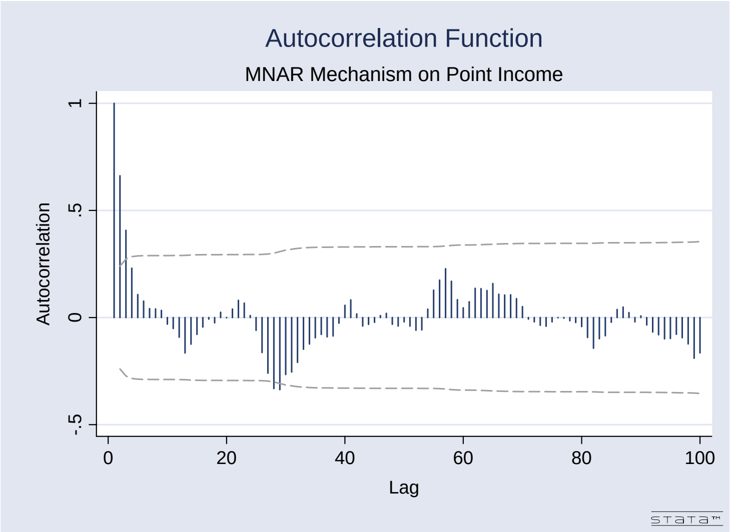

Figure 8: Autocorrelation function for mean earnings in a single chain under the MNAR mechanism. Dashed line corresponds to the 5% critical values for the test of significant serial dependence.

dence in the chain dies out more rapidly than would appear to be the case for the MAR mechanism, with insignificant correlations beyond lag 3. However, significant autocorrelation is again observed after 27 and 28 lags, in this case a significantly negative correlation coefficient. Schafer (1997) notes that, in general, one would not expect negative autocorrelations and attributes such estimates to fluctuations due to finite sample size. Indeed, when this exercise was repeated by choosing different starting seeds for the chain, the significant negative autocorrelation observed in Figure 8 disappeared entirely. Consequently, this phenomenon may be attributed purely to sampling variability.

At this point, it is worth noting that the above analyses were conducted for five chains of length 200 for each of the three missingness mechanisms. The results presented here are typical of those obtained in repeated runs of this nature, unless otherwise specified. In all simulations, it would appear that a minimum of eight runs through the Markov chain is necessary to achieve pointwise convergence of mean earnings across missingness mechanisms. Note, however, that from the time series plots presented at the outset, this does not necessary imply convergence to the correct parameter value.

In addition, Schafer (1997) notes that if the observed-data posterior distribution is oddly shaped, the chain may not have adequate opportunity to wander around certain regions of the parameter space within a reasonable number of iterations. This was in fact observed when earnings bands were included as a variable in the imputation model and point earnings was imputed from these bands. The imputation model was formulated as before with the additional specification that earnings bands were imputed prior to point earnings by means of a stochastic ordered logit model. Then, conditioning on the bands in addition to all other imputed and observed variables, point earnings were imputed using separate lognormal models for each band. This resulted in a normal distribution for log earnings (imputed and observed) within bands and hence a "lumpy" overall distribution of log earnings. The chain appeared to demonstrate difficulty in convergence for all missingness mechanisms. This result is presented in Figure 9 for the MAR mechanism below. In this case, convergence is not obvious after as many as 500 iterations. Furthermore, the algorithm is moving further away from the true earnings of R2 166 as the number of iterations increase.

The kernel densities and empirical cumulative distribution functions of the logarithm of earnings were plotted for each of the three missingness mechanisms at selected points in the chain. The distributions of the imputed missing values (red) relative to that of the true values that were set to missing (navy blue) are presented as Figures 10 and 11 on pages 44 and 45 respectively for the MAR missingness mechanism. Similar plots for the

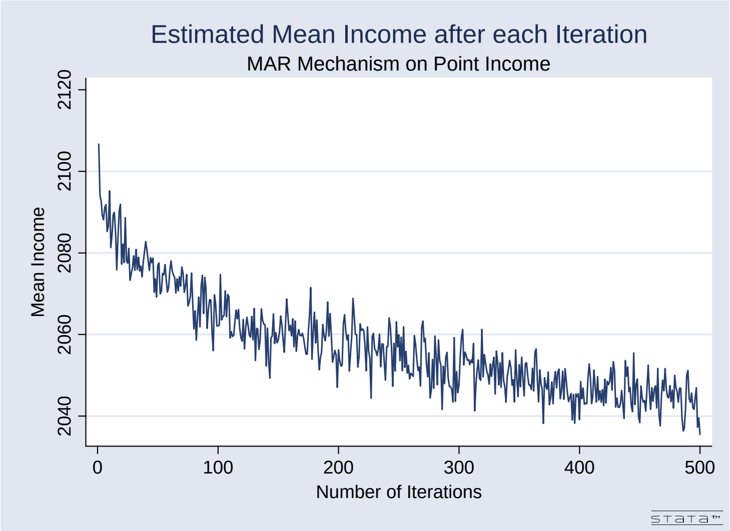

Figure 9: Non-convergence in mean earnings when imputing from bands under MAR mechanism.

MCAR and MNAR missingness mechanisms are presented as Figures B1 to B4 in Appendix B.

In all cases, there does not appear to be any major improvement in the distributional shape of the imputed values relative to the true distribution as the number of iterations increase. Indeed, the distributional form after, say, ten iterations is no worse than that obtained after 200 iterations. This observation would appear to confirm the results obtained earlier and suggests that convergence is rapid with little benefit to be reaped from increasing iterations.

As with the previous results, however, the mere fact that the distribution has stabilised does not imply that it has converged to the correct distribution. In the case of the MCAR mechanism, this would not appear to be a problem. The imputed values provide an extremely accurate approximation of the true density after as few as five iterations. Convergence under the MAR mechanism also does not raise any particular concern. Although the distribution of imputed values does not mirror the "dip" that is observed toward the centre of the true density, the approximation is fairly close for five to ten iterations. With the exception of this abnormality in the true density, the distribution of the imputed values lies very close to the true distribution over most of its range and in particular near the tails.

The distribution of imputed values under the MNAR mechanism is, however, a cause for concern. In this case, the imputed values have certainly not converged to the true density and do not even provide a reasonable approx-

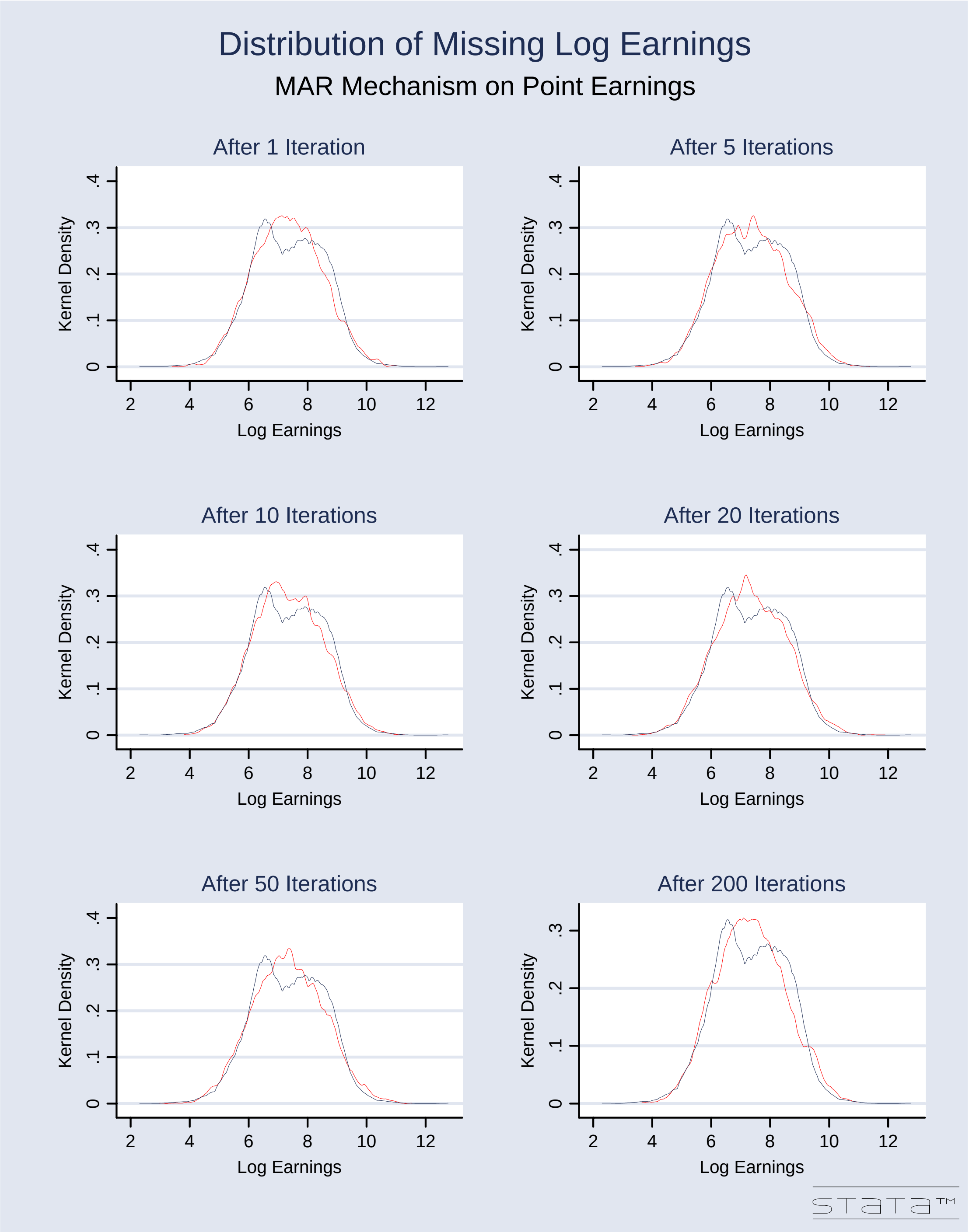

Figure 10: Distribution of missing log earnings after various iterations of the Markov chain under the MAR mechanism. Red density indicates imputed earnings and blue density indicates true earnings.

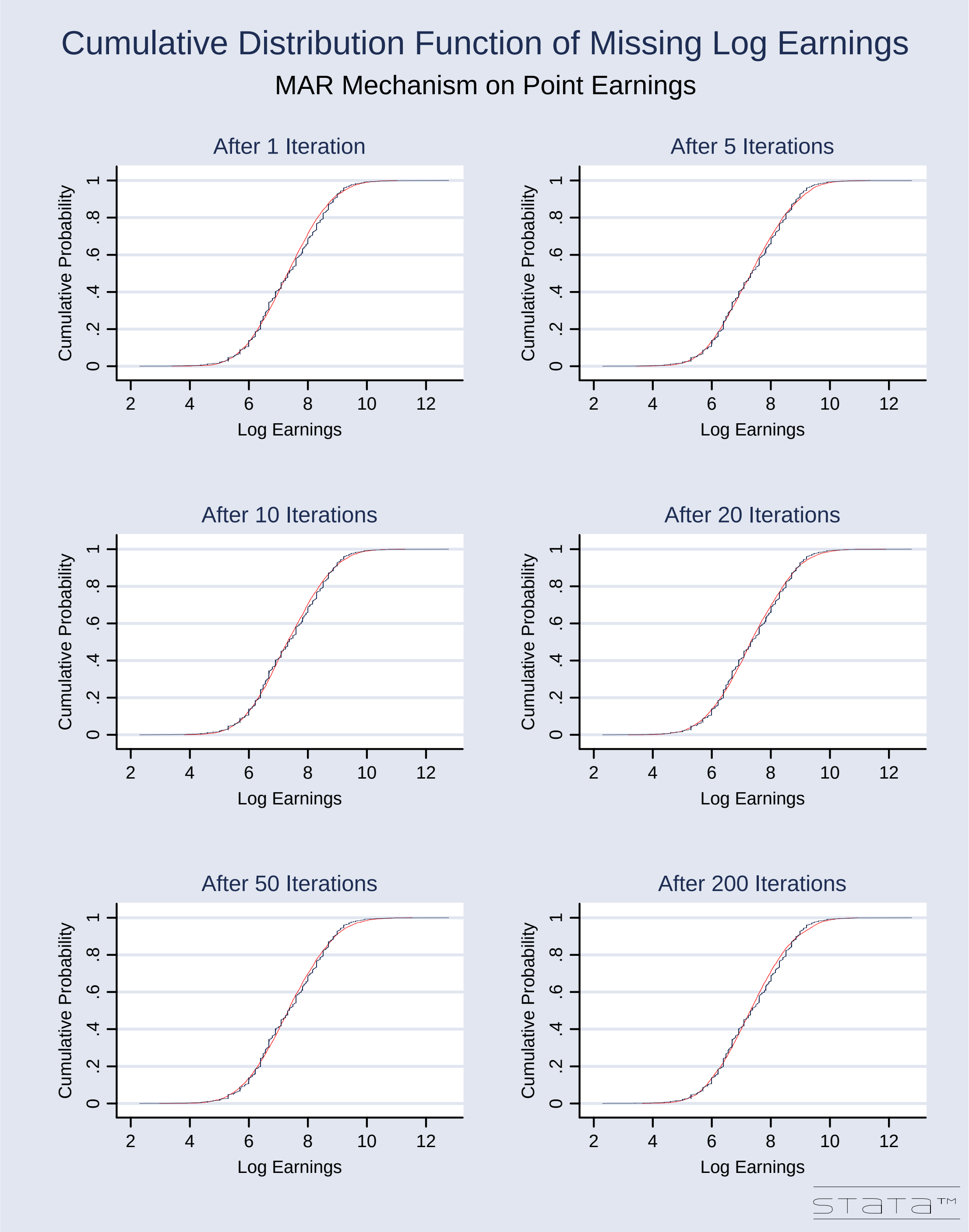

Figure 11: Cumulative distribution function of missing log earnings after various iterations of the Markov chain under MAR mechanism. Red curve is imputed earnings and blue curve is true earnings.

imation after 200 iterations. The true distribution of missing values under MNAR is skewed to the left, reflecting the fact that wealthier individuals were more likely to be set missing. The imputations, however, do not reflect this reality. Instead, the distribution of imputed values is almost symmetric and is less dense at both ends of its distribution relative to the true distribution. This lower density is particularly notable toward the right of the distribution, where the empirical distribution function with imputed values lies well above the true curve for middle to upper earnings levels and falls below the true curve toward the left of the distribution. The lower density at the tails is thus balanced by the large gain in density toward the centre of the distribution. Given that most of this shift in density is from the upper tail to the middle of the distribution, mean earnings will be biased downwards, as was observed in the time series plot for the MNAR mechanism presented earlier. These empirical results confirm the fact that the imputation model is not suited to data characterised by an MNAR missingness mechanism. Indeed, the model can only explain missingness that is dependent upon the sampled observations, whilst an MNAR mechanism clearly requires additional information beyond that which is contained within the sample.

The discussion thus far has been concerned with the distribution of the imputed values and its approximation of the true distribution of the missing values $\mathbf{Y}_{\text{MISS}}$. In the case of the MCAR and MAR missingness mechanisms, it was noted that the distribution of imputed values converges to the true distribution. By contrast, this was observed not to the case under the MNAR mechanism. This empirical result is in fact consistent with the theory presented in the literature review. Recall from Equation 2.9 that imputations are generated from the predictive posterior distribution of $\mathbf{Y}_{\text{MISS}}$, conditional on the observed data $\mathbf{Y}_{\text{OBS}}$, given by $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$. Recall further that this distribution is only valid for imputation purposes when inferences concerning $\boldsymbol{\theta}$ are unrelated to the distribution of missingness. This is the case when the missingness mechanism is either MCAR or MAR. Consequently, it should be unsurprising that the draws from $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$ provide a good approximation to the true distribution of missing values under these two mechanisms. On the other hand, since inferences on $\boldsymbol{\theta}$ do rely on the distribution of missingness under MNAR, likelihood-based inferences on $\boldsymbol{\theta}$ cannot be treated independently of the missingness mechanism, as was the case in Equation 2.4. As a result, the predictive posterior $P(\mathbf{Y}_{\text{MISS}}|\mathbf{Y}_{\text{OBS}})$ does not coincide with the actual distribution of the missing data. It therefore follows that draws from the former distribution will not be of the same distributional form as the true values where the missingness mechanism is MNAR.

Figures 12 to 14 present the kernel densities for the actual logarithm of earnings (blue), imputed log earnings after ten iterations (red) and log earnings ignoring the missing data (green) for each of the missingness mechanisms. Further kernel densities and empirical distribution functions at various iterations in the Markov chain for each of the three missingness mechanisms are presented as Figures B5 to B10 in the appendix.

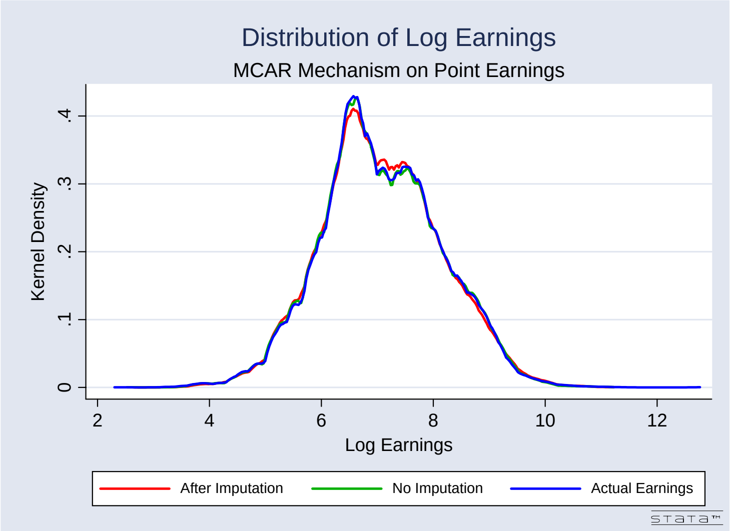

Figure 12: Distribution of log earnings with imputation after ten iterations and without imputation relative to the true distribution under an MCAR mechanism.

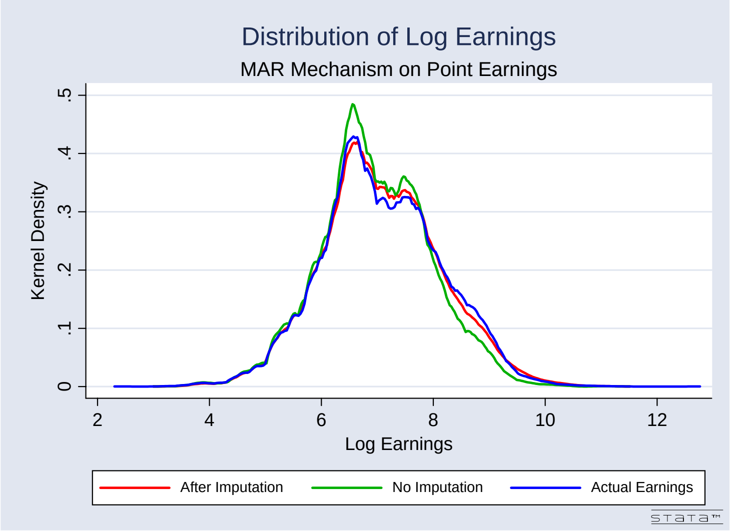

Figure 13: Distribution of log earnings with imputation after ten iterations and without imputation relative to the true distribution under an MAR mechanism.

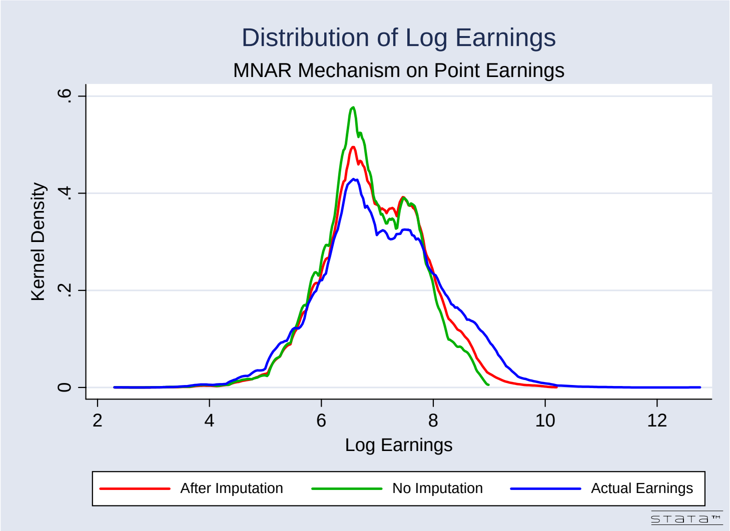

Figure 14: Distribution of log earnings with imputation after ten iterations and without imputation relative to the true distribution under an MNAR mechanism.

In terms of the MCAR mechanism, it is clear from Figure 12 that imputation is not necessary to preserve distributional form. The distribution of log earnings ignoring missing data coincides almost exactly with the true distribution, as is to be expected since the missing units do not differ systematically from respondents. Imputing missing values has, however, provided a better approximation of the true distribution relative to ignoring the missing data for both the MAR and MNAR mechanisms. For the MAR mechanism, ignoring missing values has resulted in a distribution which understates the density to the right and overstates it in the centre. Imputation produces a remarkable improvement to this by increasing the proportion of wealthier individuals in the sample. Consequently, the density for earnings after imputation lies above that which ignores missing data and only marginally below the actual distribution on the right. This results in a lower observed density in the centre of the distribution relative to the distribution ignoring missing data. The imputation model has therefore performed well under the MAR missingness mechanism.

On examination of Figure 14, it would also appear that imputation can provide some value where the missingness mechanism is MNAR. Although the distribution of log earnings after imputation under MNAR does not provide an accurate reflection of the true distribution, the graph would suggest that imputing in the presence of an MNAR missingness mechanism is better than not imputing at all. Indeed, the density of imputed earnings does lie closer to the true distribution (particularly in the middle and to the right) relative to the density without imputations. Despite this observation, the imputed density does still lie quite far away from the true distribution and hence inferences based on the imputed data would still need to be treated with caution.

The analysis of convergence has thus far been largely output-based and it would seem appropriate to formalise this analysis by recourse to an explicit monitoring statistic. Gelman and Rubin's monitoring statistic presented earlier will suffice in this regard. The monitoring statistic was computed for mean monthly earnings at each iteration in chains of length $k = 200$. Since calculation of the statistic requires M parallel sequences, $m = 5$ such chains were constructed. This value of m was informed by the preferred choice given in the literature on multiple imputation. The monitoring statistic computed at each iteration is presented in Figures 15 to 17 for each of the three missingness mechanisms. The red vertical line denotes ten iterations.

Under the MCAR mechanism, the monitoring statistic drops below 1.2 after eight iterations and is very close to 1 after just ten iterations. A similar pattern is observed under the MAR mechanism with the monitoring statistic also dipping below the threshold after just eight iterations. Convergence was extremely rapid in the presence of an MNAR missingness mechanism with the monitoring statistic falling below the threshold 1.2 after as few as three iterations. However, once below 1.2, the monitoring statistic takes at least 100 iterations to tend to 1.

In summary, the empirical evidence would suggest that convergence of the

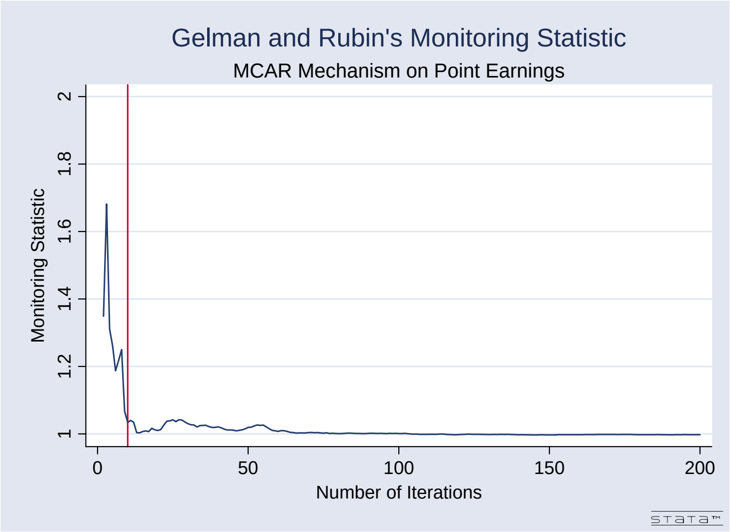

Figure 15: Gelman and Rubin's monitoring statistic at each iteration under an MCAR mechanism. Red vertical line corresponds to 10 iterations.

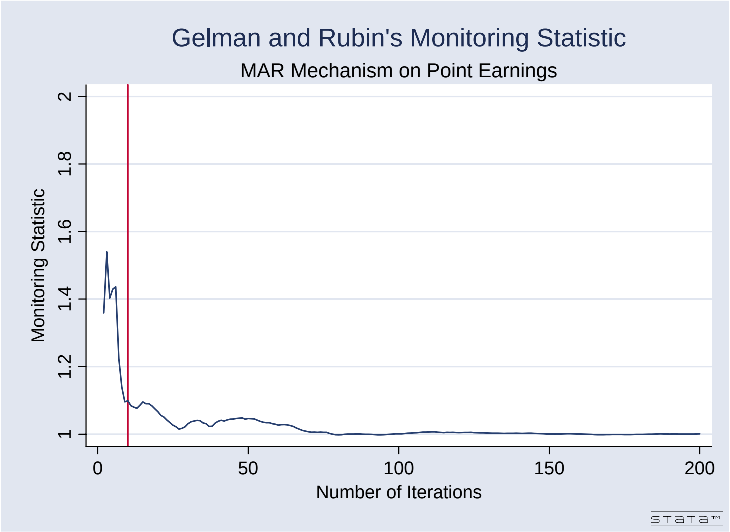

Figure 16: Gelman and Rubin's monitoring statistic at each iteration under an MAR mechanism. Red vertical line corresponds to 10 iterations.

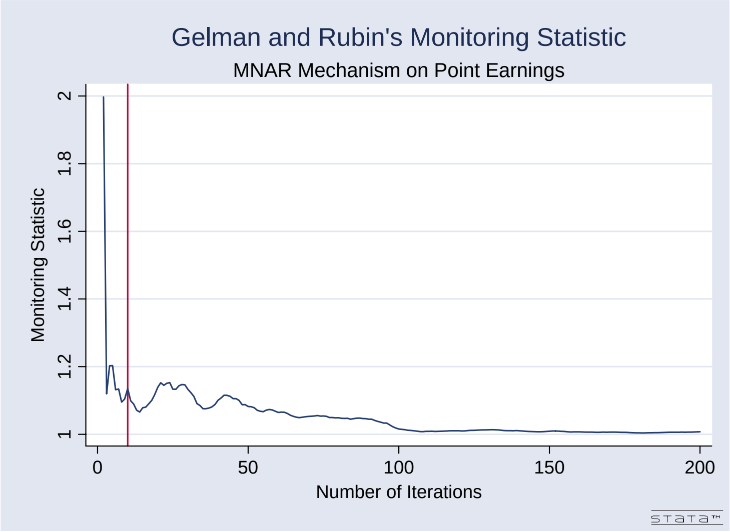

Figure 17: Gelman and Rubin's monitoring statistic at each iteration under an MNAR mechanism. Red vertical line corresponds to 10 iterations.

Markov chain to the posterior distributions of $\boldsymbol{\theta}$ and $\mathbf{Y}_{\mathrm{MISS}}$ given $\mathbf{Y}_{\mathrm{OBS}}$ is quite rapid for 30% missing data on the earnings variable. This would appear to be the case for all three missingness mechanisms. It was noted, however, that although convergence to the predictive posterior $P(\mathbf{Y}_{\mathrm{MISS}}|\mathbf{Y}_{\mathrm{OBS}})$ is achieved readily in the presence of an MNAR missingness mechanism, this distribution is not appropriate for imputing under MNAR. When the missingness mechanism is ignorable, the empirical evidence reveals that around eight iterations will suffice for convergence to the correct posterior distribution. This supports Schafer's assertion that eight to ten iterations are generally sufficient for most practical problems [13]. Indeed, ten iterations would appear to be a reasonably conservative choice in light of the empirical evidence and the additional computational power necessary to achieve this would typically be regarded as negligible.

4.2 Assessing the Optimal Number of Imputations

Whilst the preceding section was concerned with the length of a single chain, this section will focus on the number of chains that are necessary to produce accurate estimates of $\boldsymbol{\theta}$ that reflect the inherent uncertainty of the imputation process. According to the literature on this topic, a small value of m such as three or five is usually sufficient for the purposes of multiple imputation. In order to assess whether or not the empirical evidence supports this argument, the bias, standard error of estimate and root mean squared error of mean earnings were computed for various values of m. These estimates were obtained by applying Rubin's rules for mean and variance estimation and replicated for the three missingness mechanisms using ten iterations of the chain in each case. The results are presented in the Table 2.

From the table, it is noted that mean earnings is biased downwards for all

No.	**MCAR**			**MAR**			**MNAR**		
Imps	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**
0	-62.97	29.84	69.68	-444.40	24.63	445.08	-848.57	10.94	848.64
1	-53.10	25.26	58.80	-41.09	27.14	49.24	-588.36	13.61	588.51
3	-60.23	26.51	65.80	-22.18	37.45	43.52	-595.71	15.39	595.91
5	-57.08	26.84	63.07	-16.60	40.08	43.38	-595.03	17.02	595.27
10	-56.33	27.22	62.56	-9.60	45.30	46.31	-599.24	16.20	599.46
15	-56.16	27.12	62.37	-4.16	41.93	42.13	-595.50	20.62	595.86
20	-55.49	27.33	61.86	-5.48	39.69	40.07	-596.24	19.38	596.56
50	-55.12	28.25	61.94	-5.65	40.78	41.17	-594.95	19.45	595.27
100	-56.66	28.51	63.43	-7.40	40.93	41.59	-594.63	18.77	594.92

Table 2: Bias, standard error of estimate and root mean squared error for mean monthly earnings after m imputations for the three missingness mechanisms.

three missingness mechanisms. The most notable observation is the large reduction in the absolute bias after just a single imputation under the MAR and MNAR mechanisms. More specifically, the absolute bias after one imputation under an MAR mechanism is less than 10% of that observed with no imputation at all. The absolute bias continues to decline as the number of imputations is increased to fifteen, although the marginal reduction in absolute bias also decreases for successive imputations. For example, doubling computational effort from five to ten imputations reduces absolute bias by only R7 (42.17%), compared to the R403 (90.57%) reduction in absolute bias observed when imputing once as oppose to not imputing at all. Indeed, even the R19 (46.02%) reduction in absolute bias experienced when increasing the number of imputations from one to three would seem negligible for most real world problems.

With respect to the MNAR mechanism, the reduction in absolute bias from imputing once is proportionately less than that achieved under an MAR mechanism; the absolute bias after a single imputation is still more than 65% of that observed by ignoring the missing data all together. The decline in absolute bias that is observed may be attributed to the multivariate relationships between the earnings variable and the other variables in the dataset. These associations induce an indirect relationship between the propensity to respond and the observed variables. Since the imputation model is able to account for such a relationship, bias may be partially reduced via this mechanism. Nonetheless, the absolute magnitude of the remaining bias is clearly much larger than the bias observed under the MCAR and MAR mechanisms respectively. Interestingly, absolute bias under the MNAR mechanism does not decline for successive imputations, but instead remains fairly constant. Consequently, there is little to be gained from imputing more than once under an MNAR missingness mechanism. Moreover, the bias is quite substantial even after imputation. Accounting for this bias would rely on information beyond that which is contained within the sample and consequently an imputation model based solely on the sample information is inappropriate.

Under the MCAR mechanism, it is noted that imputation does not result in reduced absolute bias. Indeed, even the R10 reduction in absolute bias obtained by a single imputation is probably attributable to sampling variability, rather than the imputation model itself. This result should not be surprising given the nature of the missingness mechanism. The missing data is unrelated to the observed variables and hence any biases that exist prior to imputation will persist post-imputation. Note, however, that the bias of R63 was induced purely due to the sampling procedure used to set the data to missing and is thus a random phenomenon. In repeated sampling of this nature, one would expect zero bias. This point was noted earlier when considering the convergence of point earnings under MCAR in Figure 3.

The standard error of estimate rises initially with the number of imputations and then stabilises under the MAR and MNAR missingness mechanisms. From the literature review, this pattern is to be expected. When missing data is ignored or handled by means of a single imputation, the uncertainty surrounding the missing values is also ignored in the computation of the standard error. As a result, the standard errors will be downwardly biased and the coverage of confidence intervals will be overstated. The empirical evidence presented here would appear to support this notion. The standard error of mean earnings under the MAR and MNAR mechanisms is well below its true value of R41.20 when missing data is ignored or singly imputed. The approximation is, however, particularly close after five imputations under the MAR mechanism, with little to no improvement observed from increasing the number of imputations beyond this point.

In terms of the MNAR missingness mechanism, the standard error of estimate does not come close to its true value after as many as 100 imputations. Instead, the standard error fluctuates close to R20 (around half of the actual value) from five imputations onwards. Hence, in addition to the large biases observed under the MNAR mechanism, the variation in the estimate of mean earnings will also be substantially understated. Confidence intervals for the mean earnings will therefore be too narrow and centred around a biased estimate, regardless of the number of imputations employed.

As is the case in terms of bias, the standard error of estimate under an MCAR missingness mechanism does not appear to be altered significantly by multiple imputation. A standard error of just below R30 is observed with or without imputation. The contribution of between-imputation variation to the total variance in monthly earnings is therefore offset by the increase in the sample size relative to a complete-case analysis. The result is thus a relatively constant standard error of estimate that cannot be improved through imputation. This provides further evidence to the effect that MCAR is the most restrictive missingness mechanism from an imputation perspective.

Finally, the root mean squared error (RMSE) provides a composite measure of how close the estimated mean earnings is to its true value (bias), as well as how widely dispersed these estimates are about their mean (standard error). A low RMSE is clearly desirable since this would imply that the estimate is a good approximation of the true value and is narrowly dispersed about this value in repeated sampling. Since the bias and standard errors differ only marginally for different values of m under the MCAR mechanism, the RMSE too does not seem to exhibit any vast improvements as the number of imputations is increased. Under the MAR and MNAR mechanisms, however, a large reduction in the RMSE of mean earnings is observed for a single imputation relative to no imputations, brought about largely due to the substantial decline in absolute bias at this point. In addition, a small reduction in RMSE is observed when the number of imputations is increased from one to three. Increasing m beyond this point under either of the MAR or MNAR missingness mechanisms does not appear to be particularly beneficial in terms of the RMSE measure.

As a last observation, it should be noted that the RMSE after imputation is consistently lower where the missingness mechanism is MAR, followed by MCAR and finally MNAR. This finding is consistent with intuition. When imputing under MCAR, the imputed values should not differ systematically from the observed values and therefore any random deviations away from the true mean and variance in the dataset with missing values will be preserved in the imputed dataset. Under an MNAR mechanism, it is impossible to account for the systematic differences between the observed and missing data without additional knowledge beyond the sample information. Furthermore, the biases observed under an MNAR missingness mechanism will be amplified by the fact that the propensity to respond depends on the earnings variable itself. Consequently, large RMSEs are only to be expected. By contrast, the imputation model is specifically constructed to account for the fact that missingness may depend upon the observed variables. Since this is the very definition of MAR, it is therefore not surprising that estimates of mean earnings after imputation are the most accurate (lowest RMSE) under this missingness mechanism.

The above evaluation considered the bias and efficiency of mean monthly earnings. Indeed, the accuracy of such point estimates is an important indicator of a good imputation model. Another desirable property of an imputation model is that it should preserve the multivariate relationships between variables. In order to assess this feature, returns to education will be employed as a proxy for such relationships as described in the methodology section. This estimate and its standard error were recorded for various numbers of imputations and combined across imputations using Rubin's rules. The resulting bias, standard error of estimate and RMSE for returns to education are presented in Table 3 for various numbers of imputations, multiplied by a factor of 100 for clarity. Note that the true rate of return to education is 7.9537% and the true standard error of estimate is 0.1650% based on the hypothetical complete dataset.

Interestingly, Table 3 does not reveal any obvious associations between the bias or efficiency of the estimated rate of return to education and the number of imputations employed, irrespective of the missingness mechanism. Indeed, there would appear to be no evidence in support of the claim that multiple imputations improve the bias of this estimate. Under all mechanisms, the bias observed without imputation is of approximately the same magnitude as that observed with imputations. As expected, the absolute bias is largest under the MNAR mechanism with the rate of return to education typically

No.	**MCAR**			**MAR**			**MNAR**		
Imps	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**
0	0.0897	0.1997	0.2189	-0.4939	0.2039	0.5343	-1.9874	0.1776	1.9953
1	-0.1234	0.1645	0.2056	-0.5223	0.1620	0.5468	-1.8637	0.1459	1.8694
3	-0.0717	0.1738	0.1880	-0.5381	0.2133	0.5788	-1.8847	0.1550	1.8911
5	-0.0023	0.2090	0.2090	-0.5667	0.2033	0.6021	-1.9844	0.2132	1.9958
10	0.0660	0.2018	0.2123	-0.5179	0.2146	0.5605	-2.0016	0.1910	2.0107
15	0.0582	0.2012	0.2094	-0.5134	0.2105	0.5549	-2.0235	0.1889	2.0323
20	0.0545	0.2038	0.2110	-0.5210	0.2246	0.5673	-2.0405	0.1864	2.0490
50	0.0437	0.1950	0.1999	-0.5282	0.2064	0.5671	-2.0216	0.1812	2.0297
100	0.0469	0.1975	0.2030	-0.5373	0.2109	0.5772	-2.0283	0.1741	2.0357

Table 3: Bias, standard error of estimate and root mean squared error for returns to education after m imputations for the three missingness mechanisms. All figures multiplied by a factor of 100.

being underestimated by around 2%. A downward bias on the rate of return to education of approximately a half a percentage point is observed for the MAR missingness mechanism irrespective of the number of imputations. Since returns to education is a positive quantity, the consistent downward biases under the MAR and MNAR missingness mechanisms would imply that multivariate relationships are attenuated. The bias observed under the MCAR mechanism is negligible in magnitude and varies in direction for different numbers of imputations. In the long run, one would therefore expect the rate of return to education to be unbiased under this missingness mechanism.

The standard error of estimate is typically larger than its true value of 0.1650% and does not seem to be related to the number of imputations. Furthermore, the standard error is of much the same magnitude across the three missingness mechanisms. It therefore follows that the RMSE measure is also unrelated to the number of imputations and is worse for the MNAR mechanism, followed by the MAR mechanism and finally the MCAR mechanism. One may therefore conclude that increasing the number of imputations does not necessarily improve the estimation of multivariate relationships within the dataset.

Another means of assessing the optimal number of imputations is Rubin's relative measure of efficiency presented as Equation 2.16. This measure explicitly accounts for the fraction of missing information present in the dataset, as well as the finite number of imputations. The relative efficiency measure was computed for successive values of m with ten iterations, the results of which are displayed in Figure 18 for each of the three missingness mechanisms.

After only five imputations, efficiency is observed to equal or exceed 95% across all missingness mechanisms. More specifically, efficiency is computed

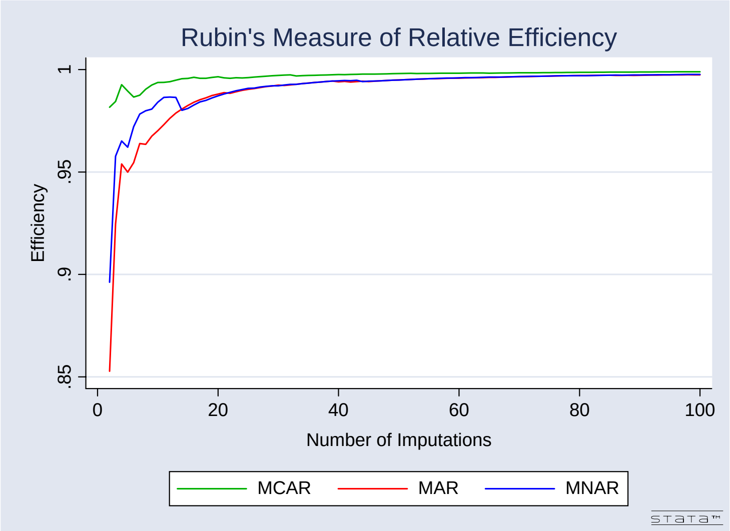

Figure 18: Rubin's relative efficiency measure after successive imputations for the MCAR, MAR and MNAR missingness mechanisms.

as 98.95%, 95.00% and 96.21% after five imputations for the datasets with MCAR, MAR and MNAR missing mechanisms respectively. This implies that the standard error of estimated mean earnings after five imputations is only 1.01, 1.05 and 1.04 times larger relative to that which could be obtained with $m = \infty$ under the MCAR, MAR and MNAR missingness mechanisms respectively. In the case of an MCAR mechanism, efficiency is almost 100% after just ten imputations. For the MAR and MNAR mechanisms, achieving full efficiency would require around 40 imputations, which is clearly not worth the computational effort. Indeed, the respective efficiencies obtained after only five imputations is more than adequate for most practical settings.

Surprisingly, the relative efficiency of the MNAR dataset is larger than that of the MAR dataset for smaller values of m. This can only be the case if the fraction of missing information is lower under the MNAR missingness mechanism relative to that of the MAR mechanism. The fraction of missing information γ for each missingness mechanism is illustrated graphically in Figure 19 for successive values of m. This graph does indeed confirm that the percentage of missing information is less for the MNAR mechanism relative to the MAR mechanism over the initial stretch of imputations, before both converge to around 50% from 20 imputations onwards.

The percentage of missing information under the MCAR mechanism is around 20%, which is somewhat less than that of both the MAR and MNAR mechanisms, as is to be expected. Recall that the rate of missing observations for the earnings variable is 30%, which does not necessarily equate

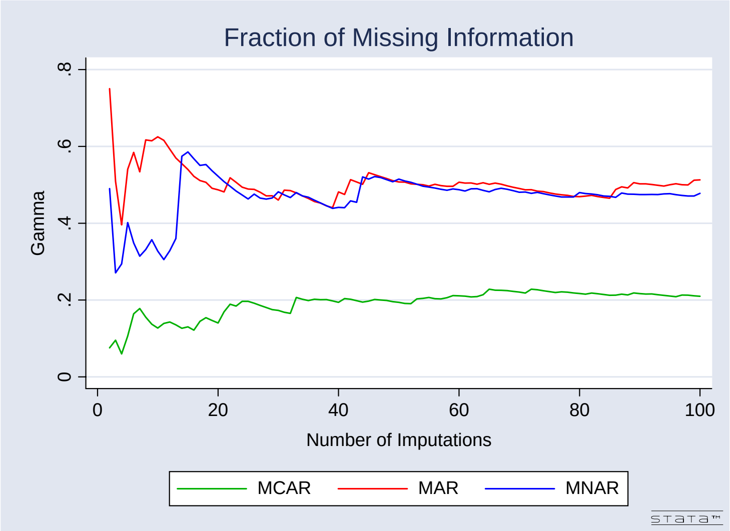

Figure 19: Fraction of missing information γ after successive imputations for the MCAR, MAR and MNAR missingness mechanisms.

to the rate of missing information, as is the case here. The exact reason as to why missing information is higher for MAR relative to MNAR for small numbers of imputations is likely to be attributed to the multivariate relationships within these specific samples. The rate of missing information depends upon the strength of the correlations between the variable of interest and the other variables that are more fully observed [13]. Since the multivariate distribution of missingness differs between the two datasets, this is suggested as the most probable cause of the observed phenomenon. It is expected that in repeated sampling the dataset with an MNAR mechanism will contain a larger amount of missing information relative to a dataset with an MAR mechanism. Indeed, the sample does contain most of the information that is necessary to impute under MAR, whilst imputing under MNAR requires information beyond that which is contained within the sample. Consequently, one would expect *a priori* that the latter missingness mechanism would result in more missing information than the former.

Despite the aforementioned idiosyncrasy, the empirical results thus far suggest that no more than five imputations are necessary to obtain a relatively efficient estimate of mean earnings. It was noted earlier that the major flaw of single imputation methods is that they do not adequately reflect the uncertainty of the imputation procedure in the standard error of estimate. This was confirmed in Table 2 where the standard errors are observed to be consistently below the true value of R41.20 for all thee missingness mechanisms. Consequently, one would expect coverage to be below its nominal value in repeated runs of a single imputation model.

Multiple imputation seeks to overcome the aforementioned problem by introducing a between-imputation component into the overall standard error of estimate, thereby resulting in wider confidence intervals and hence greater coverage. In order to assess whether actual coverage is in fact representative of its nominal value for low values of m, 95% confidence intervals were constructed using Rubin's rules for each of 100 runs of the imputation model with ten iterations and one, three and five imputations respectively. The results obtained under the MAR missingness mechanism are presented in Figures 20 to 22.

As expected, under-coverage is observed in the case of single imputation. In the 100 repeated imputations, only 91% of the confidence intervals include the true mean monthly earnings relative to the nominal value of 95%. By contrast, the 95% confidence intervals constructed for the multiple imputation models with three and five imputations respectively include the true mean in all 100 repeated imputations. This would imply that multiple imputation does provide a remedy to the problem of under-coverage which typifies single imputation methods for as few as three imputations. This finding is consistent with that of Van Buuren, Boshuizen and Knook (1999) whose simulation studies revealed that accurate results may be obtained with m as low as three for 20% missing data. Indeed, the results obtained here suggest that this conclusion still holds where 30% of entries are missing.

Although coverage would appear to be adequate for three to five imputa-

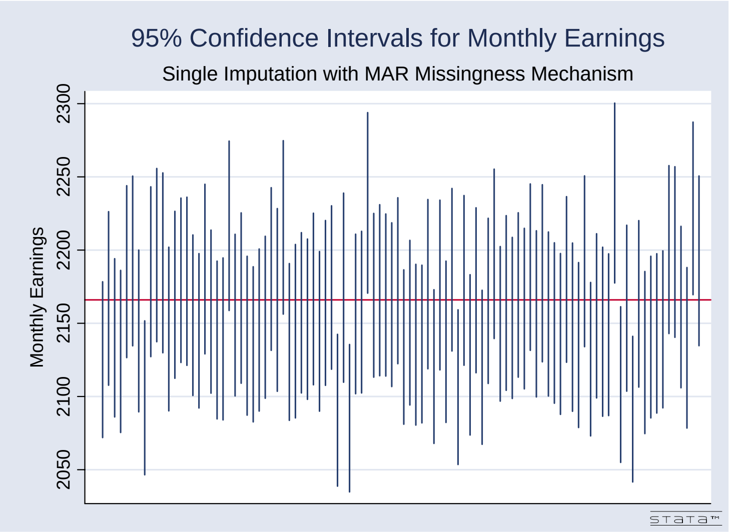

Figure 20: 95% Confidence intervals for a single imputation with MAR missingness mechanism. Red horizontal line denotes the true mean monthly earnings.

]

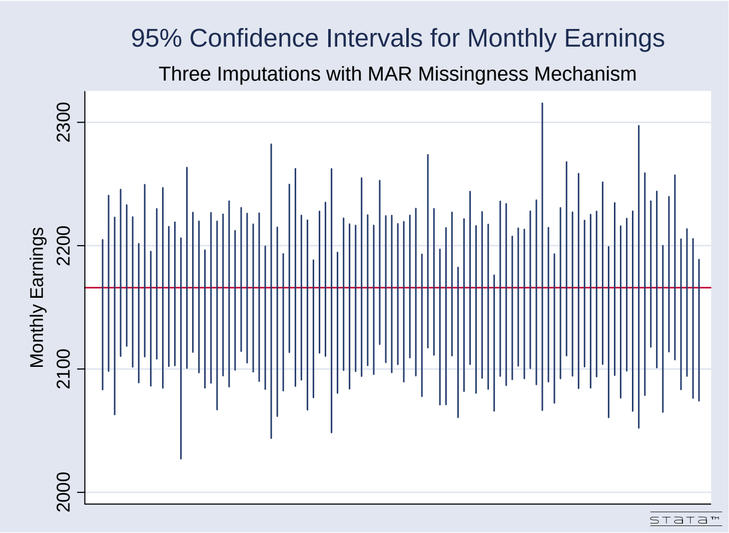

Figure 21: 95% Confidence intervals for three imputations with MAR missingness mechanism. Red horizontal line denotes the true mean monthly earnings.

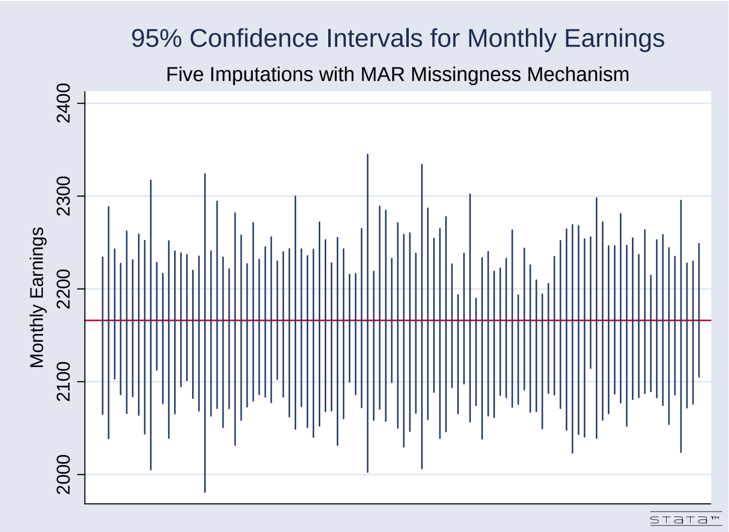

Figure 22: 95% Confidence intervals for five imputations with MAR missingness mechanism. Red horizontal line denotes the true mean monthly earnings.

tions, Royston (2004) contends that the confidence coefficient is highly variable for such low values of m, resulting in unstable confidence intervals for the scalar parameter of interest. The author's empirical work shows that the relationship between the coefficient of variation on the confidence coefficient and the number of imputations is convex to the origin, with the coefficient of variation declining gradually for successive imputations. Royston (2004) suggests that the coefficient of variation (defined as the standard deviation divided by the mean multiplied by 100) may be doubled in order to provide a rough measure of uncertainty in confidence intervals for the parameter of interest. For five imputations, his empirical study produced a coefficient of variation in the region of 13%, corresponding to an unacceptably large measure of uncertainty of approximately 26%. The author proposes that m be chosen such that the coefficient of variation is of the order of 5% or equivalently that uncertainty in the confidence intervals is roughly 10% or less. In his study, this rule of thumb would require m to be at least 20 or possibly more [10].

The 100 simulated confidence coefficients in the present study resulted in coefficients of variation of 19.65% and 26.15% for $m = 3$ and $m = 5$ respectively. This corresponds to uncertainty levels of 39.30% and 52.30% for the two models. These results would appear to be particularly concerning in light of Royston's rule of thumb. Indeed, the figures obtained here are even larger than those obtained in his study. These results therefore do provide evidence in support of the notion that smaller values of m may result in unreliable confidence intervals based on Royston's measure of uncertainty. However, Royston (2004) does not justify why doubling the coefficient of variation should be an appropriate measure of uncertainty in the confidence intervals and one may certainly question this approach. From the figures presented earlier, it is noted that the upper 95% confidence limit seldom peaks above R2 300, whilst the lower limit is largely in excess of R2 000. The confidence intervals are therefore found to fluctuate mostly within this range, which is certainly no cause for concern from a practical perspective. Indeed, when the problem is contextualised in this manner, one would not be led to question the reliability of the confidence coefficient.

4.3 Assessing the Robustness of the Imputation Model

The above findings suggest that a sequential regression multivariate imputation model with ten iterations and five imputations is a conservative choice for the imputation of missing values. Indeed, it would appear that eight iterations are sufficient for convergence and as few as three imputations are desirable in terms of bias, efficiency and coverage. However, these results were obtained for a fixed percentage of missing data (30%) and by including all variables from the artificial dataset in the imputation model. It would therefore seem appropriate to assess the robustness of the imputation model to changes in these factors. For the analyses to follow, the "conservative" model will be employed.

4.3.1 Fraction of Missing Data

It is expected *a priori* that the imputation model will produce superior estimates of mean monthly earnings for lower percentages of missing data on the earnings variable. Indeed, it was noted at the outset of this section that the convergence of a single Markov chain will depend upon the fraction of missing data, where convergence tends to be slower for larger percentages. In order to assess the conditions under which the imputation model performs well, it is useful to revisit the concepts of bias, standard error of estimate and root mean squared error as presented in Table 2. The model employed to generate that table was augmented by varying the percentage of missing data on the earnings variable to produce Table 4. Note that all other variables in the hypothetical dataset have been utilised in this model and that the same values were set to missing on the covariates for each round of imputations. The bias, standard error and root mean squared error pertain to mean monthly earnings, as computed using Rubin's rules, relative to the true mean monthly earnings of R2 166.24 in the artificial dataset.

Table 4 does indeed confirm that better estimates might be obtained for lower fractions of missing data under the MAR and MNAR missingness mechanisms. Examination of the RMSE measure for these two mechanisms provides a clear indication that the accuracy of the estimate worsens considerably as the percentage of missing data increases. Under the MNAR mechanism, absolute bias rises consistently with the fraction of missing data and is almost solely responsible for the increase in RMSE. Indeed, there does

%	**MCAR**			**MAR**			**MNAR**		
Missing	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**
10%	3.18	42.06	42.18	-29.56	37.62	47.85	-298.68	19.02	299.29
20%	-56.53	26.56	62.45	-34.14	36.01	49.62	-430.51	25.55	431.26
30%	-57.08	26.84	63.07	-16.60	40.08	43.38	-595.03	17.02	595.27
40%	-68.39	26.61	73.38	31.78	80.42	86.48	-771.51	13.02	771.62
50%	-75.13	32.16	81.73	-77.11	46.16	89.87	-907.72	20.99	907.96
60%	-37.65	29.20	47.65	-108.08	46.78	117.77	-1011.60	11.38	1011.67
70%	-66.26	38.93	76.85	-138.17	132.73	191.60	-1041.80	12.65	1041.83

Table 4: Bias, standard error of estimate and root mean squared error in mean monthly earnings for various percentages of missing data on the monthly earnings variable for the three missingness mechanisms (5 imputations, 10 iterations).

not appear to be any immediately obvious relationship between the fraction of missing data and the standard error of estimate.

Under the MAR missingness mechanism, the smallest absolute bias is actually observed with 30% missing data and mean earnings is upwardly biased for 40% missing observations whereas elsewhere the direction of this bias is downward. These two phenomena are likely to be attributed to sampling variability encountered in setting the data to missing and do not detract from the positive relationship observed between the rate of missing data and the absolute bias. As with the MNAR missingness mechanism, the standard error of estimate does not appear to be influenced in any predictable manner by the percentage of missing data. However, it is noted that the standard error of estimate is substantially larger than its true value of R41.20 for 40% and particularly 70% missing data.

Finally, in terms of the MCAR missingness mechanism, it is unclear as to whether or not the percentage of missing data has any systematic influence on the accuracy of the mean earnings estimate. It is noted that this estimate and its standard error are extremely close to their respective true values when missing data is at its lowest value of 10%. However, for higher rates of missing data, the bias and standard error of estimate appear to be somewhat random, with the best and worst estimates in terms of RMSE observed for 60% and 40% missing data respectively. One may therefore infer that the number of missing observations does not affect the efficacy of the imputation model with respect to the accuracy of point estimates in the presence of an MCAR missingness mechanism. From the findings observed thus far in this section, such a result should not be unexpected.

As was noted earlier, a sound imputation model should not only provide accurate point estimates, but also preserve the relationships between the variables in the dataset. In order to assess how the sequential regression multivariate imputation model with five imputations and ten iterations performs in this respect, the bias, standard error of estimate and RMSE for returns to education were computed as before for various rates of missing data on the earnings variable. The results, multiplied by 100, are presented in Table 5 on page 63.

From Table 5, there is a clear increase in the absolute bias across all missingness mechanisms as the percentage of missing data on the earnings variable rises. In all instances, the coefficient estimates tend to zero as the rate of missing data increases, implying that the strength of multivariate relationships are attenuated for high fractions of missing values. The observed increase in absolute bias is most notable for the MNAR missingness mechanism, followed by the MAR mechanism and finally the MCAR mechanism. For the MNAR missingness mechanism, the rate of return to education is underestimated by approximately 4% where the percentage of missing data

%	MCAR			MAR			MNAR		
Missing	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**
10%	-0.0115	0.1756	0.1760	-0.2020	0.1803	0.2708	-0.7887	0.1656	0.8059
20%	0.0112	0.1735	0.1739	-0.4222	0.1753	0.4571	-1.2277	0.1985	1.2437
30%	-0.0023	0.2090	0.2090	-0.5667	0.2033	0.6021	-1.9844	0.2132	1.9958
40%	-0.0834	0.2546	0.2679	-0.5108	0.2522	0.5697	-2.9840	0.2465	2.9942
50%	-0.0343	0.2084	0.2112	-0.8042	0.2320	0.8369	-3.6585	0.2177	3.6650
60%	-0.2501	0.2946	0.3865	-1.1131	0.2054	1.1319	-4.3601	0.1857	4.3640
70%	-0.2283	0.3042	0.3803	-1.2770	0.3082	1.3137	-4.4308	0.1258	4.4326

Table 5: Bias, standard error of estimate and root mean squared error in returns to education for various percentages of missing data on the monthly earnings variable for the three missingness mechanisms (5 imputations, 10 iterations). All figures multiplied by a factor of 100.

is 50% or more. The magnitude of this bias may be sufficient to affect policy decisions. Under an MAR mechanism, this bias is only of the order of 1%, whereas a downward bias in the region of only a quarter of a percent is observed for the MCAR mechanism with 60% to 70% missing data.

Interestingly, there does appear to be a positive relationship between the percentage of missing data and the absolute bias on the coefficient estimate for years of education under the MCAR mechanism. Recall that no such relationship could be established between the rate of missing data and the absolute bias on the point estimate of earnings under this mechanism. Hence, whilst increasing the number of missing entries does not influence the accuracy of point estimates under an MCAR mechanism, it does adversely attenuate the multivariate relationships in the dataset.

The standard error of estimate exceeds its true value of 0.1650% in all cases, except under an MNAR missingness mechanism with 70% missing data. Under the MAR mechanism, the standard error is almost double its true value with 70% missing data, whilst this is the case for 60% and 70% missing data under the MCAR missingness mechanism. In other instances, the departure for the true value does not raise any cause for concern. Nonetheless, the standard error of estimate does appear to increase as the percentage of missing data increases under the MCAR and MAR mechanisms. This may be desirable in light of the added uncertainty that results from larger fractions of missing data. Interestingly, however, the estimated standard errors are surprisingly close to their true values for very low rates of missing data.

As is to be expected from the results discussed above, the RMSE measure increases as the fraction of missing data rises. In the cases of the MCAR and MAR missingness mechanisms, this increase is driven by both the rise in absolute bias and the increase in the standard error of estimate. Under the MNAR mechanism, the increase in RMSE results largely from the enormous biases associated with the large rates of missing data.

4.3.2 Number of Covariates in the Imputation Model

Recall that monthly earnings values were set to missing under the MAR missingness mechanism based on the age, years of education, racial group, province and gender variables. In addition to these variables, hours worked per week, skills training, employment type and the occupation and sector variables were all used in imputing for monthly earnings. In practice, however, it may be difficult to ascertain exactly which variables should be included in the imputation model. Furthermore, fewer variables may be desirable in terms of computational efficiency. Indeed, most household surveys contain many variables and it would not be feasible to include all these variables in the imputation model. Consequently, it would be useful to assess the impact of the inclusion and exclusion of certain covariates in the imputation model in terms of the bias and efficiency of estimates.

Table 6 presents the bias, standard error of estimate and root mean squared error in the estimated mean monthly earnings under each of the three missingness mechanisms for various numbers of covariates. More specifically, the covariates utilised were either the *same* as those used to set the data to missing under the MAR mechanism (that is, age, years of education, racial group, province and gender), *less* than those used to set the data to missing (excluding the years of education and gender variables) or *more* than those used to induce missingness (including all variables in the artificial dataset). Five imputations and ten iterations with 30% missing data on the earnings variable were utilised.

It is noted that the absolute bias is substantially larger for both the MAR and MNAR mechanisms where fewer covariates were utilised. The standard error of estimate is well below its true value of R41.20 under the MNAR mechanism and therefore fails to provide an accurate reflection of the inherent uncertainty in the imputed values. Accordingly, the RMSE measure is much larger for both the MAR and MNAR mechanisms when less covariates are included in the imputation model than those which actually induced the missingness. This result is particularly interesting for the MAR missingness mechanism. It illustrates that even where such an ignorable

No.	**MCAR**			**MAR**			**MNAR**		
Covariates	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**
Less	-36.97	30.27	47.78	-111.44	36.82	117.37	-734.72	16.47	734.90
Same	-55.01	27.01	61.28	-50.23	42.07	65.52	-666.19	18.23	666.44
More	-57.08	26.84	63.07	-16.60	40.08	43.38	-595.03	17.02	595.27

Table 6: Bias, standard error of estimate and root mean squared error in mean monthly earnings for various numbers of covariates for the three missingness mechanisms (5 imputations, 10 iterations).

missingness mechanism persists, an imputation model that omits relevant predictors may still result in large biases and potentially under-represent uncertainty.

When all the covariates used to set the data to missing under the MAR mechanism were included in the imputation model, the absolute bias under the MAR mechanism was reduced by 54.93%. Furthermore, the standard error of estimate of R42.07 is a far more accurate reflection of its true value relative to the case with less covariates. By contrast, the observed reduction in absolute bias under the MNAR mechanism when the same number of covariates is utilised is only 9.33% and the standard error of estimate is still less than half of its true value. Hence, although there is a large improvement in RMSE for the MAR mechanism, the decrease is not nearly as dramatic under the MNAR missingness mechanism.

Finally, it would appear that including more covariates than that which actually induced missingness improves bias under both the MAR and MNAR mechanisms. Absolute bias decreases by 66.95% and 10.68% under the MAR and MNAR missingness mechanisms respectively, which is proportionally more than that observed in the preceding paragraph. Standard errors of estimates remain relatively unchanged in comparison to the case where the number of covariates equal that which induced missingness. The reduction in RMSE is therefore attributable to the decrease in absolute bias. This result implies that one should include all variables in the imputation model that are able to explain a reasonable proportion of the variation in the target variable, confirming the arguments of Van Buuren, Boshuizen and Knook (1999). Although some of these variables may not be directly related to missingness, there are likely to be indirect associations induced by the complex multivariate relationships between the variables in the dataset. Furthermore, the explanatory power of these independent variables with respect to the variation in the response variable will facilitate better predictions of the latter.

As was the case in assessing the influence of the fraction of missing data on the accuracy of estimates, it is difficult to ascertain whether or not the inclusion or exclusion of various covariates actually affects bias and efficiency under an MCAR missingness mechanism. From Table 6, it would appear that absolute bias increases as the number of covariates increase, which is clearly not consistent with the results under the MAR and MNAR mechanisms or simple intuition. In addition, small decreases in the standard error of estimate are observed as the number of covariates increase, becoming less representative of the true value. In general, one would not expect such relationships to exist on theoretical grounds and hence the small increases in RMSE observed as the number of covariates increase under the MCAR missingness mechanism are likely to be attributed to the randomness of the imputation procedure.

As before, it would also seem appropriate to consider the impact of the number of covariates on the multivariate relationships within the dataset. Table 7 presents the bias, standard error of estimate and RMSE for returns to education under the three missingness mechanisms.

Table 7 clearly suggests that absolute bias in the relationship between education and monthly earnings may be reduced by increasing the number of variables. This result is applicable for all three missingness mechanisms. In particular, the large biases in returns to education observed when imputing with less covariates is indicative of the consequences of omitting a variable of interest in post-imputation analyses (in this case, years of education) from the imputation model. Downward biases of the order of 3.5% (MAR) or 4.5% (MNAR) in returns to education can have serious consequences for policy decisions. Even the bias of approximately 2% under the MCAR mechanism may be non-trivial. Including years of education and gender in the imputation model reduces absolute bias by 36.14%, 80.11% and 58.91% under the MCAR, MAR and MNAR missingness mechanisms respectively. The reduction in absolute bias observed when including additional variables beyond those directly related to missingness is, however, not as significant. Indeed, a small (random) increase in absolute bias is observed under the MNAR mechanism. This is not entirely unexpected since the inclusion of further variables is unlikely to have a substantial impact on the relationship between monthly earnings and years of education.

The fluctuations in the standard error of estimate do not appear to be related to the number of covariates included in the imputation model. Hence, the RMSE is driven down by the reduction in absolute bias for larger numbers of covariates across all missingness mechanisms. Note that although increasing the number of covariates appeared to have no systematic influence on the bias of the point estimate of mean monthly earnings under the MCAR mechanism, this is certainly not the case when evaluating multivariate relationships under the same conditions. Increasing the number of covariates to at least include all those that are related to response probabilities is therefore justified from this perspective for the MCAR mechanism.

No.	**MCAR**			**MAR**			**MNAR**		
Covariates	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**	**Bias**	**Std Err**	**RMSE**
Less	-2.2392	0.2474	2.2528	-3.5465	0.2236	3.5535	-4.4572	0.1990	4.4617
Same	1.4300	0.2263	1.4478	0.7053	0.1957	0.7319	-1.8313	0.1612	1.8384
More	-0.0023	0.2090	0.2090	-0.5667	0.2033	0.6021	-1.9844	0.2132	1.9958

Table 7: Bias, standard error of estimate and root mean squared error in returns to education for various numbers of covariates for the three missingness mechanisms (5 imputations, 10 iterations).

5 Conclusions

The empirical findings of this paper largely support the current literature on incomplete multivariate data analysis. Irrespective of the missingness mechanism, it was established that only a few iterations are required in order to achieve convergence to the posterior predictive distribution $P(\mathbf{Y}_{\mathrm{MISS}}|\mathbf{Y}_{\mathrm{OBS}})$. In the worst-case scenario, convergence was attained after eight iterations. Ten iterations is recommended as a conservative choice, especially since the increase in computational power necessary to produce ten iterations as oppose to only eight is negligible. It should, however, be noted that although convergence may be achieved, the stationary distribution may not necessarily coincide with the actual distribution of the missing data. This will be the case when likelihood-based inferences concerning $\boldsymbol{\theta}$ are invalidated by failing to consider the missingness mechanism. Hence, sequential regression multivariate imputation is not appropriate when the missingness mechanism is non-ignorable for inferences on $\boldsymbol{\theta}$.

The empirical findings revealed that as few as three imputations are necessary when imputing under an MAR missingness mechanism. Following Rubin's (1987) advice, it is therefore recommended that five imputations be utilised as the cautious choice. After five imputations under an MAR mechanism, the bias of point estimates was found to be negligible, whilst the standard error was large enough to adequately account for the uncertainty associated with the imputation procedure. Consequently, coverage in repeated sampling was found to accurately reflect the nominal percentage under an MAR missingness mechanism. The imputation model is, however, unable to account for the biases that may arise due to random (rather than systematic) differences between respondents and non-respondents under the MCAR mechanism, although one would expect zero bias in repeated sampling under this mechanism. When the missingness mechanism is MNAR, a single imputation was found to reduce bias, although by no means eliminate it. Multiple imputations did not result in further improvements in bias and the standard error remained well below its true value. This finding further attests to the imputation model's inability to deal effectively with non-ignorable missingness. Interestingly, the number of imputations was not found to influence the multivariate relationships within the dataset.

The absolute bias increases with the percentage of missing values, becoming quite severe where this figure exceeds 50%. There does not appear to be any clear relationship between the fraction of missing data and the standard error of estimate. These conclusions are applicable to both the MAR and MNAR missingness mechanisms. As expected, neither the bias nor the standard error of point estimates appear to be influenced by the rate of missing data under an MCAR missingness mechanism. Multivariate relationships, on the other hand, were distorted for higher fractions of missing data with coefficient estimates becoming attenuated under all three missingness mechanisms.

Under the MAR and MNAR missingness mechanisms, it was established that increasing the number of covariates to include more predictors than the variables that actually induced missingness reduces the absolute bias in point estimates. Under the MAR missingness mechanism, including fewer covariates than those which were accountable for the missing values not only resulted in a large absolute bias, but also produced a standard error of estimate that was well below its true value. As with the rate of missing data, the point estimates obtained under an MCAR mechanism did not exhibit sensitivity to the number of covariates. However, the relationships between variables are affected by the number of covariates in the model under all three missingness mechanisms, with severely attenuated coefficient estimates observed when the imputation model did not include the independent variable of interest. These results indicate the importance of including all variables in the imputation model that are likely to be subjected to post-imputation statistical analyses.

It is therefore recommended that the choice of independent variables to be included in an imputation model be informed by the following three considerations. Firstly, variables that are known to influence the occurrence of missing data should clearly be included. These might be identified through tabulation or a logistic regression model with a response indicator as the dependent variable. Secondly, predictor variables should be able to explain a significant proportion of the variation in the response variable. A correlation analysis or regression model of the observed data may prove useful in this respect. Finally, in order to avoid biases in the subsequent statistical analyses to be performed on the multiply imputed data, it is necessary to include all variables that may be utilised in such analyses [18]. This latter consideration is likely to be the most arduous of the three where survey data forms part of a large database serving many users with various interests. Even if all the observed variables are included in the imputation model (which is clearly not feasible in practice), one may never by able to account for all the possible variations of interaction terms and higher-order variables that multiple users may wish to evaluate.

5.1 Further Research

Multiple imputation methods, facilitated by Markov chain Monte Carlo, provide a valuable and flexible approach to statistical inference with incomplete multivariate data. The theory surrounding these methods and their applications to missing data and related problems is among the most rapidly developing areas of modern statistical science [13]. Some of the promising current and future areas of research in this field are outlined below, along with some useful references for further reading.

5.1.1 Non-Ignorable Methods

The assumption of ignorable missingness is computationally convenient as it allows the analyst to construct an imputation model without explicitly specifying the distribution of missingness $P(\mathbf{R}|\boldsymbol{\xi})$. In many situations, however, this assumption is questionable and it would therefore be worthwhile to investigate non-ignorable alternatives [13]. Two broad approaches have been addressed in the literature and are certainly not exhaustive of all the attempts made to tackle this problem. The first of these approaches is the use of selection models, which seek to explicitly model the sample selection process that determines why some values are observed and others not. Adaptations of the Tobit model (the so-called Type II Tobit model) and Heckman's two-step method have been suggested in this respect. The second approach to non-ignorable missingness is the adoption of pattern-mixture models. Such models do not attempt to describe the individuals' response probabilities, but instead classify individuals by their missingness and use the observed data within each missingness group to extrapolate aspects of this behaviour to unseen portions of the data [15]. Chapter 15 of Little and Rubin (2002) provides a rigorous discussion of these non-ignorable models. The construction and evaluation of such models are an important area for future study.

5.1.2 Models for Complex Survey Data

The sequential regression multivariate imputation technique assumes that the dataset arises from a simple random sample. However, most surveys, including the *Labour Force Survey* utilised here, employ complex sample designs involving stratification, clustering and weighting [9]. Although the complex sample design was taken into account when analysing the multiply imputed data, the imputation technique itself does not account for the important features of the sample design. Raghunathan *et al* (2001) suggest that when used in conjunction with an appropriate variance estimation technique, such as jackknife repeated replication, Taylor series linearisation or balanced repeated replication, the sequential regression multivariate imputation model may have more appealing design-based properties. Other popular alternatives to account for the complex sample design include random effects models and generalised linear mixed models. Such models may be quite complex in practice and further research needs to be conducted to formulate more flexible models and algorithms for imputation in a complex design setting [13].

5.1.3 Models for Semicontinuous Variables

Semicontinuous variables arise in a wide variety of contexts. They are defined as variables that take on a specific value (usually zero) with a positive probability, but otherwise assume values that can be modelled by a continuous distribution [4]. An example of such a variable is earnings, which takes on a value of zero for unemployed persons and typically follows a lognormal distribution for employed individuals. Since this study was concerned only with the employed, the logarithmic transformation to normality necessary for ordinary least squares regression did not prove problematic, since very few employed persons have zero earnings. Consequently, the earnings variable was treated as if it were continuous. However, in other contexts, the normalising transformation may be infeasible due to the point mass at zero [4]. When missing values occur on semicontinuous variables of this nature, it is necessary to apply missing data methods that are explicitly designed for them. Ad hoc approaches, such as imputing the variables as if they were normally distributed and then truncating negative values to zero, do not work well in practice [13]. More sophisticated methods, such as the blocked general location model, are required for imputing these variables. Javaras and Van Dyk (2003) provide a thorough discussion of such methods.

The aforementioned areas of enquiry are by no means exhaustive and provide only a taste of what is likely to transpire from research in this field in the coming years. Indeed, the problem of non-response is here to stay and the development of more sophisticated statistical methods for dealing with incomplete datasets is likely to occupy statisticians for many years into the future.

References

[1] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**: 1 – 38.

[2] Dobson, A. J. (2002). *An Introduction to Generalized Linear Models.* 2nd ed. Chapman and Hall, United States of America.

[3] Gill, J. (2002). *Bayesian Methods: A Social and Behavioural Sciences Approach.* Chapman and Hall, United States of America.

[4] Javaras, K. N. and Van Dyk, D. A. (2003). Multiple Imputation for Incomplete Data with Semicontinuous Variables. *Journal of the American Statistical Association,* **98**, 463: 703 – 715.

[5] Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications, United States of America.

[6] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* 2nd ed. John Wiley and Sons, New York.

[7] Madow, W. G., Nisselson, J. and Olken, I. (1983). *Incomplete Data in Sample Surveys: Vol. 1. Report and Case Studies.* Academic Press, New York.

[8] Press, W. H. (1992). *Numerical Recipes in Fortran.* Cambridge University Press, New York.

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

[10] Royston, P. (2004). Multiple Imputation for Missing Values. *Stata Journal,* **4**, 3: 227 – 241.

[11] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika,* **63**: 581 – 592.

[12] Rubin, D. B. (1987). *Multiple Imputation for Non-Response in Surveys.* John Wiley and Sons, New York.

[13] Schafer, J. L. (1997). *Analysis of Multivariate Incomplete Data.* Chapman and Hall, London.

[14] Schafer, J. L. (1999). Multiple Imputation: A Primer. *Statistical Methods in Medical Research,* **8**: 3 – 15.

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

[16] STATA Corporation. (2003). *STATA Survey Data Reference Manual Release 8.* Stata Press, Texas.

[17] Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of American Statistical Association,* **82**: 528 – 550.

[18] Van Buuren, S., Boshuizen, H. and Knook, D. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine,* **18**: 681 – 694.

[19] Zhang, P. (2003). Multiple Imputation: Theory and Method. *International Statistical Review,* **71**, 3: 581 – 592.

Appendices

A Rubin's Measure of Relative Efficiency

	Fraction of Missing Information γ								
m	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
1	95	91	88	85	82	79	77	75	73
2	98	95	93	91	89	88	86	85	83
3	98	97	95	94	93	91	90	89	88
5	99	98	97	96	95	94	94	93	92
∞	100	100	100	100	100	100	100	100	100

Table A1: Large sample relative efficiency (in percentage units of standard deviations) when using a finite number of imputations, rather than an infinite number of imputations, as a function of the fraction of missing information γ (Rubin, 1989, p. 114).

B Assessing Distributional Convergence

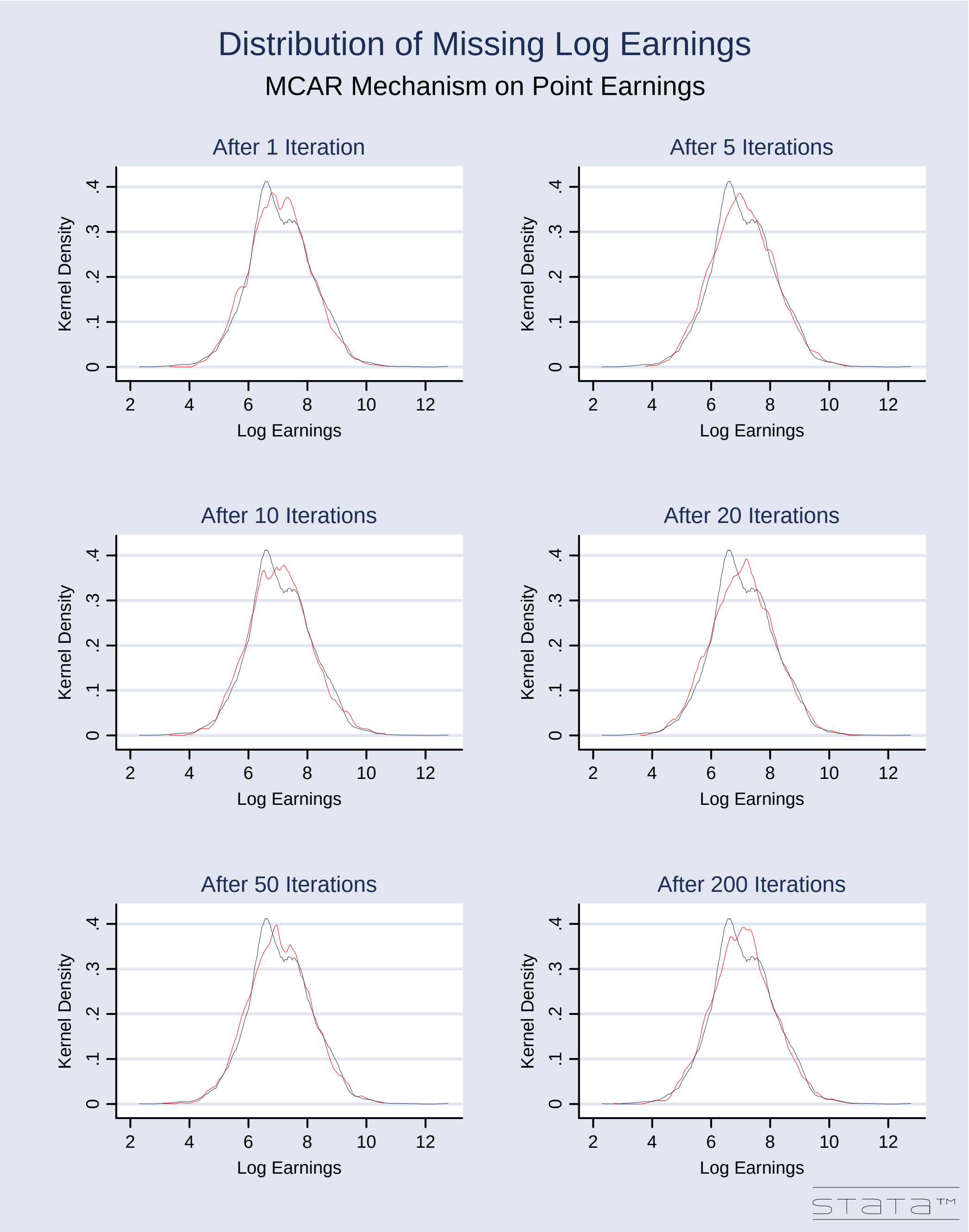

Figure B1: Distribution of missing log earnings after various iterations of the Markov chain under the MCAR mechanism. Red density indicates imputed earnings and blue density indicates true earnings.

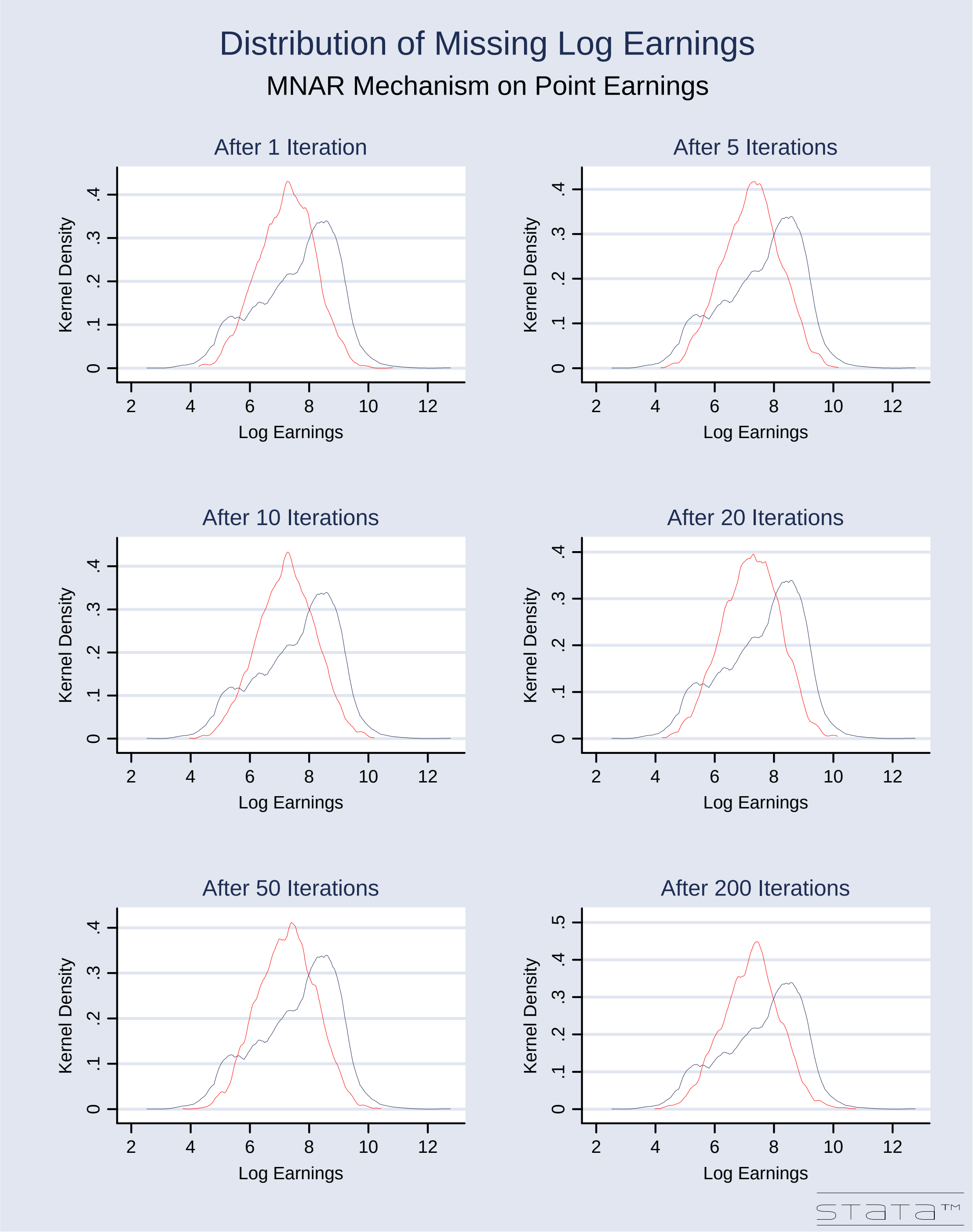

Figure B2: Distribution of missing log earnings after various iterations of the Markov chain under the MNAR mechanism. Red density indicates imputed earnings and blue density indicates true earnings.

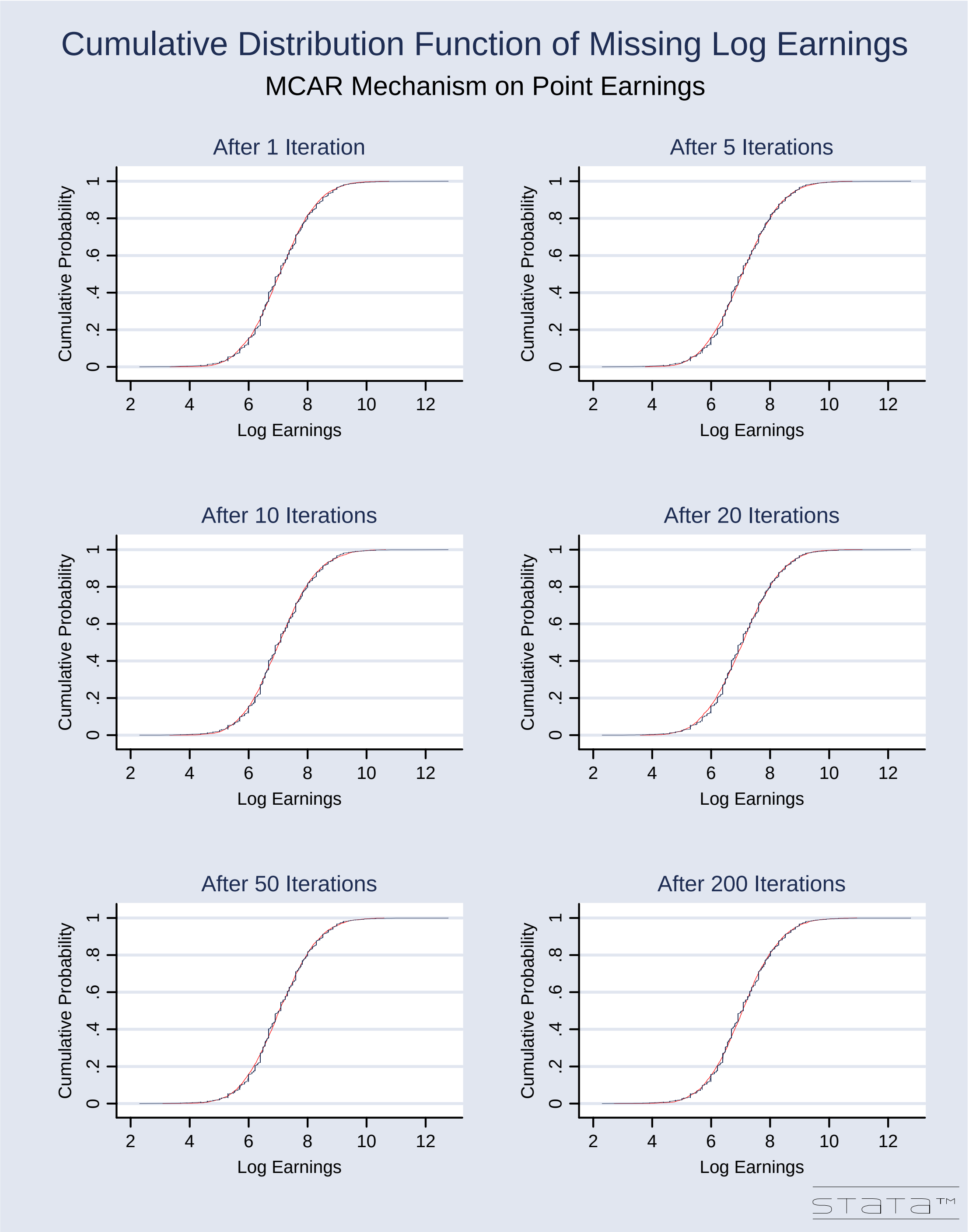

Figure B3: Cumulative distribution function of missing log earnings after various iterations of the Markov chain under MCAR mechanism. Red curve indicates imputed earnings and blue curve indicates true earnings.

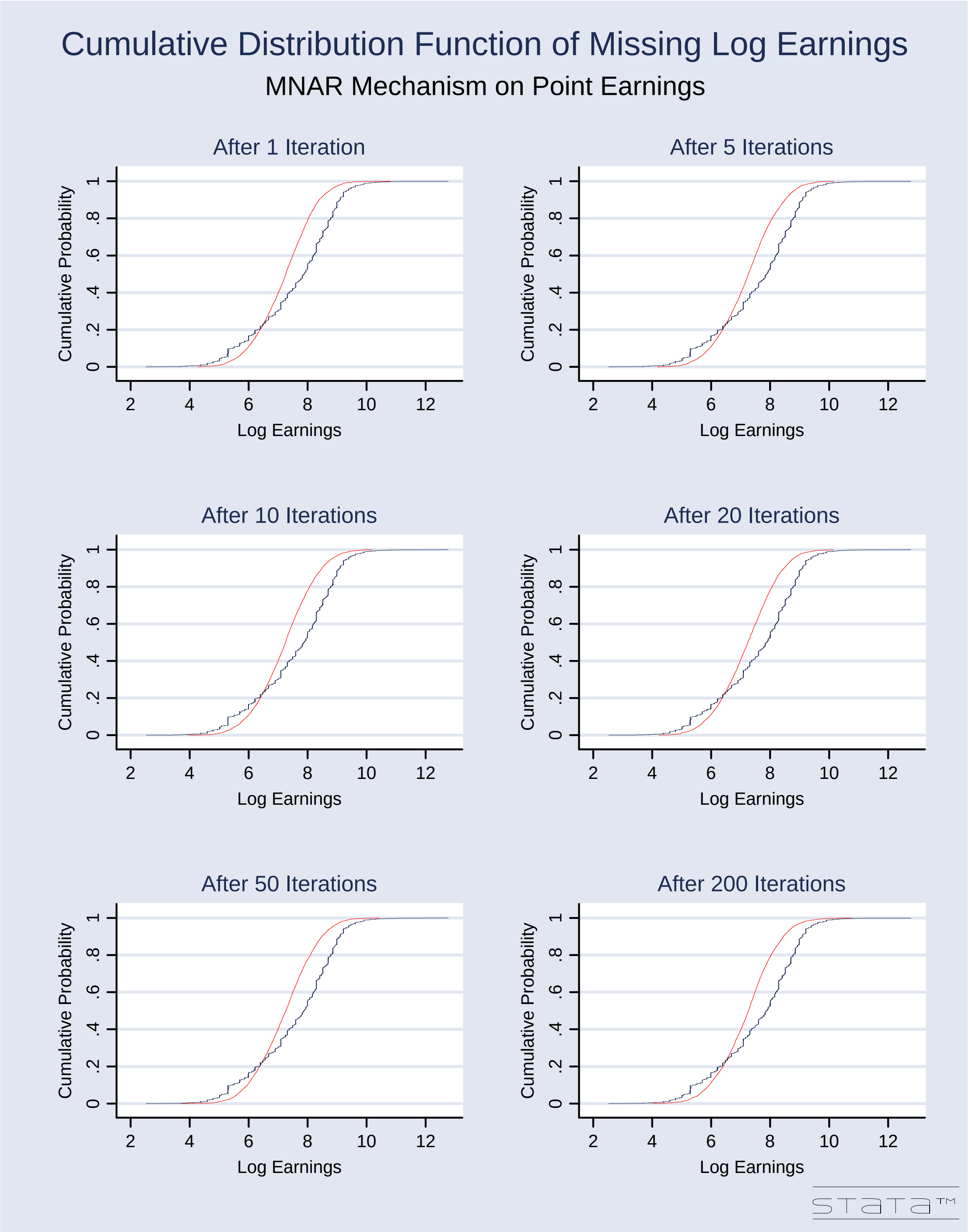

Figure B4: Cumulative distribution function of missing log earnings after various iterations of the Markov chain under MNAR mechanism. Red curve indicates imputed earnings and blue curve indicates true earnings.

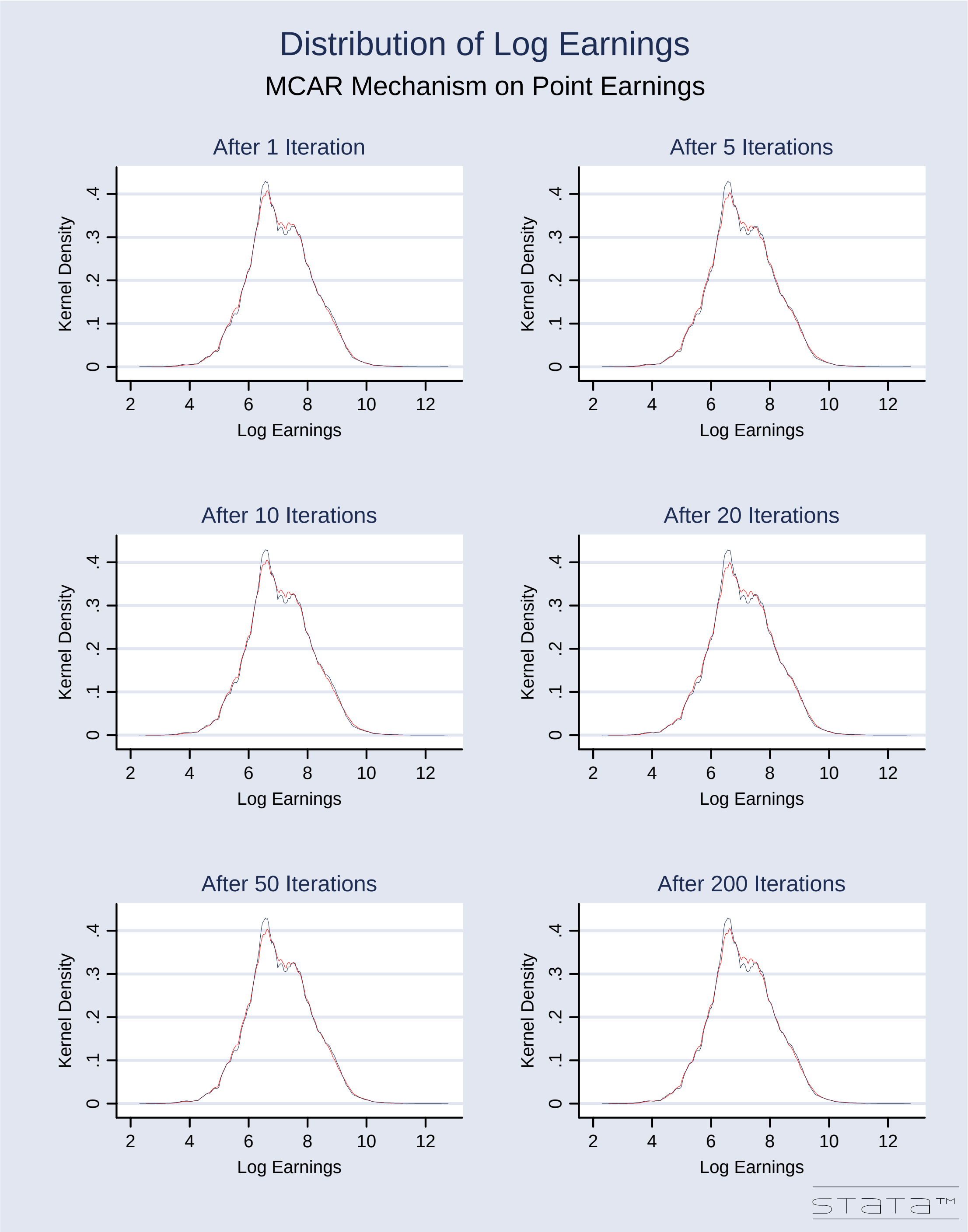

Figure B5: Distribution of missing log earnings after various iterations of the Markov chain under the MCAR mechanism. Red density indicates imputed earnings and blue density indicates true earnings.

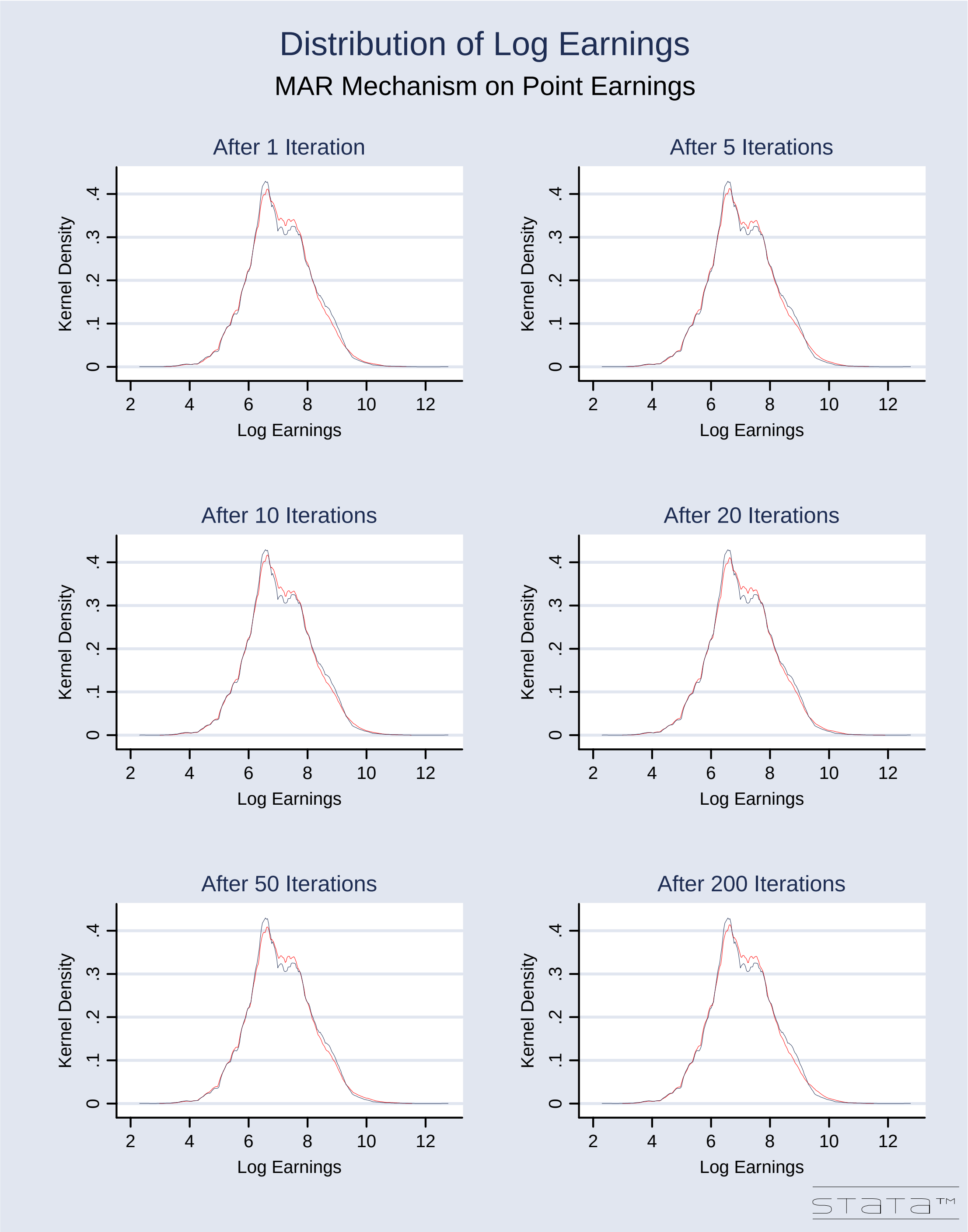

Figure B6: Distribution of missing log earnings after various iterations of the Markov chain under the MAR mechanism. Red density indicates imputed earnings and blue density indicates true earnings.

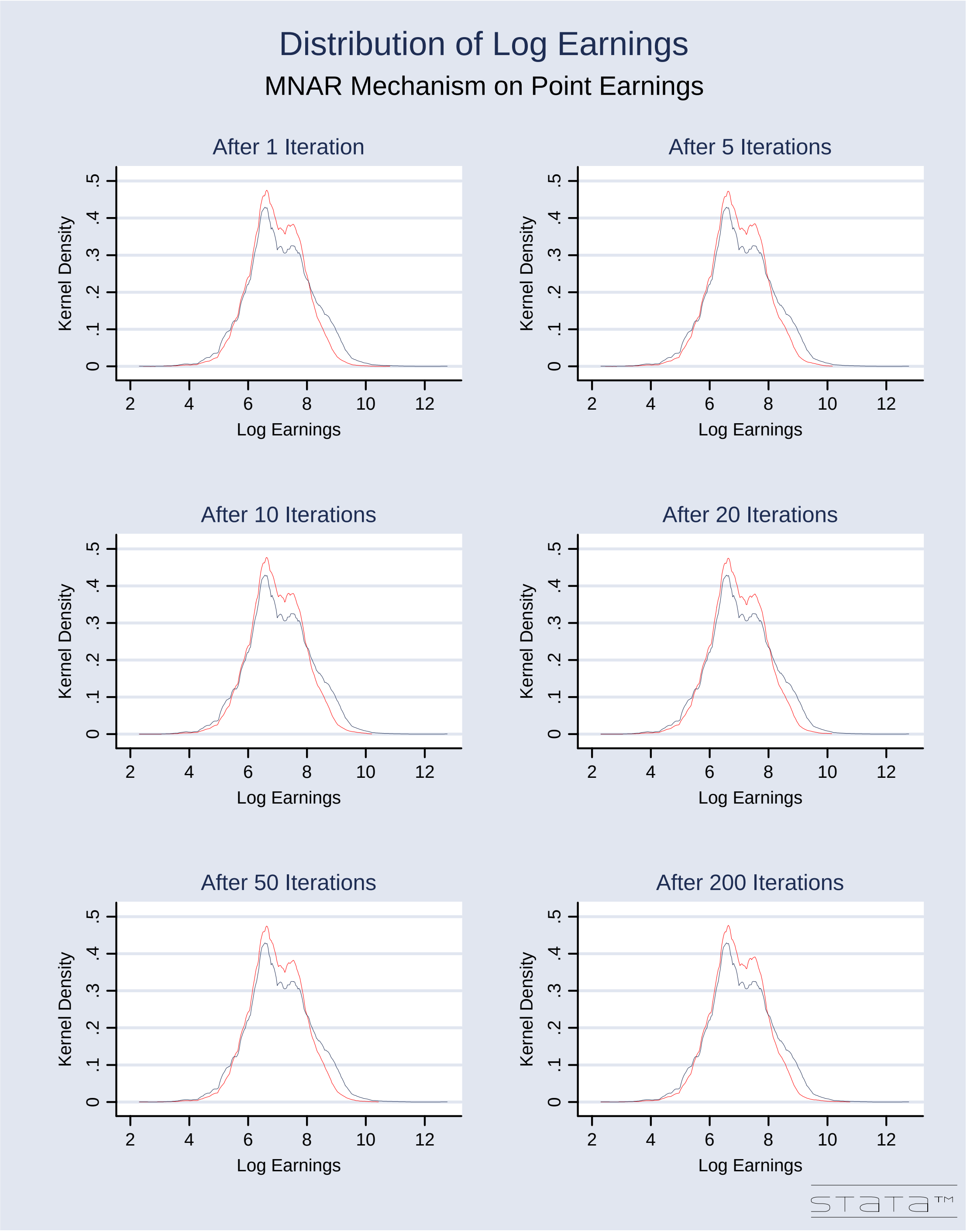

Figure B7: Distribution of missing log earnings after various iterations of the Markov chain under the MNAR mechanism. Red density indicates imputed earnings and blue density indicates true earnings.

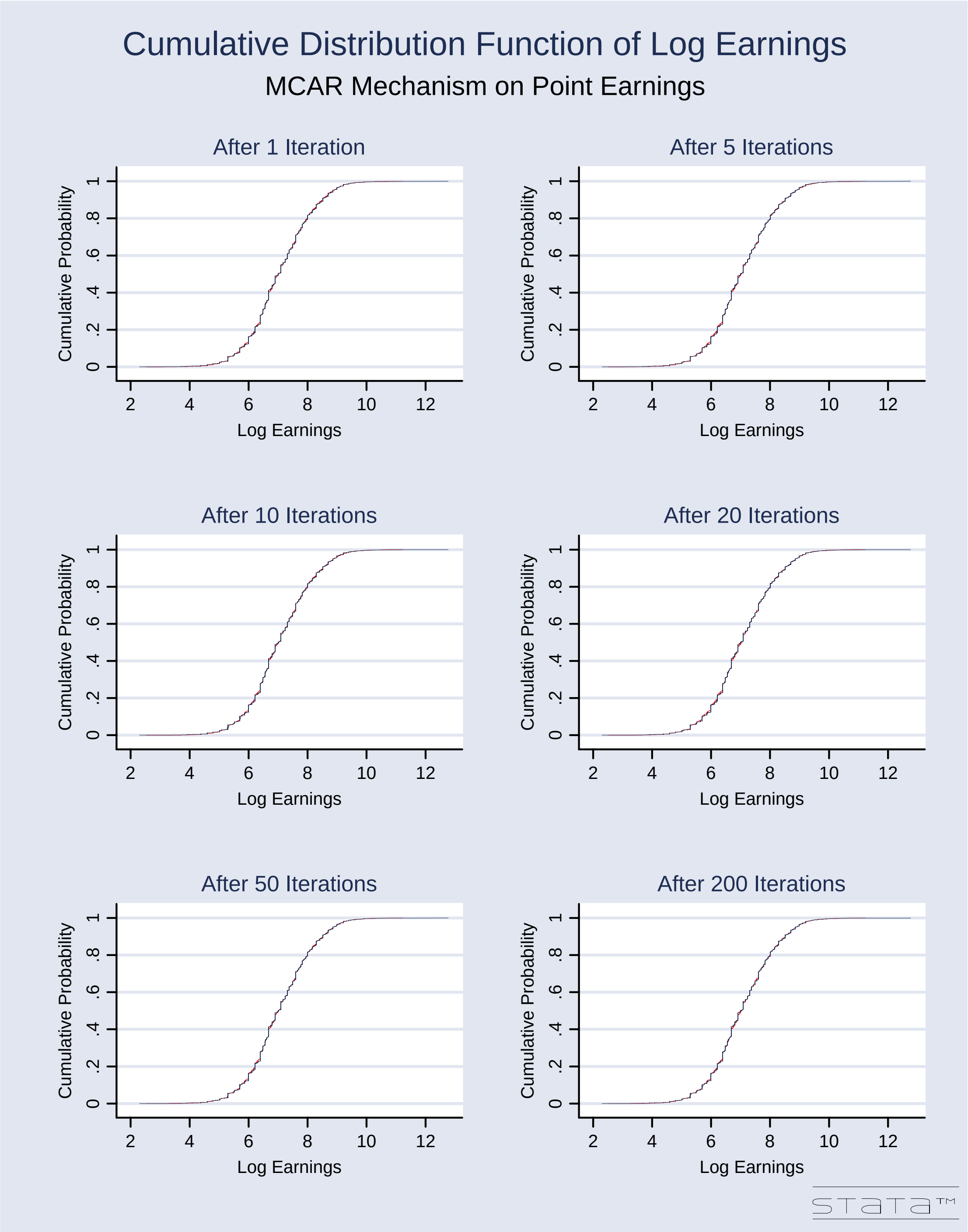

Figure B8: Cumulative distribution function of log earnings after various iterations of the Markov chain under MCAR mechanism. Red curve indicates imputed earnings and blue curve indicates true earnings.

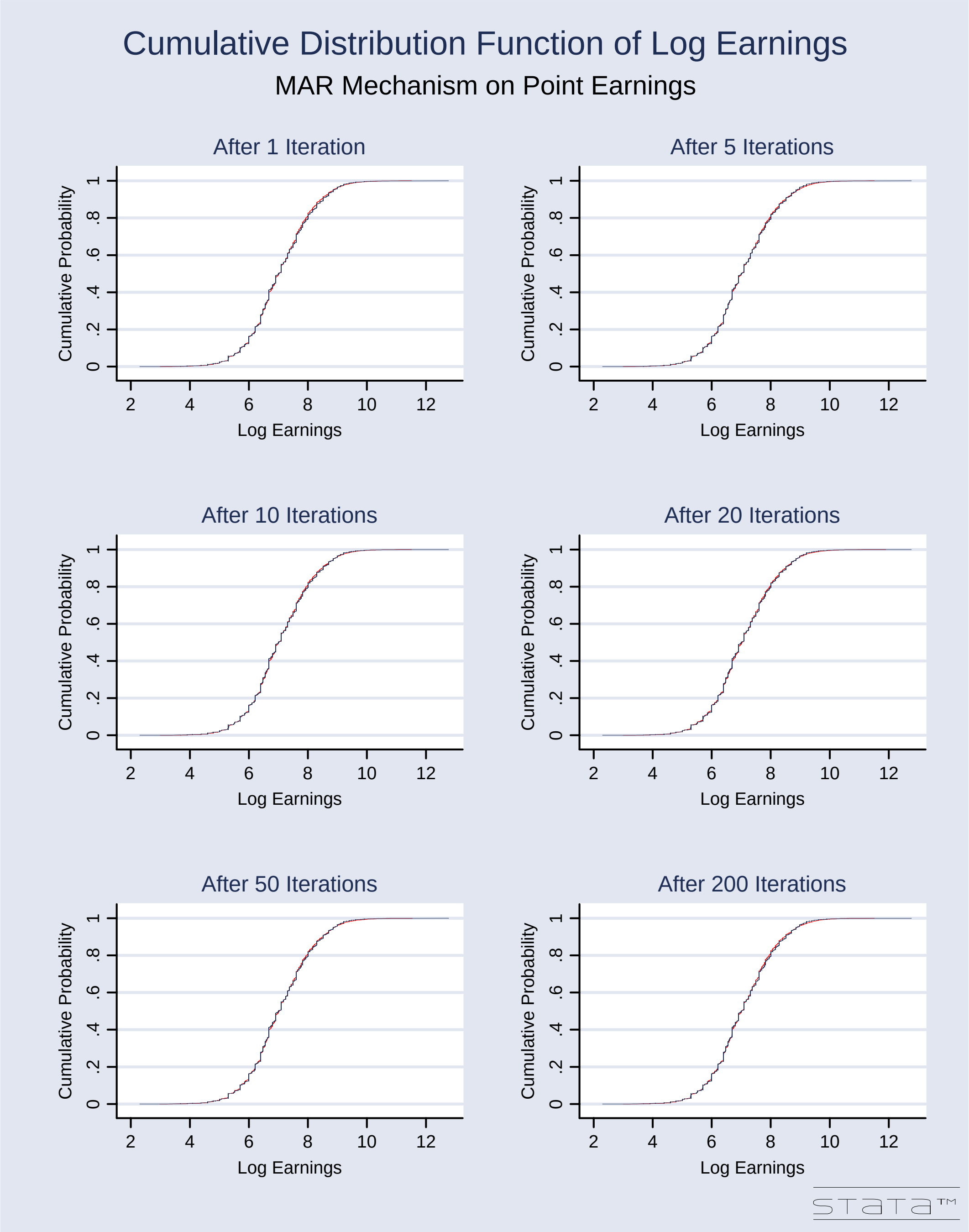

Figure B9: Cumulative distribution function of log earnings after various iterations of the Markov chain under MAR mechanism. Red curve indicates imputed earnings and blue curve indicates true earnings.

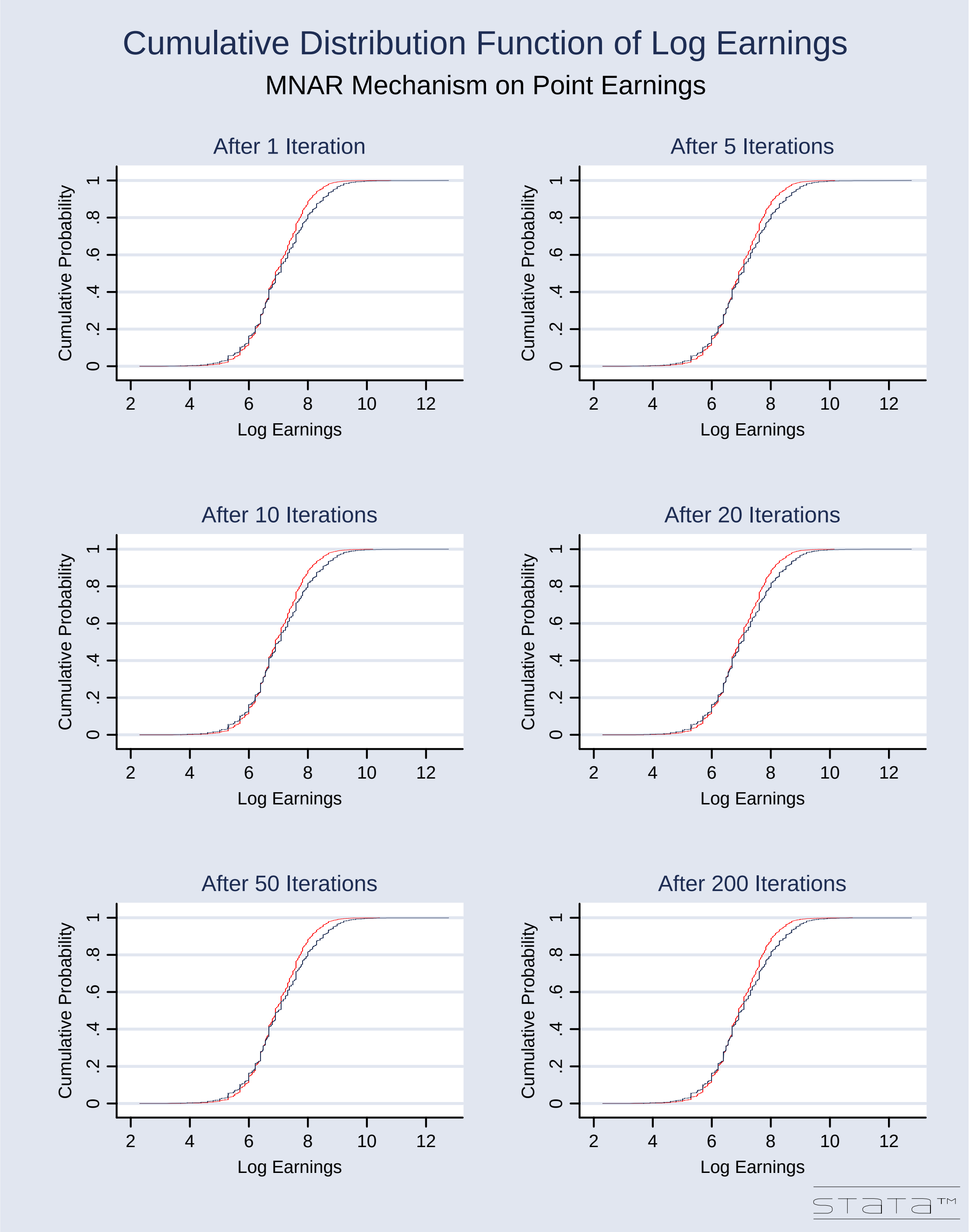

Figure B10: Cumulative distribution function of log earnings after various iterations of the Markov chain under MNAR mechanism. Red curve indicates imputed earnings and blue curve indicates true earnings.

About DatatFirst

DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
- research to improve the quality of African survey data
- training of African data managers for better data curation on the continent
- training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad
SPES BONA