

UNIVERSITY *of the* WESTERN CAPE



Imputation Techniques for Non-ordered Categorical Missing Data



A thesis submitted in fulfilment for the
degree of Doctor of Philosophy in Statistics

at the
Faculty of Natural Sciences
Department of Statistics and Population Studies

Supervisor: Prof. Danelle Kotze
Co-Supervisor: Prof. Renette Blignaut

February 2016

Declaration of authorship



I declare that **Imputation Techniques for Non-ordered Categorical Missing Data** is my own work, that it has not been submitted for any degree or examination in any other university and that all the sources I used or quoted have been indicated and acknowledged by appropriate references.

Innocent Karangwa

Signed:

Date:

Keywords

Missing data

Missing at random

Multiple imputation

Multivariate normal imputation

Multiple imputation by chained equations

Categorical data



Abstract

Missing data are common in survey data sets. Enrolled subjects do not often have data recorded for all variables of interest. The inappropriate handling of missing data may lead to bias in the estimates and incorrect inferences. Therefore, special attention is needed when analysing incomplete data. The multivariate normal imputation (MVNI) and the multiple imputation by chained equations (MICE) have emerged as the best techniques to impute or fill in missing data. The former assumes a normal distribution of the variables in the imputation model, but can also handle missing data whose distributions are not normal. The latter fills in missing values taking into account the distributional form of the variables to be imputed. The aim of this study was to determine the performance of these methods when data are missing at random (MAR) or completely at random (MCAR) on unordered or nominal categorical variables treated as predictors or response variables in the regression models. Both dichotomous and polytomous variables were considered in the analysis. The baseline data used was the 2007 Demographic and Health Survey (DHS) from the Democratic Republic of Congo. The analysis model of interest was the logistic regression model of the woman's contraceptive method use status on her marital status, controlling or not for other covariates (continuous, nominal and ordinal). Based on the data set with missing values, data sets with missing at random and missing completely at random observations on either the covariates or response variables measured on nominal scale were first simulated, and then used for imputation purposes. Under MVNI method, unordered categorical variables were first dichotomised, and then $K - 1$ (where K is the number of levels of the categorical variable of interest) dichotomised variables were included in the imputation model, leaving the other category as a reference. These variables were imputed as continuous variables using a linear regression model. Imputation with MICE considered the distributional form of each variable to be imputed. That is, imputations were drawn using binary and multinomial logistic regressions for dichotomous and polytomous variables respectively. The performance of these methods was evaluated in terms of bias and standard errors in regression coefficients that were estimated to determine the association between

the woman's contraceptive methods use status and her marital status, controlling or not for other types of variables. The analysis was done assuming that the sample was not weighted first, then the sample weight was taken into account to assess whether the sample design would affect the performance of the multiple imputation methods of interest, namely MVNI and MICE. As expected, the results showed that for all the models, MVNI and MICE produced less biased smaller standard errors than the case deletion (CD) method, which discards items with missing values from the analysis. Moreover, it was found that when data were missing (MCAR or MAR) on the nominal variables that were treated as predictors in the regression model, MVNI reduced bias in the regression coefficients and standard errors compared to MICE, for both unweighted and weighted data sets. On the other hand, the results indicated that MICE outperforms MVNI when data were missing on the response variables, either the binary or polytomous. Furthermore, it was noted that the sample design (sample weights), the rates of missingness and the missing data mechanisms (MCAR or MAR) did not affect the behaviour of the multiple imputation methods that were considered in this study. Thus, based on these results, it can be concluded that when missing values are present on the outcome variables measured on a nominal scale in regression models, the distributional form of the variable with missing values should be taken into account. When these variables are used as predictors (with missing observations), the parametric imputation approach (MVNI) would be a better option than MICE.

Acknowledgements

Writing this thesis has been an extremely good experience full of challenges and drawbacks I could not have navigated through alone. Fortunately, I was blessed to have many wonderful people around me whose support, insight, and knowledge of statistics have made this journey so easy.

First and foremost, I owe this dissertation to my supervisors, Professor Danelle Kotze and Professor Renette Blignaut who were always there for me when I needed them most. Their patient support was the strongest inspiration to persevere in the face of various obstacles. They were not only helpful in giving me advice in both composition and content, but also in providing me with financial support.

I would like to thank the entire Department of Statistics and Population Studies at the University of the Western Cape for providing an incredibly stimulating learning environment and statistical knowledge.

My friends and colleagues at this university were so dear to me, having been there through trying times. I would like to recognise their valuable opinions and experience. In particular, I would like to acknowledge Dr Siaka Lougue, Mr Aristide Bado, Mr Yasser Bucyana, Mr Justin Rutikanga and Mr Adam Andani for their extremely valuable conversations, feedback and any other kind of contribution to my thesis. I wish them all the future success they deserve.

Special thanks and recognition are also owed to the DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) that funded this research during the final year of my studies.

Finally, I also thank everyone who gave me emotional support.

Contents

Declaration of authorship	i
Keywords	ii
Abstract	iii
Acknowledgements	v
List of figures	x
List of tables	xxii
1 Introduction	1
1.1 Background on missing data	1
1.1.1 Sampling and nonresponse	1
1.1.2 Missing data	3
1.1.3 Missing data patterns	5
1.1.4 Missingness mechanisms	6
1.1.5 Testing for missingness mechanisms	9
1.2 Motivation for the study	11
1.3 Significance of the study	14
1.4 Research objectives	14
1.5 Research questions	15
1.6 Hypotheses	16
1.7 Research Design	16
1.8 Thesis overview	16
2 Markov Chain Monte Carlo process	19
2.1 Introduction	19
2.2 Monte Carlo integration	20
2.3 Importance sampling	23



2.4	Markov Chain Monte Carlo	26
2.4.1	Introduction	26
2.4.2	Metropolis-Hastings algorithm	28
2.4.3	Gibbs sampling	30
2.5	Markov Chain Monte Carlo methods in the presence of missing data	35
2.5.1	Introduction	35
2.5.2	Markov Chain Monte Carlo versus missing data	35
2.6	Summary of the chapter	36
3	Literature review on missing data methods	38
3.1	Introduction	38
3.2	Single-based imputation methods	39
3.2.1	Mean imputation	39
3.2.2	Hot-deck imputation	39
3.2.3	Cold-deck imputation	40
3.2.4	Regression imputation	40
3.2.5	Imputation using interpolation	41
3.3	Model-based methods	41
3.3.1	Expectation maximisation	41
3.3.2	Maximum likelihood method	42
3.4	Multiple imputation-based methods	43
3.4.1	Introduction	43
3.4.2	Description of multivariate normal imputation	45
3.4.3	Description of multiple imputation by chained equations	48
3.4.4	Multivariate normal imputation versus multiple imputation by chained equation: a practical example using a survey data set to impute missing values of continuous variables	52
3.4.4.1	Introduction	52
3.4.4.2	Data	52
3.4.4.3	Analysis method	54
3.4.4.4	Findings	54
3.4.4.5	Conclusion	59
3.4.5	Multivariate normal imputation versus multiple imputation by chained equations of categorical data	59
3.5	Summary of the chapter	62
4	Methodology	63
4.1	Description of data set and variables used in the study	63
4.2	Simulation of the data sets with missing values	65
4.3	Missing data models	65
4.4	Analysis method	68
4.4.1	Imputation of missing values	68
4.4.2	Model development and computation of the performance measures	73
4.4.3	Imputation models' diagnostics	73

4.5	Summary of the chapter	74
5	Results	76
5.1	Introduction	76
5.2	Scenario 1: Logistic regression models with missing values on the covariates	77
5.2.1	Model 1.1: Binary logistic regression model with missing values on a single covariate measured on a nominal scale	78
5.2.1.1	Results when 50% of data are missing at random or completely at random on the covariate	78
	Descriptive statistics	78
	Performance measures	83
	Model diagnostics	86
5.2.1.2	Results when 30% of data are missing at random or completely at random on the covariate	87
	Descriptive statistics	87
	Performance measures	90
	Model diagnostics	93
5.2.1.3	Results when 10% of data are missing at random or completely at random on the covariate	94
	Descriptive statistics	94
	Performance measures	97
	Model diagnostics	100
5.2.2	Model 1.2: Binary logistic regression model with more than two covariates in which two are measured on a nominal scale containing missing values	101
5.2.2.1	Model 1.2.1: Model with two nominal covariates both with 50% of their values missing at random or completely at random	101
	Descriptive statistics	101
	Performance measures	103
	Model diagnostics	107
5.2.2.2	Model 1.2.2: Model with various covariates with 50% missing values at random or completely at random on only the unordered categorical ones	108
	Descriptive statistics	108
	Performance measures	109
	Model diagnostics	116
5.2.3	Scenario 1: Summary of findings	116
5.3	Scenario 2: Logistic regression models with missing variables on the response variables	118
5.3.1	Model 2.1: Binary logistic regression model with missing values on the response variable	118
5.3.1.1	Description of data sets with missing values	118
5.3.1.2	Performance measures	122

5.3.1.3	Model diagnostics	126
5.3.2	Model 2.2: Multinomial logistic regression model with missing values on the response variable	126
5.3.2.1	Description of data sets with missing values	126
5.3.2.2	Computation of the performance measures	130
5.3.2.3	Model diagnostics	137
5.3.3	Scenario 2: Summary of findings	137
5.4	Summary of the chapter	138
6	Discussion and conclusion	141
Appendix A		155
Appendix B		157
Appendix C		182
Appendix D		212



List of Figures

1.1	Classification of survey errors (Bethlehem, 2009, page 180)	3
1.2	Types of missing data patterns (Enders, 2010, page 4). The shaded areas symbolize the missing data	6
1.3	Types of missing data patterns (Enders, 2010, page 12)	9
2.1	Plot of the function $h(x)$ in Equation (2.9)	22
2.2	Approximation of the integral of the function $h(x)$ by Monte Carlo method when f is a normal density: mean \pm two standards errors against iterations for the single sequence of simulations	22
2.3	Plot of the function in Equation (2.17)	25
2.4	Convergence of the importance sampling approximation of the function $(h(x))^* = h(x)p(x)$ using a sequence of samples generated from a uniform distribution: mean \pm two standard errors against iterations for the single sequence of simulations	25
2.5	20000 MCMC samples produced by a Be(3,7) distribution. Histogram from a Metropolis-Hastings algorithm and a Be(3,7) distribution.	30
2.6	Plot of the bivariate normal distribution of random variables X and Y simulated by iteratively sampling from the conditional distributions of these two variables using 1000 runs, different starting values of the chain and a correlation coefficient of 0 between X and Y	33
2.7	Plot of the bivariate normal distribution of random variables X and Y simulated by iteratively sampling from the conditional distributions of these two variables using 10000 runs, different values of the correlation coefficients between X and Y as well as a starting value of the chain of 0 for both X and Y	34
3.1	Estimates of slopes for age when the CD, MVNI and MICE methods are used at different rates of missingness	57
3.2	Estimates of slopes for education when the CD, MVNI and MICE methods are used at different rates of missingness	57
3.3	Estimates of standard errors for age when the CD, MVNI and MICE methods are used at different rates of missingness	57
3.4	Estimates of standard errors for education when the CD, MVNI and MICE methods are used at different rates of missingness	58
3.5	P-values of the models estimated using the CD, MVNI and MICE methods at different rates of missingness	58

4.1	Missing data models considered in the analysis.	67
5.1	Scenario 1: Logistic regression models with missing values on the covariates.	78
5.2	Model 1.1: Plot of bias and standard errors when 50% data are MAR or MCAR on marital status for unweighted data sets.	85
5.3	Model 1.1: Plot of bias and standard errors when 50% data are MAR or MCAR on marital status for weighted data sets.	85
5.4	Model 1.1: Plot of bias and standard errors when approximately 30% data are MAR and MCAR for unweighted data sets.	92
5.5	Model 1.1: Plot of bias and standard errors when approximately 30% data are MAR and MCAR for weighted data sets.	92
5.6	Model 1.1: Plot of bias and standard errors when 10% of the data are MAR and MCAR for unweighted data sets.	99
5.7	Model 1.1: Plot of bias and standard errors when 10% of the data are MAR and MCAR for weighted data sets.	99
5.8	Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for unweighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively.	107
5.9	Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for weighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively.	107
5.10	Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for unweighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively, 16-17 refer to variables age and education respectively, and 18-21 refer to levels of wealth index.	114
5.11	Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for weighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively, 16-17 refer to variables age and education respectively, and 18-21 refer to levels of wealth index.	115
5.12	Scenario 2: Logistic regression models with missing variables on the response variables.	118
5.13	Model 2: Plot of bias and standard errors when 50% of the data are MAR and MCAR for unweighted data sets.	125
5.14	Model 2: Plot of bias and standard errors when 50% of the data are MAR or MCAR for weighted data sets.	125
5.15	Model 2.2: Plot of bias and standard errors of the traditional method use category when 50% data are MAR and MCAR for unweighted data sets.	135
5.16	Model 2.2: Plot of bias and standard errors of the traditional method use category when 50% data are MAR and MCAR for weighted data sets.	135

5.17	Model 2.2: Plot of bias and standard errors of the modern method use category when 50% data are MAR and MCAR for unweighted data sets.	136
5.18	Model 2.2: Plot of bias and standard errors of the modern method use category when 50% data are MAR and MCAR for weighted data sets.	136
6.1	Model 1.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	158
6.2	Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	159
6.3	Model 1.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	160
6.4	Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	161
6.5	Model 1.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	162
6.6	Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	163
6.7	Model 1.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	164
6.8	Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	165
6.9	Model 1.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption under MAR assumption.	166
6.10	Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	167
6.11	Model 1.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	168
6.12	Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	169
6.13	Model 1.2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	170

6.14	Model 1.2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	171
6.15	Model 1.2.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	172
6.16	Model 1.2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.	173
6.17	Model 2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	174
6.18	Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	175
6.19	Model 2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	176
6.20	Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	177
6.21	Model 2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	178
6.22	Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	179
6.23	Model 2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	180
6.24	Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.	181
6.25	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	213
6.26	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	213
6.27	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	214
6.28	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	214

6.29	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	215
6.30	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	215
6.31	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	216
6.32	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	216
6.33	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	217
6.34	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	217
6.35	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	218
6.36	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	218
6.37	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	219
6.38	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	219
6.39	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	220
6.40	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	220
6.41	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	221
6.42	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set	221
6.43	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set	222

6.44	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set	222
6.45	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	223
6.46	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set	223
6.47	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set	224
6.48	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set	224
6.49	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	225
6.50	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	225
6.51	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	226
6.52	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set	226
6.53	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	227
6.54	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	227
6.55	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	228
6.56	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set	228
6.57	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	229
6.58	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	229

6.59	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	230
6.60	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	230
6.61	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	231
6.62	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	231
6.63	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	232
6.64	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	232
6.65	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	233
6.66	Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	233
6.67	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	234
6.68	Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	234
6.69	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	235
6.70	Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	235
6.71	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	236
6.72	Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	236
6.73	Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	237

6.74	Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	237
6.75	Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	238
6.76	Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	238
6.77	Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	239
6.78	Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	239
6.79	Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	240
6.80	Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	240
6.81	Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	241
6.82	Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	241
6.83	Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	242
6.84	Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	242
6.85	Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	243
6.86	Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	243
6.87	Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	244
6.88	Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	244

6.89	Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	245
6.90	Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	245
6.91	Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	246
6.92	Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	246
6.93	Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	247
6.94	Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	247
6.95	Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.	248
6.96	Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.	248
6.97	Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set	249
6.98	Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	249
6.99	Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	250
6.100	Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	250
6.101	Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set	251
6.102	Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	251
6.103	Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.	252

6.104	Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.	252
6.105	Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	253
6.106	Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	253
6.107	Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	254
6.108	Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	254
6.109	Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	255
6.110	Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	255
6.111	Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	256
6.112	Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	256
6.113	Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	257
6.114	Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	257

6.115	Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	258
6.116	Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	258
6.117	Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	259
6.118	Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	259
6.119	Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	260
6.120	Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	260
6.121	Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	261
6.122	Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	261
6.123	Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	262
6.124	Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	262
6.125	Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	263

6.126	Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	263
6.127	Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.	264
6.128	Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.	264
6.129	Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	265
6.130	Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	265
6.131	Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	266
6.132	Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	266
6.133	Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	267
6.134	Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	267
6.135	Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.	268
6.136	Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.	268

List of Tables

3.1	Imputation of categorical variables with more than two levels	47
3.2	Parameter estimates of a set of logistic regression models for predicting the contraceptive methods use status by women of reproductive age in Democratic Republic of Congo in 2007, using age (1 st covariate) and education (2 nd covariate) in years as explanatory variables	56
4.1	Description of the variables used in the study	64
5.1	Model 1.1: Frequency distribution of missingness on marital status under MAR assumption.	79
5.2	Model 1.1: Distribution of missingness across selected categorical variables when 50% data are MAR on marital status if a woman is not using any contraceptive method.	80
5.3	Model 1.1: Frequency distribution of missingness when 50% data are MCAR on marital status.	81
5.4	Model 1.1: Distribution of missingness by selected categorical variables when 50% data are MCAR on marital status.	82
5.5	Model 1.1: Estimates of bias when approximately 50% of data are MAR on marital status if a woman is not using any contraceptive method.	83
5.6	Model 1.1: Estimates of standard errors when approximately 50% of data are MAR on marital status if a woman is not using any contraceptive method.	84
5.7	Model 1.1: Estimates of bias when approximately 50% of data are MCAR on marital status.	84
5.8	Model 1.1: Estimates of standard errors when approximately 50% of data are MCAR on marital status.	84
5.9	Model 1.1: Frequency distribution of missingness when approximately 50% data are MAR on marital status if a woman is not using any contraceptive method.	87
5.10	Model 1.1: Distribution of missingness across selected categorical variables when approximately 30% data are MAR on marital status if a woman is not using any contraceptive method.	88
5.11	Model 1.1: Frequency distribution of missingness when approximately 30% data are MCAR on marital status.	88
5.12	Model 1.1: Distribution of missingness by selected categorical variables when approximately 30% data are MCAR on marital status.	89

5.13	Model 1.1: Estimates of bias when approximately 30% of data are MAR on marital status if a woman is not using any contraceptive method.	90
5.14	Model 1.1: Estimates of standard errors when approximately 30% of data are MAR on marital status if a woman is not using any contraceptive method.	91
5.15	Model 1.1: Estimates of bias when approximately 30% of data are MCAR on marital status.	91
5.16	Model 1.1: Estimates of standard errors when approximately 30% of data are MCAR on marital status.	91
5.17	Model 1.1: Frequency distribution of missingness when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.	94
5.18	Model 1: Distribution of missingness across selected categorical variables when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.	95
5.19	Model 1.1: Frequency distribution of missingness when approximately 10% of the data are MCAR on marital status.	95
5.20	Model 1: Distribution of missingness by selected categorical variables when approximately 10% of the data are MCAR on marital status.	96
5.21	Model 1.1: Estimates of bias when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.	97
5.22	Model 1.1: Estimates of standard errors when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.	98
5.23	Model 1.1: Estimates of bias when approximately 10% of the data are MCAR on marital status.	98
5.24	Model 1.1: Estimates of standard errors when approximately 10% of the data are MCAR on marital status.	98
5.25	Model 1.2.1: Frequency distribution of missingness when 50% of the data are MAR on marital status and region if a woman is not using any contraceptive method.	101
5.26	Model 1.2.1: Distribution of missingness across selected categorical when 50% of the data are MAR on marital status and region if a woman is not using any contraceptive method.	102
5.27	Model 1.2.1: Frequency distribution of missingness when 50% of the data are MCAR on marital status and region.	102
5.28	Model 1.2.1: Distribution of missingness when 50% of the data are MCAR on marital status and region.	103
5.29	Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the unweighted data set.	104

5.30	Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the weighted data.	105
5.31	Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the unweighted data set.	105
5.32	Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the weighted data set.	106
5.33	Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the unweighted data set.	110
5.34	Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the weighted data set.	111
5.35	Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the unweighted data set.	112
5.36	Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the weighted data set.	113
5.37	Model 2.1: Frequency distribution of missingness when 50% of the data are MAR on contraceptive method use status if a woman is aged at least 35 years.	119
5.38	Model 2.1: Distribution of missingness across categorical variables when 50% of the data are MAR on contraceptive method use status if a woman is aged at least 35 years.	120
5.39	Model 2.1: Frequency distribution of missingness when 50% of the data are MCAR on contraceptive method use status.	121
5.40	Model 2.1: Distribution of missingness across marital status when 50% of the data are MCAR on contraceptive method use status.	122
5.41	Model 2.1: Estimates of bias when approximately 50% of the data are MAR on contraceptive method use status if a woman is aged at least 35 years.	123
5.42	Model 2.1: Estimates of standard errors when approximately 50% of the data are MAR on contraceptive method use status if a woman is aged 35 years or more.	123
5.43	Model 2.1: Estimates of bias when approximately 50% of the data are MCAR on contraceptive method use status.	124
5.44	Model 2.1: Estimates of standard errors when approximately 50% of the data are MCAR on contraceptive method use status.	124

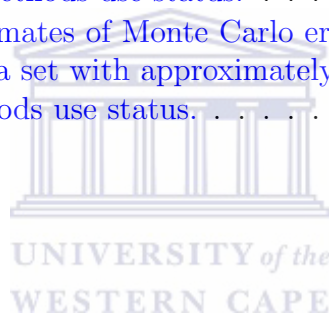
5.45	Model 2.2: Frequency distribution of missingness when 50% of the data are MAR on contraceptive method use status if a woman is at least 35 years old.	127
5.46	Model 2.2: Distribution of missingness by marital status when approximately 50% of the data are MAR on contraceptive method use status if a woman is at least 35 years old.	128
5.47	Model 2.2: Frequency distribution of missingness when 50% of the data are MCAR on contraceptive method use status	128
5.48	Model 2.2: Distribution of missingness by selected categorical variables when 50% data are MCAR on contraceptive methods use status.	129
5.49	Model 2.2: Estimates of bias in the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MAR on contraceptive method use status if a woman is aged at least 35 years.	131
5.50	Model 2.2: Estimates of standard errors of the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MAR on contraceptive method use status if a woman is aged at least 35 years.	132
5.51	Model 2.2: Estimates of bias in the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MCAR on contraceptive method use status.	133
5.52	Model 2.2: Estimates of standard errors of the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MCAR on contraceptive method use status.	134
6.1	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.	183
6.2	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.	183
6.3	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.	184
6.4	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.	184
6.5	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on marital status.	184
6.6	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on marital status.	185

6.7	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on marital status.	185
6.8	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on marital status.	185
6.9	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.	186
6.10	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.	186
6.11	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.	186
6.12	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.	187
6.13	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 30% MCAR data on marital status.	187
6.14	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 30% MCAR data on marital status.	187
6.15	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 30% MCAR data on marital status.	188
6.16	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 30% MCAR data on marital status.	188
6.17	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.	189
6.18	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.	189
6.19	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.	189
6.20	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.	190
6.21	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 10% MCAR data on marital status.	190

6.22	Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 10% MCAR data on marital status.	190
6.23	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 10% MCAR data on marital status.	191
6.24	Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 10% MCAR data on marital status.	191
6.25	Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.	192
6.26	Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.	193
6.27	Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.	193
6.28	Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.	194
6.29	Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on marital status and region.	194
6.30	Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on marital status and region.	195
6.31	Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on marital status and region.	195
6.32	Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on marital status and region.	196
6.33	Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% of data missing at random on marital status and region if a woman is not using any contraceptive method.	197
6.34	Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% of data missing at random on marital status and region if a woman is not using any contraceptive method.	198

6.35	Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on marital status and region is a woman is not using any contraceptive method.	199
6.36	Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on marital status and region is a woman is not using any contraceptive method.	200
6.37	Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on marital status and region.	201
6.38	Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on marital status and region.	202
6.39	Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on marital status and region.	203
6.40	Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on marital status and region.	204
6.41	Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	205
6.42	Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	205
6.43	Model 2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	206
6.44	Model 2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	206
6.45	Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on contraceptive method use status.	206
6.46	Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on contraceptive method use status.	207
6.47	Model 2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on contraceptive method use status.	207
6.48	Model 2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR on contraceptive method use status.	207
6.49	Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	208

6.50	Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	208
6.51	Model 2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	209
6.52	Model 2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.	209
6.53	Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on contraceptive method use status.	210
6.54	Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on contraceptive method use status.	210
6.55	Model 2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on contraceptive methods use status.	211
6.56	Model 2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on contraceptive methods use status.	211



Abbreviations

MCAR	Missing completely at random
MAR	Missing at random
NMAR	Not missing at random
MVN	Multivariate normal
MVNI	Multivariate normal imputation
MICE	Multiple imputation by chained equations
MCMC	Markov Chain Monte Carlo
MCE	Monte Carlo error
FCS	Fully conditional specification
CD	Case deletion
BD	Baseline data
EM	Expectation maximisation
ML	Maximum likelihood
DHS	Demographic health survey
WLF	Worst linear function
SE	Standard error
PV	P-value

Chapter 1

Introduction

1.1 Background on missing data

In this chapter, the fundamental reasoning behind sampling and nonresponse is discussed. Furthermore, missing data and mechanisms that generate them are reviewed. The motivation of the study, research questions and objectives, hypotheses, research design and thesis overview are presented.

1.1.1 Sampling and nonresponse

Researchers are constantly faced with problems of limited resources and time to take measurements on populations of interest. They are often bound by circumstances to measure some population units or samples using various techniques commonly known as sampling methods. The information resulting from these methods or procedures are analysed and the results are extrapolated to the whole population. Probability sampling, a method that takes into account the variability among items when selecting samples, is one of these techniques. It reduces the risk of a distorted view of the population and allows valid statistical inferences to be made.

A host of scholars provide a comprehensive review of sampling methods (Cochran, 1977; Kalton, 1983; Kish, 1965). These assessments reveal a number of errors identified with statistics based on sample survey estimates. These errors including biased estimates and large standard errors amongst others (Bethlehem,

2009) commonly known as total error, have an impact on survey estimates. Since survey estimates are never equal to the population parameters, by implication errors are involved in these estimates. The causes of such errors are numerous. A classification of the possible causes of these errors is shown in Figure 1.1 as suggested by Kish (1965) and Bethlehem (2009). The figure indicates that the total error is due to sampling and nonsampling errors.

Sampling errors arise when a researcher surveys only a subset of the population of interest instead of completely enumerating the whole population. These errors can disappear if and only if the whole population is enumerated. Sampling errors can be split into two categories: (1) selection errors, which denote effects resulting from the use of probability samples; and (2) estimation errors, which occur when the wrong selection probabilities are used to compute estimators (Bethlehem, 2009).

Nonsampling errors occur at anytime, even when the whole population is surveyed. According to Lessler and Kalsbeek (1992), nonsampling errors arise mainly during the data capturing process. They can be divided into four categories, namely the frame error, measurement error, processing error and nonresponse error.

The frame errors are errors that result from the divergences or differences between the frame and actual population. Such types of errors are observed when for instance units that are not part of the population of interest are sampled. The former situation is referred to as overcoverage and the latter as undercoverage. The measurement error refers to the difference between the reported value from the sample and the true value of the population of interest. High rates of nonresponses are an indication of such type of errors. The processing error arises when data are being processed (coded, weighted, etc) after data collection. On the other hand, the nonresponse error occurs when the entire data collection fails for reasons such as respondents not being at home, refusing to participate, amongst others, or when only partial data are available. In other words the respondents participate but do not respond to all individual items. Examples of such situations are mostly found for example in household surveys where respondents tend to refuse to answer questions on income, in health surveys where questions on sexual behaviour are not fully answered and in opinion surveys where respondents fail to express their choice or preference of individuals over others (Little & Rubin, 2002). The former situation is referred to as the *unit nonresponse* and latter as the *item nonresponse*. These two types of nonresponses are the only source of missing data. Throughout

this study, only *item nonresponse* is considered.

Kish (1965) and Bethlehem (2009) on the other hand, split nonsampling errors into two categories: observation and nonobservation errors. Observation errors cover the overcoverage and measurement errors previously stated. Processing errors occur when data are being processed, such as during the data capturing process. Nonobservation errors occur due to the fact that estimations that the researcher planned to perform can no longer be done. They include both undercoverage and nonresponse errors described earlier.

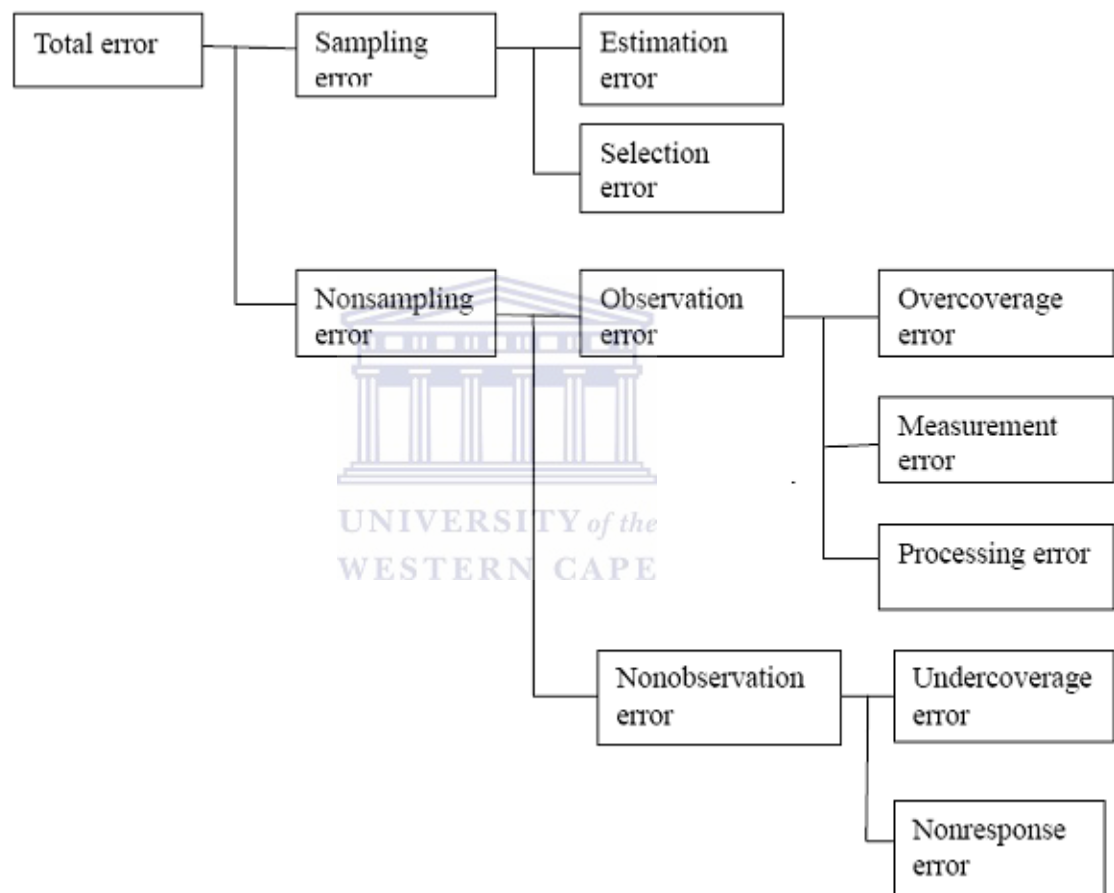


FIGURE 1.1: Classification of survey errors (Bethlehem, 2009, page 180)

1.1.2 Missing data

Missing data are common and a major problem in different fields of research, such as in operation management (Tsiriktsis, 2005), psychology (Graham, 2009) and epidemiology (Cattle et al., 2011) amongst others. Missing data are not always

given much attention by some researchers especially those who are not methodologists or statistical experts. This is due mainly to the lack of familiarity with the existing statistical literature on missing values or the ignorance of the impact that they can have on statistical inferences (McKnight et al., 2007). A traditional way of dealing with missing data is to eliminate them from the analysis, a strategy that is provided by default in most of the statistical software packages such as SPSS, SAS and STATA. This may substantially affect data analysis especially when dealing with chunks of missing data. Discarding missing observations from analysis reduces the sample size and as a result, a sample that is not representative of the population is obtained, leading to a lower power of the statistical test, biased parameter estimates and large standard errors, especially when the proportion of missing data is high (De Leeuw et al., 2003; Enders, 2010; McKnight et al., 2007).

Most researchers are not always willing to discard data that they spend a great amount of money and time on, they often try to find ways of rescuing missing data in order to make valid inferences. On this view, various methods have been developed to handle missing data (Little & Rubin, 2002; Schafer & Graham, 2002; Tsikriktsis, 2005). They are discussed in Chapter 2 of this study. The primary goal of these methods is to obtain valid and efficient statistical inferences about the population of interest but not to recover missing data or to obtain what would have been obtained if data were complete (Schafer & Graham, 2002).

The seriousness of the problems caused by missing data depends on amongst other things the amount of missing data, although there is no stated rule concerning how much is too much missing data. According to Cohen (1983), missing data are considered to be small if 5 to 10% of the data are missing and high when at least 40% of the data are missing (Raymond & Roberts, 1987). The degree of missingness impacts negatively on the data analysis when missing values are excluded from the analysis (when case deletion is used). Monte Carlo simulation studies have shown for instance that if 2% of the data are missing at random and a researcher deletes entire cases with missing data (which can result in up to 18.3% loss of the total data set) (Kim & Curry, 1977), this can affect the statistical power and lead to incorrect results and conclusions.

In general, no matter how much the degree of missingness is, problems associated with missing data will always arise. The only way to completely remedy these problems is to avoid missing data in data sets, which can be done during the data collection process or before, otherwise whatever approach used to handle missing data will be to reduce bias or any other associated problem.

1.1.3 Missing data patterns

A missing data pattern refers to the configuration or classification of observed missing data values that describes the location of holes in data (Baraldi & Enders, 2010). Six types of missing data patterns can be distinguished: the univariate pattern, unit nonresponse pattern, monotone missing data pattern, general pattern, planned missing pattern and latent variable pattern. A detailed description of these patterns is given by Little and Rubin (2002), Schafer and Graham (2002) and Baraldi and Enders (2010).

To understand missing data patterns, we consider an example in Figure 1.2 that was proposed by Baraldi and Enders (2010). A missing data pattern containing missing values that are isolated to a single variable is called a univariate pattern (panel A) which occurs mostly in experimental studies. A unit nonresponse (panel B) occurs in surveys, when one or two respondents refuse to answer a particular questionnaire. A monotone missing data pattern (panel C) occurs in longitudinal studies when participants drop out and never come back. Monotone missing data look like a staircase in such a way that cases with missing data on a particular consideration are always missing in successive measurements. The general missing pattern (panel D) is a pattern that has missing data that are disseminated all the way through the data matrix in a haphazard way. It is the most common configuration of missing data (Baraldi & Enders, 2010).

The planned missing data pattern corresponds to the three-form questionnaire design that is used to distribute questionnaires across different forms and administer a subset of the forms to each respondent as suggested by (Graham et al., 1996). For example, in Panel E, four questionnaires are distributed across three forms in such a way that Y_1 is included in each form but Y_2 to Y_4 are missing. This type of missing data pattern is very helpful when a researcher wants to collect a large amount of data and reduce the respondent burden. Lastly, a latent variable pattern (panel F) is a pattern for which the values of the latent variables are missing for the entire sample. From a practical point of view, distinguishing among missing data patterns is no longer regarded as important because newly developed missing data techniques such as the Maximum Likelihood and Multiple Imputation are not sensitive to missing data patterns (Baraldi & Enders, 2010). In this thesis, missing patterns are ignored since the missing observations methods that will be used are based on the multiple imputation approach.

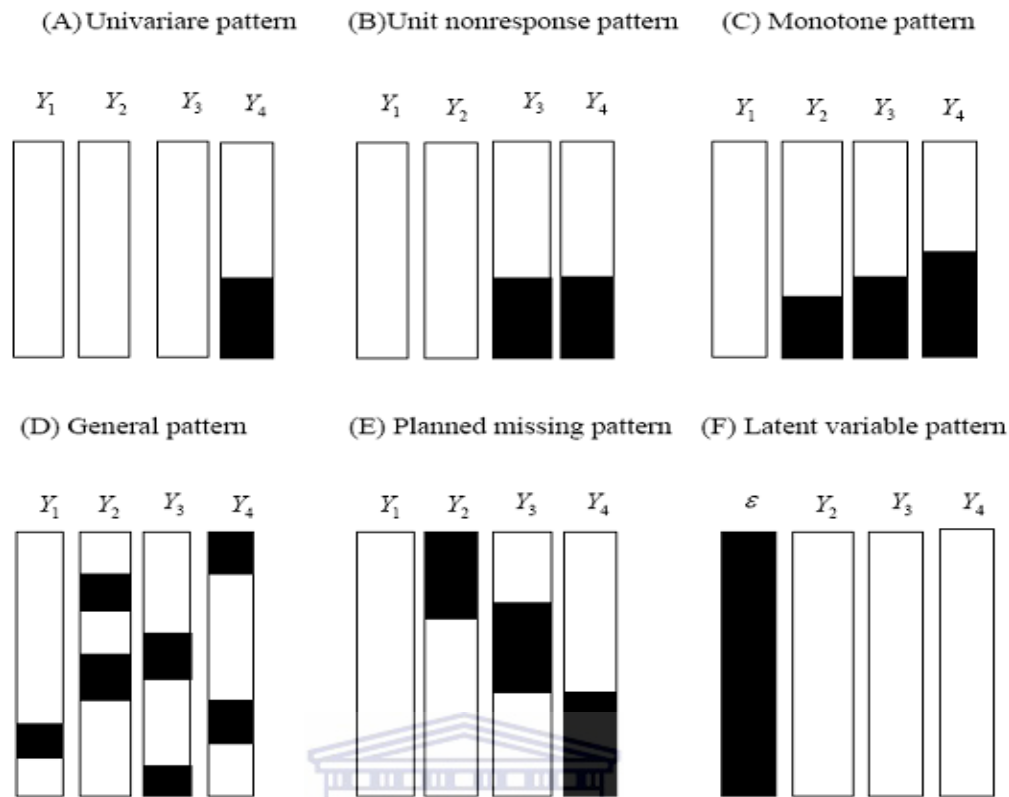


Figure 2: Types of missing data patterns (Enders, 2010).

FIGURE 1.2: Types of missing data patterns (Enders, 2010, page 4). The shaded areas symbolize the missing data

1.1.4 Missingness mechanisms

Statisticians and methodologists classify missing data into three categories: (1) Missing Completely at Random (MCAR), (2) Missing at Random (MAR) and (3) Missing Not at Random (MNAR) (Schafer & Olsen, 1998; Schafer & Yucel, 2002). This classification is termed *missing mechanism* and refers to the possible relationship between variables in the data set and the probability of missing data or missingness (missing or not missing). The classification was initially introduced by Rubin (1976) and it is currently used to facilitate communication, diagnose and identify the proper techniques for handling missing data (McKnight et al., 2007).

Data are MCAR if the probability that a particular value is missing is not related to the value itself or any other observed values in the data set. An example of such case is, for instance, in health surveys where subjects are randomly selected to undergo more extensive physical examination. When the probability of a particular value being missing depends on observed values in the data set, the

missing mechanism is referred to as MAR. This happens for example in surveys when more women tend to answer some questions than men.

These two mechanisms are termed ignorable, because conditional on the observed data set, one can draw valid inferences without explicitly modelling the missing mechanism. If the missing information depends on unobserved values, the missing mechanism is called non-ignorable. In this case, even conditioning on observed data does not yield valid inferences. The missing data mechanism needs to be modelled to estimate the parameter vectors (Allison, 2002). Data with such missing mechanism is known as NMAR (Schafer & Graham, 2002). Income is an example of such a case. People with very low income tend to answer questions about their income differently when compared to people with very high income.

In terms of probability, the above assumptions can be explained as follows: Let $Y = (Y_{obs}, Y_{mis})$ be a partition of the dataset Y in an observed part, Y_{obs} and a missing part, Y_{mis} . Let also R_1, R_2, \dots, R_N represent response indicators that indicate which survey items are missing and which are not. In the case of item nonresponse, R_i is a binary variable indicating for each sample element i whether survey items are observed ($R_i = 1$) or missing ($R_i = 0$). The distribution of the missingness is characterized by the conditional distribution of R which is given by:

$$P(R|Y) = P(R|Y_{obs}, Y_{mis}, \phi) \quad (1.1)$$

where P is a general symbol for a probability distribution and ϕ is a parameter or a set of parameters that describes the relationship between missingness (R) and the data. This form of missing mechanism is referred to as MNAR and it says that the probability that R takes on values 1 (observed) or 0 (missing) can depend on both Y_{obs} and Y_{mis} via some parameter or set of parameters ϕ . The data are said to be MAR when the following equation holds:

$$P(R|Y) = P(R|Y_{obs}, \phi). \quad (1.2)$$

This indicates that the conditional probabilities of missingness depend on the observed portion of data via some parameter or a set of parameters that relate Y_{obs} to R . When the conditional probabilities of missingness do not depend on the data at all, the data are said to be MCAR. This can be mathematically written as:

$$P(R|Y) = P(R|\phi). \quad (1.3)$$

A graphical representation of Rubin's (1976) missing data mechanisms is provided in Figure 1.3. Assuming that X represents variables that are completely observed, Y denotes variables that are partly missing, Z represents the component of causes of missingness not related to X and Y , and R is the missingness (missing or not). As Figure 1.3 indicates, MCAR requires that the causes of missingness be entirely contained within the unrelated part Z . MAR allows some causes of missingness to be related to X , whereas MNAR allows the causes of missingness to be related to Y after a relationship between X and Y is considered (Enders, 2010; Schafer & Graham, 2002).

Understanding this classification can help in choosing the appropriate methods for handling missing data. If not modelled or fixed, all these missingness mechanisms may lead to serious consequences. Discarding cases with missing data from the analysis for instance, may lead to efficiency or greater variability in the obtained results. Not modelling MAR and NMAR data leads to bias and efficiency problems. When modelling MCAR and MAR data to look like non-missing data, observed data are used to impute or fill in missing values. As a result, bias and efficiency problems are reduced.

A number of methods to model MAR and MCAR data have been developed. These include single-based imputation methods such as the mean imputation, regression imputation, interpolation (for panel data), multiple imputation such as the multivariate normal imputation (Schafer, 1997) and the multiple imputation by chained equations (Raghunathan et al., 2001; Van Buuren, 2007). These methods are discussed in detail in Chapter 3 of this thesis.

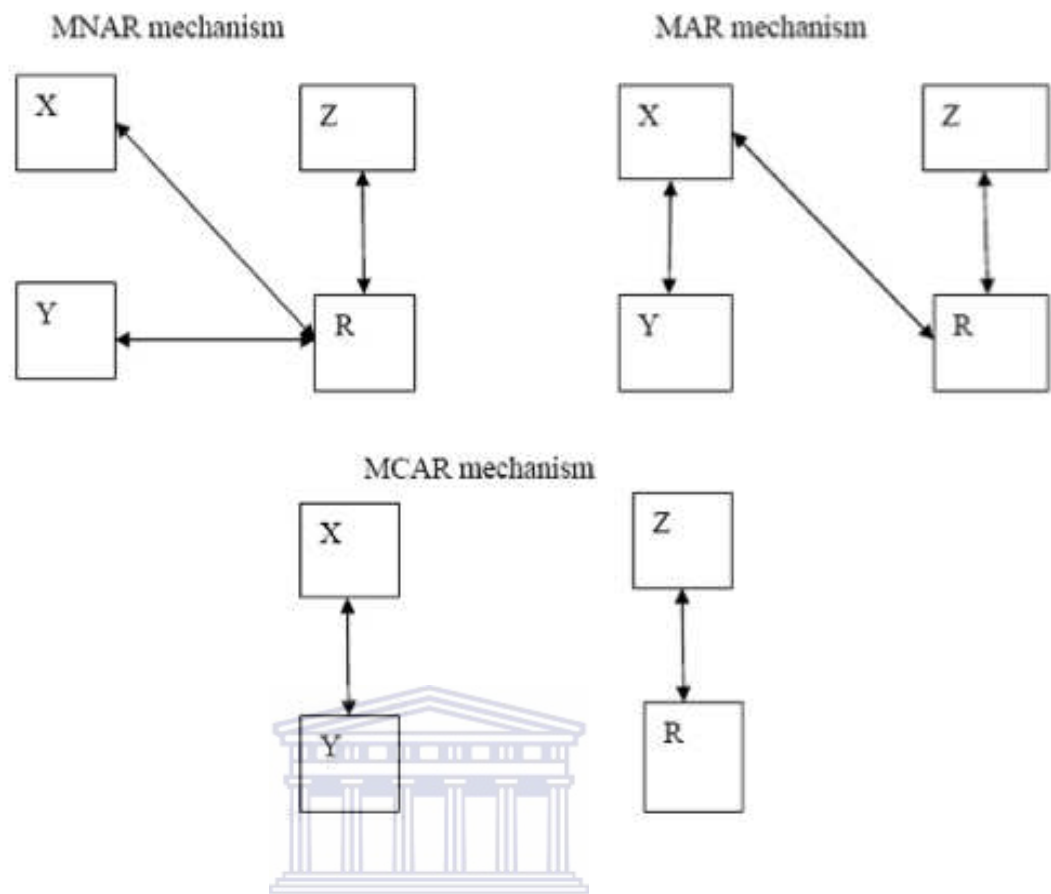


FIGURE 1.3: Types of missing data patterns (Enders, 2010, page 12)

1.1.5 Testing for missingness mechanisms

Before handling missing data, researchers need to know why data are missing (Carpenter & Kenward, 2012; Little & Rubin, 1989). There are many reasons that cause missing data. These include amongst others; data entry errors, failure to complete the entire questionnaire which occur at random, and reasons such as refusal to respond to certain questions such as income level, etc, which are not random responses. It is thus important to establish the mechanism to use and how item nonresponse should be treated in the statistical analysis (De Leeuw et al., 2003). Missing data mechanisms need to be random (MCAR or MAR); otherwise there is no statistical means to simplify the problem (Tsikriktsis, 2005). By implication all the methods that are used to deal with missing data should assume that the pattern of data loss is random (either MCAR or MAR). Identifying the underlying missing mechanism is very important because this influences how missing data will be handled. MCAR data have been found less likely to introduce

serious bias, regardless of the methods used to deal with missing data (Graham, 2009; Musil et al., 2002), whereas NMAR data remain the most difficult to identify and handle because true values of missing values are not known (Little & Rubin, 2002).

Testing for the missing mechanism is equivalent to testing for randomness of missing data. Several methods have been developed for this matter (Baraldi & Enders, 2010; McKnight et al., 2007; Schafer & Graham, 2002). Previous work has confirmed that the MCAR mechanism is the only testable mechanism (Enders, 2010; Schafer, 1997; Schafer & Graham, 2002). The traditional way of diagnosing the MCAR mechanism is to use t-tests to assess whether missing data are MCAR when one or few data are missing (McKnight et al., 2007). Normally these tests consist of creating dummy codes of missing variables and two groups for each variable of interest: those missing and those with complete data. Then t-tests are conducted to compare the means of each of the two groups on some or all of the remaining variables in the data set to see if there is a difference between variables with or without missing data in the data set. A statistically significant difference between the two groups indicates the departure from the MCAR mechanism. However, when there is a large number of variables in the data set, this test becomes problematic; the analyst will have to conduct a maximum number of t-tests equal to the number of variables in the data set, say q minus one ($q - 1$), with missing data on which the two groups are based. This can lead to a large alpha and Type 1 error, which consists of incorrectly rejecting the null hypothesis that there is no difference between the variables with missing and no missing data, in favour of the alternative hypothesis that the variables differ (Enders, 2010; McKnight et al., 2007).

To avoid this problem, Little (1988) has proposed a method based on a chi-square distributed variable to test for MCAR in large data sets. The observed variable means or averages for each pattern of missing data are compared with expected population means for which an overall weighted squared deviation is computed. If data are MCAR, each subsample meeting the requirements of a specific pattern of missing data will yield the same mean for each variable as the variables computed for the entire data set using any robust method for parameter estimation. The aim here is generally to compare what would appear to be missing at random and what is observed. If data indicate a departure from the completely random process (i.e., missing and non-missing cases differ for all observed data), then the chi-square test becomes statistically significant, indicating

that data are not MCAR. This test works appropriately for continuous variables but can produce biased estimates for categorical variables. Alternatively, Allison (2005) proposes running a logistic regression of R_i on a set of explanatory variables to test for the MCAR mechanism in data sets. Significant coefficients will indicate a departure from MCAR, which suggest the MAR or MNAR mechanism.

On the other hand, researchers investigate missing mechanisms by means of the test for correlation between missingness and other variables in the dataset. Low correlation coefficients will reflect more randomness or MCAR data, whereas high correlation coefficients will indicate nonrandomness which is associated with MAR (Musil et al., 2002). A detailed description of some of the methods for testing for the MCAR mechanism in data sets is given amongst others by Kim and Bentler (2002) and Chen and Little (1999).

When the MCAR mechanism test is rejected in the data set, MAR or MNAR are assumed. These two mechanisms have no statistical procedure, either numeric or graphic, to detect them (Little & Rubin, 2002; Schafer & Graham, 2002). Researchers rely on the logic and sound understanding of the study design and domain in order to decide whether data are MAR or MNAR (McKnight et al., 2007; Schafer, 1997). To determine whether data are MAR for instance, the researcher needs to look at the sources of the data outside their studies (eg., previous findings) or to the follow-up with respondents, or if double sampling was done during the study (McKnight et al., 2007). Data are assumed to be MNAR if for instance there is no follow-up to understand the sources of missingness in sample surveys or if observations in the data set were missed unintentionally. Alternatively, when the prevention of data from missing is beyond the researcher's control, the MNAR mechanism is assumed (Enders, 2010; McKnight et al., 2007; Schafer, 1997).

1.2 Motivation for the study

Missing data are common in survey research. Enrolled subjects do not often have data recorded for all variables of interest. As previously stated, this may be due to data entry errors or refusal by the respondents to answer some items from a survey. As a result, missing values are created in data sets, and if they are not modelled properly, it can lead to incorrect inferences. The common way of handling missing values is to discard them from the analysis. This approach is referred to as case deletion or complete case analysis and can lead to low power of the statistical test

and biased parameter estimates when the proportion of missing values is high and data are missing in a systematic manner or at random (Graham, 2009).

To reduce these problems, various methods of rescuing missing data have been developed (Graham, 2009; Schafer & Graham, 2002; Tsiriktsis, 2005). Items with no observations at all are directly discarded from the analysis because they do not provide any particular information about the data. However, if data are partially missing on variables of interest, the latter should not be discarded as they still contain some information that can be used to draw useful inferences.

Estimating a model without doing any kind of processing when data are missing is difficult. For example, if a linear regression has to be run, say Y as a function of X_1 and X_2 , but some of the values of X_1 and X_2 are missing, it is still possible to fit regression coefficients to the independent variables. One way of doing this is to get rid of the missing information and use the available data, which is sometimes problematic. But when the researcher is forced to use the data set with missing data without discarding cases, it has to be done in a way that minimizes the damage in the inferences to be drawn.

The first thing to do is to identify the missingness mechanisms in the data set or the reasons why data are missing. These include the MCAR, MAR and NMAR mechanisms. Data are MCAR if the probability that a particular value is missing is not related to the value itself or any other observed values in the data set. When the probability that a particular value is missing depends on observed values in the data set, the missing mechanism is referred to as MAR data. These two mechanisms are termed ignorable as mentioned before, because conditional on the observed data set, one can draw valid inferences. If missingness is related to unobserved values in the data set, the missing mechanism is called non-ignorable. In this case, even conditioning on observed data does not lead to valid inferences. Data sets with such missing mechanism is known as not missing at random or NMAR (Graham, 2009; Schafer & Graham, 2002).

If not fixed, all these missingness mechanisms may lead to serious consequences. Discarding cases with missing data from the analysis for instance, leads to inefficiency or greater variability in the obtained results. Not modelling MAR and NMAR data lead to bias and efficiency problems. Modelling MCAR and MAR data to look like non-missing data, observed data are used to impute missing values. As a result, bias and efficiency problems are reduced.

A number of methods have been developed to model MAR and MCAR data. These include single-based imputation methods such as the mean imputation,

regression imputation, interpolation (for panel data), multiple imputation based methods such as the multivariate normal imputation (MVNI) and the multiple imputation by chained equations (MICE) (Raghunathan et al., 2001; Van Buuren, 2007) also known as imputation by fully conditional specification (Van Buuren et al., 1999), conditional model (Carpenter & Kenward, 2012) or sequential regression multiple imputation (Raghunathan et al., 2001; Van Buuren, 2007).

These last two multiple imputation-based methods are increasingly being used and have been made popular in almost all the main statistical software packages such as SAS, STATA and R. They are considered the best as they account for the statistical uncertainty in the imputations, which is not the case when single-based imputation methods are used (Lee & Carlin, 2010). The description of these methods is provided in Chapter 3.

Despite the popularity of MVNI and MICE, there is still no clear guidance on which method to choose between the two when the multiple imputation needs to be done on continuous, binary and categorical (polytomous with more than two categories) variables containing missing values. It is against this backdrop that this study attempted to explore mainly the performance of these methods when data are missing at random or missing completely at random on unordered or nominal variables treated as predictors in regression models rather than outcome variables that were explored by (Kropko et al., 2014). As the performance of these two methods is still ongoing research in different fields, researchers always recommend the use of other data sets to compare the obtained results with the previous ones (Kropko et al., 2014). In this regard, this study considered also the case where missing values are observed on the outcome variables (binary and polytomous variables).

The ignorability of missing data was assumed throughout this study. That is, based on the available data, estimates of missing observations were obtained. Therefore, the MCAR and MAR assumptions were only considered. This was done to investigate whether or not the missingness mechanisms have an impact on the performance of MVNI and MICE techniques. Thus, simulated data sets with missing values at random or completely at random on the variables of interest were used to assess the performance of these methods.

This study also used a 2007 Democratic Republic of Congo Demographic Health Survey (DHS) data set, which is a complex survey with a complex sampling design and weighting procedure that need to be taken into consideration during the analysis. Various studies have demonstrated that when survey data

sets contain weight variables, weighted results are preferred as they produce less bias in the estimates than unweighted results (Korn & Graubard, 1995). This issue is also addressed by Reiter et al. (2006), Schenker et al. (2006), He et al. (2009) and Molenberghs et al. (2014) amongst others. Therefore, the results of this study were based on both the regular data sets (without taking into account the randomization distribution due to the sample selection procedure) and the weighted data sets to investigate whether the performance of MVNI and MICE may be influenced by this issue as data analysts have to always deal with both weighted and unweighted data sets.

1.3 Significance of the study

Non-ordered categorical missing data are common in survey data sets and the inadequate handling of such data may lead to incorrect results and conclusions. Hence, understanding and mastering how missing data should be treated is important to any researcher or survey data set user. This study is significant to academics, researchers and other users of survey data sets in filling gaps in knowledge and understanding of how non-ordered categorical missing data should be handled in order to obtain unbiased statistical estimates from incomplete data, thereby leading to valid inferences.

1.4 Research objectives

The primary objective of this study was to determine the performance of MVNI and MICE methods when data are missing at random or missing completely at random on unordered categorical variables treated as predictors in the regression models. As the performance of these two methods is still ongoing research, this study explored also their performance in the situation where missing values were found on the outcome variables, either binary (with two outcomes) or polytomous (more than two outcomes). Although this case has partly been explored before (Kropko et al., 2014), the findings from this study could strengthen existing knowledge about these methods as the author used only one data set and suggested (in the study limitation section) that other data sets should be used to look at the

performance of these two methods.

Other specific objectives were:

1. To review the literature on MVNI and MICE methods and illustrate their performance when data are missing on continuous variables.
2. To show that as expected, multiple imputation of interest produce less biased estimates than the case deletion which discard missing values from the analysis.
3. To investigate whether the rates of missing values in the data sets can impact on the performance of the multiple imputation methods of interest, namely MVNI and MICE.
4. To determine whether the sample design can impact on the performance of MVNI and MICE.
5. To draw relevant conclusions on how specifically non-ordered or nominal categorical data containing MCAR or MAR data should be imputed under different circumstances, especially when missing values are present on the outcome or predictor variables in the regression models.

1.5 Research questions

The study thought to answer the following questions about non-ordered or nominal categorical data:

1. What is the performance (in terms of bias and standard errors) of MVNI and MICE methods when data are MCAR or MAR on non-ordered categorical variables with more than two levels or categories treated as predictors in regression models?
2. What is the performance (in terms of bias and standard errors) of MVNI and MICE methods when data are MCAR or MAR on non-ordered categorical variables with three or more levels treated as outcome or response variables in the regression models?

3. What is the performance (in terms of bias and standard errors) of MVNI and MICE methods when data are MCAR or MAR on non-ordered dichotomous variables treated as outcome or response variables in the regression models?

1.6 Hypotheses

The following hypotheses were tested:

1. The MVNI method which assumes a normal distribution for the variables in the imputation model and MICE which fills in missing values taking into consideration the distributional form of the variables with missing values, yield similar parameter estimates for specifically non-ordered categorical variables containing missing data, which are treated either as predictors or outcome variables in the regression models.
2. The performance of MVNI and MICE is not affected by the survey design.
3. Missing data mechanisms (MAR and MCAR) have no impact on the performance of MVNI and MICE.

1.7 Research Design

The research design used in this thesis conforms to a quantitative paradigm. This is chosen because of its ideologies and compatibility with numerical data, which is relevant for addressing the thesis research questions and hypotheses.

1.8 Thesis overview

This thesis is structured as follows. Chapter 1 provides a general introduction and background of the study including the introduction to missing data and mechanisms that generate them. The motivation, significance of the study, research objectives and questions, hypotheses, research design and motivation underlying the study are highlighted. In Chapter 2, The Markov Chain Monte Carlo (MCMC)

process is presented. Understanding the idea behind this procedure is very important as the multiple imputation methods used in this study use a MCMC procedure to draw imputed values from their predictive distributions. Therefore, the Monte Carlo integration, importance sampling and MCMC techniques are discussed in general and in the context of missing data in particular. These discussions are illustrated by practical examples.

In Chapter 3, the literature review on missing data methods is provided in general and discussed in particular for missing data designed for categorical variables. Thus, single-based, model-based and multiple imputation methods are briefly reviewed. The literature on multiple imputation methods of interest; multivariate normal imputation (MVNI) and multiple imputation by chained equations or MICE is provided. A real data set is used to evaluate the performance of these two methods when data are missing completely at random on continuous and normally distributed variables containing missing values. The performance of these two methods is also investigated when different data sets (data sets with different rates of missing values) are considered.

In Chapter 4, the methodology used to analyse data is explained. The data sets and specific variables used for analysis are described. The missing data models as well as the analysis method (imputation of missing values, model development and computation of the performance measures and imputation diagnostics) are explained. The results are presented and discussed in Chapter 5, whilst Chapter 6 provides further discussion, draws conclusion and recommendations based on emerging findings. Areas for further research are suggested in this last Chapter.

Some sections of this thesis have already been published. These include the results on Chapter 3 where the performance of the MVNI and MICE was evaluated when data were missing completely at random on continuous and normally distributed variables in the regression model. In Chapter 5, the results on the performance of these methods when data are missing at random on nominal categorical that are treated as predictors in the regression models were also published. These papers are summarised as follows:

1. I Karangwa, D Kotze and RJ Blignaut. (2015). Multiple imputation of unordered categorical missing data: A comparison of the multivariate normal imputation and multiple imputation by chained equations. *Brazilian Journal of Probability and Statistics*.

2. I Karangwa and D Kotze. (2013). Using the Markov Chain Monte Carlo Method to Make Inferences on Items of Data Contaminated by Missing Values. *American Journal of Theoretical and Applied Statistics* 2(3):48-53.



Chapter 2

Markov Chain Monte Carlo process

2.1 Introduction

In statistics, the MCMC process is used to estimate parameters of interest under difficult conditions such as when data are missing or when underlying distributions do not meet the assumptions of the maximum likelihood process (Enders, 2010). The main objective of this process is to find a probability distribution known as a posterior distribution in Bayesian statistics that can be used to estimate target parameters. Robert and Casella (2010) amongst others provide a comprehensive description of MCMC methods.

The objective of this chapter is to review the theory behind this process and its link with the estimation of missing values in data sets. In the second section (Section 2.2), the idea behind the Monte Carlo integration is described. Section 2.3 reviews the importance sampling, which is a method used to increase the accuracy of the Monte Carlo estimates. In Section 2.4, the MCMC process is explained and two of its algorithms, namely the Metropolis-Hastings and Gibbs sampling algorithms, are discussed. The use of the MCMC procedure in the presence of missing values in the data set is explained in Section 2.5. The last section (Section 2.6) summarises the main points of this chapter.

2.2 Monte Carlo integration

Monte Carlo integration is a technique used to evaluate complex integrals by sampling randomly in the domain of integration from the function to be integrated as mentioned by [Robert and Casella \(2010\)](#). To illustrate this technique, suppose we have to estimate the following integral of a function h over some domain G :

$$I(h) = \int_G h(x)dx. \quad (2.1)$$

Let the function $h(x)$ be broken down into two functions $l(x)$ and the probability density function (p.d.f) $p(x)$ defined over the domain G such that $h(x) = l(x)p(x)$. Then the following equation holds:

$$\int_G h(x)dx = \int_G l(x)p(x)dx = E[l(X)]. \quad (2.2)$$

This shows that the integral $I(h)$ in equation (2.1) can be expressed as an expectation of $l(x)$ with respect to the probability density function $p(x)$. If a large number of independent samples say X_1, X_2, \dots, X_n are drawn from the probability density function $p(x)$, then

$$\int_G h(x)dx = E[l(X)] \approx \frac{1}{n} \sum l(X_i) = \hat{I}(h). \quad (2.3)$$

This approach is referred to as the Monte Carlo method and is described in detail by [Robert and Casella \(2010\)](#), [Gentle \(2009\)](#) and [Metropolis and Ulam \(1949\)](#) amongst others.

For large samples, the estimate of $I(h)$ will converge to the correct answer. That is,

$$\hat{I}(h) = Pr\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum l(X_i)\right) = I(h). \quad (2.4)$$

The variance V of $\hat{I}(h)$ is given by:

$$V(\hat{I}(h)) = V\left(\frac{1}{n} \sum l(X_i)\right) = \frac{\sigma^2}{n} \quad (2.5)$$

where $\sigma^2 = V(l(X))$ and is obtained from the data as follows:

$$\hat{V}(l(X)) = \frac{1}{n-1} \sum l(X_i)^2 - \frac{1}{(n-1)n} \left(\sum l(X_i)\right)^2. \quad (2.6)$$

For the multivariate case, the following is obtained:

$$I(h) = \int \dots \int_G h(x_1, x_2, \dots, x_k) dx_1 \dots dx_k \quad (2.7)$$

for G a given k -dimension region. Therefore, given a probability density function $p(x)$ on the dimension G the following equation holds:

$$I(h) = \int \dots \int_G h(X) dX = \int \dots \int_G l(x) p(x) dX = E[l(X)] \approx \frac{1}{n} \sum l(X_i) = \hat{I}(h).$$

The variance V of $\hat{I}(h)$ in this case is given by:

$$V(\hat{I}(h)) = V\left(\frac{1}{n} \sum l(X_i)\right) = \frac{1}{n} V(l(X_i)). \quad (2.8)$$

To illustrate the Monte Carlo integration, consider the function $h(x)$ defined as follows:

$$h(x) = \exp\left(\frac{-(x-2)^2}{2}\right) + \exp\left(\frac{-(x-4)^2}{2}\right). \quad (2.9)$$

To evaluate the integral of this function over the domain D using the Monte Carlo Integral, the following is done:

$$I(h(x)) = \int_D h(x) p(x) dx = E_p[h(X)] \quad (2.10)$$

where $h(x)$ is the function in (2.9), $p(x)$ is the normal probability function that was chosen arbitrarily and $E_p[h(X)]$ is an expectation with respect to the density p . The Monte Carlo method for approximating the equation in (2.10) consists of generating a sample (X_1, X_2, \dots, X_n) from the density function p which is in this case a normal distribution and then compute the following empirical average:

$$\bar{h}_n = \frac{1}{n} \sum h(x_j) \quad (2.11)$$

which is an approximated estimate that converges to $E_p[h(X)]$. In this example, $h(x)$ is first plotted in Figure 2.1 and then a sample $X_i = (X_1, X_2, \dots, X_n)$ of random variables is generated from the normal distribution and used to compute the average $\bar{h}_n = \frac{1}{n} \sum h(x_j)$ which converges to $E_p[h(X)]$ by the law of large numbers known as the Central Limit Theorem (CLT). The plot of this quantity is

also shown in Figure 2.2, together with its confidence bands (mean ± 2 standard errors against iterations for a single sequence of iterations).

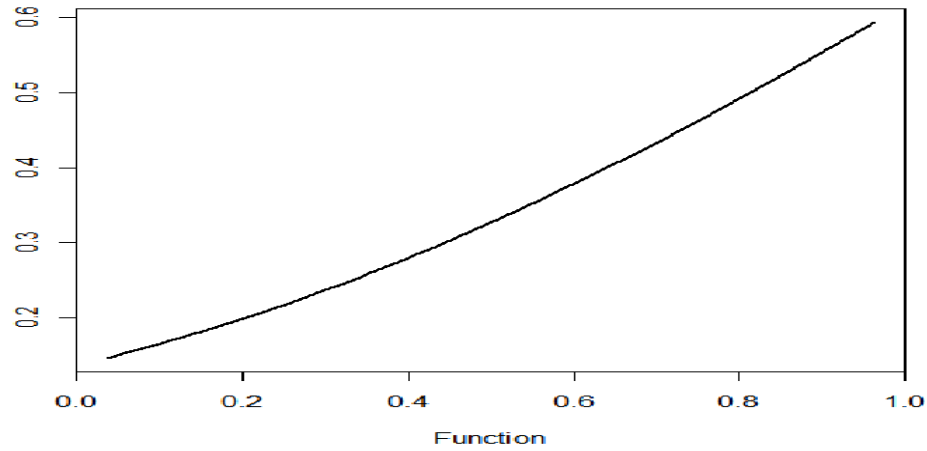


FIGURE 2.1: Plot of the function $h(x)$ in Equation (2.9)

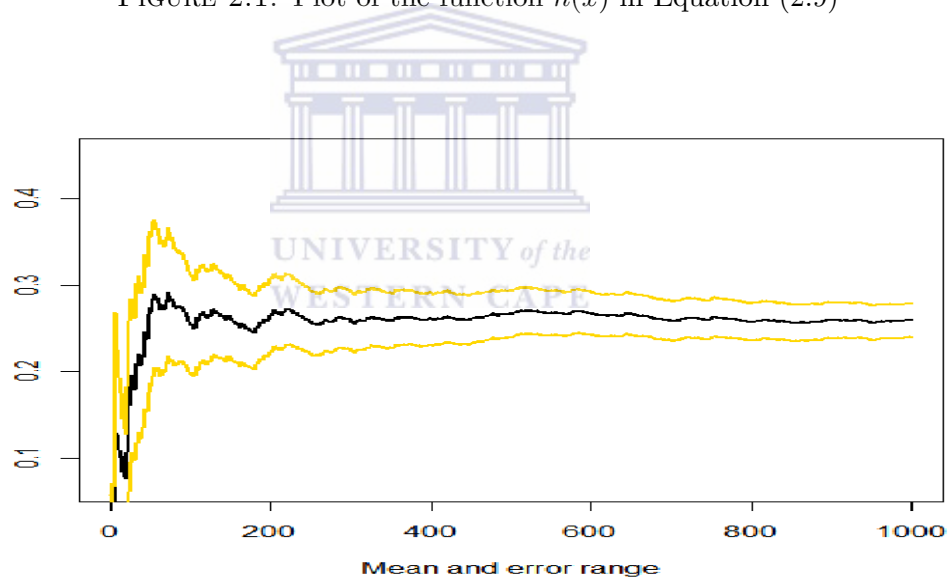


FIGURE 2.2: Approximation of the integral of the function $h(x)$ by Monte Carlo method when f is a normal density: mean \pm two standards errors against iterations for the single sequence of simulations

The variance of the approximation \bar{h}_n is given by:

$$V(\bar{h}_n) = \frac{1}{n} \int (h(x) - E_p[h(x)])^2 p(x) dx \quad (2.12)$$

which can be estimated from the sample through the following equation:

$$V_n = \frac{1}{n^2} \sum [h(x_j) - \bar{h}_n]^2. \quad (2.13)$$

The quantity $\frac{\bar{h}_n - E_p[h(x)]}{\sqrt{V_n}}$ is distributed as a normal random variable with a mean of 0 and a variance of 1, and hence a standard error of 1 according to the Central Limit Theorem (CLT) for large numbers. This allows to test for convergence and confidence bounds on the approximation $E_p[h(x)]$, the expectation with respect to the density $p(x)$.

2.3 Importance sampling

An alternative mathematical representation of equation (2.1) can be presented as:

$$\int_D h(x) dx = \int_D \frac{h(x)f(x)}{f(x)} dx = E\left[\frac{h(x)}{f(x)}\right] = E_f\left[\frac{h(x)p(x)}{f(x)}\right] \quad (2.14)$$

which is an expectation under the density f . In the above cases, the density $f(x)$ is arbitrary and positive ($f(x) > 0$) on condition that $h \times f \neq 0$.

The estimate of $I(h)$ is

$$\hat{I}(h) = \frac{1}{n} \sum \frac{h(X_i)}{f(X_i)} \quad (2.15)$$

where the X_i 's denote random samples from the density function f and $f(X_i) \neq 0$ for all X_i in the domain D for which $h(X_i) \neq 0$.

The Monte Carlo technique in equations (2.14) and (2.15) consists of computing the average of the quantity $\frac{h(X_i)}{f(X_i)}$ for a number of samples. If p is very small for a given sample, the ratio $\frac{h(X_i)}{f(X_i)}$ will be arbitrarily large. This large sample skews the sample mean away from the true mean and increases the sample variance. In order to cancel out these negative effects, one needs to increase the number of samples. To avoid cases like these, it is necessary to choose the values of f as close to h as possible, so that the variance and hence the error can be reduced. This method of choosing a probability density function that corresponds to the integrand h is referred to as *importance sampling* and is used to increase the accuracy of the Monte Carlo estimates.

To illustrate the importance sampling technique, consider the function in equation (2.14), which is evaluated as in equation (2.15) by the Monte Carlo method. Assume that $p(x)$ is a normal distribution with mean 0 and variance 1.

This equation can be rewritten as:

$$\int \frac{h(x)p(x)g(x)}{g(x)} dx = E_g\left[\frac{h(x)p(x)}{g(x)}\right] \quad (2.16)$$

where g corresponds to the uniform distribution $U(-9,-2)$. To approximate the quantity in (2.9), a sample (U_1, U_2, \dots, U_n) must be generated from the density g and used to compute the empirical average $\frac{1}{n} \sum \frac{p(x_j)h(x_j)}{g(x_j)}$, which converges to the average in (2.9) according to the law of large numbers. In this example, the function $h(x)p(x)$ must be used as a new $h(x)$ in the previous example. It can be evaluated as follows:

$$\begin{aligned} h(x)^* = h(x)p(x) &= \exp\left[\left(\frac{-(x-2)^2}{2}\right) + \exp\left(\frac{-(x-4)^2}{2}\right)\right] \frac{1}{\sqrt{2\Pi}} \exp\left(-\frac{1}{2}x^2\right) \\ &= \frac{1}{\sqrt{2\Pi}} \exp\left(-\frac{1}{2}x^2\right) \left[\exp\left(-\frac{1}{2}(x^2 - 4x + 4)\right) \right. \\ &\quad \left. + \exp\left(-\frac{1}{2}(x^2 - 8x + 16)\right) \right] \\ &= \frac{1}{\sqrt{2\Pi}} \left[\exp(-x^2 + 2x - 2) + \exp(-x^2 + 4x - 8) \right]. \end{aligned} \quad (2.17)$$

A sample (U_1, U_2, \dots, U_n) is generated from the density g and used to approximate the expectation with respect to this density. The plot of the function in (2.17) is shown in Figure 2.3 and the convergence of the importance sampling approximation of this function based on a sequence generated from a uniform distribution is shown in Figure 2.4.

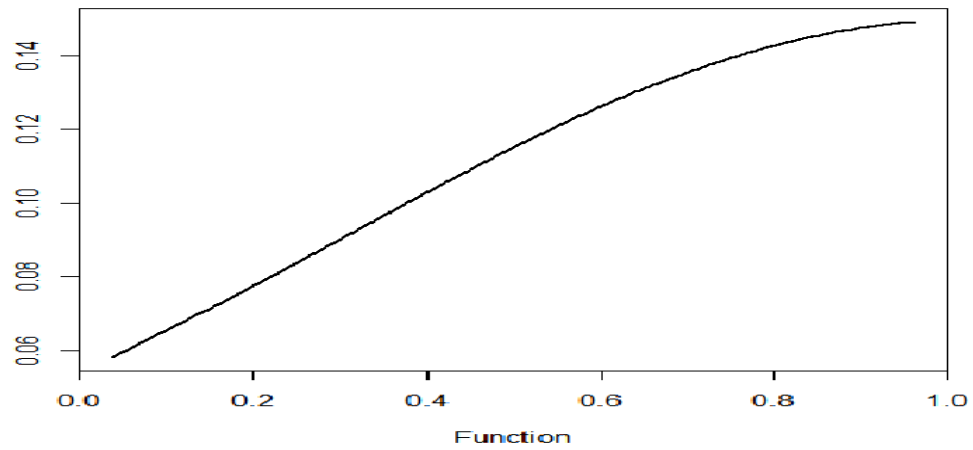
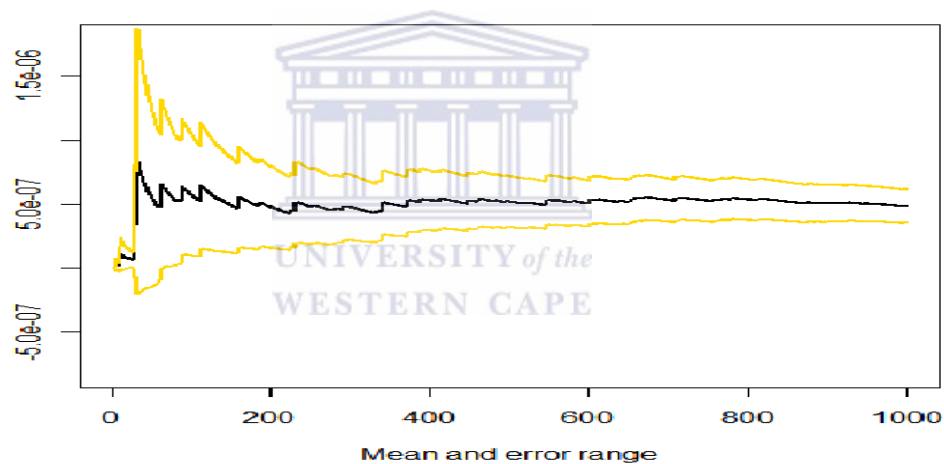


FIGURE 2.3: Plot of the function in Equation (2.17)

FIGURE 2.4: Convergence of the importance sampling approximation of the function $(h(x))^* = h(x)p(x)$ using a sequence of samples generated from a uniform distribution: mean \pm two standard errors against iterations for the single sequence of simulations

As shown by Figure 2.4, the accuracy in this figure seems to be better (close to a straight line) than the accuracy in Figure 2.2 where the original function is evaluated by the Monte Carlo method and improved by the importance sampling.

2.4 Markov Chain Monte Carlo

2.4.1 Introduction

Nowadays, information on many events goes into assessing their probability distributions. This information is often divided into two types; the general background knowledge and the information specific to the situation at hand. When these two sources of information are combined, an overall distribution of the parameters of interest, say θ , is obtained.

To illustrate this, let $p(\theta)$ be the marginal distribution of θ , which represents the background information and $p(Y = Y_1, \dots, Y_n | \theta)$ is the conditional distribution of the data Y given the parameter θ , which represents the available information. The combination of these two distributions yields $p(\theta | Y = Y_1, \dots, Y_n)$, which is the state of knowledge about a particular event when the background information and the data specific to the problem at hand are taken into consideration. This situation is referred to as Bayesian statistics, where $p(\theta)$ stands for the prior distribution, $p(Y = Y_1, \dots, Y_n | \theta)$ is the likelihood function and $p(\theta | Y = Y_1, \dots, Y_n)$ is the posterior distribution. In this case, the parameter θ represents all unknown quantities in the model, which include for instance missing data and model parameters such as the mean and covariance matrix amongst others (Deltour et al., 1999).

A practical example in real life is when weather forecasters for instance want to determine the probability of rainfall at a particular time during the year. Suppose that from their past experience, they know that normally, during that particular time, the probability of rainfall is $p(\theta)$, which represents the general background knowledge about rainfall during that particular time. To be sure about this past information, they collect data on rainfall to obtain $p(Y = Y_1, \dots, Y_n | \theta)$; the information specific to rainfall. The combination of these two pieces of information gives $p(\theta | Y = Y_1, \dots, Y_n)$, which is the probability of raining given the evidence or data on rainfall.

In Bayesian statistics, the posterior distribution is mathematically written as follows:

$$p(\theta | y) = \frac{p(y | \theta)p(\theta)}{\sum p(y | \theta)p(\theta)} \quad (2.18)$$

where the summation in the denominator represents the accumulation across all possible outcomes of θ and therefore can be taken as the probability of Y . When

parameters are continuous values, equation (2.18) becomes

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta} \quad (2.19)$$

which is referred to as a joint posterior density of the model parameter given the outcome or data Y . In this case, the use of $f(\cdot)$ and \int in place of p and Σ respectively accounts for the continuous nature of the parameter values in the Bayesian theorem. The parameter θ and the outcome Y which represent a single event in equation (2.18), combine multiple parameters and outcomes respectively in equation (2.14).

In Bayesian theory, equation (2.19) leads to the following results:

$$\begin{aligned} f(\theta|y) &= \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta} \\ &= \frac{f(\theta)l(\theta)}{\int f(\theta)l(\theta)d\theta} \\ &= \frac{f(\theta)l(\theta)}{C} \end{aligned} \quad (2.20)$$

where $l(\theta)$ is the likelihood function of θ and $C = \int f(\theta)l(\theta)d\theta$ is a constant that is independent of θ . In fact, by integrating with respect to θ , the results will not depend on θ , which means that θ is automatically eliminated after the integration (Lavine, 2005). C is known as a normalizing constant which has a role of rescaling the function in the numerator so that it can integrate to one, that is, $\int f(y|\theta)d\theta = 1$.

Therefore equation (2.20) is always written as:

$$f(\theta|y) \propto f(\theta)l(\theta) \quad (2.21)$$

which states that the joint posterior density function of θ given Y is proportional to the prior density function of the model parameter or the likelihood of particular parameter values before the collection of data and the likelihood of the response data given all parameters (Lavine, 2005).

Some posterior distributions can be analytically intractable and therefore need to be integrated numerically. MCMC methods are powerful techniques used to evaluate these kinds of integrals in Bayesian analysis.

The term Markov Chain Monte Carlo consists of two parts; Monte Carlo and

Markov Chain. The former means evaluating the integral using random draws from given distributions, whereas the latter refers to how these draws or samples are produced (Lavine, 2005). This technique is explained in the following example. Suppose that the following equation needs to be evaluated:

$$p(\theta_1|y) = \int \dots \int p(\theta_1, \dots, \theta_k|y) d\theta_2 \dots d\theta_k. \quad (2.22)$$

Let $\vec{\theta} = (\theta_1, \dots, \theta_k)$. We can generate or draw many samples $\vec{\theta}_1, \dots, \vec{\theta}_M$ of $\vec{\theta}$ from its posterior distribution and then evaluate equation (2.22). These draws are produced using transition densities.

In the Markov Chain technique, there exists a transition density also known as a transition kernel $K(\vec{\theta}_i|\vec{\theta}_{i-1})$ which is a density for generating $\vec{\theta}_i$ given $\vec{\theta}_{i-1}$. The first sample $\vec{\theta}_1$ is normally chosen arbitrarily and then $K(\vec{\theta}_2|\vec{\theta}_1)$, $K(\vec{\theta}_3|\vec{\theta}_2)$, ... samples are generated in as many steps as needed. Each $\vec{\theta}_i$ is associated with density $p \equiv p(\vec{\theta}_i)$ which depends on $\vec{\theta}_1$ and the transition kernel. Under some conditions the sequence p_i will converge to a limiting or stationary distribution that does not depend on $\vec{\theta}_1$ and the transition kernel $K(\vec{\theta}_i|\vec{\theta}_{i-1})$ can be chosen in such a way that the stationary distribution p equals $p(\vec{\theta}|y)$.

With the MCMC method, when a target density f or a posterior distribution in the Bayesian statistics term is given, the primary goal is to build a Markov kernel K with a stationary distribution f and then, generate a Markov chain $\vec{\theta}_i^{(t)}$ using this kernel in such a way that the limiting distribution of the drawn samples $\vec{\theta}_i^{(t)}$ is f . Next, integrals can be used to approximate $\frac{1}{T} \sum f(\vec{\theta}_i^{(t)})$.

To construct the kernel K associated with an arbitrary density f is a difficult task (Lavine, 2005). Luckily, MCMC algorithms can be used to derive such kernels. These include the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) and the Gibbs sampling (Geman & Geman, 1984) amongst others. A detailed description of these methods is given in Robert and Casella (2010) and Chib and Greenberg (1995) and many others. The next two sections provide a brief description of these two methods as they are the most frequently used algorithms of MCMC.

2.4.2 Metropolis-Hastings algorithm

Metropolis et al. (1953) initiated the first MCMC process and later on, Hastings (1970) generalized the process; giving rise to the term Metropolis-Hastings.

A Markov chain having a kernel K normally satisfies the following so-called balance condition or equation

$$f(\vec{\theta}_i)K(\vec{\theta}_i|\vec{\theta}_{i-1}) = f(\vec{\theta}_{i-1})K(\vec{\theta}_{i-1}|\vec{\theta}_i) \quad (2.23)$$

If this equation holds, then f is a stationary distribution of the chain $\vec{\theta}^{(t)}$ and the chain is reversible. Therefore, finding a Markov chain with a stationary distribution is the same as deriving a transition density from equation (2.23) which in case of the Metropolis-Hastings algorithm can be constructed as follows:

$$K(\vec{\theta}_i, \vec{\theta}_{i-1}^*) = q(\vec{\theta}_i, \vec{\theta}_{i-1}^*)\alpha(\vec{\theta}_i, \vec{\theta}_{i-1}^*) \quad (2.24)$$

where q stands for the proposal or candidate distribution and α denotes the acceptance probability (Robert & Casella, 2010). Therefore, the new value has to be generated using the proposal distribution q and then accepted with the probability α given by:

$$\alpha(\vec{\theta}_i, \vec{\theta}_{i-1}^*) = \begin{cases} \min \left\{ \frac{f(\vec{\theta}_{i-1})q(\vec{\theta}_{i-1}, \vec{\theta}_i)}{f(\vec{\theta}_i)q(\vec{\theta}_i, \vec{\theta}_{i-1})}, 1 \right\} & \text{if } f(\vec{\theta}_i)q(\vec{\theta}_i, \vec{\theta}_{i-1}) > 0, \\ 1 & \text{elsewhere} \end{cases} \quad (2.25)$$

In general, the Metropolis-Hastings algorithm works as follows:

1. Choose a starting value $\vec{\theta}_1$
2. Choose a proposal density $q(\vec{\theta}^*|\vec{\theta}_{i-1})$
3. For $i = 2, 3, \dots$ generate a proposal $\vec{\theta}^*$ from $q(\vec{\theta}^*|\vec{\theta}_{i-1})$ and set

$$\alpha \equiv \left\{ \min \left\{ \frac{f(\vec{\theta}_{i-1})q(\vec{\theta}_{i-1}, \vec{\theta}_i)}{f(\vec{\theta}_i)q(\vec{\theta}_i, \vec{\theta}_{i-1})}, 1 \right\} \right\} \quad (2.26)$$

set $\vec{\theta}_i = \vec{\theta}^*$ with probability α and $\vec{\theta}_i = \vec{\theta}_{i-1}$ with probability $1 - \alpha$.

The last step of the algorithm defines the transition density or Kernel K .

To illustrate how the Metropolis-Hasting algorithm works, consider Figure 2.5 which was produced by generating 20000 random samples from a Beta distribution $\text{Be}(3,7)$ using a proposal density $q(\theta^*|\theta) = U(\theta - 0.02, \theta + 0.02)$ and an arbitrary chosen initial value $\theta_1 = 0.5$. The $\text{Be}(3,7)$ distribution is indicated by the curve around the histogram that was produced using the Metropolis-Hastings samples. The figure shows that they match closely, which is an indication that

the Metropolis-Hastings algorithm performed well. That is, the MCMC chains converged and delivered samples from the target distribution which is the $\text{Be}(3,7)$ distribution.

Early samples of the chain are not normally considered as they are accused of being non-representative of the state of the chain (Deltour et al., 1999). Indeed, these samples are influenced by the distribution of the initial sample $\theta_1^{(0)}$, which is not seemingly drawn from the posterior distribution $p(\theta|y)$. In the $\text{Be}(3,7)$ distribution example, we dropped the first 1000 samples to attain convergence.

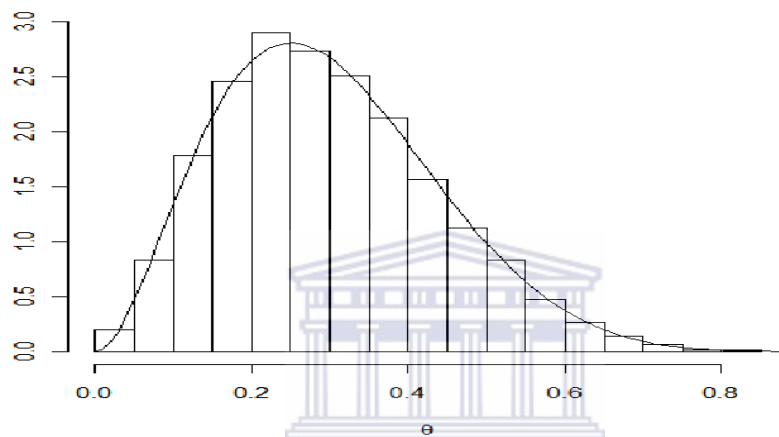


FIGURE 2.5: 20000 MCMC samples produced by a $\text{Be}(3,7)$ distribution. Histogram from a Metropolis-Hastings algorithm and a $\text{Be}(3,7)$ distribution.

2.4.3 Gibbs sampling

An alternative method to the Metropolis-Hastings is the Gibbs sampling developed by Geman and Geman (1984) and later demonstrated by Gelfand and Smith (1990).

This method works as follows. Let $\theta^{(t)} = \{\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_p^{(t)}\}'$ and $f(x)$ be a vector of items that constitute $\theta^{(t)}$ and a joint distribution of θ respectively. The Gibbs sampling method consists of the transition from $\theta^{(t)}$ to $\theta^{(t+1)}$, where the items in the vector are updated one after the other using their conditional distributions given the other. The value of $\theta^{(t+1)}$ will be consecutively obtained as follows:

Draw $\theta_1^{(t+1)}$ from $f_1\{\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_p^{(t)}\}$

Draw $\theta_2^{(t+1)}$ from $f_2\{\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_p^{(t)}\}$

Draw $\theta_3^{(t+1)}$ from $f_3\{\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, \dots, \theta_p^{(t)}\}$

⋮

Draw $\theta_p^{(t+1)}$ from $f_p\{\theta_p|\theta_1^{(t+1)}, \theta_2^{(t+2)}, \theta_3^{(t+3)}, \dots, \theta_{p-1}^{(t+1)}\}$.

The obtained sample elements form $\theta^{(t)}$ and the densities f_1, f_2, \dots, f_p are referred to as "full conditionals" and are only used for simulation purposes in the Gibbs sampling method (Robert & Casella, 2010). This process is iterated in order to get a sequence $\{\theta^{(t)}\}$ (where the iteration t takes place over $t = 1, 2, \dots, M$) that forms the Markov chain.

The Gibbs sampling is a special case of a Metropolis-Hastings where the proposal distribution is the target distribution. Thus, the acceptance probability in equation (2.26) becomes

$$\alpha = \frac{f(\vec{\theta}_{i-1})q(\vec{\theta}_i|\vec{\theta}_{i-1})}{f(\vec{\theta}_i)f(\vec{\theta}_{i-1}|\vec{\theta}_i)} = \frac{f(\vec{\theta}_{i-1}, \vec{\theta}_i)}{f(\vec{\theta}_i, \vec{\theta}_{i-1})} = 1 \quad (2.27)$$

which means that every move or sampled value is accepted (Robert & Casella, 2010). The Gibbs sampler normally draws samples from full conditionals. If it is not possible, the Metropolis-Hastings is used.

To illustrate the Gibbs sampler procedure, consider the bivariate normal distribution of two independent random variables X and Y which are correlated. In this case, X and Y are normally distributed with means of 0 (0,0) and a variance and correlation matrix of

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

where ρ denotes the correlation between X and Y . The conditional distributions of X and Y for the bivariate normal case are given by:

$$P(X|Y = y) \sim N(\rho y, 1 - \rho^2) \quad (2.28)$$

and

$$P(Y|X = x) \sim N(\rho x, 1 - \rho^2). \quad (2.29)$$

In Figure 2.6, the Gibbs sampler is used to simulate a bivariate normal distribution by iteratively sampling from these conditionals. This works properly because it is a Markov chain. Given that the starting value of the chain is 0 for instance

for the random variable X , if $X^{(0)} = x_0$, then the distribution of $X^{(n)}$ becomes $N(\rho^{2n}x_0, 1 - \rho^{4n})$, which converges to a standard normal distribution as n tends to infinity (Seefeld & Linder, 2007). Thus, no matter what the starting values of the chain are, X and Y will be normally distributed with a mean and standard deviation of 0 and 1 respectively after enough runs or iterations. The joint distribution of these two random variables were plotted (Figure 2.6), with different starting values of the chain for X and Y respectively; (0,0), (5,5), (-5,5), (10,10), (-10,10), (15,15). A correlation of 0 between X and Y and 1000 runs were used. As shown by the figure, after 1000 iterations, the joint distribution of X and Y simulated using the Gibbs algorithm seems to be the same as if the chain started at (0,0). However, a difference is observed if the values of the correlation coefficients differ (Figure 2.7) and 10000 runs are used. As shown by the figure, the larger the value of the correlation coefficient, the more the distribution tends to a straight line.



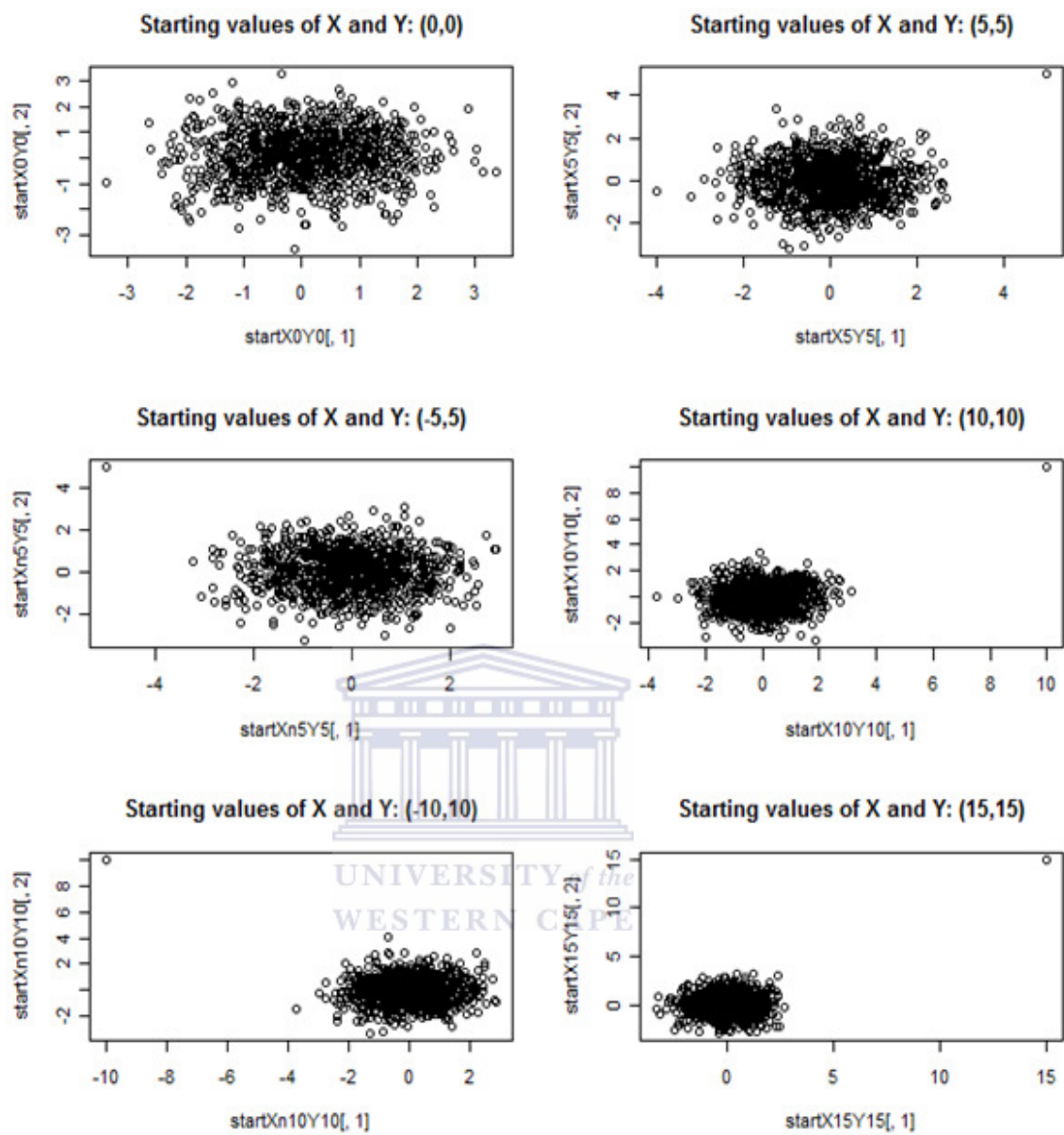


FIGURE 2.6: Plot of the bivariate normal distribution of random variables X and Y simulated by iteratively sampling from the conditional distributions of these two variables using 1000 runs, different starting values of the chain and a correlation coefficient of 0 between X and Y

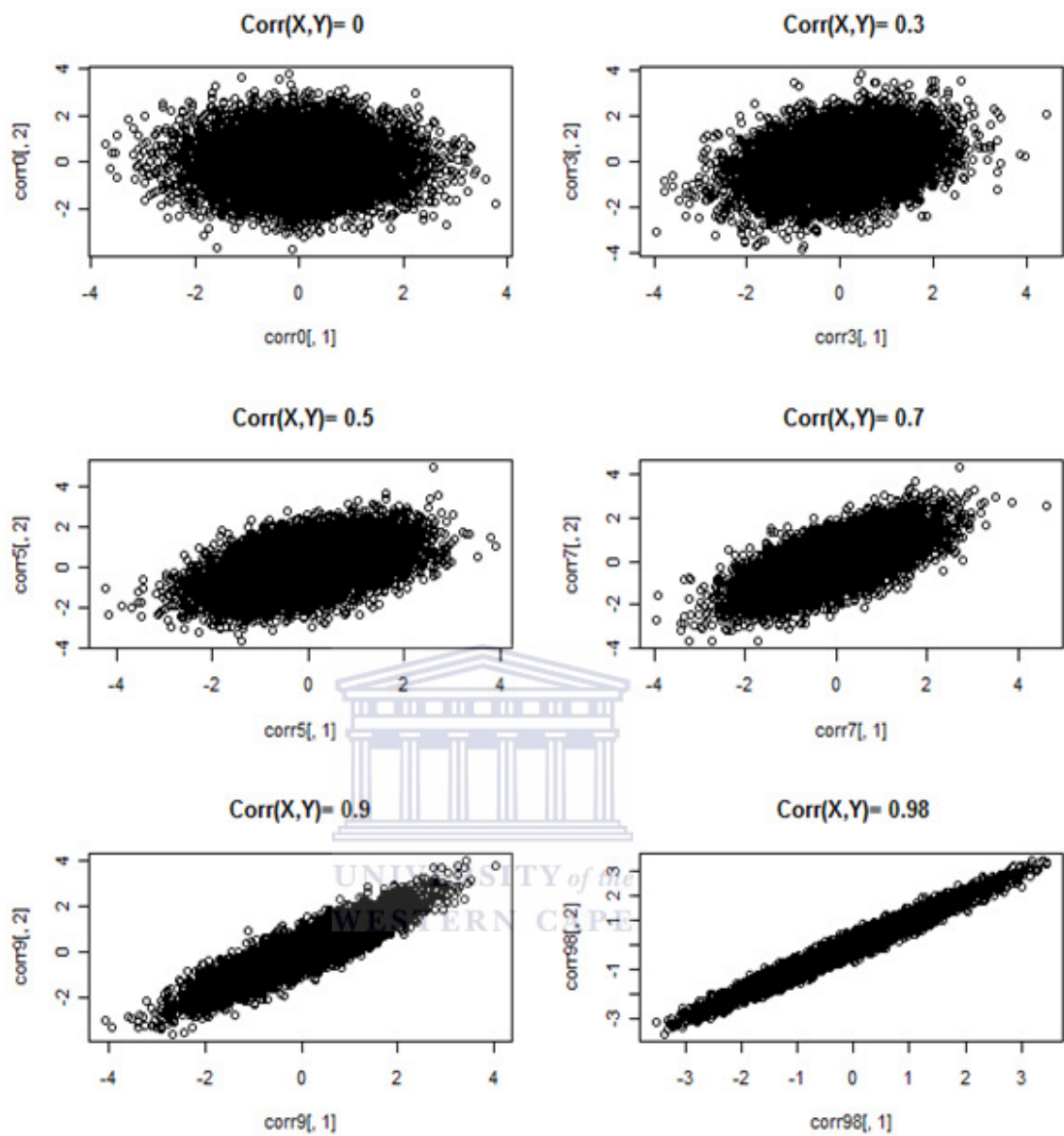


FIGURE 2.7: Plot of the bivariate normal distribution of random variables X and Y simulated by iteratively sampling from the conditional distributions of these two variables using 10000 runs, different values of the correlation coefficients between X and Y as well as a starting value of the chain of 0 for both X and Y

2.5 Markov Chain Monte Carlo methods in the presence of missing data

2.5.1 Introduction

When data are missing, the primary goal of a researcher is to generate unbiased estimates in order to make good inferences. This is not generally easy when the available data (observed data after discarding missing items) are used. With MCMC methods, the available data need to be augmented with simulated values of the missing data in order to obtain good parameter estimates. This section explains the idea behind the MCMC method and how it is used to draw imputation values from the desired distributions.

2.5.2 Markov Chain Monte Carlo versus missing data

In the presence of missing values in the data set, the target variable Y becomes $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} are the observed and missing part respectively. In this case, the posterior distribution is

$$p(\theta|Y_{obs}) = \int p(\theta|Y_{obs}, Y_{mis})p(Y_{mis}|Y_{obs})dY_{mis}. \quad (2.30)$$

To evaluate this distribution, one needs to condition on Y_{mis} and then estimate its values, which is done by integrating over the density of Y_{mis} , $p(Y_{mis}|Y_{obs})$, which averages over more or less the values of Y_{mis} (Lavine, 2005). Equivalently, using the Monte Carlo method, one needs to draw n independent copies of Y_{mis} ($Y_{mis}(1), Y_{mis}(2), \dots, Y_{mis}(n)$) from the conditional distribution $p(Y_{mis}|Y_{obs})$, and then compute the average $\frac{1}{n} \sum p(\theta|Y(j))$ as an approximation of $p(\theta|Y_{obs})$, where $Y(j)$ denotes the augmented dataset $(Y_{obs}, Y_{mis}(j))$ for $j = 1, 2, \dots, n$ and $Y_{mis}(j) = (Y_{mis})_1(j), (Y_{mis})_2(j), \dots, (Y_{mis})_n(j)$.

In the MCMC context, the above mentioned idea can be simply done using the Imputation-Parameter (IP) algorithm suggested by Schafer (1997) which works as follows. Assuming multivariate normally distributed data, at the t^{th} iteration one needs to draw $Y_{mis}^{(t+1)}$ from $p(Y_{mis}|Y_{obs}, \theta^{(t)})$, and then draw $\theta^{(t+1)}$ from $p(\theta|Y_{obs}, Y_{mis}^{(t+1)}, \theta^{(t)})$. The former step is referred to as the Imputation (I) step and the latter as the Parameter (P) step. The resulting sequence forms the following

Markov chain:

$\{Y_{mis}^{(1)}, \theta^{(1)}\}, \{Y_{mis}^{(2)}, \theta^{(2)}\}, \{Y_{mis}^{(3)}, \theta^{(3)}\}, \dots, \{Y_{mis}^{(t+1)}, \theta^{(t+1)}\}$, which must converge to the distribution $p(Y_{mis}|Y_{obs}, \theta)$ (Horton & Lipsitz, 2001) and then used in the Multiple Imputation of missing values. In words, at the I-step missing values (Y_{mis}) are simulated for each observation independently by using the observed data (Y_{obs}) and the estimates of the mean vector and covariance matrix represented by $\theta^{(t)}$. The P-step uses the complete data set (full data with generated missing values) from the I-step to generate new estimates of the mean vector and covariance matrix, which are to be used in the next I-step to simulate new values. The repetition of these two steps (I-step and P-step) creates a Markov chain (sequence of random variables in which the distribution of each element is related to the values of the previous one) whose role is to generate a distribution of values from which random samples of simulated missing values are obtained and used in Multiple Imputation methods to estimate the parameters of interest. The chain needs to be long enough for the distribution of the elements to stabilize to a common distribution referred to as the stationary distribution (Schafer, 1997).

The advantage of the MCMC methods over the maximum likelihood methods is their efficiency and flexibility. Indeed, they allow researchers to estimate parameters when the underlying distributions are unknown or not normally distributed (Allison, 2002).

2.6 Summary of the chapter

Multiple imputation methods used in this study rely on the MCMC process to estimate missing values in the data sets. This chapter discussed the theory behind this process in general and its use in estimating missing data in particular. It was explained that in the Bayesian framework, missing values are considered as a set of other parameters to estimate using the MCMC to draw imputation values that must converge to the desired distributions.

In the next chapter, the literature review on missing data methods is provided. Multiple imputation methods of interest, namely MVNI and MICE, are discussed and a practical example using a real data set is given to illustrate the performance of these methods when data are missing completely at random on

continuous variables. The aim is to get a better understanding of these techniques and explore their performance when data are missing on continuous data as background theory and literature.



Chapter 3

Literature review on missing data methods

3.1 Introduction

As previously stated, a common way of handling missing values is to discard them from the analysis; a technique that is provided by default in many statistical packages such as the statistical package for social sciences (SPSS), Stata, and SAS amongst others. This approach is referred to as case deletion or complete case analysis, and it leads to low power of the statistical test and biased parameter estimates, especially when the proportion of missing values is high and data are missing in a systematic manner or missing at random (Graham, 2009). Although its problems are well known, it is still the most popular missing data method as a high number of researchers still use it (Kropko et al., 2014). To reduce problems associated with complete case analysis, various methods of rescuing missing data have been implemented (Carpenter & Kenward, 2012; Graham, 2009; Little & Rubin, 2002; Schafer & Graham, 2002; Tsikriktsis, 2005). Items with completely missing data are directly discarded from the analysis because they do not provide any particular information about the data. However, if data are partially missing on items of interest, these items should not be discarded as they still contain some information that can be used to draw inferences on the variables of interest.

This chapter reviews the literature related to developed methods that are used to treat missing data. The chapter covers five main sections. The first section introduces the content of the chapter. In Section 3.2, a brief review of single

imputation methods is provided. Section 3.3 discusses model-based imputation methods. In Section 3.4, the idea behind multiple imputation is explained and two multiple imputation techniques (MVNI and MICE) that are currently considered as the best are presented. A practical example using real data is also provided to illustrate the performance of these two methods when data are missing on continuous data. The last section (Section 3.5) provides a summary of the content of the chapter.

3.2 Single-based imputation methods

3.2.1 Mean imputation

The mean imputation technique replaces missing values with the observed mean of the available data on the variable containing missing data. With this technique, the efficiency problem is solved, however, standard errors of the estimates are underestimated (Carpenter & Kenward, 2012; Graham, 2009; Schafer, 1997). In addition, the estimate of the mean is treated as true whereas it is not the case, and the method does not even attempt to recover the existing relationship between variables. This is not optimal because the key objective of doing imputation is to try to recover or preserve an initially existing relationship between variables. However, imputing missing values using the mean of the observed data is a good guess, better than not doing anything at all if there are no other options or the researcher does not have any knowledge about other missing data handling methods, especially when there is a large amount of missing values in the data set to be used for a particular study. This technique is discussed by many researchers including (Graham, 2009).

3.2.2 Hot-deck imputation

The hot-deck imputation method replaces missing values with the actual values from respondents or donors in the available sample. The advantage of this technique is that it is a nonparametric method and hence it does not need strong modelling assumptions to be made in order to estimate individual values, apart from the fact that data need to be missing at random (MAR) with regards to the

auxiliary variables in data sets. Three common hot-deck imputations can be distinguished: (1) Sequential hot-deck imputation, (2) Random hot-deck imputation and (3) Nearest neighbour imputation.

The sequential hot-deck imputation consists of replacing a missing item value with a value from the last responding unit preceding it in the data file. Before using this method, the data file needs to be sorted first if auxiliary variables are quantitative or divided into subclasses if these variables are categorical. The shortcoming of this method is that for a large amount of missing data, the accuracy of the survey parameter estimates is reduced and its variance is underestimated, which therefore leads to incorrect statistical inference.

The random hot-deck imputation consists first of allocating respondents into imputation classes based on auxiliary data in order to be able to consider elements in the same class as similar. Then an item value of a randomly selected respondent within an imputation class is assigned to the missing item value.

The nearest neighbour imputation is a kind of hot-deck imputation method in which donors are selected from the neighbours (i.e., the complete cases) in such a way that they minimize some similarity measure. Unlike the mean substitution in which replacement values are influenced by all values, replacement values in this method depend on the most similar or related cases. For more details about these approaches, see [Batista and Monard \(2001\)](#) amongst others.

3.2.3 Cold-deck imputation

Cold-deck imputation consists of substituting missing item's values with values from an external source such as administrative data, which must be matched to the survey in order to recover missing information. [Ford \(1983\)](#) and [Sande \(1983\)](#) provide a detailed description of the hot-deck and cold-deck imputation techniques.

3.2.4 Regression imputation

Regression imputation uses some selected prediction of a missing value on a variable of interest. For instance, to predict a missing value for the variable say, X_1 , this variable is used as a function of other variables, say, X_2 and X_3 , in a model that could even include the dependent variable, say Y . As an illustration, suppose

that the initial model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2). \quad (3.1)$$

To obtain the best guess of X_1 , the following prediction is proposed:

$$X_1 = \hat{\phi}_0 + \hat{\phi}_1 X_2 + \hat{\phi}_2 X_3 + \hat{\phi}_3 Y + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma_{\varepsilon_1}^2). \quad (3.2)$$

Just like the mean imputation, the uncertainty is not incorporated very well because the estimates are random variables. Therefore, there is uncertainty in $\hat{\phi}$ that should be incorporated in the model of Y . That is, if the estimated X_1 is substituted in the model of Y , the uncertainty on how the $\hat{\phi}$ coefficients were obtained should fit into uncertainty in β . However, the problem of this method is that it yields small standard errors (Carpenter & Kenward, 2012).

3.2.5 Imputation using interpolation

In panel data, interpolation is used to impute missing values (Norazian et al., 2008). For instance, suppose that a variable X is measured at times $t = 1, 2, 3$ (X_1, X_2 and X_3) and some of the values are missing at time = 2 (X_2). With this method, the quantity $X_2 = \frac{X_1 + X_3}{2}$ is computed and then substituted into the missing values. As highlighted by Norazian et al. (2008), this technique creates bias and large confidence intervals.

3.3 Model-based methods

3.3.1 Expectation maximisation

The Expectation maximisation (EM) method is an iterative process in two stages: Expectation step or E-step and maximisation-step or M-step (Dempster et al., 1977; Schafer & Graham, 2002), which estimates values for missing data. In the E-step, expected values based on available data are calculated. In the M-step, missing values are replaced with values generated in the E-step and then new expected values are recomputed. This two-step process iterates until similar values are obtained (Gelman et al., 2005; Graham, 2009) or changes in expected values from

iteration to iteration become insignificant (Hedderley, 1995). The EM technique works well under MCAR compared to other single imputation approaches (Acock, 2005; Graham, 2009; Schafer & Graham, 2002). However, the method yields low standard errors which can affect some test statistics such as the t-test (Allison, 2002).

3.3.2 Maximum likelihood method

The maximum likelihood (ML) method estimates parameters based on available data and uses these estimates to estimate missing data. It is considered to be better than all the missing data methods discussed previously because it satisfies all three criteria for a good missing data technique mentioned earlier (Graham, 2009). It works well under the MAR assumption, and produces good estimates especially when large samples are used. The ML method for missing data is mathematically described as follows (Graham, 2009):

The likelihood function that expresses the probability of the data as a function of unknown parameters needs to be specified first. Let X , Z and $p(x, z|\theta)$ be discrete variables and a joint probability function respectively, where $p(x, z|\theta)$ refers to the probability that $X = x$ and $Z = z$. In the absence of missing data and if observations are independent, the following likelihood is obtained:

$$L(\theta) = \prod_{i=1}^n p(x_i, z_i|\theta). \quad (3.3)$$

To obtain the maximum likelihood, it is necessary to find the value of the parameter θ that maximises this function or the value of the parameter θ for which the observed data are most likely. In the presence of missing values in the data set, assume that data are missing at random or MAR on Z for the first r cases and then on X for the next s cases. The following holds:

$$g(x|\theta) = \sum_z p(x_i, z_i|\theta) \quad (3.4)$$

and

$$h(z|\theta) = \sum_x p(x_i, z_i|\theta) \quad (3.5)$$

which are the marginal distributions of X and Z respectively. The likelihood is then given by:

$$L(\theta) = \prod_{i=1}^r g(x_i|\theta) \prod_{i=r+1}^{r+s} h(z_i|\theta) \prod_{i=r+s+1}^n p(x_i, z_i|\theta) \quad (3.6)$$

which can be factored into parts corresponding to different missing data patterns. To find a likelihood of each pattern, the joint distribution over all possible values of the variables with missing data is summed. Summation signs must be replaced by integral signs when the variables used are continuous.

To implement ML for missing data, one needs a model for the joint distribution of all relevant variables and a numerical method for maximizing the likelihood. In case all variables are categorical, the unlimited multinomial model or log-linear model should be used.

When all variables are continuous, a multivariate normal model is assumed. This means that each variable is normally distributed and can be expressed as a linear combination of other variables, with homoscedastic errors and mean of zero. Under this assumption, the maximum likelihood can be obtained using the Expectation maximisation (EM) or the direct maximum likelihood (Allison, 2003). The disadvantages associated with the ML method for missing data is that it is often difficult to specify the joint distribution for all variables. As noted by Graham (2009), only linear and log-linear modelling cases are provided in commercial software but nothing is provided for ML with missing data in Poisson, Cox or Logistic regressions.

3.4 Multiple imputation-based methods

3.4.1 Introduction

Single-based imputation methods mentioned in the previous section constitute an improvement over the case deletion method, but they do not account for uncertainty in the imputations as imputed values are treated as true rather than estimates of the missing values. This leads to the underestimation of the variance of the estimates and the distortion of relationships among variables (Stuart et al., 2009).

Currently, many researchers view multiple imputation as a better way of filling in missing values (Schafer & Graham, 2002; Stuart et al., 2009). The goal of this method is to impute missing values in such a way that the uncertainty in the imputed values is accounted for. That is, imputed values are estimates rather than known values of missing observations, thus leading to the appropriate standard errors of the estimates.

The multiple imputation methods use a selected model such as the regression model to predict missing values based on observed data. Instead of picking one value for the missing value, many values are chosen and the uncertainty is represented in the variance covariance matrix (VCV) of β estimates used to predict missing values. As an example, suppose that a regression model of Y on X_1 and X_2 is estimated but the variable X_1 contains missing values. The following imputation model is specified:

$$X_1 = \phi_1 + \phi_2 X_2 + \phi_3 Y + \varepsilon_2, \quad \varepsilon_2 \sim N(0, \sigma_{\varepsilon_2}^2). \quad (3.7)$$

In this case, there is a VCV matrix of the ϕ_i estimates that incorporates and measures uncertainty in extent to which Y and X_2 can be used to plug in the values of X_1 . This can be done by just picking many copies of ϕ from its asymptotic distribution (for example a multivariate normal distribution for this regression model), and use the estimates of ϕ and the $VCV(\Sigma)$ to fill in the mean and VCV of the distribution $\Phi(\hat{\phi}, \hat{\Sigma})$.

Consider the following substantive model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_3, \quad \varepsilon_3 \sim N(0, \sigma_{\varepsilon_3}^2). \quad (3.8)$$

The values of X_1 are imputed using the imputation model in (3.7) and m copies of $\hat{\phi}$ are drawn from the asymptotic distribution of $\hat{\phi}$. Now m copies of the data that gives m copies of the β estimates are created when those m data sets are plugged back into the original model of Y . Therefore, m estimates of β for each data set are obtained, and from there, the final estimate of $\hat{\beta}$ is calculated. In other words, all the estimates of β are combined by taking the mean of the m estimates of β . The variance V_β of the new (combined) estimate of β is a function of the within (W) data set variance, S_m^2 , which is an ordinary least square (OLS) estimate of the normal σ^2 , and B is the between data set variance, which is a variance due to uncertainty in the imputation of X_1 .

The above mentioned quantities can be technically presented as: $\hat{\beta} = \sum_{m=1}^M \hat{\beta}_m$,

$V_{\hat{\beta}} = W + (1 + \frac{1}{m})B$ where $W = \frac{1}{m} \sum_{m=1}^M S_m^2$ and $B = \frac{1}{m-1} \sum_{m=1}^M (\hat{\beta}_m - \hat{\beta})^2$. The factor $(1 + \frac{1}{m})B$ corresponds to the inflation in the standard errors (SEs) of $\hat{\beta}$ which is done in order to correct for imputation (Rubin, 1978). These quantities are not computed manually; many statistical software packages do that. The multivariate normal imputation or MVNI and multiple imputation by chained equations or MICE are among the best ways of combining these estimates or implementing these procedures (Carpenter & Kenward, 2012). Both techniques are currently available in many statistical packages. A brief description of MVNI and MICE is given in the next two sections of this thesis.

Although this technique is applicable to different types of data and produces more reliable estimates, it is still an underused approach (Helenowski, 2015). It was developed taking into account the missing data mechanisms, a fact which is generally ignored by other missing data approaches (Helenowski, 2015). These mechanisms consist of the missing at random (MAR) and missing completely at random (MCAR) mechanisms (Demirtas, 2004; Helenowski, 2015), as well as the missing not at random (MNAR) mechanism (Demirtas, 2005; Siddique et al., 2008, 2012). In this thesis, only the first two mechanisms are assumed.

3.4.2 Description of multivariate normal imputation

As stated in the previous section, the multivariate normal imputation or MVNI assumes that all the variables in the imputation model are normally distributed. Furthermore, it implies that each variable used in the imputation process can be expressed as a linear function of all other variables, plus a normal homoscedastic error term. However, in practical settings, the MVNI can be used to impute missing values of the variables whose distributions deviate from the normal distribution (Schafer, 1997).

Under this method, a linear regression of each variable with missing data is estimated. In this case, the regression parameters in these models are random draws from the posterior distribution as in Bayesian statistics. The predicted values for the cases with missing values are then generated using the estimated regression equations, and a random draw of the residual normal distribution of that variable is added to each predicted value. The most difficult task during this process is to randomly draw samples from the posterior distribution of the regression parameters. Luckily, algorithms that perform that task have been implemented in software packages such as SAS, STATA and R, amongst others. These include

the data augmentation (DA) developed by [Schafer \(1997\)](#).

Under MVNI, the data augmentation process is a type of MCMC algorithm that is used to construct a posterior distribution in Bayesian statistics. Prior to starting DA, all the variables in the imputation model of interest need to be specified. Then, the following steps are followed during the DA process ([Schafer, 1997](#)).

1. The starting values for the parameters (means and covariance matrix of the multivariate normal model) are chosen. These values can be obtained using the case deletion method or the maximum likelihood estimation method.
2. The current values of the means and covariances are used to obtain the estimates of the regression coefficients for equations in which the variables with missing values are regressed on all observed variables.
3. The regression estimates are used to produce predicted values for missing values and a random draw from the residual normal distribution of that variable is added to each predicted value.
4. The completed data set (observed plus imputed values) is utilised to compute the means and covariance matrix using the standard formulas.
5. The estimated means and variances in (4) are used to make a random draw from the posterior distribution of the means and covariances.
6. Using the randomly drawn means and covariances in (5), return to (2) and keep iterating through the steps until stable estimates are obtained or convergence is reached. The imputation values obtained at the last iteration are the ones that are used to form a complete data set.

To obtain a posterior distribution in (5), a noninformative prior (prior that has little or no information about the parameters) is used. MVNI generally uses a uniform prior.

In mathematical terms, the DA or MCMC procedure is used to obtain imputed values from the estimated multivariate distribution, allowing appropriately for uncertainty in the estimated model parameters, which is a requirement for proper imputation ([Rubin, 1978](#)). Assuming multivariate normally distributed data, at the t^{th} iteration missing values $Y_{\text{mis}}^{(t+1)}$ are drawn from $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$, which is the distribution of missing data given the observed data Y_{obs} and the model

parameters $\theta^{(t)}$ (such as regression coefficients and covariance matrix) of the previous iteration. Then new model parameters $\theta^{(t+1)}$ are drawn from $p(\theta|Y_{obs}, Y_{mis}^{t+1}, \theta^{(t)})$; the posterior distribution of the unknown parameters given the observed data, the estimated missing values and previously estimated model parameters. The resulting sequence forms a Markov chain $\{Y_{mis}^1, \theta^1; Y_{mis}^2, \theta^2; \dots; Y_{mis}^{t+1}, \theta^{t+1}\}$, which must converge to the conditional distribution $p(Y_{mis}|Y_{obs}, \theta)$ and is used to impute missing values (Horton & Lipsitz, 2001; Jackman, 2000).

MVNI works properly under the MAR assumption and can handle both continuous and categorical missing data whose distributions are not normal (Allison, 2001; Graham, 2009; Lee et al., 2012). According to Allison (2001), dichotomous variables, which are normally represented by dummies (0 or 1) can be imputed as continuous variables and the imputed values rounded to the nearest integer (0 or 1). When categorical variables contain more than two levels, they are dichotomised first before being imputed. Then $K - 1$ variables, where K is the number of variables, are included in the imputation model, leaving the other category as a reference. To illustrate the procedure, let no method (NM), traditional (T) and modern (M) be the categories forming contraceptive method use status respectively. To impute this variable using MVNI, dummies for NM, T and M are created, and T and M new variables are included in the imputation model, leaving NM as a reference category. Imputed values (of T and M) are used to produce the final coding as shown in Table 3.1.

TABLE 3.1: Imputation of categorical variables with more than two levels

Imputed values		Reference	Final values	
N	M	1-N-M	N	M
0.3	0.4	0.3	0	1
0.6	0.4	0	1	0
-0.1	0.3	0.8	0	0

Suppose that the values in Table 3.1, represent the imputed values of the dummies T and M. To produce the final imputed values, the values of these dummies are subtracted from 1 (1-T-M) where the value (1-T-M) is considered as a reference, and the following rule is used to determine the final coding of the imputed values.

1. Determine the category with the highest imputed value.

2. If the highest value corresponds to the reference category (1-T-M), assign a value of 0 to each dummy variable (T and M), otherwise assign a value of 1 to the dummy with the highest value (a value greater than the values of the other dummy and reference category) and 0 for the other dummy variable.

3.4.3 Description of multiple imputation by chained equations

As any other imputation method, the MICE technique discards observations with no information at all. This makes sense because if there is no information provided on the variables to be used, regression coefficients for instance cannot be fitted. However, when information is partially missing on these variables, the procedure works as follows: (1) For all missing observations in the data set, missing values are filled in with random draws from the observed values first or a simple imputation such as the mean imputation is done for every missing value in a data set ([Azur et al., 2011](#)). (2) By moving through the columns of variables, a single variable imputation is performed using a method such as regression imputation. The obtained new guess is temporally used to fill in the missing value of the variable on which the regression was performed. Note that as we go along, previous guesses are used in the regressions of other variables to be imputed until the whole data set is imputed. (3) The new fitted values are used as replacements to the original inputs in stage (1). (4) The process is repeated until a certain number of cycles is completed or until convergence is attained (or until the distribution of parameters governing the imputations becomes stable). By repeating steps 1 – 4 above m times, m imputed data sets are generated and analyzed using simple rules ([Little & Rubin, 2002](#)). According to [Raghunathan et al. \(2001\)](#), ten cycles are generally performed. However, as suggested by [Azur et al. \(2011\)](#), research is needed to determine the best possible number of cycles required to impute data under different conditions. On the other hand, the number of m imputed data sets depends on the size of the data set and the amount of missing information in the data set.

When generating imputations, a linear regression model is used for continuous data, a logistic regression is applied for binary variables, multinomial logistic and Poisson regressions are utilised for polytomous and count variables respectively. With this method, the choice of regression models depend on the nature of the variables to be imputed. [White et al. \(2011\)](#) has suggested the way imputation

of variables should be done. As indicated these variables are imputed as follows:

Suppose that data are missing at random on the random variable Y whose missing values have to be imputed from other variables, say $X = (X_1, X_2, \dots, X_k)$. To make the imputation simple, assume that the variable X includes the column of ones, so that one can be able to estimate the k number of regression coefficients. Let also n_{obs} be the number of individuals with available observations of Y . When Y is a normally distributed continuous variable, a linear regression model is used to impute missing values. That is,

$$Y|X; \beta \sim N(\beta X, \sigma^2). \quad (3.9)$$

Suppose that $\hat{\beta}$ is the estimated parameter which is a row vector of length k from fitting this model to individuals with observed values of Y . Let V and $\hat{\sigma}$ be the estimated covariance matrix of $\hat{\beta}$ and the estimated root mean-square respectively. The imputation parameters σ^* and β^* are drawn from the joint distribution of σ and β . In the first case, σ^* is drawn as:

$$\sigma^* = \hat{\sigma} \sqrt{\frac{n_{obs} - k}{g}} \quad (3.10)$$

where g is a random draw from a chi-square (χ^2) distribution with $n_{obs} - k$ degrees of freedom. In the second case, β^* is drawn as

$$\beta^* = \hat{\beta} \frac{\sigma^*}{\hat{\sigma}} u_1 V^{\frac{1}{2}} \quad (3.11)$$

where u_1 refers to the row vector of k independent draws from the standard normal distribution and $V^{\frac{1}{2}}$ is the Cholesky decomposition of V . The imputed values Y_i^* are then obtained as

$$Y_i^* = \beta^* X_i + u_{2i} \sigma^* \quad (3.12)$$

where u_{2i} is randomly drawn from a standard normal distribution.

When Y is a binary variable, a logistic regression model is used to impute it given X . Technically, the model is given by

$$\text{logit}P(Y = 1|X; \beta) = \beta X. \quad (3.13)$$

Suppose that $\hat{\beta}$ is the estimated parameter that is obtained from fitting the logistic regression model above to the observed values of Y , with estimated variance–covariance matrix V . Suppose also that β^* is a draw from the posterior distribution of β ,

which is approximated by $MVN(\hat{\beta}, V)$ according to (Little & Rubin, 1989). For each missing observation Y_i , an imputed value Y_i^* is drawn as

$$Y_i^* = \begin{cases} 1 & \text{if } u_i < p_i^* \\ 0 & \text{otherwise} \end{cases}$$

where p_i^* is estimated by

$$p_i^* = [1 + \exp(-\beta^*)X_i]^{-1} \quad (3.14)$$

and u_i is random sample from a uniform distribution over the interval $(0, 1)$ or $U(0,1)$. When Y is an unordered or nominal variable with more than two levels or categories, say $L > 2$, a multinomial logistic regression model is used to impute missing values. In this case, each of the categories forming this variable, has a logistic regression equation that compares it with the baseline category:

$$P(Y = l|X; \beta) = \left[\sum_{l'=1}^n \exp(\beta_{l'} X) \right]^{-1} \quad (3.15)$$

where β_l is a vector that has dimension $k = \dim(X)$ and $\beta_1 = 0$. Suppose that β^* is a random draw from the normal approximation of $\beta = (\beta_2, \dots, \beta_L)$, which is a vector of length $k(L-1)$. For each missing value Y_i , assume that $p_{il}^* = P(Y_i = l|X_i; \beta^*)$ ($l = 1, \dots, L$) is the drawn class membership probabilities and $c_{il} = \sum_{l'=1}^l p_{il'}^*$. The imputed values are given by

$$Y_i^* = 1 + \sum_{l=1}^{L-1} I(u_i > c_{il}) \quad (3.16)$$

where u_i is randomly drawn from the uniform distribution $U(0, 1)$ and $I(u_i > c_{il}) = 1$ if $u_i > c_{il}$, and 0 otherwise.

MICE Bayesian data augmentation (Little & Rubin, 1989) is another technique that is considered the best in terms of drawing imputations from their predictive distributions, and can be compared to a Markov chain. It starts with imputation values (obtained from the mean imputation for example) and update each imputation based on the state of the rest of the imputed values. For instance, given variables Y , X_1 , X_2 and X_3 with initial values chosen randomly. To impute X_3 the values of Y , X_1 and X_2 are used to generate imputation values. To impute X_2 , the imputed values of X_3 are used together with the values of Y and

X_1 and so on. This is in fact how the Markov chain is updated using the Gibbs sampling method (Gelfand et al., 1990; Geman & Geman, 1984). In other words, the previous states of the Markov chains are utilised plus any update that was already made about this particular iteration to create a new link in the chain for the variable of interest.

As this method is compared to a Markov chain, the aim is to build Markov chains as part of the Bayesian estimates to draw samples from the posterior distribution in order to derive inferences about that posterior. To build a missing data model that fits the Bayesian approach, missing values are treated as other parameters to estimate by drawing them from their posterior distribution. The model is as follows:

$$f(\beta, Y_{mis}|Y_{obs}) \propto f(Y_{obs}|\beta, Y_{mis})f(\beta, Y_{mis}) \quad (3.17)$$

where β denotes the model parameters, Y_{mis} and Y_{obs} are the missing part and observed data respectively. Equation (3.17) says that $f(\beta, Y_{mis})$ is a function of the data at hand rather than $f(\beta)$. It is written as $f(\beta, Y_{mis}|Y_{obs})$ and is proportional to the distribution of observed data conditional on model parameters β and the missing values, Y_{mis} , times some prior about β and Y_{mis} . This proportional relationship between the right and the left of the equation (3.17) constitutes a key to the Bayes law (Rubin, 1978). In this case, what is done (which is an augmentation process) is sampling not only the model parameters β , but also the missing values, Y_{mis} , out of the posterior distribution using the MCMC procedure. By doing so, many samples of Y_{mis} and β are obtained. The draws or samples of Y_{mis} serve as imputations or filled in missing values.

Despite the growing popularity of MICE, it lacks theoretical justification (Raghunathan et al., 2001). One concern is the incompatibility among the conditional models; that is, the possibility that there is no joint distribution with the conditionals of the assumed forms (He, 2010). However, as suggested by Brand (1999) and Schafer and Graham (2002), this should not be a big problem in applied settings. A number of researchers continuously use this technique as they believe that it is the right method to handle any missing data given its flexibility and capability to be used in a broad range of settings (Azur et al., 2011; Lee & Carlin, 2010; Twisk et al., 2013). MICE works under the assumption that data are missing at random (MAR) and unbiased results can only be obtained when this assumption is met.

3.4.4 Multivariate normal imputation versus multiple imputation by chained equation: a practical example using a survey data set to impute missing values of continuous variables

3.4.4.1 Introduction

Prior to investigating the performance of MVNI and MICE on unordered data, these methods were compared in terms of parameters' estimation and standard errors of the regression models estimated when data are missing completely at random on normally distributed continuous random variables. Previous studies have already compared these two techniques and have indicated that these two methods produce similar results when data are missing on continuous and normally distributed data ([Kropko et al., 2014](#); [Raghunathan et al., 2001](#)). Based on a complex data set for the current study contained in the subsequent chapter, the comparison of the two methods was verified using simulated data sets with different rates of missing completely at random data (5%, 10%, 15%, 20%, 25%, 30%, 35% and 40%). This was done to assess whether the amount of missing values can have an impact on the performance of these two methods, a fact that was not investigated by these authors. Furthermore, it was also intended to determine the impact of the amount of missing values on the inferences through the p-values of the models estimated using data sets generated using different rates of missingness (5%, 10%, 15%, 20%, 25%, 30%, 35% and 40%). The findings from this analysis were published in [Karangwa and Kotze \(2013\)](#) and could strengthen the existing knowledge about multiple imputation of missing values on continuous variables using MVNI and MICE.

3.4.4.2 Data

Throughout this thesis, the DHS conducted in the Democratic Republic of Congo (DRC) in 2007 was used. It consists of a household and women's questionnaire where a sample of women between 15 and 49 years of age were interviewed regardless of their marital status in each sampled household. Information was collected on fertility and family planning in addition to socio-demographic and economic data. The sample of women in the analysis included women of reproductive age

who were not pregnant at the time of interview and who were sexually active. Respondents were asked about their knowledge and use of contraception methods amongst other things. Information on whether they have ever used contraception was first obtained and then the types of contraceptive methods used were asked. Contraceptive methods used included the modern (i.e. pill, injections and other), traditional (i.e. abstinence and other) and folkloric (i.e. herbal plant and other) methods. The purpose was to determine whether and to what extent certain covariates such as her marital status, are associated with the woman's use of contraception. Thus, slopes, standard errors and results of hypothesis tests were considered as outcomes of interest to be analysed.

In order to assess the performance of the MVNI and MICE techniques, data sets were created with partially missing completely at random (MCAR) data on variables age and education that were statistically significant in the regression model estimated with the data set with no missing values. This assumption means that missingness probabilities are not related at all to any variable in the data set.

Eight data sets were generated with different rates of missingness; 5%, 10%, 15%, 20%, 25%, 30%, 35% and 40% on variables age and education. Therefore, based on the variables of interest (age and education) with no missing data, a 0 – 1 random generator if the observation was missing (1) or not (0) was constructed if the observation was missing or not. This means that missing data are random samples from the Bernoulli distribution with the parameter p that represents the percentage of missing values of interest. Technically, this can be represented as follows: let Y_i be a complete data vector for respondent i . Then Y_i can be partitioned into $Y_{i,obs}$ and $Y_{i,mis}$, the observed and missing parts respectively. That is, $Y_i = (Y_{i,obs}, Y_{i,mis})$. Let also $R_i = (r_{ij})$ be the missing data indicator, where $r_{ij} = 1$ if a value is missing and $r_{ij} = 0$ otherwise. Given some parameter θ , the MCAR assumption states that

$$P(r_{ij}|Y_{i,obs}, Y_{i,mis}, \theta_i) = P(r_{ij}) \quad (3.18)$$

which in words means that the distribution of missingness does not depend on the data at all. This can be seen as a Bernoulli distribution with the following probability density function:

$$\prod p(r_{ij}|\theta_{r_j}) = \theta_j^{r_j}(1 - \theta_j^{r_j})^{1-r_j} = p_i^{r_j} q^{1-r_j} \quad (3.19)$$

for $q = 1 - p_i^{rj}$ and $rj \subset (0, 1)$.

The purpose of creating the different data sets with different rates of missing values was to investigate whether the amount of missing data can have an impact on the multiple imputation methods of interest.

3.4.4.3 Analysis method

For each simulated data set with missing values on the variables of interest (age and education in years), the MVNI method which assumes normality of the variables in the imputation model and the MICE technique which uses a sequence of regression models to impute missing values, were applied. The regression models were estimated using the data set with no missing values, data sets with missing values (incomplete data), as well as imputed (observed + imputed) data sets. The results were compared in terms of slopes, standard errors and p-values. The MVNI method was performed using the STATA implementation of Schafer's NORM program (Galati & Carlin, 2009) whereas the MICE was carried out using the mice command in STATA (Van Buuren et al., 1999).

3.4.4.4 Findings

Table 3.2 presents the results of the parameters' estimation, standard errors and p-values of the two slopes from the binary regression of women's contraceptive use status (dependent variable) on their age and education in years (independent variables). The results of the data set with no missing data and data sets with missing values (8 datasets with 5%, 10%, 15%, 20%, 25%, 30%, 35% and 40% missing completely at random observations on age and education respectively). As expected, the results indicate that multiple imputation-based methods, namely MVNI and MICE, produce less biased estimates than the case deletion (CD) method, which discard the items with missing values from the analysis. It is also shown that the MVNI and MICE yield similar parameter estimates (Figures 3.1 and 3.2) when applied to continuous and normally distributed variables. Furthermore, the results indicate that all the missing data methods considered in the analysis overestimate the standard errors of the models (Figures 3.3 and 3.4). In Figures 3.3 and 3.4, it is observed that the CD produces larger standard errors than the MVNI and MICE, and the larger the percentage of missing data, the more inflated standard errors are obtained. In Figure 3.5, the p-values of education are reported. It is shown

that at lower percentages of missing values (at most 15%), the three missing data methods produced similar and unbiased p-values, otherwise the CD yields higher p-values. The findings indicate also that the higher the percentage of missing data, the more the relationship between the dependent and the independent variables (which were all statistically significant in the regression model of the data set with no missing data) is distorted when missing observations are excluded from the analysis. Indeed, the results show that when at least 20% of observations are missing on independent variables, some of them lose their statistically significant relationship with the dependent variable. Finally, it is also observed that at some stage (when at least 25% of the data are missing), neither the imputation methods used nor the CD can help to maintain the relationship that exists between the dependent and independent variables when the analysis is done using the data set with no missing values (see Figure 3.2). This shows that at some stage, the missing values techniques may not be successful and therefore the data users may be forced to give up on the data set that was intended to be used.



TABLE 3.2: Parameter estimates of a set of logistic regression models for predicting the contraceptive methods use status by women of reproductive age in Democratic Republic of Congo in 2007, using age (1st covariate) and education (2nd covariate) in years as explanatory variables

Proportion of missing data	Slopes			Std errors			P-values		
	CD	MVNI	MICE	CD	MVNI	MICE	CD	MVNI	MICE
0.00	0.0261	NA	NA	0.0029	NA	NA	0.000*	NA	NA
	0.0067	NA	NA	0.0026	NA	NA	0.011*	NA	NA
0.05	0.0272	0.0267	0.0272	0.0032	0.0030	0.0030	0.000*	0.000*	0.000*
	0.0067	0.0069	0.0073	0.0028	0.0027	0.0027	0.017*	0.010*	0.007*
0.10	0.0251	0.0269	0.0271	0.0033	0.0032	0.0031	0.000*	0.000*	0.000*
	0.0065	0.0067	0.0071	0.0029	0.0027	0.0028	0.028*	0.014*	0.011*
0.15	0.0302	0.0296	0.0284	0.0037	0.0034	0.0034	0.000*	0.000*	0.000*
	0.0065	0.0078	0.0076	0.0031	0.0031	0.0028	0.021*	0.006*	0.006*
0.20	0.0293	0.0290	0.0279	0.0041	0.0037	0.0035	0.000*	0.000*	0.000*
	0.0045	0.0059	0.0052	0.0033	0.0030	0.0028	0.172	0.045*	0.066
0.25	0.0275	0.0268	0.0279	0.0041	0.0033	0.0036	0.000*	0.000*	0.000*
	0.0027	0.0045	0.0038	0.0035	0.0028	0.0031	0.451	0.105	0.075
0.30	0.0252	0.0263	0.0262	0.0041	0.0038	0.0037	0.000*	0.000*	0.000*
	0.0029	0.0043	0.0038	0.0037	0.0031	0.0032	0.433	0.169	0.228
0.35	0.0262	0.0264	0.0265	0.0047	0.0034	0.0038	0.000*	0.000*	0.000*
	0.0023	0.0043	0.0041	0.0040	0.0032	0.0032	0.564	0.177	0.211
0.40	0.0262	0.0248	0.0250	0.0050	0.0041	0.0039	0.000*	0.000*	0.000*
	0.0008	0.0047	0.0043	0.0044	0.0033	0.0037	0.855	0.155	0.249

*: Significant at 5% level.

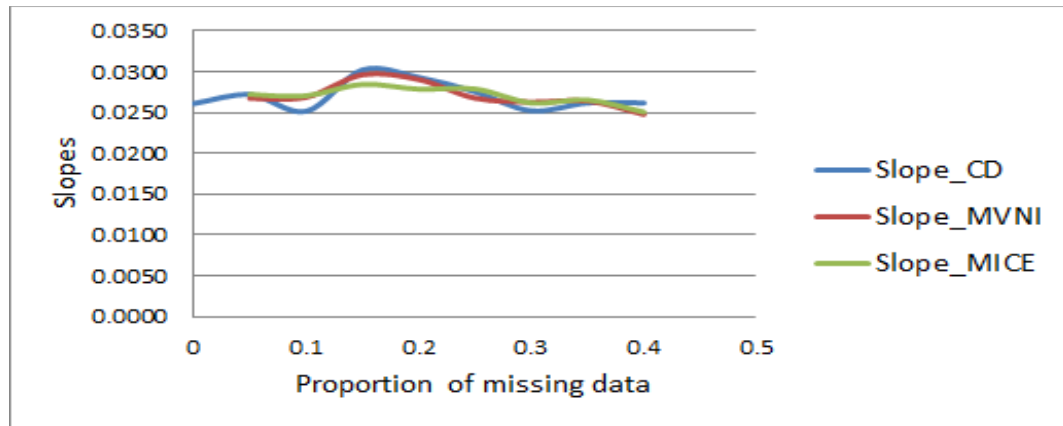


FIGURE 3.1: Estimates of slopes for age when the CD, MVNI and MICE methods are used at different rates of missingness

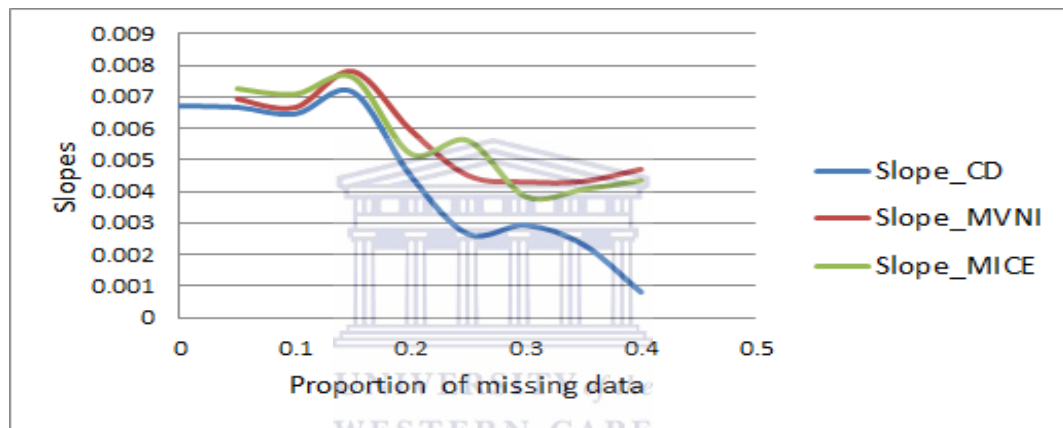


FIGURE 3.2: Estimates of slopes for education when the CD, MVNI and MICE methods are used at different rates of missingness

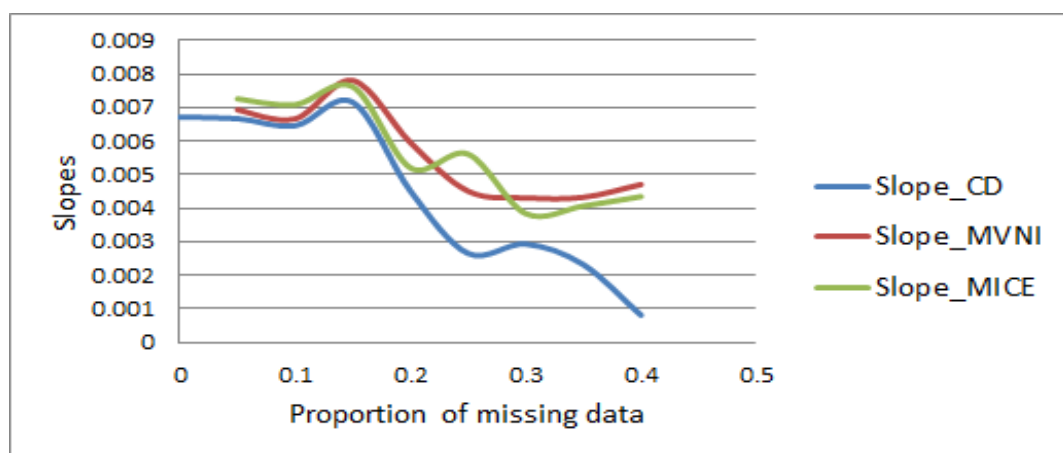


FIGURE 3.3: Estimates of standard errors for age when the CD, MVNI and MICE methods are used at different rates of missingness

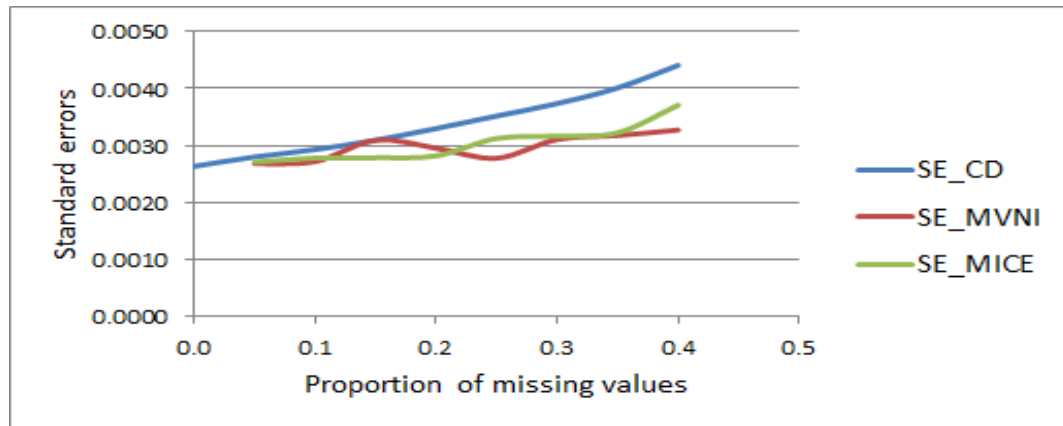


FIGURE 3.4: Estimates of standard errors for education when the CD, MVNI and MICE methods are used at different rates of missingness

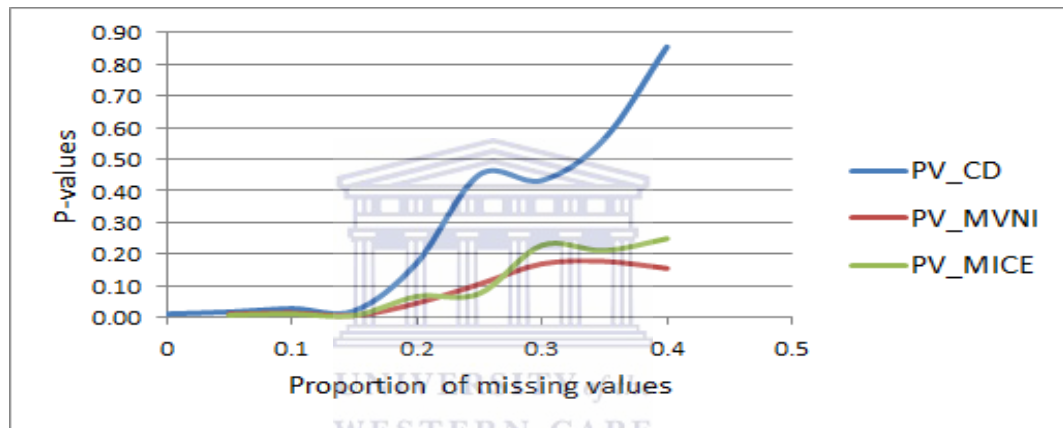


FIGURE 3.5: P-values of the models estimated using the CD, MVNI and MICE methods at different rates of missingness

3.4.4.5 Conclusion

In general, the results indicated that the higher the proportion of missing data, the more the relationship between variables is distorted when missing data techniques (CD, MVNI and MICE) are used to impute continuous variables containing missing values. As expected, it was also shown that the multiple imputation methods of interest; MVNI and MICE, yield less biased estimates than the CD method which discard items with missing values from the analysis. Furthermore, the findings indicated that the MVNI and the MICE produce similar parameter estimates as was noted by [Van Buuren \(2007\)](#) but the MVNI is better in terms of preserving an existing relationship between variables at higher rates of missing values (at least 25% missing data). Finally, it was found that at some stage (when the proportion of missing data becomes high), neither the imputation methods nor the case deletion method can help to maintain the existing relationship between variables in the models estimated using the data set with no missing values.

It is important to note at this stage that although continuous variables were used in Section 3.4.4, the investigation into the behaviour of MVNI and MICE for different rates of missingness (5%, 10%, 15%, 20%, 25%, 30%, 35% and 40%) ([Karangwa et al., 2015](#)) contributed to the new knowledge with respect to these methods in the continuous case. These continuous comparisons were the starting point of the exploration for the scenario where unordered categorical data and variables were used in a similar exercise (Model 1.1 of Section 5.2)

3.4.5 Multivariate normal imputation versus multiple imputation by chained equations of categorical data

A categorical variable is a variable that has a measurement scale consisting of a set of categories ([Agresti, 2001](#)). For instance, the marital status of an individual is frequently measured as single, married, divorced, widow, living together and not living together. The wealth status of respondents in demographic health surveys is categorized as poorest, poorer, poor, rich, richer and richest. These kinds of variables are ubiquitous in survey data sets in fields such as epidemiology and public health (eg. types of contraceptive methods used), demography (eg. region where the respondent comes from), biostatistics (eg. types of diseases), marketing (eg. brands of cars such as BMW, VW, TOYOTA, etc.), and so on.

There are two types of scales of categorical variables, namely nominal and ordinal scaled variables. Nominal variables have no natural ordering whereas ordinal variables are characterized by ordered categories. For nominal variables, examples include the person's religious affiliation (eg. catholic, muslim, etc.), food choice (fruits, vegetables, cereals, etc), means of transport (bicycle, car, train, etc) and region or province of origin. Ordinal data include the Likert-scale (strongly disagree, disagree, agree and strongly agree) which is used to study perceptions, attitudes or opinions of people, educational attainment (no schooling, primary, secondary and tertiary education) amongst others.

As stated by [Agresti \(2002\)](#), when listing categories of nominal variables, the order does not matter and statistical tests performed on these variables do not take into consideration this order. These methods can be used for ordinal data by just ignoring the order of the categories but not vice versa. Ordinal data can also be statistically analysed using methods designed for continuous data as they are believed to have some important quantitative features; that is, each category has a smaller or a bigger magnitude of characteristic than another category that is not easy to measure but in some way has a continuous nature.

In missing data analysis, methods designed to handle missing data of continuous variables have also been applied to missing data of ordinal and binary variables ([Finch, 2010](#); [Lee & Carlin, 2010](#)). As suggested by these authors, there is a need to look at the performance of these methods when data are missing on non-ordered or nominal variables with more than two categories.

As mentioned earlier, the two multiple imputation methods; MVNI and MICE, are increasingly being used to fill in missing values of both continuous and categorical variables. They have been made popular in almost all the main statistical software packages such as SAS, STATA, etc. They are currently considered the best as they account for the statistical uncertainty in the imputations, which is not the case when single-based imputation methods are used ([Lee & Carlin, 2010](#)).

A number of studies compared these methods ([Lee & Carlin, 2010](#); [Van Buuren, 2007](#); [Yu & Schaid, 2007](#)). These studies concentrated on different aspects and applied the two methods to data with a mixture of variables (continuous, discrete and semi-continuous). Mixed results were obtained; some concluded that the MVNI was better than the MICE ([Lee & Carlin, 2010](#); [Van Buuren, 2007](#); [Yu & Schaid, 2007](#)) and others found the opposite ([Demirtas et al., 2008](#); [Kropko et al., 2014](#)).

The MVNI was initially designed to handle missing data of continuous and normally distributed variables, but it was later used to impute missing values of categorical data which do not assume normality (Allison, 2001). It works properly under the MAR assumption and can handle both continuous and categorical missing data although the latter do not assume normality (Allison, 2001; Graham, 2009). Given for instance a binary or a two-level categorical variable coded as 1 and 0, the proportion of responses with 1s will be the same as the mean of that variable. Therefore, unbiased estimates for the variables are obtained even if multiple imputation-based models that assume normality are used. When a two-level categorical variable is used as a covariate or independent variable in regression analysis, the imputed values should be used without rounding. If this variable is to be used in the analysis as a discrete binary variable, then rounding should be done to the nearest value (0 or 1) as suggested by Bernaards et al. (2007). For categorical variables with more than two levels, these need to be dummy-coded first and $K - 1$ (where K is the number of categories) dummy variables are included in the imputation model (Allison, 2001). For example, if a variable such as marital status with six categories (never married, married, divorced, widow, living together and not living together) contains missing values and therefore needs to be imputed, it has to be dichotomized to obtain dummies for never married, married, divorced, widow and living together respectively. The imputation is done with only these five variables and filled-in values are used to produce final coding, while the sixth category (not living together) is treated as a reference category.

On the other hand, MICE fills in missing values sequentially, taking into account the distributional form of the variables to be imputed. That is, a linear regression is used for continuous variables, a binary logistic regression is suitable for binary variables, whereas the ordinal and multinomial regressions are used for ordinal and unordered or nominal variables respectively. The detail about this method is also given in Van Buuren (2007), White et al. (2011), Carpenter and Kenward (2012) and Kropko et al. (2014) amongst others.

Several studies have compared MVNI and MICE in terms of parameter estimation and standard errors and have indicated that these two methods produce similar results when data are missing on continuous and normally distributed data (Kropko et al., 2014; Raghunathan et al., 2001). The multivariate normal imputation outperformed the multiple imputation by chained equations when data were missing on ordinal data (Finch, 2010; Lee & Carlin, 2010) and on binary variables Lee and Carlin (2010). As suggested by these two authors, an empirical study is

still needed to determine the performance of these two methods when data are missing on non-ordered or nominal categorical variables. This study was designed based on the recommendations from these authors in 2011. Three years later, [Kropko et al. \(2014\)](#) attempted to compare the performance of these methods when data were missing at random on continuous, binary, ordinal and unordered categorical variables that were used as outcome variables in the regression models. Their findings indicated that MICE performed better than MVNI in terms of regression coefficients' accuracy.

This study considered the suggestion by [Finch \(2010\)](#) and [Lee and Carlin \(2010\)](#) and extended the analysis by [Kropko et al. \(2014\)](#) to missing at random or missing completely at random data on unordered or nominal variables treated as either predictors or response variables in the regression models.

3.5 Summary of the chapter

In this chapter, the literature related to methods that are used to fill in missing values of items with partially missing data were reviewed. These included single-based, model-based methods and multiple imputation methods. A brief description of the two multiple imputation methods that have been made popular in most of the statistical software packages, and constitute a subject of an ongoing research regarding their performance in terms of parameter estimates' accuracy were discussed. Using a practical example with a real survey data set, the performance of these methods was compared in terms of regression coefficients and standard errors. The results indicated that MVNI and MICE produce similar results when data are missing completely at random on continuous data. The main objective here is to illustrate the performance of these methods when data are missing on continuous variables and to address the knowledge gap of these methods when it comes to treating missing values of categorical data measured on nominal scales and are used as either response or independent variables in the regression models.

In the next chapter, the methodology used for the analysis is presented. The data sets used for analysis are described and the stochastic models used to impute missing values of the variables of interest are specified.

Chapter 4

Methodology

This chapter outlines the methodology used in this study. The initial section (Section 4.1) describes the data set and variables used. Section 4.2 explains how data sets with missing data were generated. The missing data models considered in the analysis are highlighted in Section 4.3, whilst the analysis methods used are covered in Section 4.4. The stochastic models used to impute variables with missing values are specified in this section. The model development, computation of the performance measures for the CD, MVNI and MICE, as well as the imputation models diagnostics are also explained in this section. A summary of the whole chapter is provided in Section 4.5.

4.1 Description of data set and variables used in the study

The data set described in Chapter 3 was used for analysis. The variables of interest were the woman's contraceptive method use status, marital status, region or origin, age in single years, education in completed years as well as her wealth index. The woman's contraceptive method use status was measured in two ways; (1) as a dichotomous variable that was measured as any contraceptive method used by including all women who reported using modern, traditional and folkloric methods. In this case, it was coded as 1 if a woman has ever used any contraceptive method and 0 otherwise. This variable was used as a dependent variable in the binary logistic regression models that were estimated to investigate the effect of marital

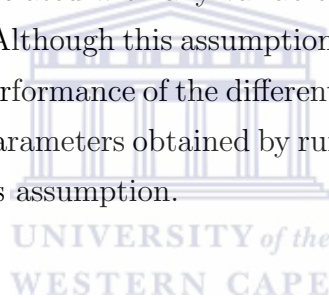
status on contraceptive method use status, controlling or not for other variables. (2) Contraceptive method use status was also used as an outcome measure coded 1 if a woman has not used any contraceptive method, 2 if she has used a traditional or folkloric method and 3 if she has used a modern contraceptive method. This variable was used to determine the association between contraceptive method use and marital status through the multinomial logistic regression models. Marital status has six categories, namely: never married, married, divorced, widow, living together and not living together, while wealth index has five levels, which are: poorest, poorer, middle, richer, and richest. The variable region on the other hand contains eleven categories; Kinshasa, Bas Kongo, Bandundu, Equateur, Oriental, Nord Kivu, Maniema, Sud Kivu, Katanga, Kasai Oriental and Kasai Occidental. Age and education completed in years are both continuous variables. Table 4.1 provides the description of these variables. These variables were used either in the estimation of the regression models or multiple imputation models of interest.

TABLE 4.1: Description of the variables used in the study

Variable	Description	Type
Contraceptive method use status 1	Do not use any contraceptive method, use a contraceptive method	Nominal (dichotomous)
Contraceptive method use status 2	Do not use any contraceptive method, use a traditional or folkloric method, use a modern method	Nominal (multinomial)
Marital status	Never married, married, divorced, widow, living together and not living together	Nominal (multinomial)
Region	Kinshasa, Bas Kongo, Bandundu, Equateur, Oriental, Nord Kivu, Maniema, Sud Kivu, Katanga, Kasai Oriental, Kasai Occidental	Nominal (multinomial)
Wealth index	Poorest, poorer, middle, richer and richest	Ordinal
Age	Age in single years	Continuous
Education	Number of years spent at school	Continuous

4.2 Simulation of the data sets with missing values

Throughout this study, missing values are assumed to be ignorable, which means that based on the available data, they can be estimated. Therefore, to obtain data sets with missing values, a data set with no missing values or baseline data set was used to simulate data sets containing MAR or MCAR values on the variables of interest. Data sets with values missing at random (MAR) were arbitrary created in such a way that missingness was related to variables of interest but not on the values of the variables that had missing data. Concerning the MCAR mechanism, data are a random sample where missing values are a result of random reasons, independent of observed and unobserved variables. Therefore, to obtain data sets with missing values according to this assumption, values were deleted such that missingness was not associated with any variable in the data set that was used as explained in Chapter 3. Although this assumption is accused of being unrealistic, it can give insight on the performance of the different imputation methods concerning point estimation of the parameters obtained by running different models using data sets generated under this assumption.



4.3 Missing data models

This study considered two scenarios (see Figure 4.1). The first scenario contains two binary logistic regression models (Models 1.1 and 1.2 in Figure 4.1). In the first model (Model 1.1), a single covariate (marital status) containing missing values is considered. This was done to assess the performance of the multiple imputation methods of interest on the unordered categorical variable alone, with no influence of other covariates. This model was also used to investigate whether the rates of missingness can impact on the performance of the multiple imputation techniques used in this study. This fact is true according to the literature (Karangwa & Kotze, 2013), but we believe that an empirical study is needed to confirm it when missing values are MAR or MCAR on unordered categorical variables. Therefore data sets with 50%, 30% and 10% missing values were considered for analysis for only this model, whereas only 50% rate of missing values was considered for other models. The second model or Model 1.2 in Scenario 1 is split into two models,

namely Models 1.2.1 and 1.2.2 as shown in Figure 4.1. Model 1.2.1 considers a binary logistic regression model with two covariates measured on nominal scale, whereas in Model 1.2.2 various covariates (nominal, continuous and ordinal) are used, with missing values on only the nominal or unordered categorical ones. The aim of considering these models was to assess the performance of MICE and MVNI when missing values are present on unordered categorical variables that are used as independent variables alone or with other covariates to see if the introduction of other covariates can impact on the performance of these methods. The second scenario contains two regression models as well, but in this case with missing values on the outcome variables (binary and polytomous) that have no natural order. The first model or Model 2.1 in Figure 4.1 is a binary logistic regression model with missing values on the binary outcome variable. The second model or Model 2.2 is a multinomial logistic regression model of the response variable (contraceptive method use status) with more than two levels measured on nominal scale. Both models considered a single covariate to investigate the performance of the multiple imputation methods of interest. All the models regress contraceptive method use status on marital status to investigate the effect of marital status on contraceptive method use status, in order to fully observe the performance of the multiple imputation methods of interest (MVNI and MICE) when missing values are present on the nominal categorical variables alone first and then in the presence of the other covariates in regression models.

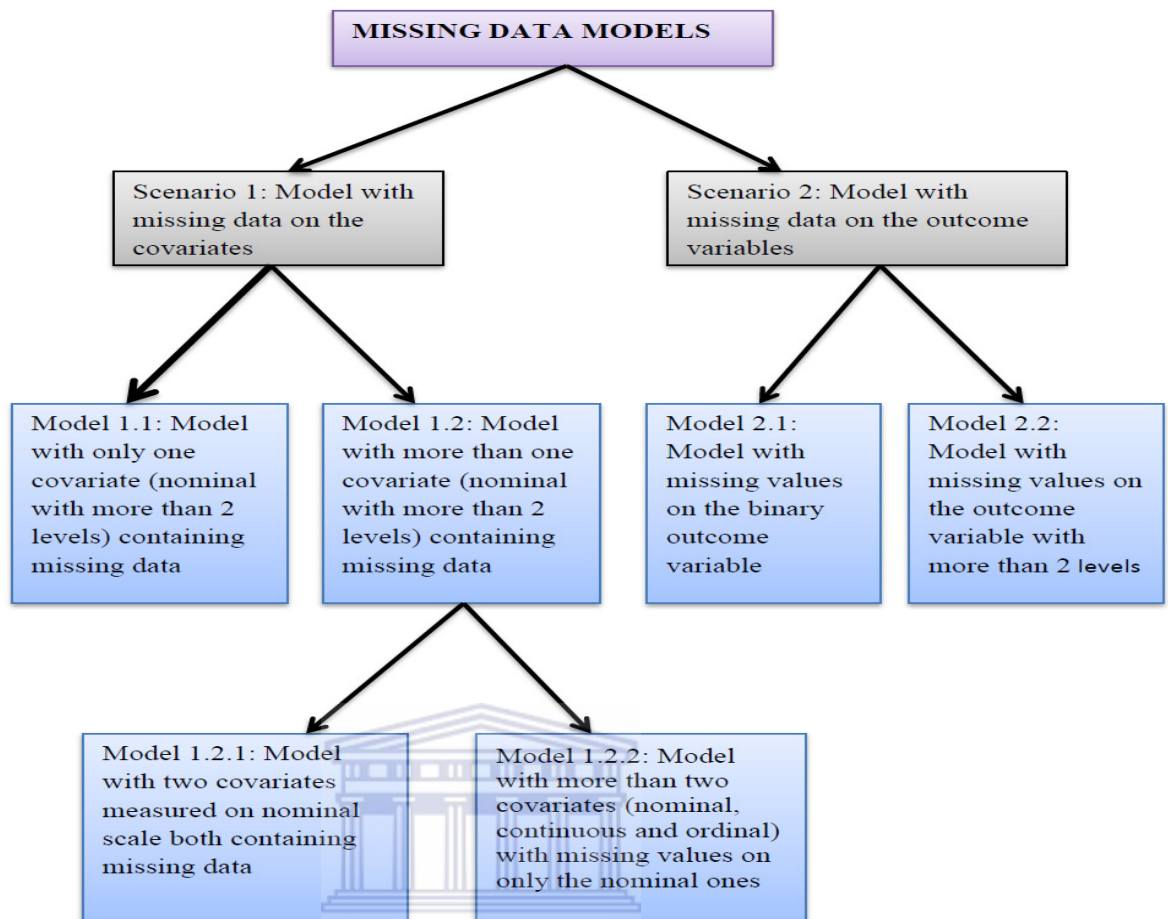


FIGURE 4.1: Missing data models considered in the analysis.

As previously explained, under the MAR assumption, data sets with missing data on the covariates were arbitrary and randomly deleted if a woman was not using any contraceptive method, thus allowing missingness to depend on contraceptive method use status. The same percentage of missing data was imposed on the data set with no missing values to simulate the MCAR data set, in this case with missingness not related to any variable used in the regression model of interest or data set used. On the other hand, missing values on the outcome variables were arbitrary deleted at random if a woman was at least 35 years old, thus allowing missingness on the outcome variables to depend on the woman's age. MCAR data were also generated such that no variables in the data set was related to missingness on the response variables. The descriptive statistics of missingness for all these cases is provided in Chapter 5.

4.4 Analysis method

4.4.1 Imputation of missing values

Existing multiple imputation methods assume that data are MAR, but as suggested by [Rubin and Schenker \(1986\)](#) and [White et al. \(2011\)](#), other missing data mechanisms such as MCAR can also be assumed if the objective is to compare the performance of multiple imputation methods. This study aimed to determine the performance of the multivariate normal model (MVNI) and the multiple imputation by chained equations (MICE) in terms of bias in the estimated regression coefficients and standard errors of the regression coefficients, when data are missing at random or missing completely at random on either the response variables or the covariates in the regression models.

Multiple imputation methods normally replace each missing value by an array of $m > 1$ pseudo random values generated by a computer algorithm ([Little & Rubin, 2002](#)) included in many statistical software packages such as SPSS, STATA and SAS. As explained in Chapter 3, these values are drawn from the posterior predictive distributions of the variables to be imputed, which allows the multiple imputations to capture the sampling variability as well as the uncertainty in the imputation model. As highlighted by [White et al. \(2011\)](#), various studies on multiple imputations state that 3 to 5 imputations are enough. However, according to these authors, larger numbers of imputations are required if the objective is to compare imputation methods, or to obtain stable and less unbiased estimates. To obtain sufficient accuracy while comparing these methods, this study used 100 imputations for each data set with missing values, which resulted in 100 different imputed simulated versions of complete data sets. Each imputed data set was analysed separately using standard statistical techniques, and the point estimates as well as standard errors were combined (averaged) to produce single estimates that account for uncertainty due to missing data as suggested by [Little and Rubin \(2002\)](#).

The imputation models included the analysis model covariates and dependent variables of interest, as well as the auxiliary variables or variables that were associated with missingness on the variables to be imputed although they were not part of the substantive or analysis model. This was done to improve the imputation quality as suggested by [White et al. \(2011\)](#) and [Enders \(2010\)](#). The regression models (binary and multinomial) with no missing data were fitted and the results

were compared to the models estimated using data sets with missing data (case deletion method) and the completed (observed + imputed) data sets using MICE and MVNI.

The stochastic models of the variables with missing values on either the outcome or the predictor variables are as follows. As previously stated, Model 1.1 is a binary logistic regression of contraceptive method use status (measured as 1 if a woman use any contraceptive method and 0 otherwise) on marital status. Marital status is an unordered or nominal variables with 6 categories; never married, married, living together, widowed, divorced and not living together.

To impute this variable using MVNI, the Allison (2001) approach was utilized. That is, it was first dichotomised before being imputed. Thus, five dummy variables of marital status (never married, living together, widowed, divorced and not living together) were included in the imputation model, treating married as a reference category. Assuming that MSL denotes the levels of marital status to be imputed, the following linear regression model was estimated for each marital status dichotomised level or category:

$$MSL = \beta_0 + \beta_1 C + \beta_2 A + \beta_3 E + \beta_4 W + \beta_5 R + \varepsilon \quad (4.1)$$

where β_i denote the regression coefficients and C, A, E, W, R and ε denote contraceptive method use status, age, education in completed years, wealth index, region and random variation respectively. This means that MSL is imputed taking in account information on contraceptive method use status, age, education in completed years, wealth index and region of origin. Recall that under the MAR assumption, missing values were created such that missingness depended on whether a woman was using a contraceptive method or not. Thus, variables that are not part of the analysis model but are related to missingness such as age, education in completed years, wealth index and region were all included in the imputation model as auxiliary variables to improve the imputation quality. The same imputation model was also used when data were MCAR as all these variables are associated with the dependent variable (contraceptive method use status) that was used in the substantive model. The predicted values from the regression estimates were used to impute missing values of the dichotomised categories forming marital status. Imputed values were treated as explained in Chapter 3 (imputation of categorical variables with more than 2 categories using MVNI).

Under the MICE, marital status was imputed using the following multinomial logistic regression model of marital status on the same variables as in equation

4.1:

$$MS = \beta_0 + \beta_1 C + \beta_2 A + \beta_3 E + \beta_4 W + \beta_5 R + \varepsilon \quad (4.2)$$

where MS denotes marital status and other terms are as explained in equation 4.1.

Models 1.2.1 in Scenario 1 regress contraceptive method use status on two nominal variables containing missing values, namely marital status and region. Under the MVNI method, marital status and region were dichotomised before being imputed. Marital status was dichotomised as in Model 1.1, whereas dummy variables of region were Bas Kongo, Bandundu, Equateur, Oriental, Nord Kivu, Maniema, Sud Kivu, Katanga, Kasai Oriental and Kasai Occidental, leaving Kinshasa a reference category. The stochastic imputation model of these variables is similar to the one specified in equation 4.1, except that they were imputed conditional on contraceptive method use status, age, education and wealth index. Under MICE technique, these variables were imputed using a multinomial logistic regression conditional on contraceptive method use status, age, education and wealth index. Model 1.2.2 was imputed as Model 1.2.1, as missing values were observed on the same variables.

Imputation methods of interest were also compared when data were missing on nominal variables that were treated as response variables in the regression models. For the first case (Model 2.1), missing values were deleted on the binary outcome variable (contraceptive method use status) coded 1 if a woman has used a contraceptive method and 0 otherwise. As suggested by several authors, variables with binary outcomes can be imputed using the parametric-based imputation method MVNI (Catellier et al., 2005; Efron, 1994). Therefore, under the MVNI method, this variable was imputed taking into account the suggestions of these authors. The following linear regression model was used to impute contraceptive method use status:

$$C = \beta_0 + \beta_1 MS + \beta_2 A + \beta_3 E + \beta_4 W + \beta_5 R + \varepsilon \quad (4.3)$$

where C, MS, A, E, W, R and ε denote contraceptive method use status, marital status, age, education in completed years, wealth index, region and random variation respectively. The generated imputations were rounded to the nearest integer (0 or 1) to keep the dichotomy nature of the imputed variable, as it had to be used as a dependent variable in the substantive binary logistic regression model.

Under MICE, imputation values of contraceptive method use status (dichotomous variable) were drawn using the following binary logistic regression model:

$$C = \beta_0 + \beta_1 MS + \beta_2 A + \beta_3 E + \beta_4 W + \beta_5 R + \varepsilon \quad (4.4)$$

where the terms are defined as in equation 4.3.

The performance of the multiple imputation methods of interest was also evaluated when missing values were missing on unordered or nominal variables with more than two categories (three categories in this case), treated as outcome variables in the regression models (see Model 2.2). For this particular case, contraceptive method use status, coded 1 if a woman did not use any contraceptive method, 2 if she used a traditional or folkloric method 3 if she utilised a modern method, was used. Under the MICE method, the following multinomial logistic regression model was utilised to impute this variable:

$$C^* = \beta_0 + \beta_1 MS + \beta_2 A + \beta_3 E + \beta_4 W + \beta_5 R + \varepsilon \quad (4.5)$$

where C^* denotes contraceptive method use status with 3 categories and other terms are defined as in equations 4.3 and 4.4.

Under the MVNI technique, contraceptive method use status was dichotomised and two of its levels (traditional and modern contraceptive method use) were imputed as continuous variables using a linear regression model, leaving the other category (use no contraceptive method) as a reference. The model is as follows:

$$C_L = \beta_0 + \beta_1 MS + \beta_2 A + \beta_3 E + \beta_4 W + \beta_5 R + \varepsilon \quad (4.6)$$

where C_L stands for contraceptive method levels or categories to be imputed.

As mentioned previously, this study used a DHS data set, which is a complex survey with a complex sampling design and weighting procedure that need to be taken into consideration during the analysis. In survey sampling, all the units do not have the same probabilities to be included in the sample. With complex survey data sets, these probabilities are computed and then used to calculate the sample weights. As an example, consider a population for which the estimator of the total X is as follows:

$$\hat{X}_w = \sum_{i=1}^n w_i x_i \quad (4.7)$$

where x_i represent the observed values of X and w_i is the weight that depends on the probability that the unit i will be included in the sample. The weight in this case is the inverse selection probability and it represents the number of individuals in the target population represented by sample unit i (Levy & Lemeshow, 2013). As an example, a weight of 1000 means that the sample unit represents itself and the other 999 units which are in the population of interest. In most of the cases, weights are greater than one, since a unit should at least represent itself and they vary even when all the sample units have equal inclusion probability. As noted by Levy and Lemeshow (2013), ignoring the sample weights during the analysis results in more biased estimators than weighted estimators. In fact, a sample is a small subset of the respondent population which when weighted, is enlarged to the level of the target population, and therefore, improved estimators will be obtained if weights are accounted for. Various studies have demonstrated that when survey data sets contain weight variables, weighted results are preferred as they produce less bias in the estimates than unweighted results (Korn & Graubard, 1995). This issue was also addressed by Reiter et al. (2006), Schenker et al. (2006), He et al. (2009), Horvitz and Thompson (1952), as well as Molenberghs et al. (2014) amongst others. However, in some cases, the non-use of weights may be justified in situations such as when the software packages used for analysis does not accommodate the survey weighting procedure (Levy & Lemeshow, 2013).

The results of this study were based on both the regular data sets (without taking into account the randomization distribution due to the sample selection procedure) and the weighted data sets to obtain estimators that account for the unequal probabilities of selection for each sample unit. The objective of doing this was not to compare the estimators from unweighted and weighted data sets, but to investigate whether weighting or not may have an impact on the performance of the multiple imputation methods of interest, namely MVNI and MICE. Thus, besides the results from the weighted data sets, the analysis was also done using unweighted data sets, assuming that there was no sample weight variable in the data set that was used (which is not true).

The MVNI was performed using the STATA 13 command "impute mvn" with a uniform prior distribution whereas the mice command in the same Stata version was used to perform imputation with MICE. SPSS 22 (SPSS, 2013) and R were used for data preparation. Data analysis was carried out using STATA 13. Most of the graphs were produced in Microsoft Excel.

4.4.2 Model development and computation of the performance measures

The regression models with the baseline data set or data set with no missing values were first estimated to get the values of the regression coefficients (true coefficients) and their corresponding standard errors. The results from this model were considered as true results that served as a benchmark of results from the imputed data sets and data sets with missing values. Then regression models with the data sets with missing values and imputed data sets using MVNI and MICE were estimated and the results (in terms of bias and standard errors' estimates) were recorded.

To judge the performance of the multiple imputation methods of interest (MVNI and MICE), these estimates were considered. They were compared for each data set to assess the performance of the CD, MVNI and MICE when data were arbitrary MAR or MCAR on unordered or nominal categorical variables, treated as predictors or outcome measures in the regression models.

4.4.3 Imputation models' diagnostics

Current results based on MCMC computations require the reporting of the Monte Carlo error (MCE), which is a measure of the accuracy of the resulting estimates such as the mean, standard errors of the estimates and test statistics. According to [Hoaglin and Andrews \(1975\)](#), it is essential to report these measures to allow the reader to make objective assessments of the numerical quality of the results obtained after MCMC calculations. In the context of missing values analysis, the objective of reporting the MCE is to show that similar results across repeated uses of the same imputation procedures are obtained. That is, the simulation error related to the results obtained after multiple imputations need to be minimized. This error is assessed using the MCE of the multiple imputation results such as parameter estimates, p-values and confidence intervals ([Lee et al., 2012](#); [White et al., 2011](#)). According to [White et al. \(2011\)](#), MCE estimates of coefficients should be less than 10% of the standard errors of the coefficients; MCE estimates of t-test statistic should be approximately less than or equal 0.1; and MCE estimates of p-values should be approximately 0.01 when the true p-value is 0.05 and 0.02 when the true p-value is 0.1. The analysis results based on 100 imputations were

assessed to see if they satisfy these conditions, so that one can be reasonably sure about their statistical reproducibility.

The convergence check was done through the MCMC sequence to investigate whether imputation values with MVNI and MICE converged to the desired distributions. This process is always accompanied by the investigation of the serial dependence among the MCMC draws to obtain independent imputations. In fact, at each iteration, say T^{th} for instance, the imputation model is first estimated using the observed data and the imputed data from the previous iteration, say T^{t-1} and so on. New imputed values are then drawn from their distributions. Consequently, each iteration is correlated with the previous imputation. The first iteration is normally known to be atypical or different from other iterations and because iterations are correlated, it can make other following iterations atypical too. To avoid this problem, the algorithm of the multiple imputation methods goes through the 10 first iterations and save only the results of the 10th. The convergence of imputations is discussed amongst others by [Allison \(2001\)](#) and [Schafer \(1997\)](#).

4.5 Summary of the chapter

To determine the performance of the multiple imputation methods of interest, namely MVNI and MICE, the methodology used to impute missing values on the variables of interest is explained in this chapter. Using a real data set (the 2007 Democratic Republic of Congo demographic health survey or DHS), the regression models in which missing values were observed on either the response or predictor variables measured on a nominal scale, were estimated. Two scenarios were considered for analysis. The first scenario consisted on regression models with missing values on the covariates, whereas the second scenario contained models with missing values on the outcome variables measured on nominal scale. In Scenario 1, two types of models were used. The first model or Model 1.1 regressed contraceptive method use status on marital status alone, with missing values on the independent variable (marital status). The second model or Model 1.2 regressed contraceptive method use status on marital status, controlling for other variables (nominal, continuous and ordinal). Model 1.2 was split into two models; Models 1.2.1 and 1.2.2. In Model 1.2.1, two nominal covariates (marital status and region) were considered for analysis. Model 1.2.2 regressed contraceptive method

use status on various variables (continuous, nominal and ordinal) with missing values on only nominal variables. The second scenario contained two models with missing values on the response variables. The first model or Model 2.1 in this scenario regressed contraceptive method use status (a dichotomous variable) on marital status, whereas the second model or Model 2.2 regressed contraceptive method use status (measured as a polytomous variable). Throughout this thesis, the rate of missingness that was considered is 50%. However, to verify whether the rates of missingness could have an effect on the performance of the multiple imputation methods of interest, 50%, 30% and 10% missing data were considered for only Model 1.1.

Missing values on the variables of these models were imputed using the parametric imputation technique (MVNI) and the MICE method which takes into account the distributional form of the variables to be imputed. Stochastic imputation models of the variables to be imputed were developed and explained. The imputation models diagnostics to assess whether or not imputation values were drawn from the desired distributions or similar results across repeated uses of the same imputation procedures were obtained, were indicated.

Under MVNI, dichotomous nominal variables were imputed as a continuous variable using linear regression models and imputed values were rounded to the nearest integer (0 or 1) to keep the dichotomy nature of the imputed variables that had to be used as dependent variables in the regression models. On the other hand, polytomous variables (marital status and contraceptive method use status with 3 categories) with missing data were dichotomised first and $K - 1$ (where K is the number of categories forming the variable) dichotomised variables were included in the imputation models, leaving one of the categories as a reference. Imputation using MICE was done such that the distributional form of the variables with missing values was accounted for. That is, imputation values were drawn from the binary and multinomial logistic regressions for the dichotomous and polytomous variables respectively. The analysis using both the weighted and unweighted data sets and under MAR and MCAR assumptions, was considered.

In the next chapter, the results (in terms of bias and standard errors) are presented and used to compare the multiple imputation methods of interest, namely MVNI and MICE.

Chapter 5

Results

5.1 Introduction

This chapter presents the results based on the regular data sets (without taking into account the sample weights) and the weighted data sets (in which the sample selection procedure is accounted for), when data are missing at random or missing completely at random. Four sections are covered. The first section (Section 5.1) provides an overview of the chapter. The second section (Section 5.2) presents the results of the first scenario. The results of Model 1.1 are presented when 50%, 30% and 10% data are missing at random or completely at random on a single covariate. Moreover, the findings on Models 1.2.1 and 1.2.2, on which missing values are observed on more than one covariate with missing values, are also presented. The third section (Section 5.3) presents the results for the second scenario, which contains two models (Models 2.1 and 2.2) with missing values on the outcome variables; binary and polytomous. For each model in each scenario, the descriptive statistics of the data sets with missing values is provided. The estimates of bias and standard errors in the regression coefficients obtained using the case deletion or CD, MVNI and MICE techniques are reported. The results on the imputation models' diagnostics are also provided in this section. In Section 4 (Section 5.4), the summary of the chapter is given.

5.2 Scenario 1: Logistic regression models with missing values on the covariates

As mentioned in Chapter 4, Scenario 1 contains the regression models with missing values on the covariates. In Model 1.1, a single covariate with missing values is considered. This model is used to investigate whether the rates of missing values in the data sets may have an impact on the imputation methods of interest, namely MVNI and MICE. The rates of missingness considered for this case are 50%, 30% and 10%. For all other models in the study, only 50% rate of missingness was used for analysis. In Model 1.2, regression models with at least two covariates containing missing values on only the nominal ones are considered for analysis. A graphical representation of these models is given in Figure 5.1.

The objective of using all these models was to explore the behaviour of MVNI and MICE when missing values were observed on the unordered categorical variables alone first, then in the presence of other types of variables (continuous and ordinal variables).



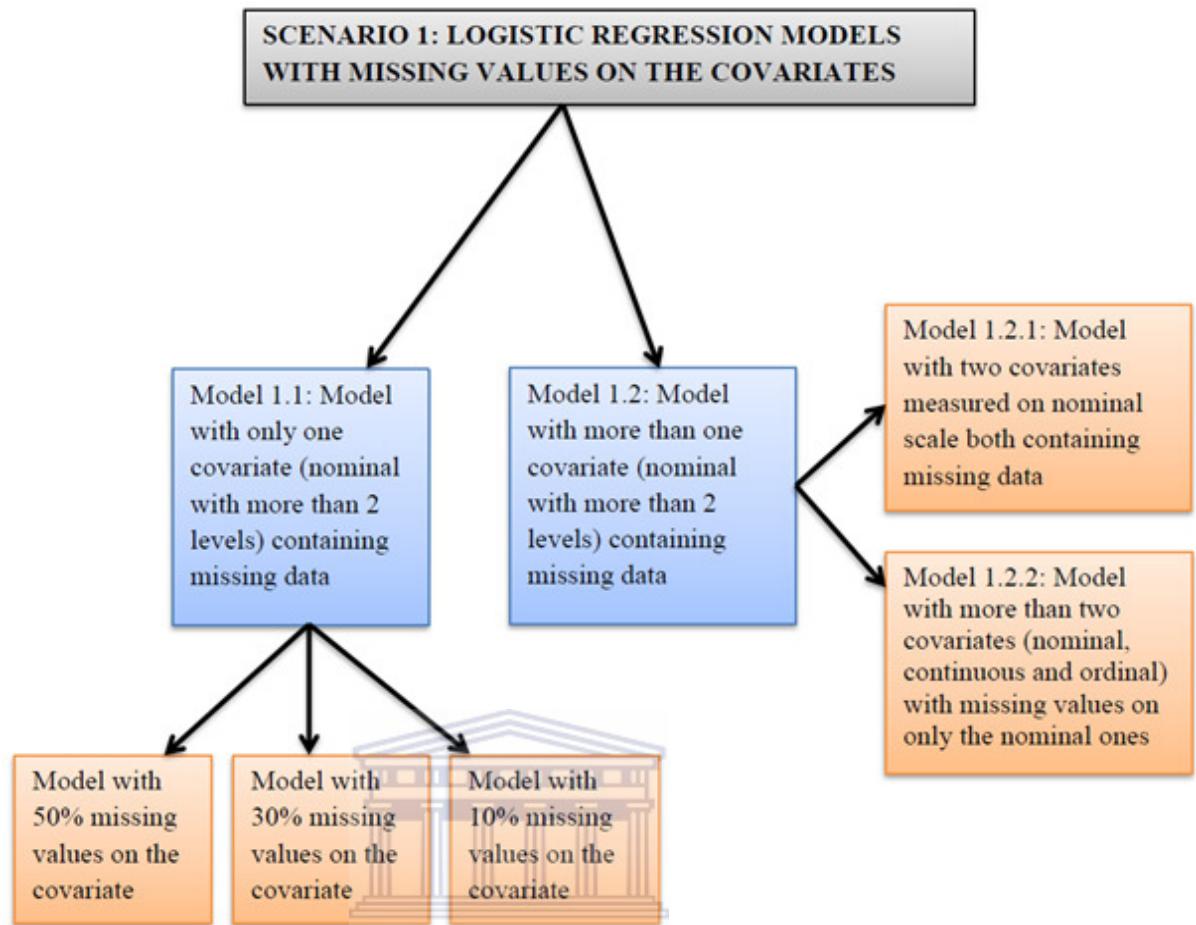


FIGURE 5.1: Scenario 1: Logistic regression models with missing values on the covariates.

5.2.1 Model 1.1: Binary logistic regression model with missing values on a single covariate measured on a nominal scale

5.2.1.1 Results when 50% of data are missing at random or completely at random on the covariate

Descriptive statistics As stated in Chapter 4, Model 1.1 regresses contraceptive method use status on marital status, to determine the performance of the multiple imputation methods of interest when data are missing at random or completely at random on nominal variables (marital status) treated as predictors in the regression models. Recall that under MAR assumption, missing values were deleted in a random manner such that missingness depended on the target variables in the data set that was used. In this case, 50% missing values were arbitrary

and randomly deleted on marital status if a woman was not using any contraceptive method, thus allowing the missing value indicator or missingness to depend on contraceptive method use status. As shown in Table 5.1, deleting 50% of the data at random on marital status when a woman was not using any contraceptive method resulted in approximately 39.30% missing values on the entire variable. This makes sense because values were deleted on one side (only if a woman was not using any contraceptive method).

TABLE 5.1: Model 1.1: Frequency distribution of missingness on marital status under MAR assumption.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	17937	60.70	60.70
Missing	11611	39.30	100.00
Total	29548	100.00	

To investigate whether data were missing at random, a bivariate analysis of the data was conducted to see if the proportions or means of missing values differed across the demographic particulars of the respondents. Therefore, the cross tabulations of categorical variables against missingness (present or missing) were first generated. Frequencies in every category for each categorical variable were produced to determine whether there were differences in missing values among categories of variables. The results are shown in Table 5.2. As indicated, approximately 50% of data were arbitrary and randomly deleted on marital status if a woman was not using any contraceptive method, and no value (0%) was deleted if she was using it. This significant difference in proportions of missing values across the contraceptive method use status was confirmed by the Pearson chi-square test for association with a p-value less than 5% level of significance ($\chi^2_{(1)} = 5.2e + 03$, p-value = 0.000). This indicates that missingness is associated with contraceptive method status. This is a sufficient evidence to confirm that the missing at random assumption was met, as this assumption requires missingness to be associated with at least one variable in the data set to be used.

The cross tabulation of wealth index and the missingness was also done. The results indicate that the percentages of missing values in the indicator variables appear to vary much across the wealth index levels. As observed, the percentages of missing values varied between 30% and 43% across the wealth index categories. It can be seen that the poorer the respondent, the more missing values are observed. This difference is likely due to chance and is also an indication that the missing at random assumption was met. Furthermore, the Chi-square test for

association confirms that missingness is associated to the woman's wealth status ($\chi^2_{(14)} = 239.5261$, p-value = 0.000) which is less than the level of significance of 0.05 that was used. A similar analysis was also done to assess whether missingness was related to the woman region of origin. As indicated in Table 5.2, missingness was found to be associated with region (p-value = 0.000 less than 0.05).

TABLE 5.2: Model 1.1: Distribution of missingness across selected categorical variables when 50% data are MAR on marital status if a woman is not using any contraceptive method.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.000
No method	50.10	49.90	100	
At least one method	100	0.00	100	
Total	60.70	39.30	100	
Wealth index				0.000
Poorest	57.08	42.92	100	
Poorer	57.48	42.52	100	
Middle	58.78	41.22	100	
Richer	62.36	37.64	100	
Richest	69.32	30.68	100	
Total	60.70	39.30	100	
Region				0.000
Kinshasa	69.71	30.29	100	
Bas Kongo	67.94	32.06	100	
Bandundu	63.31	36.69	100	
Eguateur	61.81	38.19	100	
Oriental	58.56	41.44	100	
Nord Kivu	58.27	41.73	100	
Maniema	58.75	41.25	100	
Sud Kivu	56.32	43.68	100	
Katanga	60.77	39.23	100	
Kasai Oriental	56.65	43.35	100	
Kasai Occidental	56.71	43.29	100	
Total	60.92	39.30	100	

The independent-samples t-test was also conducted to identify variables whose pattern of missing values might be influencing the continuous variables of interest. In this case, age and education in completed years were considered. The means of age and education in completed years for missingness were calculated. The results are found in Appendix B in Figure 6.1 (age) and Figure 6.2 (education). As indicated, there was a significant difference in means age scores for not missing (Mean = 35.0, SD = 8.0) and missing (Mean = 35.4, SD = 8.3). Similarly, a significant

difference in means education (in completed years) for the groups missing (Mean = 4.0, SD = 3.7) and not missing (Mean = 4.8, SD = 4.0) was found. These findings were supported by the t-test of equality of the means with p-values less than 5% significance level (p-value = 0.000). This indicates that missingness is associated with age and education in completed years as well, which is additional information that observations were missing at random on marital status.

Under the MCAR assumption, 50% missing values were also arbitrary deleted on marital status such that missingness was not related to the values of any variables in the data set subject to analysis. The frequency distribution of missingness on this variable is presented in Table 5.3. As indicated, approximately 50% of values were deleted on marital status.

TABLE 5.3: Model 1.1: Frequency distribution of missingness when 50% data are MCAR on marital status.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	14563	49.43	49.43
Missing	14985	50.57	100.00
Total	29548	100.00	

To ensure that these values were deleted completely at random, a bivariate analysis of the missing value indicator or missingness and other socio-demographic characteristics of the respondent was conducted. The variables considered in this case are the woman's contraceptive method use status, region and wealth index. The results are shown in Table 5.4. As indicated, missingness on marital status is not associated with any of these variables. This fact is confirmed by the Chi-square test for association with p-values greater than 0.05 for each variable, an indication that data were missing completely at random on marital status.

TABLE 5.4: Model 1.1: Distribution of missingness by selected categorical variables when 50% data are MCAR on marital status.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.280
No method	49.27	50.73	100	
At least one method	50.04	49.96	100	
Total	49.43	50.53	100	
Wealth index				0.053
Poorest	47.83	52.17	100	
Poorer	49.69	50.31	100	
Middle	49.92	50.08	100	
Richer	50.34	49.66	100	
Richest	49.57	50.43	100	
Total	49.43	50.57	100	
Region				0.989
Kinshasa	49.05	50.95	100	
Bas Kongo	48.71	51.29	100	
Bandundu	49.51	50.49	100	
Eguateur	48.71	51.29	100	
Oriental	49.84	50.16	100	
Nord Kivu	50.27	49.73	100	
Maniema	49.23	50.77	100	
Sud Kivu	49.33	50.67	100	
Katanga	49.47	50.53	100	
Kasai Oriental	49.98	50.02	100	
Kasai Occidental	49.80	50.20	100	
Total	49.43	50.57	100	

The independent-samples t-test was also conducted to examine if there was a significant difference in means of age and education in completed years across missingness. The results are shown in Appendix B in Figures 6.3 and 6.4 for age and education in completed years respectively. The results indicated that there was no significant difference in means of age and education in single years between the two groups (missing and not missing) as confirmed by the p-values associated with the test (p-values of 0.33 and 0.071 for age and education respectively), which are greater than the significance level of 5% that was used. This shows that no association exists between missingness and these two continuous variables, an additional information that the MCAR assumption was met.

Performance measures To assess the performance of the imputation methods of interest (MVNI and MICE), the bias in the regression coefficients and the standard errors of the regression coefficients were considered for analysis. The logistic regression models of contraceptive method use status on marital status were estimated using the data set with no missing values (baseline data set or BD), the data set with missing values of interest (case deletion or CD method) and the completed or imputed data sets with MVNI and MICE. Then, the bias was computed and reported along with the standards errors. In Tables 5.5 and 5.6, these estimates are presented for both weighted and unweighted data sets, when approximately 50% of data are missing at random on marital status if a woman is not using any contraceptive method. In Tables 5.7 and 5.8, the same statistics are reported when approximately 50% data are missing completely at random on marital status. The plot of bias and standard errors are shown in Figures 5.2 and 5.3 for both unweighted and weighted data sets to look at the pattern of performance of the missing data methods of interest; case deletion (CD), MVNI and MICE.

As expected, the results show that under the MAR and MCAR assumptions, multiple imputations with MVNI and MICE yields less biased (estimates of bias closer to zero) and more accurate standard errors (standards errors close to the standard errors obtained using the true data set or data set with no missing values) than case deletion, which discards items with missing values from the analysis. The results indicate also that the MVNI technique is less biased and yields more accurate standard errors than MICE, either when data are missing at random or missing completely at random for both unweighted (Figure 5.2) and weighted (Figure 5.3) data sets.

TABLE 5.5: Model 1.1: Estimates of bias when approximately 50% of data are MAR on marital status if a woman is not using any contraceptive method.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.030	0.014	0.008	0.012	-0.060
CD-weighted	-0.012	0.026	0.007	0.026	-0.072
MVNI-unweighted	0.017	0.009	0.003	0.012	-0.043
MVNI-weighted	-0.009	0.014	0.002	0.016	-0.064
MICE-unweighted	0.027	0.013	0.006	0.010	-0.057
MICE-weighted	-0.010	0.025	0.003	0.024	-0.069

TABLE 5.6: Model 1.1: Estimates of standard errors when approximately 50% of data are MAR on marital status if a woman is not using any contraceptive method.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.137	0.050	0.149	0.121	0.074
CD-weighted	0.211	0.072	0.170	0.179	0.104
MVNI-unweighted	0.130	0.048	0.141	0.118	0.074
MVNI-weighted	0.179	0.067	0.155	0.179	0.103
MICE-unweighted	0.137	0.049	0.149	0.120	0.074
MICE-weighted	0.200	0.069	0.162	0.179	0.104

TABLE 5.7: Model 1.1: Estimates of bias when approximately 50% of data are MCAR on marital status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.052	0.042	-0.113	-0.046	0.087
CD-weighted	-0.058	0.072	0.049	-0.021	-0.034
MVNI-unweighted	0.028	0.036	-0.083	-0.028	0.077
MVNI-weighted	-0.051	0.041	0.024	0.014	-0.006
MICE-unweighted	0.051	0.040	-0.103	-0.041	0.078
MICE-weighted	-0.057	0.063	0.030	-0.018	-0.017

TABLE 5.8: Model 1.1: Estimates of standard errors when approximately 50% of data are MCAR on marital status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.137	0.050	0.149	0.121	0.074
CD-weighted	0.211	0.072	0.170	0.179	0.104
MVNI-unweighted	0.130	0.048	0.141	0.118	0.074
MVNI-weighted	0.179	0.067	0.155	0.179	0.103
MICE-unweighted	0.137	0.049	0.149	0.120	0.074
MICE-weighted	0.200	0.069	0.162	0.179	0.104

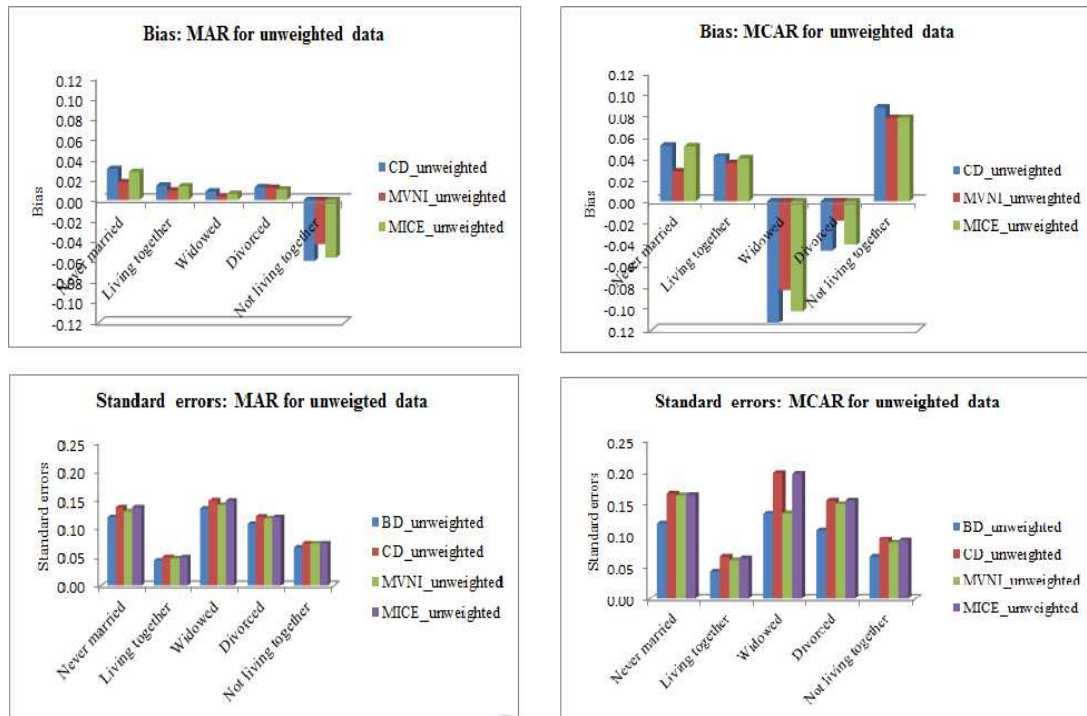


FIGURE 5.2: Model 1.1: Plot of bias and standard errors when 50% data are MAR or MCAR on marital status for unweighted data sets.

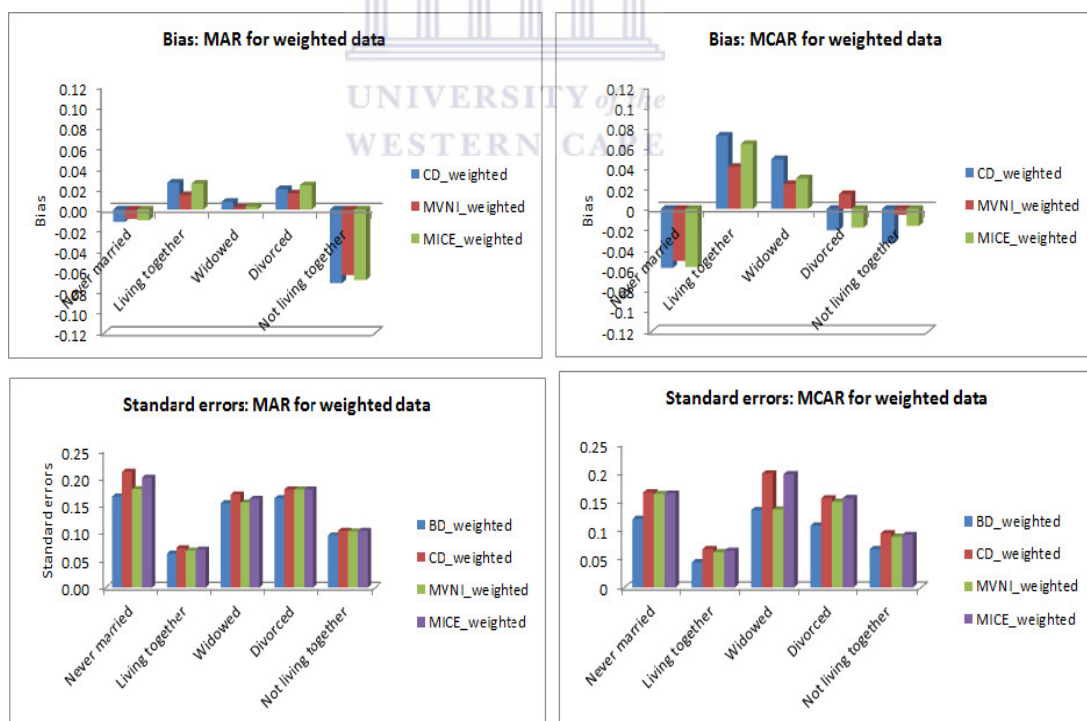


FIGURE 5.3: Model 1.1: Plot of bias and standard errors when 50% data are MAR or MCAR on marital status for weighted data sets.

Model diagnostics The MCEs after MVNI and MICE of the statistics involved in the estimation of the regression models were computed. These estimates are the regression coefficients, standard errors, t-values, p-values and confidence limits of the regression coefficients. Note that the estimates of the MCEs of the confidence limits are measures of the accuracy of these limits after 100 imputations, they have nothing to do with the estimates of the MCEs of the regression coefficients, which are also measures of accuracy of the coefficients. This means that it is not expected that the MCEs of the regression coefficients lie between the MCEs of the confidence limits, or the MCEs of the upper limits to be higher than the MCEs of the lower limits of the confidence interval or vice versa. These estimates are presented in Appendix C in Tables 6.1, 6.2, 6.3 and 6.4 when 50% data were missing at random on marital status if a woman was not using any contraceptive method, and in Appendix C in Tables 6.5, 6.6, 6.7 and 6.8 when data were missing completely at random on marital status. The results show that the suggested criteria for the Monte Carlo errors are met. In fact, the Monte Carlo errors on the coefficients are less than 10% of the standard error for unweighted and weighted data sets. The Monte Carlo errors of the p-values are also approximately less than 0.1 when 5% level of significance was used for both MVNI and MICE. The Monte Carlo errors of the t-test statistic were found to be approximately less than 0.1 for all the methods and data sets (White et al., 2011). Therefore, based on these results, one can reasonably be sure about their statistical reproducibility. This suggests that the number of imputations used (100 imputations) were enough to produce stable results.

To ensure that imputations converged to the desired distributions, the convergence was investigated for both unweighted and weighted data sets, under MAR and MCAR assumptions. The estimates of the worst linear function (WLF) were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE methods. Under MAR assumption, the results are presented in Appendix D in Figures 6.25 and 6.26 for the MVNI approach, and in Figures 6.27 and 6.28 for the MICE technique for unweighted data sets. For weighted data sets, the results are shown in Figures 6.29 and 6.30 for MVNI, and in Figures 6.31 and 6.32 for MICE. Under MCAR assumption, the same estimates are shown in Figures 6.33, 6.34, 6.35 and 6.36 for unweighted data sets, and in Figures 6.37, 6.38, 6.39 and 6.40 for weighted data. As indicated, the plots of the estimates of WLF against the iteration numbers show no visible trend, thus indicating that convergence is assured with the number of iterations used (1000 iterations). On

the other hand, the plots of WLF's estimates against the lag numbers show the autocorrelations that die off quickly, which implies that even a smaller number (of iterations) than what was used, such as 10 iterations between imputations, can be used to obtain independent samples.

5.2.1.2 Results when 30% of data are missing at random or completely at random on the covariate

Descriptive statistics The performance of the multiple imputation methods of interest was also investigated when approximately 30% of the data were deleted at random or completely at random on the covariate.

Under the MAR assumption, 30% missing values were arbitrary and randomly deleted on marital status if a woman was not using any contraceptive method. This allowed the missing value indicator or missingness to depend on contraceptive method use status. As shown in Table 5.9, deleting 30% of the data at random on marital status when a woman was not using any contraceptive method resulted in approximately 23.46% missing values on the entire variable (marital status).

TABLE 5.9: Model 1.1: Frequency distribution of missingness when approximately 50% data are MAR on marital status if a woman is not using any contraceptive method.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	22617	76.54	76.54
Missing	6931	23.46	100.00
Total	29548	100.00	

To investigate whether or not data were missing at random, a bivariate analysis of the data was conducted to assess whether the proportions or means of missing values differed across the demographic particulars of the respondents. The results show that missingness is associated with the woman's contraceptive method use status, wealth index, region, age and education in completed years as indicated by the p-values associated with the Chi-square test for association (for categorical variables) and the t-test for equality of the means (for continuous variables) that are less than the significance level of 5%. The results are shown in Table 5.10 for categorical variables (contraceptive method use status, wealth index and region) and in Appendix B in Figures 6.5 and 6.6 for continuous variables (age and education respectively).

TABLE 5.10: Model 1.1: Distribution of missingness across selected categorical variables when approximately 30% data are MAR on marital status if a woman is not using any contraceptive method.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.000
No method	70.21	29.79	100	
At least one method	100	0.00	100	
Total	76.54	23.46	100	
Wealth index				0.000
Poorest	73.65	26.35	100	
Poorer	75.65	24.35	100	
Middle	75.79	24.21	100	
Richer	77.54	22.46	100	
Richest	80.94	19.06	100	
Total	76.54	23.46	100	
Region				0.000
Kinshasa	81.49	18.51	100	
Bas Kongo	81.00	19.00	100	
Bandundu	78.79	21.21	100	
Eguateur	78.09	21.91	100	
Oriental	74.42	25.58	100	
Nord Kivu	75.18	24.82	100	
Maniema	74.36	25.64	100.00	
Sud Kivu	74.49	25.51	100	
Katanga	75.94	24.06	100	
Kasai Oriental	73.68	26.32	100	
Kasai Occidental	73.80	26.20	100	
Total	76.54	23.46	100	

Under the MCAR assumption, 30% missing values were arbitrary deleted on marital status such that missingness was not associated with any variable in the data set subject to analysis. The frequency distribution of missingness on this variable is presented in Table 5.11. As indicated, approximately 30% of values were deleted on marital status.

TABLE 5.11: Model 1.1: Frequency distribution of missingness when approximately 30% data are MCAR on marital status.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	20611	69.75	69.75
Missing	8937	30.25	100.00
Total	29548	100.00	

To ensure that these values were deleted completely at random, a bivariate analysis of the missing value indicator or missingness and other socio-demographic characteristics of the respondent was conducted. The results are shown in Table 5.12 for categorical variables and in Appendix B in Figures 6.7 and 6.8 for continuous variables; age and education respectively. As indicated, missingness on marital status is not associated with the woman contraceptive method use status, wealth index and region as confirmed by the Chi-square test for association (categorical variables) and the independent-samples t-test (for continuous variables) with p-values greater than 0.05, an indication that data were missing completely at random on marital status.

TABLE 5.12: Model 1.1: Distribution of missingness by selected categorical variables when approximately 30% data are MCAR on marital status.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.166
No method	69.56	30.44	100	
At least one method	70.47	29.53	100	
Total	69.75	30.25	100	
Wealth index				0.391
Poorest	68.77	31.23	100	
Poorer	70.10	29.90	100	
Middle	70.20	29.80	100	
Richer	70.01	29.99	100	
Richest	69.79	30.21	100	
Total	69.75	30.25	100	
Region				0.856
Kinshasa	68.80	31.20	100	
Bas Kongo	68.36	31.64	100	
Bandundu	70.31	29.69	100	
Eguateur	70.24	29.76	100	
Oriental	70.45	29.55	100	
Nord Kivu	69.99	30.01	100	
Maniema	70.17	29.83	100	
Sud Kivu	69.43	30.57	100	
Katanga	70.07	29.93	100	
Kasai Oriental	69.44	30.56	100	
Kasai Occidental	70.09	29.91	100	
Total	69.75	30.25	100	

Performance measures To assess the performance of the imputation methods of interest (MVNI and MICE), the logistic regression models of contraceptive method use status on marital status were first estimated using the data set with no missing values (baseline data set or BD), the data set with missing values of interest (case deletion or CD method) and the completed or imputed data sets with MVNI and MICE. Then, the bias was computed and reported along with the standards errors.

In Tables 5.13 and 5.14, these estimates are presented for both weighted and unweighted data sets respectively, when approximately 30% of data are missing at random on marital status if a woman is not using any contraceptive method. In Tables 5.15 and 5.16, the same statistics are reported when approximately 30% data are missing completely at random on marital status. The plot of bias and standard errors are shown in Figure 5.4 and 5.5 for both unweighted and weighted data sets to look at the pattern of performance of the missing data methods of interest; case deletion (CD), MVNI and MICE. The results show that under MAR and MCAR assumptions, multiple imputations with MVNI and MICE produce less biased and more accurate standard errors than case deletion. Furthermore, it is observed that MVNI is less biased and yields better standard errors than MICE, either when data are missing at random or missing completely at random for both unweighted (Figure 5.4) and weighted (Figure 5.5) data sets.

TABLE 5.13: Model 1.1: Estimates of bias when approximately 30% of data are MAR on marital status if a woman is not using any contraceptive method.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	-0.035	-0.023	0.024	-0.059	-0.025
CD-weighted	-0.051	-0.130	-0.041	-0.052	-0.040
MVNI-unweighted	-0.018	-0.009	0.006	-0.044	-0.013
MVNI-weighted	-0.022	-0.104	-0.014	-0.030	-0.020
MICE-unweighted	-0.023	-0.012	0.007	-0.052	-0.018
MICE-weighted	-0.026	-0.113	-0.031	-0.040	-0.026

TABLE 5.14: Model 1.1: Estimates of standard errors when approximately 30% of data are MAR on marital status if a woman is not using any contraceptive method.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.126	0.047	0.136	0.113	0.069
CD-weighted	0.177	0.067	0.157	0.167	0.099
MVNI-unweighted	0.124	0.046	0.136	0.112	0.069
MVNI-weighted	0.172	0.064	0.156	0.165	0.098
MICE-unweighted	0.125	0.047	0.136	0.112	0.069
MICE-weighted	0.173	0.065	0.157	0.168	0.099

TABLE 5.15: Model 1.1: Estimates of bias when approximately 30% of data are MCAR on marital status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.031	0.031	-0.095	-0.121	0.033
CD-weighted	0.163	-0.085	0.140	0.045	0.097
MVNI-unweighted	0.020	0.025	-0.056	-0.106	0.014
MVNI-weighted	0.102	-0.062	0.116	0.009	0.064
MICE-unweighted	0.029	0.028	-0.068	-0.116	0.026
MICE-weighted	0.132	-0.063	0.136	0.016	0.071

TABLE 5.16: Model 1.1: Estimates of standard errors when approximately 30% of data are MCAR on marital status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.149	0.056	0.157	0.128	0.081
CD-weighted	0.187	0.079	0.180	0.193	0.118
MVNI-unweighted	0.147	0.053	0.155	0.126	0.073
MVNI-weighted	0.182	0.077	0.167	0.186	0.116
MICE-unweighted	0.148	0.054	0.157	0.126	0.078
MICE-weighted	0.184	0.078	0.172	0.187	0.116

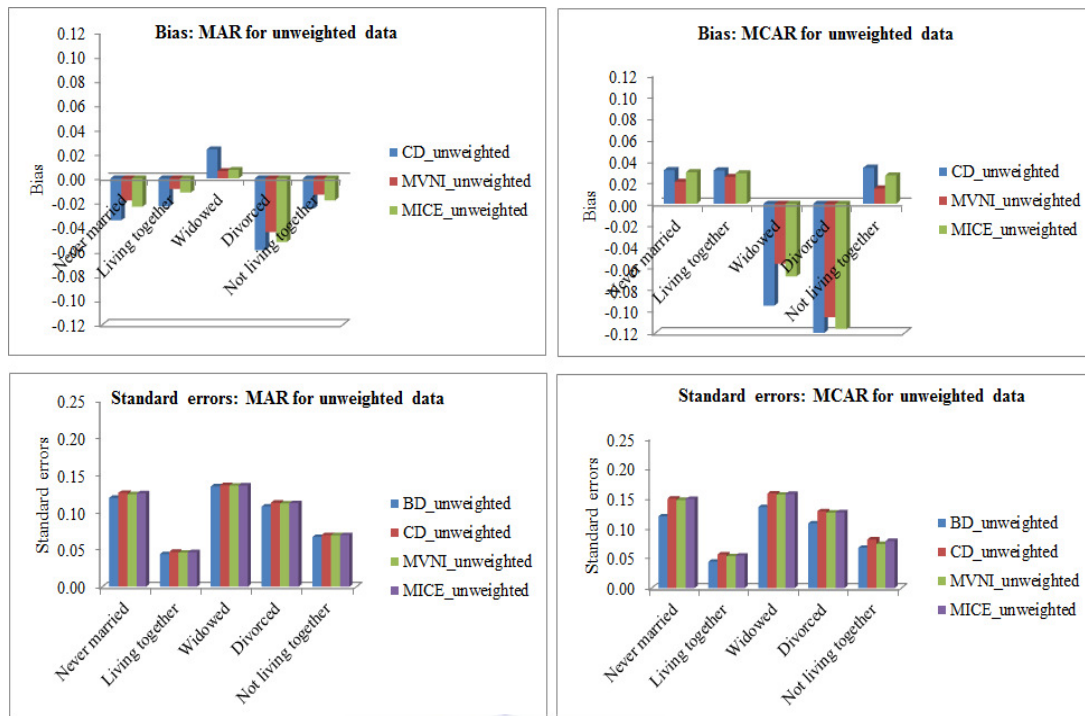


FIGURE 5.4: Model 1.1: Plot of bias and standard errors when approximately 30% data are MAR and MCAR for unweighted data sets.

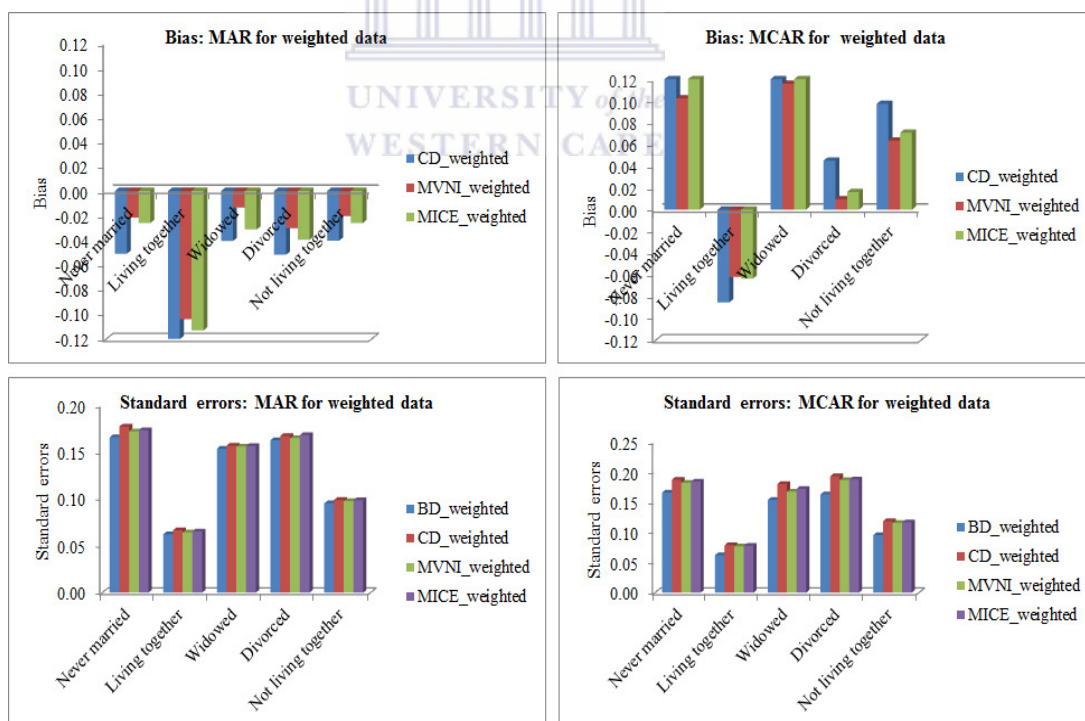


FIGURE 5.5: Model 1.1: Plot of bias and standard errors when approximately 30% data are MAR and MCAR for weighted data sets.

Model diagnostics The estimates of Monte Carlo errors (MCE) after MVNI and MICE of the statistics involved in the estimation of the regression models were also estimated. These estimates are presented in Appendix C in Tables 6.9, 6.10, 6.11 and 6.12 when 30% data were missing at random on marital status if a woman was not using any contraceptive method, and in Appendix C in Tables 6.13, 6.14, 6.15 and 6.16 when data were missing completely at random on marital status. The results indicate that the suggested criteria for the Monte Carlo errors are met. Indeed, the Monte Carlo errors on the coefficients are less than 10% of the standard error for unweighted and weighted data sets. The Monte Carlo errors of the p-values are also approximately less than 0.1 when 5% level of significance was used for both MVNI and MICE. The Monte Carlo errors of the t-test statistic were found to be approximately less than 0.1 for all the methods and data sets. Therefore, based on these results, one can reasonably be sure about their statistical reproducibility. This indicates that the number of imputations used (100 imputations) were enough to produce stable results.

To ensure that imputations converged to the desired distributions, the convergence check was done for both unweighted and weighted data sets, under MAR or MCAR assumptions. The estimates of the worst linear function (WLF) were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE methods. Under the MAR assumption, the results are presented in Appendix D in Figures 6.41 and 6.42 for the MVNI approach, and in Figures 6.43 and 6.44 for the MICE technique for unweighted data sets. For weighted data sets, the results are shown in Figures 6.45 and 6.46 for MVNI, and in Figures 6.47 and 6.48 for MICE. Under the MCAR assumption, the same results are provided in Appendix D in Figures 6.49, 6.50, 6.51, 6.52, 6.53, 6.54, 6.55 and 6.56. As indicated, the plots of the estimates of WLF against the iteration numbers show no visible trend, thus indicating that convergence is attained with the number of iterations used (1000 iterations). On the other hand, the plots of WLF's estimates against the lag numbers show the autocorrelations that die off quickly, which implies that even a smaller number (of iterations) than what was used, such as 10 iterations between imputations, can be used to obtain independent samples.

5.2.1.3 Results when 10% of data are missing at random or completely at random on the covariate

Descriptive statistics Data sets with 10% missing at random or missing completely at random data were also considered for analysis, to determine the performance of MICE and MVNI. Under the MAR assumption, 10% missing values were randomly deleted on marital status if a woman was not using any contraceptive method. This allowed missingness on marital status to depend on contraceptive method use status. As shown in Table 5.17, deleting 10% of the data at random on marital status when a woman was not using any contraceptive method led to approximately 7.86% missing values on the entire variable.

TABLE 5.17: Model 1.1: Frequency distribution of missingness when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	27227	92.14	92.14
Missing	2321	7.86	100.00
Total	29548	100.00	

To investigate whether or not data were missing at random, a bivariate analysis of the data was conducted to see if the proportions or means of missing values differed across the demographic particulars of the respondents. The results show that missingness is related to contraceptive method use status, wealth index and region and education in completed years as indicated by the p-values associated by both the t-test (for continuous variables) and Chi-square test (for the categorical variable) that are less than the significance level of 5%. The results are shown in Table 5.18 for categorical variables (contraceptive method use status, wealth index and region) and in Appendix B in Figures 6.9 and 6.10 for continuous variables (age and education in completed years).

TABLE 5.18: Model 1: Distribution of missingness across selected categorical variables when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.000
No method	90.02	9.98	100	
At least one method	100	0.00	100	
Total	92.14	7.86	100	
Wealth index				0.000
Poorest	91.49	8.51	100.00	
Poorer	91.85	8.15	100.00	
Middle	91.96	8.04	100.00	
Richer	91.98	8.02	100.00	
Richest	93.71	6.29	100.00	
Total	92.14	7.86	100.00	
Region				0.000
Kinshasa	93.34	6.66	100.00	
Bas Kongo	93.73	6.27	100.00	
Bandundu	93.10	6.90	100.00	
Eguateur—	93.01	6.99	100.00	
Oriental	91.01	8.99	100.00	
Nord Kivu	91.75	8.25	100.00	
Maniema	92.34	7.66	100.00	
Sud Kivu	90.93	9.07	100.00	
Katanga	92.48	7.52	100.00	
Kasai Oriental	90.90	9.10	100.00	
Kasai Occidental	90.65	9.35	100.00	
Total	92.14	7.86	100.00	

Under the MCAR assumption, 10% missing values were arbitrary deleted on marital status such that missingness was not associated with any variables in the data set subject to analysis. The frequency distribution of missingness on this variable is presented in Table 5.19. As indicated, around 10% of values were deleted on marital status.

TABLE 5.19: Model 1.1: Frequency distribution of missingness when approximately 10% of the data are MCAR on marital status.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	26581	89.96	89.96
Missing	2967	10.04	100.00
Total	29548	100.00	

To ensure that these values were deleted completely at random, a bivariate analysis of the missing value indicator or missingness and other socio-demographic characteristics of the respondent was conducted. The results are shown in Table 5.20 for categorical variables and in Appendix B in Figures 6.11 and 6.12 for age and education respectively. As indicated, missingness on marital status is not associated with the woman's contraceptive method use status, wealth index, region, age and education in completed years as confirmed by the Chi-square test for association (categorical variables) and the independent-samples t-test for continuous variables that showed p-values greater than 0.05, which is an indication that data were missing completely at random on marital status.

TABLE 5.20: Model 1: Distribution of missingness by selected categorical variables when approximately 10% of the data are MCAR on marital status.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.593
No method	90.01	9.99	100	
At least one method	89.78	10.22	100	
Total	89.96	10.04	100	
Wealth index				0.965
Poorest	89.83	10.17	100.00	
Poorer	89.82	10.18	100.00	
Middle	89.96	10.04	100.00	
Richer	90.18	9.82	100.00	
Richest	90.03	9.97	100.00	
Total	89.96	10.04	100.00	
Region				0.805
Kinshasa	89.43	10.57	100.00	
Bas Kongo	89.30	10.70	100.00	
Bandundu	89.92	10.08	100.00	
Eguateur	90.54	9.46	100.00	
Oriental	90.56	9.44	100.00	
Nord Kivu	90.25	9.75	100.00	
Maniema	89.79	10.21	100.00	
Sud Kivu	90.54	9.46	100.00	
Katanga	90.02	9.98	100.00	
Kasai Oriental	89.47	10.53	100.00	
Kasai Occidental	89.90	10.10	100.00	
Total	89.96	10.04	100.00	

Performance measures In this section, the performance of the multiple imputation methods, MVNI and MICE, was assessed. Logistic regression models of contraceptive method use status on marital status were first estimated using the data set with no missing values, the data set with missing values and the completed or imputed data sets using MVNI and MICE. Then, the bias was computed and reported along with the standard errors of the regression coefficients. In Tables 5.21 and 5.22, these estimates are presented for both weighted and unweighted data sets, when approximately 10% of data are missing at random on marital status if a woman is not using any contraceptive method. In Tables 5.23 and 5.24, the same statistics are reported when approximately 10% data are missing completely at random on marital status. The plot of bias and standard errors are shown in Figures 5.6 and 5.7 for both unweighted and weighted data sets to look at the pattern of performance of the missing data methods of interest; case deletion (CD), MVNI and MICE. The results show that under MAR and MCAR assumptions, multiple imputations with MVNI and MICE yields less biased and more accurate standard errors than case deletion or complete case analysis. It can also be seen that the MVNI technique outperforms MICE in terms of bias in the regression coefficients and standard errors, either when data are missing at random or missing completely at random for both unweighted (Figure 5.6) and weighted (Figure 5.7) data sets.

TABLE 5.21: Model 1.1: Estimates of bias when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.002	0.001	0.001	0.001	0.001
CD-weighted	0.003	0.001	0.001	0.001	0.002
MVNI-unweighted	0.001	0.001	0.000	0.001	0.000
MVNI-weighted	0.002	0.001	0.000	0.001	0.001
MICE-unweighted	0.001	0.001	0.001	0.001	0.000
MICE-weighted	0.003	0.001	0.000	0.001	0.001

TABLE 5.22: Model 1.1: Estimates of standard errors when approximately 10% of the data are MAR on marital status if a woman is not using any contraceptive method.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.121	0.045	0.136	0.109	0.068
CD-weighted	0.169	0.063	0.154	0.164	0.097
MVNI-unweighted	0.121	0.045	0.135	0.109	0.067
MVNI-weighted	0.168	0.063	0.154	0.164	0.096
MICE-unweighted	0.121	0.045	0.135	0.109	0.067
MICE-weighted	0.169	0.063	0.154	0.164	0.097

TABLE 5.23: Model 1.1: Estimates of bias when approximately 10% of the data are MCAR on marital status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.024	0.015	-0.086	-0.080	-0.009
CD-weighted	0.077	-0.068	-0.124	0.007	-0.026
MVNI-unweighted	0.013	0.011	-0.024	-0.075	-0.005
MVNI-weighted	0.051	-0.066	-0.043	-0.002	-0.002
MICE-unweighted	0.017	0.012	-0.040	-0.077	-0.005
MICE-weighted	0.054	-0.068	-0.060	-0.004	-0.007

TABLE 5.24: Model 1.1: Estimates of standard errors when approximately 10% of the data are MCAR on marital status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.126	0.046	0.147	0.114	0.071
CD-weighted	0.174	0.068	0.171	0.170	0.108
MVNI-unweighted	0.127	0.046	0.144	0.111	0.070
MVNI-weighted	0.174	0.067	0.170	0.167	0.102
MICE-unweighted	0.127	0.046	0.145	0.113	0.070
MICE-weighted	0.174	0.068	0.170	0.168	0.105

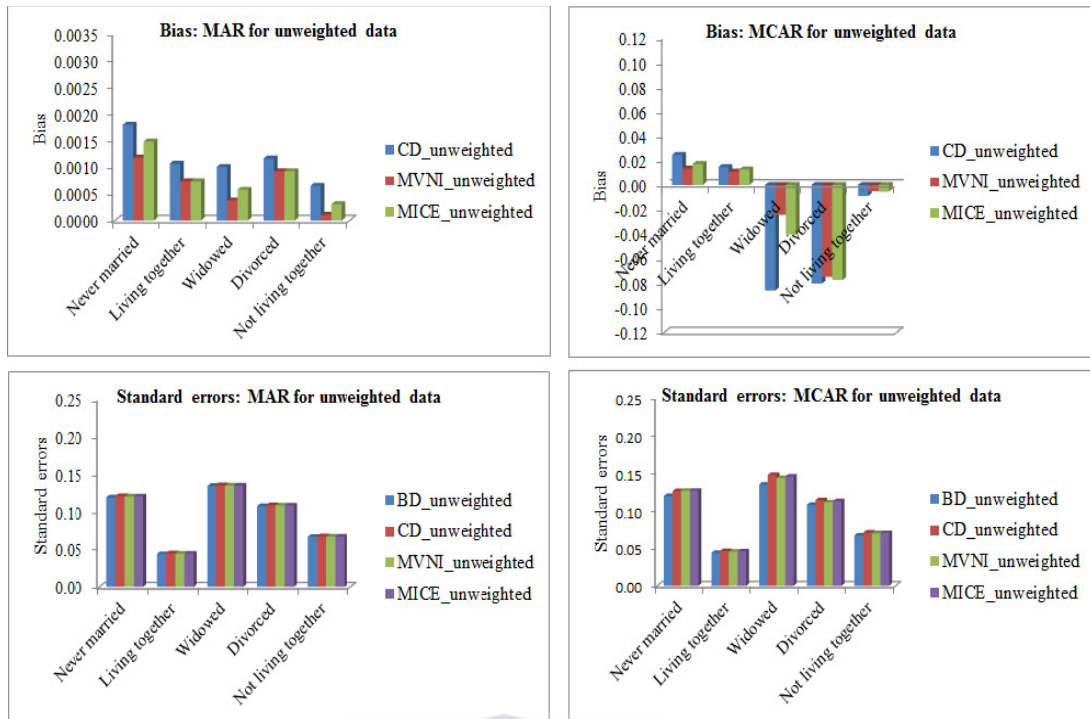


FIGURE 5.6: Model 1.1: Plot of bias and standard errors when 10% of the data are MAR and MCAR for unweighted data sets.

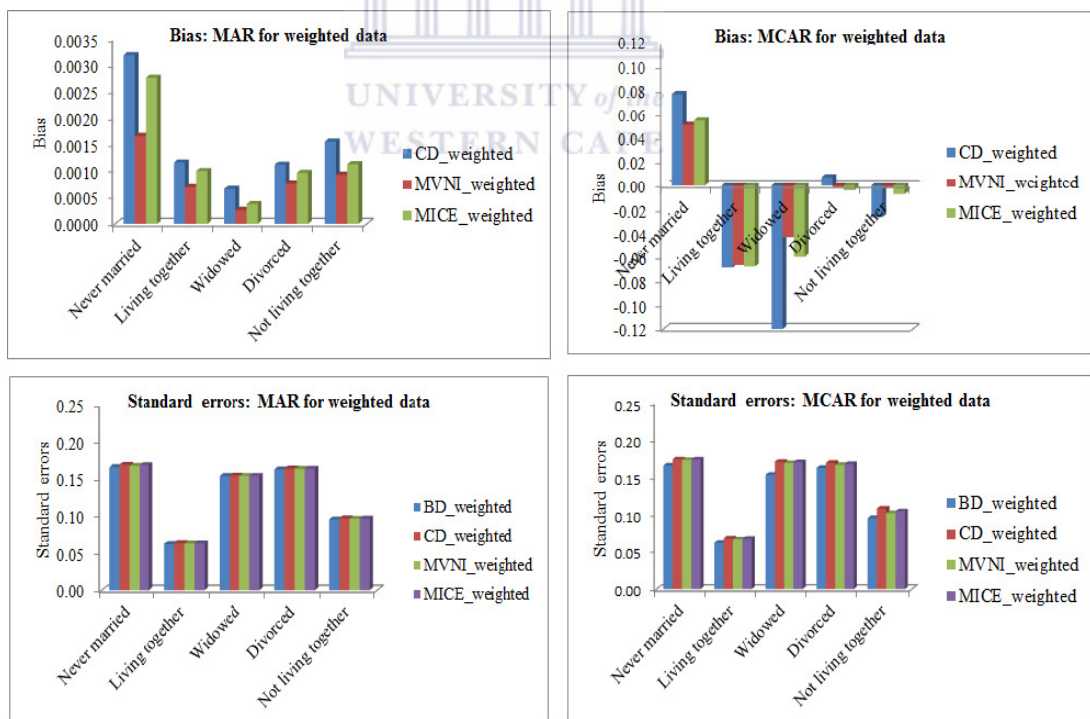


FIGURE 5.7: Model 1.1: Plot of bias and standard errors when 10% of the data are MAR and MCAR for weighted data sets.

Model diagnostics The estimates of Monte Carlo errors (MCE) after MVNI and MICE of the statistics involved in the estimation of the regression models were estimated. These estimates are presented in Appendix C in Tables 6.17, 6.18, 6.19 and 6.20 when 10% data were missing at random on marital status if a woman was not using any contraceptive method, and in Appendix C in Tables 6.21, 6.22, 6.23 and 6.24 when the same proportion of data were missing completely at random on marital status. The results show that the suggested criteria for the Monte Carlo errors are met. In fact, the Monte Carlo errors on the coefficients are less than 10% of the standard error for unweighted and weighted data sets. The Monte Carlo errors of the p-values are also approximately less than 0.1 when 5% level of significance was used for both MVNI and MICE. The Monte Carlo errors of the t-test statistic were found to be approximately less than 0.1 for all the methods and data sets. Therefore, based on these results, one can be reasonably sure about their statistical reproducibility. This suggests that the number of imputations used (100 imputations) were enough to produce stable results.

To assess whether imputations converged to the desired distributions, the convergence check was done for both unweighted and weighted data sets, under MAR and MCAR assumptions. The estimates of the worst linear function (WLF) were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE methods. Under MAR assumption, the results are presented in Appendix D in Figures 6.57 and 6.58 for the MVNI approach, and in Figures 6.59 and 6.60 for the MICE technique for unweighted data sets. For weighted data sets, the results are shown in Figures 6.61 and 6.62 for MVNI, and in Figures 6.63 and 6.64 for MICE. Under MCAR assumption, similar results are presented in Figures 6.65, 6.66, 6.67 and 6.68 for unweighted data and in Figures 6.69, 6.70, 6.71 and 6.72 for weighted data. As shown by the figures, the plots of the estimates of WLF against the iteration numbers show no visible trend, thus indicating that convergence is assured with the number of iterations used (1000 iterations). On the other hand, the plots of WLF's estimates against the lag numbers show the autocorrelations that die off quickly, which implies that even a smaller number (of iterations) than what was used, such as 10 iterations between imputations, can be used to obtain independent samples.

5.2.2 Model 1.2: Binary logistic regression model with more than two covariates in which two are measured on a nominal scale containing missing values

5.2.2.1 Model 1.2.1: Model with two nominal covariates both with 50% of their values missing at random or completely at random

Descriptive statistics Data sets with 50% missing at random or missing completely at random were considered for analysis to determine the performance of MVNI and MICE when more than one covariates with missing values are used. These covariates are the woman's marital status and region of origin, which are both categorical variables that have no natural order. Under the MAR assumption, 50% missing values were randomly deleted on marital status and region if a woman was not using any contraceptive method. This allowed missingness on these variables to depend on contraceptive method use status. As shown in Table 5.25, deleting 50% of the data at random on marital status and region when a woman was not using any contraceptive method led to approximately 39% missing values on these two variables.

TABLE 5.25: Model 1.2.1: Frequency distribution of missingness when 50% of the data are MAR on marital status and region if a woman is not using any contraceptive method.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	18002	60.92	60.92
Missing	11546	39.08	100.00
Total	29548	100.00	

To investigate whether data were missing at random, a bivariate analysis of the data was conducted to see if the proportions or means of missing values differed across the demographic particulars of the respondents. The results are shown in Table 5.26 for categorical variables (contraceptive method use status and wealth index) and in Appendix B in Figures 6.13 and 6.14 for continuous variables (age and education respectively). It is observed that missingness is related to contraceptive method use status, wealth index, age and education in completed years as indicated by the p-values that are less than the significance level of 5% that was used.

TABLE 5.26: Model 1.2.1: Distribution of missingness across selected categorical when 50% of the data are MAR on marital status and region if a woman is not using any contraceptive method.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.000
No method	50.38	49.62	100	
At least one method	100	0.00	100	
Total	60.92	39.08	100	
Wealth index				0.000
Poorest	56.16	43.84	100.00	
Poorer	58.91	41.09	100.00	
Middle	59.74	40.26	100.00	
Richer	62.31	37.69	100.00	
Richest	69.03	30.97	100.00	
Total	60.92	39.08	100.00	

Under the MCAR assumption, 50% missing values were also arbitrary deleted on marital status and region such that missingness was not associated with any variables in the data set subject to analysis. The frequency distribution of missingness on these variables is presented in Table 5.27. As indicated, around 50% of values were deleted on marital status and region.

TABLE 5.27: Model 1.2.1: Frequency distribution of missingness when 50% of the data are MCAR on marital status and region.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	14607	49.43	49.43
Missing	14941	50.57	100.00
Total	29548	100.00	

To determine whether these values were deleted completely at random, a bivariate analysis of the missing value indicator or missingness and other socio-demographic characteristics of the respondents was conducted. The results are shown in Table 5.28 for categorical variables and in Appendix B in Figures 6.15 and 6.16 for continuous variables (age and education respectively). As indicated, missingness on marital status and region is not associated with the woman contraceptive method use status, wealth index, age and education in completed years as confirmed by the chi-square test for association (categorical variables) and the independent-samples t-test (for continuous variables) with p-values greater than the significance level of

0.05. This is indication that data were missing completely at random on marital status.

TABLE 5.28: Model 1.2.1: Distribution of missingness when 50% of the data are MCAR on marital status and region.

	Missingness			
	Not Missing	Missing	Total	P-values
Contraceptive method use status				0.280
No method	49.27	50.73	100	
At least one method	50.04	49.96	100	
Total	49.43	50.57	100	
Wealth index				0.053
Poorest	47.83	52.17	100.00	
Poorer	49.69	50.31	100.00	
Middle	49.92	50.08	100.00	
Richer	50.34	49.66	100.00	
Richest	49.57	50.43	100.00	
Total	49.43	50.57	100.00	

Performance measures To determine the performance of the imputation methods of interest (MVNI and MICE), the logistic regression models of contraceptive method use status on marital status and region were first estimated using the data set with no missing values, the data set with missing values and the completed or imputed data sets with MVNI and MICE. Then, the bias was computed and reported along with the standards errors. In Tables 5.29 and 5.30, these estimates are presented for both weighted and unweighted data sets, when approximately 50% of data are missing at random on marital status if a woman is not using any contraceptive method. In Tables 5.31 and 5.32, the same parameters are reported when approximately 50% data are missing completely at random on marital status. The plots of bias and standard errors are shown in Figure 5.8 and 5.9 for both unweighted and weighted data sets to examine the pattern of performance of the missing data methods of interest; case deletion (CD), MVNI and MICE. The results show that under MAR and MCAR assumptions, multiple imputations with MVNI and MICE yields less bias and more accurate standard errors than case deletion which discards items with missing values from the analysis. It can also be seen that MVNI technique is less biased and yields more accurate standard errors than MICE, either when data are missing at random or missing

completely at random for both unweighted (Figure 5.8) and weighted (Figure 5.9) data sets.

TABLE 5.29: Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the unweighted data set.

Variable	CD	MVNI	MICE
Marital status			
Never married	0.038 (0.170)	0.015 (0.136)	0.033 (0.142)
Living together	-0.057 (0.055)	-0.028(0.050)	-0.044 (0.052)
Widowed	0.082(0.151)	0.054 (0.140)	0.062 (0.141)
Divorced	0.127 (0.130)	0.109 (0.120)	0.119 (0.124)
Not living together	0.009 (0.080)	0.002 (0.075)	-0.004 (0.076)
Region			
Bas Kongo	-0.046 (0.068)	-0.028 (0.063)	-0.040 (0.065)
Bandundu	0.041 (0.065)	0.025 (0.061)	0.031 (0.062)
Equateur	-0.057 (0.066)	-0.023 (0.064)	-0.036 (0.065)
Orientale	0.066 (0.073)	0.037 (0.071)	0.052 (0.072)
Nord Kivu	0.096 (0.073)	0.085 (0.071)	0.089 (0.072)
Maniema	0.062 (0.073)	0.039 (0.068)	0.046 (0.069)
Sud Kivu	0.071 (0.080)	0.037 (0.077)	0.050 (0.078)
Katanga	-0.037 (0.067)	-0.018 (0.065)	-0.032 (0.066)
Kasai Occidental	0.029 (0.078)	0.013 (0.073)	0.022 (0.076)
Kasai Oriental	0.042 (0.076)	0.009 (0.073)	0.034 (0.075)

TABLE 5.30: Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the weighted data.

Variable	CD	MVNI	MICE
Marital status			
Never married	0.116 (0.217)	0.074 (0.194)	0.085 (0.209)
Living together	-0.055 (0.076)	-0.035 (0.070)	-0.036 (0.071)
Widowed	0.051 (0.178)	0.026 (0.163)	0.040 (0.164)
Divorced	0.077 (0.185)	0.142 (0.177)	0.066 (0.183)
Not living together	0.054 (0.117)	0.014 (0.109)	0.034 (0.112)
Region			
Bas Kongo	-0.064 (0.088)	-0.028 (0.078)	-0.042 (0.081)
Bandundu	0.061 (0.086)	0.037 (0.079)	0.053 (0.081)
Equateur	0.071 (0.087)	0.041 (0.086)	0.052 (0.087)
Orientale	0.033 (0.101)	0.011 (0.097)	0.023 (0.099)
Nord Kivu	-0.043 (0.102)	-0.031 (0.096)	-0.034 (0.101)
Maniema	-0.030 (0.101)	-0.006 (0.096)	-0.020 (0.100)
Sud Kivu	-0.038 (0.116)	-0.022 (0.114)	-0.029 (0.114)
Katanga	0.056 (0.079)	0.038 (0.077)	0.040 (0.079)
Kasai Occidental	-0.068 (0.096)	-0.027 (0.089)	-0.037 (0.089)
Kasai Oriental	0.057 (0.093)	0.002 (0.089)	0.026 (0.092)

TABLE 5.31: Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the unweighted data set.

Variable	CD	MVNI	MICE
Marital status			
Never married	-0.087 (0.269)	-0.064 (0.182)	-0.076 (0.207)
Living together	0.078 (0.095)	0.045 (0.069)	0.065 (0.071)
Widowed	0.131 (0.271)	0.086 (0.155)	0.115 (0.189)
Divorced	0.037 (0.228)	0.018 (0.154)	0.021 (0.158)
Not living together	0.053 (0.146)	0.036 (0.102)	0.042 (0.106)
Region			
Bas Kongo	-0.096 (0.117)	-0.060 (0.077)	-0.074 (0.079)
Bandundu	-0.148 (0.114)	-0.078 (0.076)	-0.088 (0.078)
Equateur	-0.092 (0.121)	-0.052 (0.078)	-0.056 (0.079)
Orientale	-0.208 (0.140)	-0.122 (0.094)	-0.138 (0.099)
Nord Kivu	-0.013 (0.131)	-0.011 (0.086)	-0.012 (0.092)
Maniema	-0.061 (0.126)	-0.016 (0.087)	-0.037 (0.088)
Sud Kivu	-0.083 (0.149)	-0.033 (0.102)	-0.054 (0.103)
Katanga	-0.098 (0.122)	-0.036 (0.084)	-0.049 (0.088)
Kasai Occidental	-0.068 (0.130)	-0.042 (0.090)	-0.047 (0.092)
Kasai Oriental	-0.033 (0.135)	-0.020 (0.088)	-0.028 (0.099)

TABLE 5.32: Model 1.2.1: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the weighted data set.

Variable	CD	MVNI	MICE
Marital status			
Never married	-0.041 (0.329)	-0.015 (0.241)	-0.025 (0.245)
Living together	0.042 (0.135)	0.027 (0.092)	0.035 (0.094)
Widowed	0.050 (0.314)	0.006 (0.167)	0.029 (0.259)
Divorced	-0.033 (0.349)	-0.015 (0.211)	-0.020 (0.220)
Not living together	0.046 (0.209)	0.024 (0.136)	0.034 (0.142)
Region			
Bas Kongo	-0.071 (0.133)	-0.028 (0.110)	-0.054 (0.114)
Bandundu	0.058 (0.150)	0.037 (0.097)	0.049 (0.099)
Equateur	-0.053 (0.165)	-0.017 (0.103)	-0.035 (0.104)
Orientale	-0.059 (0.196)	-0.034 (0.119)	-0.044 (0.123)
Nord Kivu	-0.093 (0.179)	-0.052 (0.125)	-0.058 (0.132)
Maniema	0.053 (0.145)	0.016 (0.131)	0.030 (0.138)
Sud Kivu	-0.040 (0.213)	-0.026 (0.143)	-0.034 (0.147)
Katanga	-0.047 (0.145)	-0.025 (0.094)	-0.033 (0.112)
Kasai Occidental	-0.123 (0.161)	-0.065 (0.104)	-0.079 (0.116)
Kasai Oriental	0.127 (0.163)	0.043 (0.115)	0.095 (0.109)

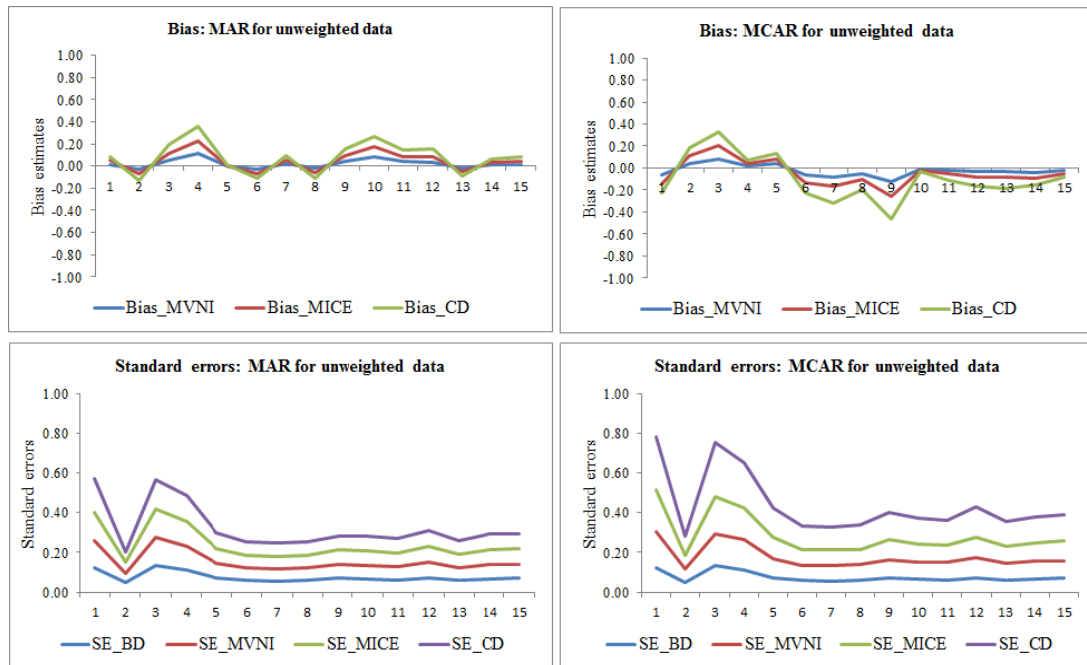


FIGURE 5.8: Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for unweighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively.

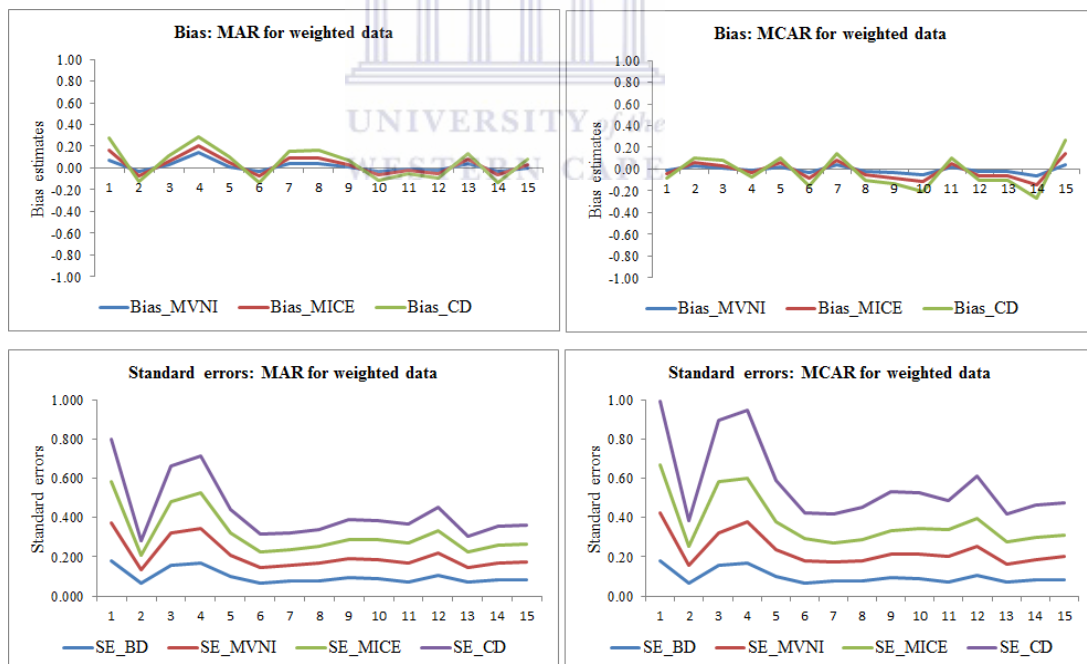


FIGURE 5.9: Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for weighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively.

Model diagnostics The estimates of Monte Carlo errors (MCE) after MVNI and MICE of the statistics involved in the estimation of the regression models

were estimated. These estimates are presented in Appendix C in Tables 6.25, 6.26, 6.27 and 6.28 when 50% data were missing at random on marital status and region if a woman was not using any contraceptive method, and in Appendix C in Tables 6.29, 6.30, 6.31 and 6.32 when data were missing completely at random on marital status. The results indicate that Monte Carlo errors on the coefficients are less than 10% of the standard error for unweighted and weighted data sets. The Monte Carlo errors of the p-values are also approximately less than 0.1 when 5% level of significance was used for both MVNI and MICE. The Monte Carlo errors of the t-test statistic were found to be approximately less than 0.1 for all the methods and data sets. Therefore, based on these results, one can reasonably be sure about their statistical reproducibility. This suggests that the number of imputations used (100 imputations) were enough to produce stable results.

To ensure that imputations converged to the desired distributions, convergence was assessed for both unweighted and weighted data sets, under MAR and MCAR assumptions. The estimates of the worst linear function (WLF) were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE methods. Under the MAR assumption, the results are presented in Appendix D in Figures 6.73 and 6.74 for the MVNI approach, and in Figures 6.75 and 6.76 for the MICE technique for unweighted data sets. For weighted data sets, the results are shown in Figures 6.77 and 6.78 for MVNI, and in Figures 6.79 and 6.80 for MICE. Under the MCAR assumption, similar results are presented in Figures 6.81, 6.82, 6.83, 6.84, 6.85, 6.86, 6.87 and 6.88. As indicated, the plots of the estimates of WLF against the iteration numbers show no visible trend, thus indicating that convergence is assured with the number of iterations that were used (1000 iterations). On the other hand, the plots of WLF's estimates against the lag numbers show the autocorrelations that die off quickly, which implies that even a smaller number (of iterations) than what was used, such as 10 iterations between imputations, can be used to obtain independent samples.

5.2.2.2 Model 1.2.2: Model with various covariates with 50% missing values at random or completely at random on only the unordered categorical ones

Descriptive statistics A binary logistic regression model of contraceptive method use status on various covariates was fitted to determine the performance of MICE and MVNI on unordered or nominal variables (marital status and region)

controlling for other types of variables (continuous and ordinal). Data sets with 50% missing at random or missing completely at random data on the woman's marital status and region were also considered for analysis.

Under the MAR assumption, 50% missing values were randomly deleted on marital status and region if a woman was not using any contraceptive method. This allowed the missing value indicator or missingness to depend on contraceptive method use status. Under the MCAR assumption, 50% missing values were also arbitrary deleted on marital status and region such that missingness was not associated with any variables in the data set subject to analysis. The descriptive statistics of missingness on these variables is the same as in Model 1.2.1, as the same rate of missing values were deleted on the same variables. The difference is on the regression models that were estimated for these two models. In fact, a regression model was estimated using only two nominal variables as independent variables in Model 1.2.1, whereas various variables; continuous, ordinal and nominal, were considered as independent variables in Model 1.2.2.

Performance measures To determine the performance of the MVNI and MICE techniques, the logistic regression models of contraceptive method use status on marital status and region were estimated using the data set with no missing values, the data set with missing values and the completed or imputed data sets with MVNI and MICE. Then, the bias was computed and reported along with the standard errors. In Tables 5.33 and 5.34, these estimates are presented for both weighted and unweighted data sets, when approximately 50% of data are missing at random on marital status if a woman is not using any contraceptive method. In Tables 5.35 and 5.36, the same statistics are reported when approximately 50% data are missing completely at random on marital status. The plot of bias and standard errors are shown in Figure 5.10 and 5.11 for both unweighted and weighted data sets to look at the pattern of performance of the missing data methods of interest; case deletion (CD), MVNI and MICE. The results show that under MAR and MCAR assumptions, multiple imputations with MVNI and MICE yields less bias in regression coefficients and more accurate standard errors than case deletion. It can also be seen that MVNI technique is less biased and yields better standard errors than MICE, either when data are missing at random or missing completely at random for both unweighted (Figure 5.10) and weighted (Figure 5.11) data sets.

TABLE 5.33: Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the unweighted data set.

Variable	CD	MVNI	MICE
Marital status			
Never married	0.066 (0.164)	0.014 (0.136)	0.015 (0.152)
Living together	-0.028 (0.054)	-0.024 (0.050)	-0.025 (0.052)
Widowed	0.078 (0.147)	0.031 (0.145)	0.038 (0.145)
Divorced	0.111 (0.128)	0.089 (0.123)	0.102 (0.123)
Not living together	0.037 (0.079)	0.017 (0.077)	0.026 (0.077)
Region			
Bas Kongo	-0.025 (0.082)	0.001 (0.076)	-0.002 (0.079)
Bandundu	0.035 (0.082)	0.026 (0.077)	0.029 (0.080)
Equateur	-0.043 (0.082)	-0.035 (0.080)	-0.037 (0.081)
Orientale	0.034 (0.088)	0.016 (0.086)	0.029 (0.087)
Nord Kivu	0.041 (0.084)	0.004 (0.082)	0.017 (0.083)
Maniema	0.032 (0.084)	0.018 (0.082)	0.026 (0.083)
Sud Kivu	0.041 (0.089)	0.030 (0.085)	0.038 (0.086)
Katanga	-0.016 (0.078)	-0.011 (0.073)	-0.014 (0.075)
Kasai Occidental	0.029 (0.091)	0.008 (0.083)	0.016 (0.089)
Kasai Oriental	-0.055 (0.091)	-0.019 (0.085)	-0.034 (0.087)
Age	-0.004 (0.006)	-0.001 (0.003)	-0.002 (0.004)
Education	0.002 (0.009)	0.001 (0.006)	0.001 (0.007)
Wealth index			
Poorer	0.023 (0.057)	0.001 (0.053)	0.020 (0.054)
Middle	0.036 (0.056)	0.022 (0.053)	0.027 (0.053)
Richer	-0.032 (0.057)	-0.019 (0.053)	-0.024 (0.054)
Richest	-0.035 (0.078)	-0.025 (0.066)	-0.032 (0.067)

TABLE 5.34: Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MAR on variables marital status and region if a woman is not using any contraceptive method: results from the weighted data set.

Variable	CD	MVNI	MICE
Marital status			
Never married	0.115 (0.209)	0.039 (0.192)	0.092 (0.204)
Living together	-0.040 (0.079)	-0.023 (0.072)	-0.034 (0.073)
Widowed	-0.073 (0.173)	-0.032 (0.166)	-0.047 (0.171)
Divorced	0.158 (0.191)	0.071 (0.181)	0.075 (0.183)
Not living together	-0.035(0.111)	-0.010(0.107)	-0.029(0.109)
Region			
Bas Kongo	-0.404 (0.101)	-0.012 (0.093)	-0.387 (0.095)
Bandundu	0.033 (0.107)	0.021 (0.100)	0.027 (0.103)
Equateur	-0.137 (0.108)	-0.075 (0.106)	-0.108 (0.107)
Orientale	-0.029 (0.114)	-0.022 (0.110)	-0.023 (0.113)
Nord Kivu	-0.065 (0.115)	-0.034 (0.112)	-0.036 (0.114)
Maniema	-0.052 (0.110)	-0.036 (0.098)	-0.045 (0.104)
Sud Kivu	-0.026 (0.125)	-0.020 (0.120)	-0.022 (0.121)
Katanga	0.060 (0.089)	0.047 (0.086)	0.058 (0.087)
Kasai Occidental	0.038 (0.110)	0.016 (0.099)	0.017 (0.100)
Kasai Oriental	0.026 (0.110)	0.012 (0.105)	0.012(0.109)
Age	0.005 (0.005)	0.002 (0.003)	0.003 (0.004)
Education	-0.008 (0.010)	-0.004 (0.008)	-0.006 (0.009)
Wealth index			
Poorer	-0.033 (0.088)	-0.012 (0.079)	-0.013 (0.084)
Middle	0.023 (0.088)	0.010 (0.082)	0.013 (0.084)
Richer	0.030 (0.084)	0.020 (0.080)	0.026 (0.081)
Richest	0.074 (0.094)	0.048 (0.089)	0.054 (0.090)

TABLE 5.35: Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the unweighted data set.

Variable	CD	MVNI	MICE
Marital status			
Never married	-0.193 (0.276)	-0.093 (0.184)	-0.097 (0.188)
Living together	0.098 (0.098)	0.069 (0.069)	0.071 (0.072)
Widowed	-0.152 (0.276)	-0.119 (0.140)	-0.129 (0.192)
Divorced	0.198 (0.232)	0.147 (0.157)	0.158 (0.158)
Not living together	-0.181 (0.148)	-0.127 (0.106)	-0.148 (0.102)
Region			
Bas Kongo	-0.078 (0.139)	-0.055 (0.100)	-0.060 (0.100)
Bandundu	-0.208 (0.142)	-0.168 (0.101)	-0.178 (0.107)
Equateur	-0.279 (0.147)	-0.147 (0.102)	-0.189 (0.108)
Orientale	-0.344 (0.160)	-0.152 (0.113)	-0.226 (0.120)
Nord Kivu	-0.247 (0.149)	-0.152 (0.102)	-0.163 (0.113)
Maniema	-0.288 (0.151)	-0.184 (0.109)	-0.199 (0.116)
Sud Kivu	-0.215 (0.165)	-0.100 (0.119)	-0.116 (0.119)
Katanga	-0.306 (0.134)	-0.225 (0.094)	-0.241 (0.102)
Kasai Occidental	-0.047 (0.149)	-0.016 (0.113)	-0.035 (0.114)
Kasai Oriental	-0.312 (0.159)	-0.280 (0.107)	-0.289 (0.123)
Age	-0.064 (0.004)	-0.039 (0.002)	-0.054 (0.002)
Education	0.070 (0.009)	0.018 (0.005)	0.043 (0.005)
Wealth index			
Poorer	-0.204 (0.110)	-0.142 (0.054)	-0.162 (0.054)
Middle	0.177 (0.105)	0.118 (0.054)	0.127 (0.055)
Richer	-0.076 (0.109)	-0.046 (0.054)	-0.055 (0.055)
Richest	-0.141 (0.131)	-0.079 (0.072)	-0.094 (0.077)

TABLE 5.36: Model 1.2.2: Estimates of bias and standard errors (SE) obtained when 50% of the data are MCAR on variables marital status and region: results from the weighted data set.

Variable	CD	MVNI	MICE
Marital status			
Never married	-0.133 (0.315)	-0.068 (0.233)	-0.080 (0.246)
Living together	0.235 (0.140)	0.142 (0.094)	0.164 (0.096)
Widowed	0.320 (0.325)	0.205 (0.171)	0.205 (0.263)
Divorced	-0.252 (0.359)	-0.221 (0.213)	-0.228 (0.218)
Not living together	0.164 (0.213)	0.101 (0.136)	0.126 (0.142)
Region			
Bas Kongo	-0.136 (0.157)	-0.081 (0.130)	-0.086 (0.133)
Bandundu	-0.236 (0.185)	-0.136 (0.127)	-0.151 (0.135)
Equateur	-0.269 (0.199)	-0.166 (0.131)	-0.176 (0.135)
Orientale	-0.156 (0.215)	-0.116 (0.142)	-0.132 (0.148)
Nord Kivu	-0.262 (0.204)	-0.168 (0.149)	-0.169 (0.149)
Maniema	-0.198 (0.178)	-0.116 (0.152)	-0.123 (0.169)
Sud Kivu	-0.222 (0.227)	-0.152 (0.157)	-0.157 (0.164)
Katanga	-0.264 (0.157)	-0.149 (0.105)	-0.226 (0.125)
Kasai Occidental	0.130 (0.178)	0.091 (0.118)	0.094 (0.148)
Kasai Oriental	-0.132 (0.190)	-0.065 (0.129)	-0.072 (0.137)
Age	0.063 (0.006)	0.022 (0.004)	0.033 (0.004)
Education	-0.128 (0.013)	-0.065 (0.007)	-0.080 (0.008)
Wealth index			
Poorer	-0.092 (0.164)	-0.052 (0.080)	-0.055 (0.082)
Middle	-0.048 (0.164)	-0.032 (0.082)	-0.039 (0.082)
Richer	-0.155 (0.161)	-0.115 (0.079)	-0.118 (0.080)
Richest	-0.189 (0.184)	-0.132 (0.097)	-0.136 (0.100)

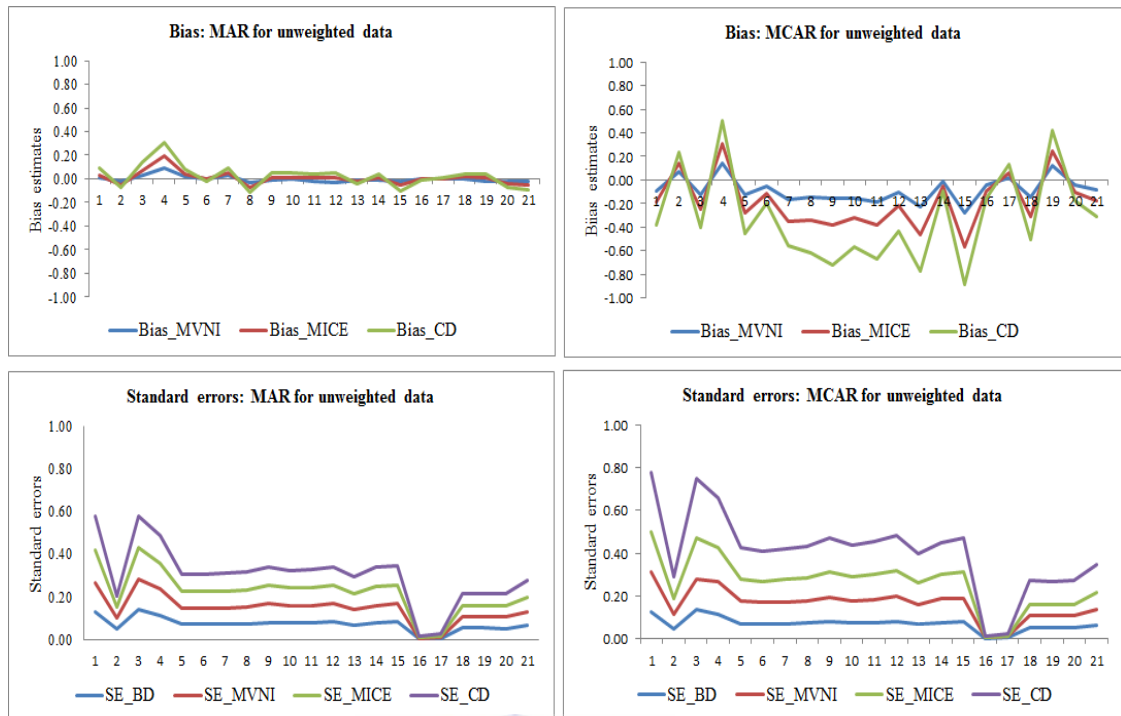


FIGURE 5.10: Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for unweighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively, 16-17 refer to variables age and education respectively, and 18-21 refer to levels of wealth index.

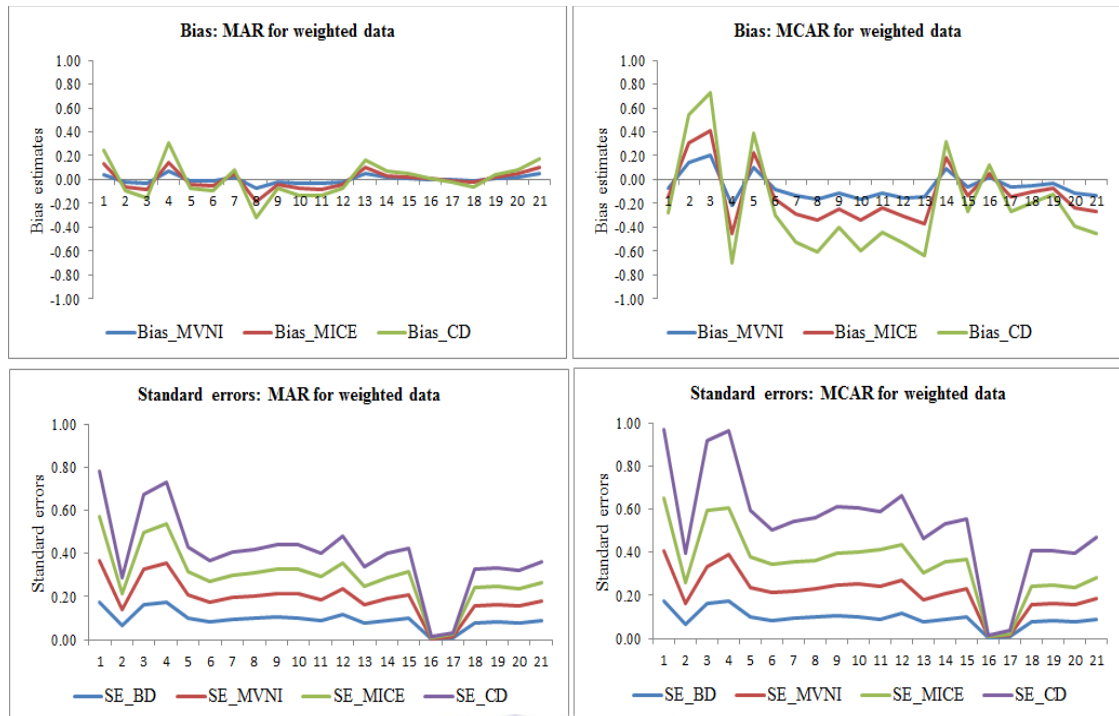


FIGURE 5.11: Model 1.2.1: Plot of bias and standard errors when 50% of the data are MAR and MCAR for weighted data sets. Numbers 1-5 and 6-15 refer to levels or categories of the variable marital status and region respectively, 16-17 refer to variables age and education respectively, and 18-21 refer to levels of wealth index.

Model diagnostics The estimates of Monte Carlo errors (MCE) after MVNI and MICE of the statistics involved in the estimation of the regression models were computed. These estimates are presented in Appendix C in Tables 6.33, 6.34, 6.35 and 6.36 when 50% data were missing at random on marital status if a woman was not using any contraceptive method, and in Appendix C in Tables 6.37, 6.38, 6.39 and 6.40 when data were missing completely at random on marital status and region. The results show that the suggested criteria for the Monte Carlo errors are met. In fact, the Monte Carlo errors on the coefficients are less than 10% of the standard error for unweighted and weighted data sets. The Monte Carlo errors of the p-values are also approximately less than 0.1 when 5% level of significance was used for both MVNI and MICE. The Monte Carlo errors of the t-test statistic were found to be approximately less than 0.1 for all the methods and data sets. Therefore, based on these results, one can reasonably be sure about their statistical reproducibility. This suggests that the number of imputations used (100 imputations) were enough to produce stable results.

To ensure that imputations converged to the desired distributions, convergence was assessed for both unweighted and weighted data sets, under MAR and MCAR assumptions. The estimates of the worst linear function (WLF) were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE methods. Under the MAR assumption, the results are presented in Appendix D in Figures 6.89 and 6.90 for the MVNI approach, and in Figures 6.91 and 6.92 for the MICE technique for unweighted data sets. For weighted data sets, the results are shown in Figures 6.93 and 6.94 for MVNI, and in Figures 6.95 and 6.96 for MICE. Under the MCAR assumption, similar results were obtained and presented in Figures 6.97, 6.98, 6.99, 6.100, 6.101, 6.102, 6.103 and 6.104. As indicated, the plots of the estimates of WLF against the iteration numbers show no visible trend, thus indicating that convergence is assured with the number of iterations used (1000 iterations). On the other hand, the plots of WLF's estimates against the lag numbers show the autocorrelations that die off quickly, which implies that even a smaller number (of iterations) than what was used, such as 10 iterations between imputations, can be used to obtain independent samples.

5.2.3 Scenario 1: Summary of findings

As explained earlier, Scenario 1 investigates the behaviour of MICE and MVNI when missing values are observed on the independent variables. Two types of

results were presented for this case. These include the results of the model with a single nominal covariate containing missing values (Model 1.1), and the results of the models with at least two covariates on which missing values are found on only the unordered categorical variables (Models 1.2.1 and 1.2.2). In Model 1.1, the behaviour of MVNI and MICE was assessed using three rates of missingness, namely 50%, 30% and 10%. The purpose of doing this was first of all to determine the performance of these methods on a single unordered categorical variable alone, with no influence of other variables. Furthermore, this model was used to investigate whether the amount of missing observations in the data sets may have an impact on the performance of MVNI and MICE. The results indicated that MVNI outperformed MICE in terms of bias and standard errors. In addition, it was found that no matter what rate of missingness used, the behaviour of MVNI and MICE did not change the direction. That is, MVNI outperformed MICE at lower and higher rates of missingness in the data sets. The results of Models 1.2.1 and 1.2.2 were also presented. In Model 1.2.1, two unordered independent variables containing missing values were considered for analysis. This model was estimated to strengthen the results in Model 1.1 that assessed the performance of MVNI and MICE when missing values were present on only unordered categorical variables that were treated as predictors in the regression models. With the results from this model, MVNI outperformed MICE as well. Model 1.2.2 is an extension of Model 1.2.1 to the model with covariates of different types (nominal, continuous and ordinal). This was done to assess whether the behaviour of MICE and MVNI on unordered categorical variables may be affected when variables of other types that have no missing values are introduced in the model. It was found that this fact did not have any impact on the behaviour of MICE and MVNI when missing values were present on only the nominal variables. Indeed, MVNI still performed better than MICE when other types of variables were introduced in the model. Thus, based on all the findings in Scenario 1, it can be concluded that when missing values are observed on unordered categorical variables that are treated as predictors in the regression models, MVNI would be a better imputation technique than MICE.

5.3 Scenario 2: Logistic regression models with missing variables on the response variables

5.3.1 Model 2.1: Binary logistic regression model with missing values on the response variable

5.3.1.1 Description of data sets with missing values

In the second scenario, logistic regression models with missing values on the outcome variables were considered for analysis. The objective of doing this was to determine the behaviour of the multiple imputation methods of interest when missing values are present on unordered categorical variables which are treated as response variables in the regression models. A graphical representation of this scenario is presented in Figure 5.12.

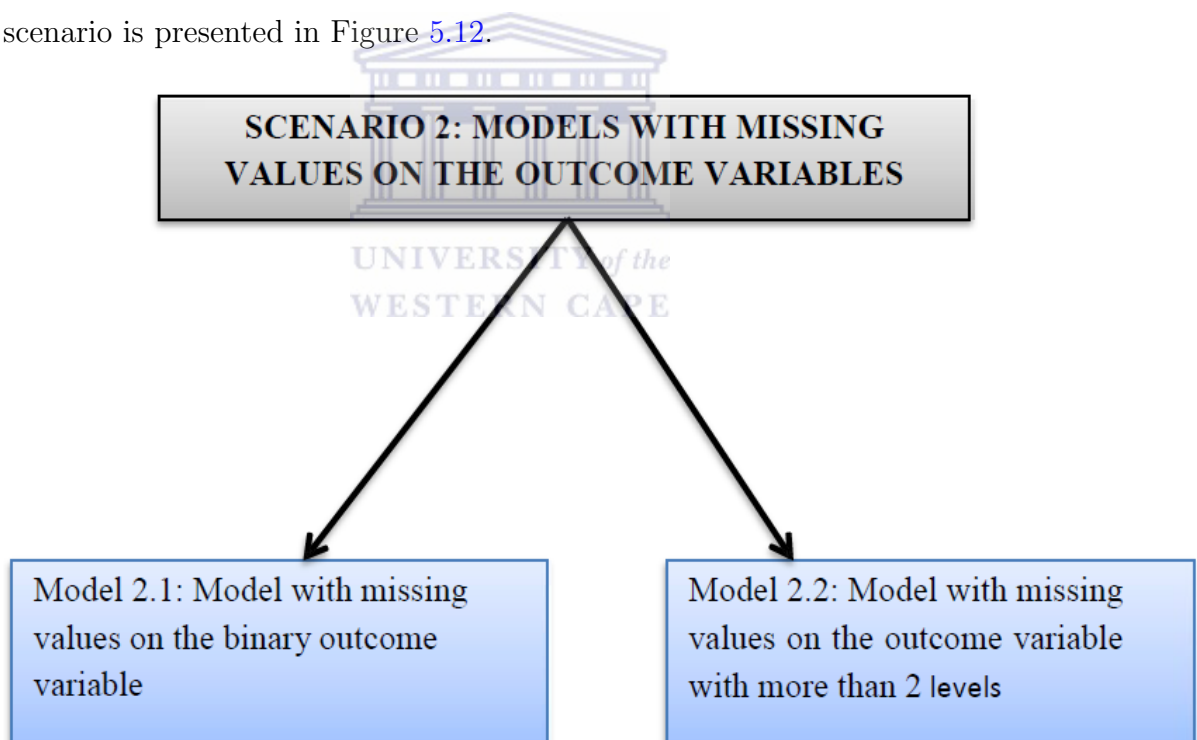


FIGURE 5.12: Scenario 2: Logistic regression models with missing variables on the response variables.

In this scenario, two logistic regression models were fitted. The first model or Model 2.1 as described in Figures 4.1 and 5.12 consists of a binary logistic regression model of contraceptive method use status on marital status. In this case, missing values were observed on the outcome variable; contraceptive method use

status, which is coded 1 if the respondent uses a contraceptive method and 0 otherwise.

Under MAR assumption, 50% missing values were randomly deleted on contraceptive method use status if a woman was aged at least 35 years, thus allowing missingness to depend on the woman's age. In Table 5.37, the frequency distribution of missingness on this variable is given. The results show that removing 50% of the values from the variable contraceptive method use status if a woman is at least 35 years old resulted in approximately 26.89% of the values missing on the whole variable.

TABLE 5.37: Model 2.1: Frequency distribution of missingness when 50% of the data are MAR on contraceptive method use status if a woman is aged at least 35 years.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	21604	73.11	73.11
Missing	7944	26.89	100.00
Total	29548	100.00	

The t-test was conducted to investigate if the means of age in single years differed across missingness. The results (see Appendix B Figure 6.17) showed that the mean age (approximately 33 years) when data are missing is different from the mean age when data are not missing (around 42 years). These findings were also supported by the t-test of equality of the means with a p-value less than 5% significance level (p-value = 0.000). This shows that missingness is associated with age of the woman, which is an indication that data are missing at random on the variable age. Similar analysis was done for the woman's education in completed years. The results indicated that this variable was associated with missingness as was confirmed by the p-value (p-value = 0.000) associated with the independent samples t-test for the equality of the means of missing and not missing (see Appendix B Figure 6.18).

Beside being related to age (a sufficient condition to meet the MAR assumption) and education in completed years, missingness was also found to be associated with the woman's marital status, wealth index and region of origin as shown in Table 5.38. As indicated, the proportions of missing values on contraceptive methods use status varied across the levels or categories of these variables. The Chi-square test for association confirmed also the existing relationship between missingness and these variables as the p-values associated with the independent samples t-test were less than 5% significance level. These results are in fact an

additional information that the MAR assumption on contraceptive method use status was met.

TABLE 5.38: Model 2.1: Distribution of missingness across categorical variables when 50% of the data are MAR on contraceptive method use status if a woman is aged at least 35 years.

	Missingness			P-values
	Not Missing	Missing	Total	
Marital status				0.000
Never married	92.26	7.74	100.00	
Married	73.42	26.58	100.00	
Living together	77.55	22.45	100.00	
Widowed	53.57	46.43	100.00	
Divorced	62.06	37.94	100.00	
Not living together	74.74	25.26	100.00	
Total	73.11	26.89	100.00	
Region				0.000
Kinshasa	70.92	29.08	100.00	
Bas Congo	72.65	27.35	100.00	
Bandundu	70.61	29.39	100.00	
Equateur	72.23	27.77	100.00	
Oriental	72.98	27.02	100.00	
Nord Kivu	75.18	24.82	100.00	
Maniema	75.31	24.69	100.00	
Sud Kivu	74.10	25.90	100.00	
Katanga	71.65	28.35	100.00	
Kasai Oriental	74.51	25.49	100.00	
Kasai Occidental	75.00	25.00	100.00	
Total	73.11	26.89	100.00	
Wealth index				0.002
Poorest	73.85	26.15	100.00	
Poorer	72.13	27.87	100.00	
Middle	71.75	28.25	100.00	
Richer	74.61	25.39	100.00	
Richest	73.20	26.80	100.00	
Total	73.11	26.89	100.00	

Under MCAR assumption, approximately 50% of the data were arbitrary deleted at random on contraceptive methods use status. The results indicate that approximately 50% of the values were deleted on marital status (see Table 5.39).

TABLE 5.39: Model 2.1: Frequency distribution of missingness when 50% of the data are MCAR on contraceptive method use status.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	14858	50.28	550.28
Missing	14690	49.72	100.00
Total	29548	100.00	

To assess whether data were missing completely at random on contraceptive methods use status, the bivariate analysis was conducted to investigate whether missingness was associated with variables in the data set. As indicated in Table 5.40, deleting 50% values on contraceptive method use status resulted in non significant differences in proportions of missingness across the marital status, wealth index and region categories. The Chi-square test for association confirmed also that there was no statistical significant difference in proportions of missing values of these groups as the p-value associated with this test were greater than the significance level of 0.05 that was used. The independent samples t-test showed also that there was no significant difference in means of age and education in completed years across missingness on contraceptive method use status. The results are found in Appendix B in Figures 6.19 and 6.20 for age and education respectively.

TABLE 5.40: Model 2.1: Distribution of missingness across marital status when 50% of the data are MCAR on contraceptive method use status.

	Missingness			
	Not Missing	Missing	Total	P-values
Marital status				0.673
Never married	51.79	48.21	100.00	
Married	50.11	49.89	100.00	
Living together	50.82	49.18	100.00	
Widowed	49.64	50.36	100.00	
Divorced	49.92	50.08	100.00	
Not living together	52.05	47.95	100.00	
Total	50.28	49.72	100.00	
Region				0.068
Kinshasa	52.44	47.56	100.00	
Bas Congo	51.01	48.99	100.00	
Bandundu	51.31	48.69	100.00	
Equateur	48.47	51.53	100.00	
Oriental	50.25	49.75	100.00	
Nord Kivu	50.19	49.81	100.00	
Maniema	49.20	50.80	100.00	
Sud Kivu	49.73	50.27	100.00	
Katanga	51.50	48.50	100.00	
Kasai Oriental	48.98	51.02	100.00	
Kasai Occidental	49.76	50.24	100.00	
Total	50.28	49.72	100.00	
Wealth index				0.061
Poorest	49.36	50.64	100.00	
Poorer	50.69	49.31	100.00	
Middle	49.93	50.07	100.00	
Richer	49.81	50.19	100.00	
Richest	51.92	48.08	100.00	
Total	50.28	49.72	100.00	

5.3.1.2 Performance measures

The binary logistic regression models of contraceptive method use status on marital status were estimated using the data set with no missing values, the data set with missing at random values and the completed or imputed data sets with MVNI and MICE. Thereafter, the bias and standard errors' estimates were reported and used to compare the multiple imputation methods of interest, namely MVNI and MICE. These estimates are reported in Table 5.41 (bias estimates) and Table 5.42 (standard errors) for MAR data and in Table 5.43 (bias) and in

Table 5.44 (standard errors) for MCAR data. In Figures 5.13 and 5.14, the estimates of bias and standard errors when data are MCAR and MAR are plotted for unweighted and weighted data sets respectively. It can be seen that MVNI and MICE produced less bias than case deletion method. The figure shows also that when data are missing either at random or completely at random on the dependent variable, MICE produces more accurate results than MVNI for both unweighted and weighted data sets respectively.

TABLE 5.41: Model 2.1: Estimates of bias when approximately 50% of the data are MAR on contraceptive method use status if a woman is aged at least 35 years.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	-0.014	0.022	0.204	0.167	0.058
CD-weighted	-0.039	0.050	0.329	0.072	0.091
MVNI-unweighted	-0.009	0.005	0.173	0.158	0.057
MVNI-weighted	-0.019	0.060	0.198	0.068	0.087
MICE-unweighted	0.000	-0.007	0.157	0.147	0.025
MICE-weighted	-0.013	0.029	0.169	0.060	0.074

TABLE 5.42: Model 2.1: Estimates of standard errors when approximately 50% of the data are MAR on contraceptive method use status if a woman is aged 35 years or more.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.135	0.053	0.172	0.138	0.078
CD-weighted	0.189	0.073	0.206	0.192	0.118
MVNI-unweighted	0.124	0.049	0.166	0.128	0.076
MVNI-weighted	0.173	0.071	0.190	0.186	0.109
MICE-unweighted	0.123	0.050	0.164	0.124	0.075
MICE-weighted	0.172	0.070	0.180	0.172	0.109

TABLE 5.43: Model 2.1: Estimates of bias when approximately 50% of the data are MCAR on contraceptive method use status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.367	-0.067	-0.094	-0.028	0.050
CD-weighted	0.327	-0.034	-0.424	-0.092	-0.080
MVNI-unweighted	0.329	-0.041	-0.039	0.014	0.037
MVNI-weighted	0.173	-0.015	-0.084	-0.083	-0.065
MICE-unweighted	0.315	-0.024	-0.019	0.011	0.019
MICE-weighted	0.125	-0.002	-0.022	-0.055	-0.018

TABLE 5.44: Model 2.1: Estimates of standard errors when approximately 50% of the data are MCAR on contraceptive method use status.

	Marital status				
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.119	0.044	0.135	0.108	0.067
BD-weighted	0.166	0.062	0.154	0.163	0.095
CD-unweighted	0.160	0.063	0.195	0.156	0.095
CD-weighted	0.221	0.089	0.227	0.219	0.136
MVNI-unweighted	0.142	0.056	0.162	0.133	0.091
MVNI-weighted	0.197	0.083	0.172	0.182	0.119
MICE-unweighted	0.139	0.054	0.148	0.118	0.082
MICE-weighted	0.191	0.078	0.168	0.174	0.118

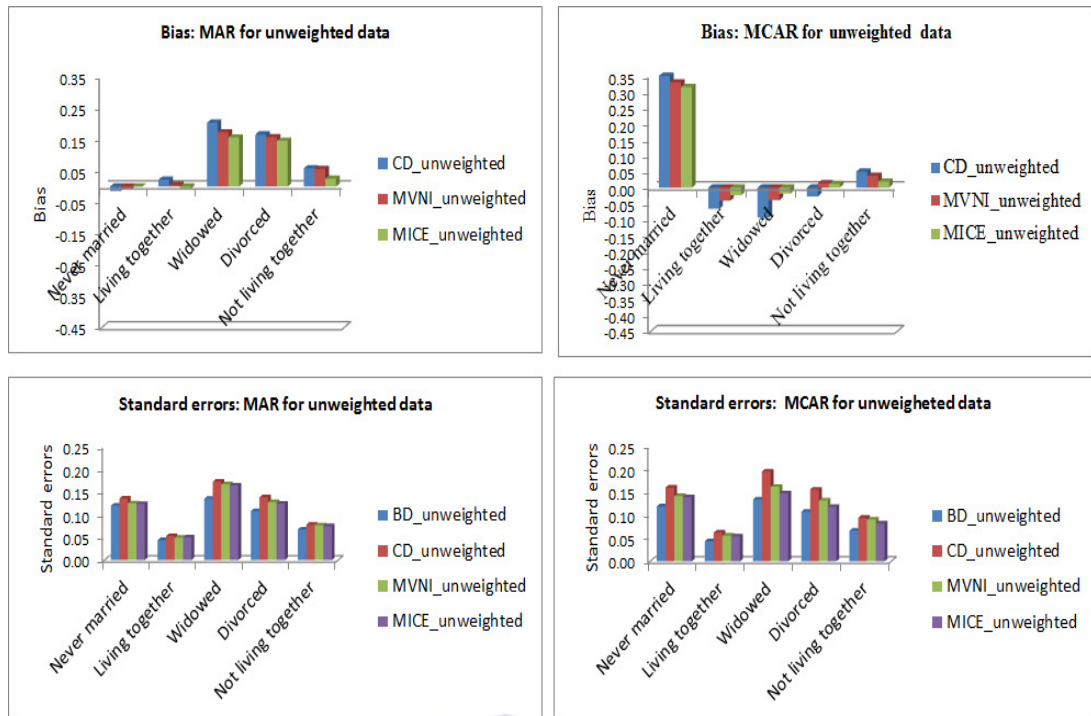


FIGURE 5.13: Model 2: Plot of bias and standard errors when 50% of the data are MAR and MCAR for unweighted data sets.

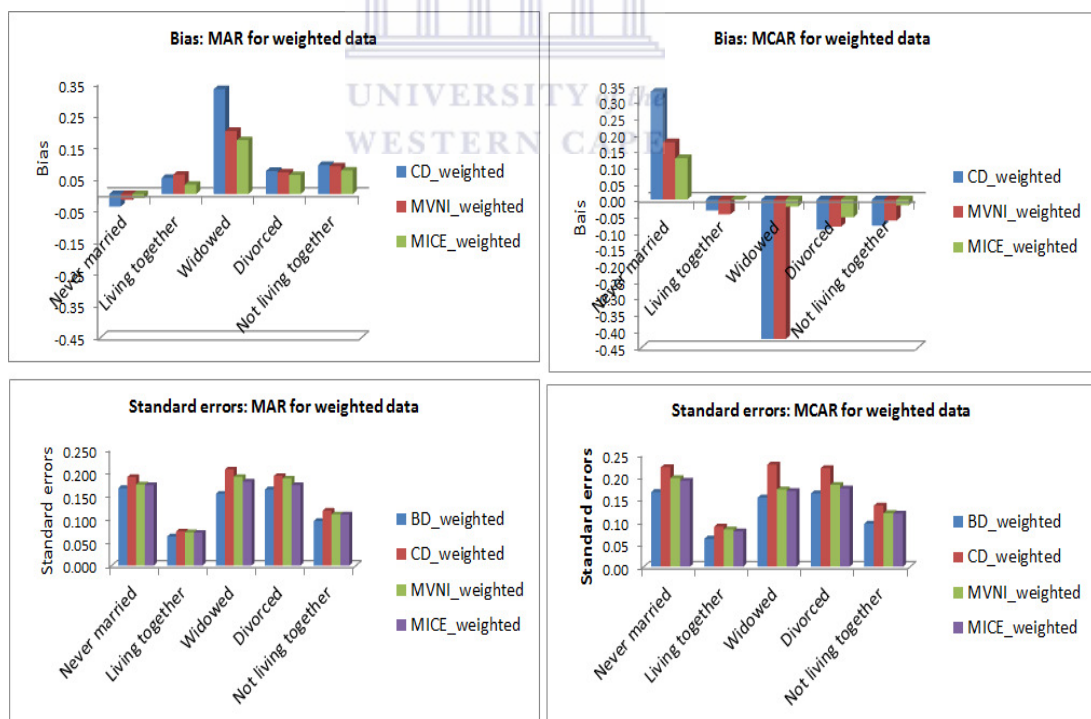


FIGURE 5.14: Model 2: Plot of bias and standard errors when 50% of the data are MAR or MCAR for weighted data sets.

5.3.1.3 Model diagnostics

The Monte Carlo errors (MCE) after MVNI and MICE for the regression models' statistics were also reported. The MCE of the coefficients were less than 10% of the standard error for the unweighted and weighted data sets, for both MVNI and MICE. Those of the p-values were also less than 0.1 when the 5% level of significance was used for both MVNI and MICE. The Monte Carlo errors of the t-test statistic were approximately less than 0.1 for all the methods and data sets. Under the MAR assumption, the results are presented in Appendix C in Tables 6.41, 6.42, 6.43 and 6.44 for MAR data and Tables 6.45, 6.46, 6.47 and 6.48 for MCAR data. Based on these results, it can be concluded that the number of imputations used (100) were sufficient to produce stable results.

The convergence of imputations to desired distributions was also investigated for both unweighted and weighted data sets, when data were MAR and MCAR on contraceptive method use status. The estimates of the WLF were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE techniques. Under the MAR assumption, the results are presented in Appendix D in Tables 6.105, 6.106, 6.107 and 6.108 for unweighted data and Tables 6.109, 6.110, 6.111 and 6.113 for weighted data. Under the MCAR assumption, similar results are provided in Figures 6.114, 6.115, 6.116, 6.117, 6.118, 6.119 and 6.120. The plots of the estimates of WLF against the iteration numbers show no visible trend for both methods, thus indicating that convergence was attained when 1000 iterations were used. On the other hand, the plots of WLF's estimates against the lag numbers showed that the autocorrelations die off quickly, which is an indication that even a smaller number of iterations than what was employed, can be used to obtain independent samples.

5.3.2 Model 2.2: Multinomial logistic regression model with missing values on the response variable

5.3.2.1 Description of data sets with missing values

As stated in the methodology chapter, the second logistic regression model in Scenario 2 is multinomial. The outcome measure is contraceptive methods use status but in this case coded 1 if a woman has not used any contraceptive method, 2 if she has used a traditional or folkloric method and 3 if she has utilised a

modern contraceptive method. It is assumed that the levels of this variable have no natural ordering. The multinomial logistic regression was used to determine the association between contraceptive methods use status and marital status.

Under MAR assumption, the frequency distribution of missingness on the dependent variable is given in Table 5.45. The results show that removing 50% of the values from the variable contraceptive method use status if a woman is at least 35 years old led to approximately 26.20% missing values on the whole variable.

TABLE 5.45: Model 2.2: Frequency distribution of missingness when 50% of the data are MAR on contraceptive method use status if a woman is at least 35 years old.

Missingness	Frequency	Percent	Cumulative frequency
Not missing	21806	73.80	73.80
Missing	7742	26.20	100.00
Total	29548	100.00	

The frequency distribution of missingness across categorical variables in the model or data set was done to investigate whether they were associated with missingness. The results indicated that beside the woman's age and education completed in years (see Appendix B in Figures 6.21 and 6.22), marital status and region were also found to be associated with missingness (see Table 5.46). Indeed, the proportions of missing values across the marital status categories varied significantly, and this was also confirmed by the Chi-square test for association with p-values less than the significance level of 5% that was used. These results indicate that the MAR assumption was met.

TABLE 5.46: Model 2.2: Distribution of missingness by marital status when approximately 50% of the data are MAR on contraceptive method use status if a woman is at least 35 years old.

	Missingness			P-values
	Not Missing	Missing	Total	
Marital status				0.000
Never married	93.15	6.85	100.00	
Married	74.13	25.87	100.00	
Living together	77.69	22.31	100.00	
Widowed	56.42	43.58	100.00	
Divorced	63.59	36.41	100.00	
Not living together	74.25	25.75	100.00	
Total	73.80	26.20	100.00	
Region				0.012
Kinshasa	72.37	27.63	100.00	
Bas Congo	73.94	26.06	100.00	
Bandundu	71.40	28.60	100.00	
Equateur	73.50	26.50	100.00	
Oriental	75.19	24.81	100.00	
Nord Kivu	74.68	25.32	100.00	
Maniema	75.82	24.18	100.00	
Sud Kivu	74.57	25.43	100.00	
Katanga	73.14	26.86	100.00	
Kasai Oriental	73.58	26.42	100.00	
Kasai Occidental	74.48	25.52	100.00	
Total	73.80	26.20	100.00	

Under the MCAR assumption, approximately 50% of the data were arbitrary deleted on contraceptive method use status such that no variable in the substantive model or data set of interest was related to missingness. In Table 5.47, the results indicate that approximately 50% of the values were deleted completely at random on contraceptive method use status.

TABLE 5.47: Model 2.2: Frequency distribution of missingness when 50% of the data are MCAR on contraceptive method use status

Missingness	Frequency	Percent	Cumulative frequency
Not missing	14833	50.20	50.20
Missing	14715	49.80	100.00
Total	29548	100.00	

To investigate whether values were deleted completely at random, the bivariate analysis was conducted to assess whether missingness was associated with variables in the data set. In Table 5.48, the results show that by deleting 50% values

completely at random on contraceptive methods use status, no significant difference was found in the proportions of missing values across the marital status, region and wealth index categories. The Chi-square test for association confirmed also that there was no association between missingness and these three categorical variables as the p-values associated with this test were bigger than the significance level of 5% that was used.

TABLE 5.48: Model 2.2: Distribution of missingness by selected categorical variables when 50% data are MCAR on contraceptive methods use status.

	Missingness			P-values
	Not Missing	Missing	Total	
Marital status				0.711
Never married	49.70	50.30	100.00	
Married	50.20	49.80	100.00	
Living together	50.51	49.49	100.00	
Widowed	51.96	48.04	100.00	
Divorced	48.39	51.61	100.00	
Not living together	49.30	50.70	100.00	
Total	50.20	49.80	100.00	
Wealth index				0.560
Poorest	50.41	49.59	100.00	
Poorer	49.34	50.66	100.00	
Middle	48.94	51.06	100.00	
Richer	51.32	48.68	100.00	
Richest	51.09	48.91	100.00	
Total	50.20	49.80	100.00	
Region				0.0.860
Kinshasa	50.23	49.77	100.00	
Bas Congo	50.83	49.17	100.00	
Bandundu	50.04	49.96	100.00	
Equateur	49.88	50.12	100.00	
Oriental	51.24	48.76	100.00	
Nord Kivu	50.57	49.43	100.00	
Maniema	49.67	50.33	100.00	
Sud Kivu	49.65	50.35	100.00	
Katanga	51.40	48.60	100.00	
Kasai Oriental	49.72	50.28	100.00	
Kasai Occidental	49.12	50.88	100.00	
Total	50.20	49.80	100.00	

The t-test was also conducted to determine the association between missingness and continuous variables in the data set; age in single years and education completed in years. The results indicated that there was no significant difference in the means of age and education of the two groups (present or missing), as the p-values

associated with the independent samples t-test are more than the significance level of 5% that was used. The results are found in Appendix B in Figure 6.23 for age and Figure 6.24 for education.

5.3.2.2 Computation of the performance measures

As previously stated, the multinomial logistic regression was used to determine the association between contraceptive methods use status and marital status. The second and third levels (traditional and modern contraceptive methods use respectively) of contraceptive method use status served as replicates of the dependent variable, representing two models that were estimated: the second level relative to the first (use no method) and the third level relative to the first. In other words, the multinomial logistic regression parameters were estimated for these two levels, relative to the first level. Marital status was treated as an independent variable and its parameters were estimated for each of its levels, leaving married as a reference category (as it is a category with the highest frequency to avoid the bias in the regression coefficients as suggested by [Wißmann et al. \(2007\)](#)). The models were estimated using the data set with no missing values (baseline data set or BD), the data set with missing values (case deletion or CD method) and the completed or imputed data sets with MVNI and MICE. The results about the estimated parameters are presented in tables and interpreted using figures. In Tables 5.49 and 5.50, the estimates of bias and standard errors when approximately 50% data are missing at random on contraceptive methods use status if a woman is aged at least 35 years, are given. The same parameters are reported in Tables 5.51 and 5.52 when data are missing completely at random on the same variable.

In Figures 5.15 and 5.16, the plots of bias in the regression coefficients and standard errors of the respondents who use traditional relative to those who do not use any contraceptive method are plotted, for both unweighted and weighted data sets, when data are MAR or MCAR. The results indicate that for both unweighted and weighted data sets, multiple imputation-based methods (MVNI and MICE) are less biased and produce more accurate standard deviations than the case deletion method, which discards items with missing values from the analysis. The figures indicate also that MICE produces better parameter estimates (less bias in coefficients and standard errors) than MVNI when data are missing on contraceptive method use status, treated as an outcome variable in the multinomial logistic regression models. In Figures 5.17 and 5.18, the bias and standard errors

for women who use modern methods relative to those who do not use any contraceptive method are plotted. As indicated, imputations with MICE and MVNI lead to unbiased results compared to case deletion that discards missing values from the analysis. The figures show also that MICE produced better estimates of bias and standard deviations than MVNI, for both unweighted and weighted data sets, when data are MAR or MCAR.

TABLE 5.49: Model 2.2: Estimates of bias in the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MAR on contraceptive method use status if a woman is aged at least 35 years.

Traditional Method					
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.050	0.032	0.277	0.153	-0.046
CD-weighted	0.059	0.090	0.482	0.172	0.091
MVNI-unweighted	0.046	0.024	0.274	0.142	-0.034
MVNI-weighted	0.066	0.058	0.470	0.067	0.083
MICE-unweighted	0.019	0.006	0.226	0.112	-0.025
MICE-weighted	0.046	0.030	0.432	0.056	0.052
Modern method					
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	-0.069	-0.045	0.130	0.021	0.027
CD-weighted	-0.083	-0.082	0.089	0.083	0.054
MVNI-unweighted	-0.050	-0.025	0.109	0.021	0.011
MVNI-weighted	-0.046	-0.065	0.067	0.053	0.025
MICE-unweighted	-0.023	-0.015	0.073	0.018	0.007
MICE-weighted	-0.038	-0.040	0.046	0.027	0.001

TABLE 5.50: Model 2.2: Estimates of standard errors of the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MAR on contraceptive method use status if a woman is aged at least 35 years.

Traditional Method					
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.151	0.051	0.232	0.133	0.078
BD-weighted	0.213	0.072	0.261	0.227	0.110
CD-unweighted	0.155	0.058	0.273	0.169	0.092
CD-weighted	0.219	0.082	0.316	0.255	0.132
MVNI-unweighted	0.154	0.057	0.272	0.158	0.089
MVNI-weighted	0.217	0.080	0.308	0.237	0.129
MICE-unweighted	0.154	0.054	0.262	0.148	0.085
MICE-weighted	0.215	0.076	0.281	0.229	0.128
Modern method					
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.161	0.069	0.164	0.164	0.112
BD-weighted	0.217	0.095	0.187	0.206	0.155
CD-unweighted	0.176	0.080	0.207	0.191	0.125
CD-weighted	0.229	0.109	0.246	0.240	0.169
MVNI-unweighted	0.170	0.079	0.198	0.186	0.120
MVNI-weighted	0.224	0.104	0.235	0.234	0.165
MICE-unweighted	0.169	0.076	0.186	0.178	0.120
MICE-weighted	0.220	0.099	0.221	0.221	0.161

TABLE 5.51: Model 2.2: Estimates of bias in the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MCAR on contraceptive method use status.

Traditional Method					
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	-0.068	0.012	0.057	-0.098	0.042
CD-weighted	-0.114	0.081	0.117	0.046	0.089
MVNI-unweighted	-0.061	0.007	0.040	-0.092	0.037
MVNI-weighted	-0.086	0.071	0.104	0.040	0.069
MICE-unweighted	-0.044	0.005	0.037	-0.089	0.028
MICE-weighted	-0.056	0.062	0.085	0.029	0.052
Modern method					
	Never married	Living together	Widowed	Divorced	Not living together
CD-unweighted	0.020	-0.042	0.064	0.057	-0.055
CD-weighted	0.069	-0.055	-0.070	0.053	-0.046
MVNI-unweighted	0.010	-0.038	0.053	0.047	-0.005
MVNI-weighted	0.049	-0.051	-0.060	0.044	-0.031
MICE-unweighted	0.001	-0.030	0.045	0.034	-0.004
MICE-weighted	0.033	-0.030	-0.050	0.022	-0.026

TABLE 5.52: Model 2.2: Estimates of standard errors of the regression coefficients of traditional and modern contraceptive methods when approximately 50% of data are MCAR on contraceptive method use status.

Traditional Method					
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.151	0.051	0.232	0.133	0.078
BD-weighted	0.213	0.072	0.261	0.227	0.110
CD-unweighted	0.214	0.072	0.320	0.197	0.109
CD-weighted	0.249	0.102	0.348	0.302	0.148
MVNI-unweighted	0.210	0.070	0.320	0.180	0.109
MVNI-weighted	0.240	0.099	0.328	0.281	0.140
MICE-unweighted	0.209	0.069	0.314	0.169	0.104
MICE-weighted	0.231	0.091	0.313	0.261	0.130
Modern method					
	Never married	Living together	Widowed	Divorced	Not living together
BD-unweighted	0.161	0.069	0.164	0.164	0.112
BD-weighted	0.217	0.095	0.187	0.206	0.155
CD-unweighted	0.223	0.098	0.231	0.217	0.149
CD-weighted	0.290	0.138	0.262	0.274	0.201
MVNI-unweighted	0.211	0.089	0.226	0.208	0.150
MVNI-weighted	0.281	0.130	0.248	0.261	0.191
MICE-unweighted	0.206	0.081	0.203	0.201	0.132
MICE-weighted	0.273	0.130	0.237	0.253	0.190

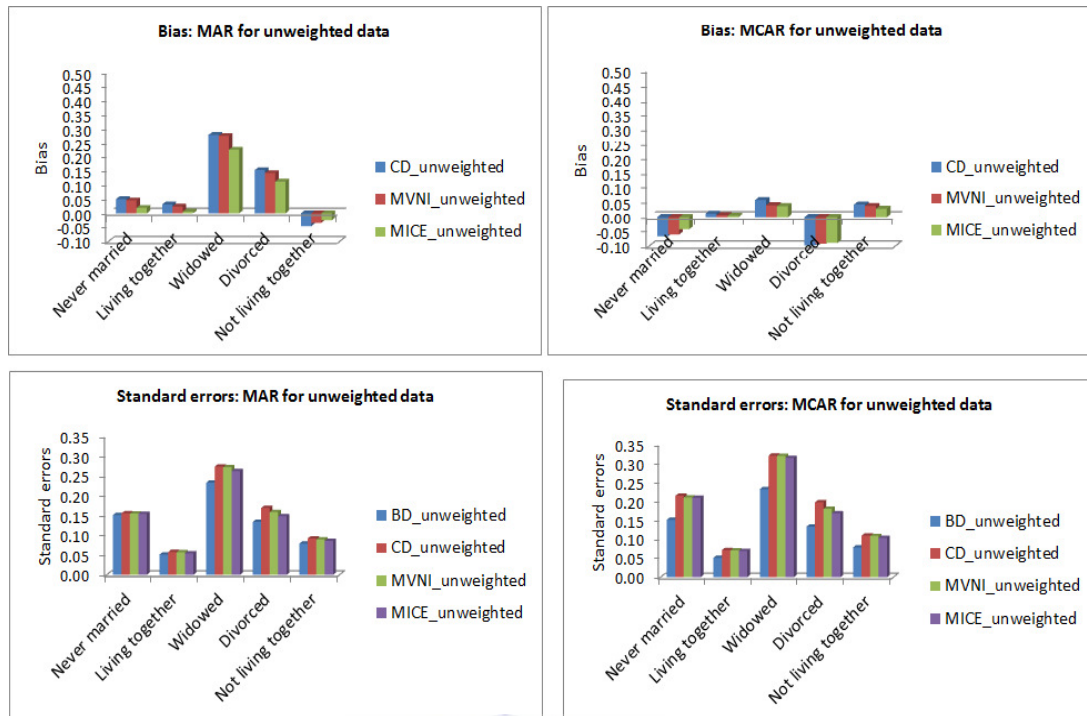


FIGURE 5.15: Model 2.2: Plot of bias and standard errors of the traditional method use category when 50% data are MAR and MCAR for unweighted data sets.

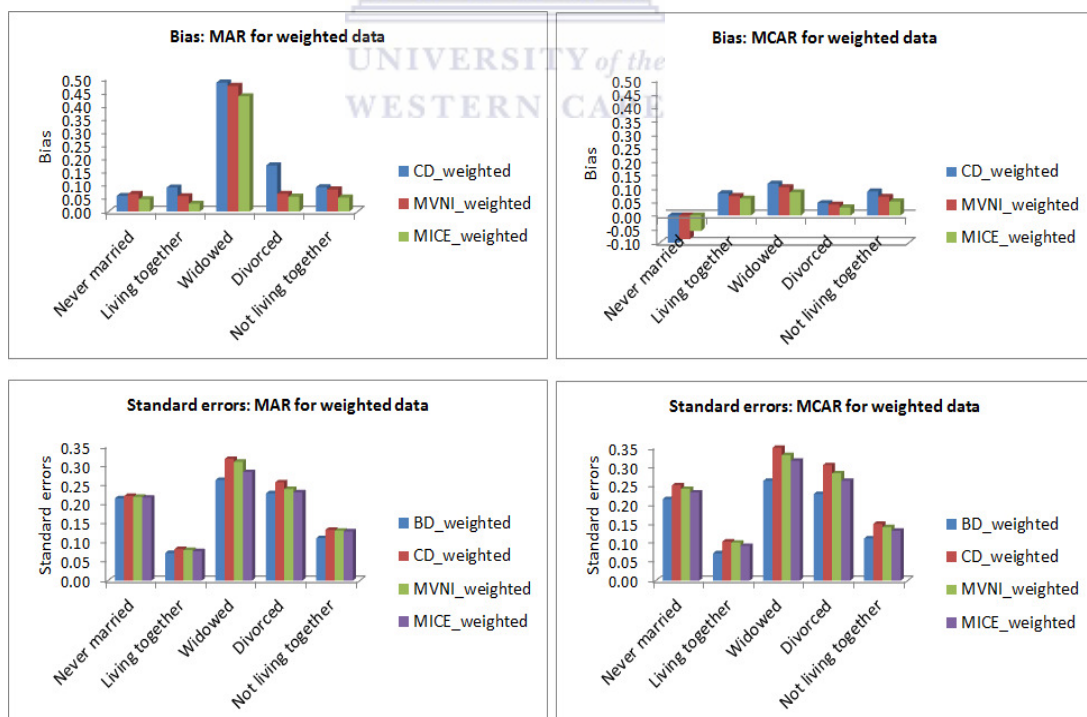


FIGURE 5.16: Model 2.2: Plot of bias and standard errors of the traditional method use category when 50% data are MAR and MCAR for weighted data sets.

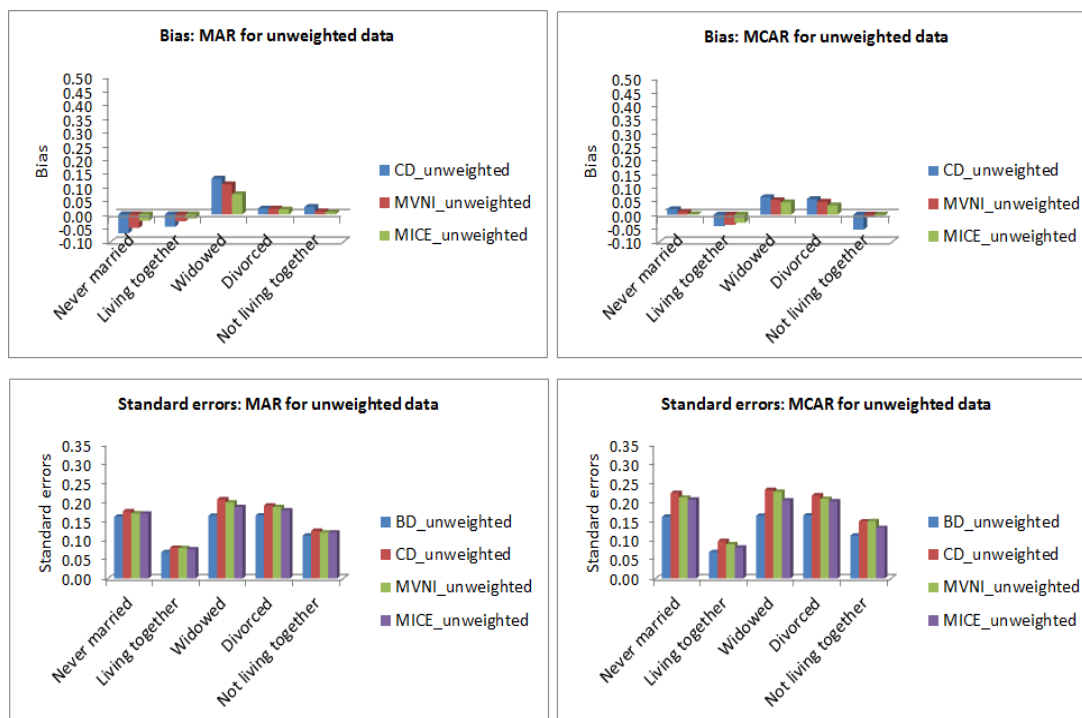


FIGURE 5.17: Model 2.2: Plot of bias and standard errors of the modern method use category when 50% data are MAR and MCAR for unweighted data sets.

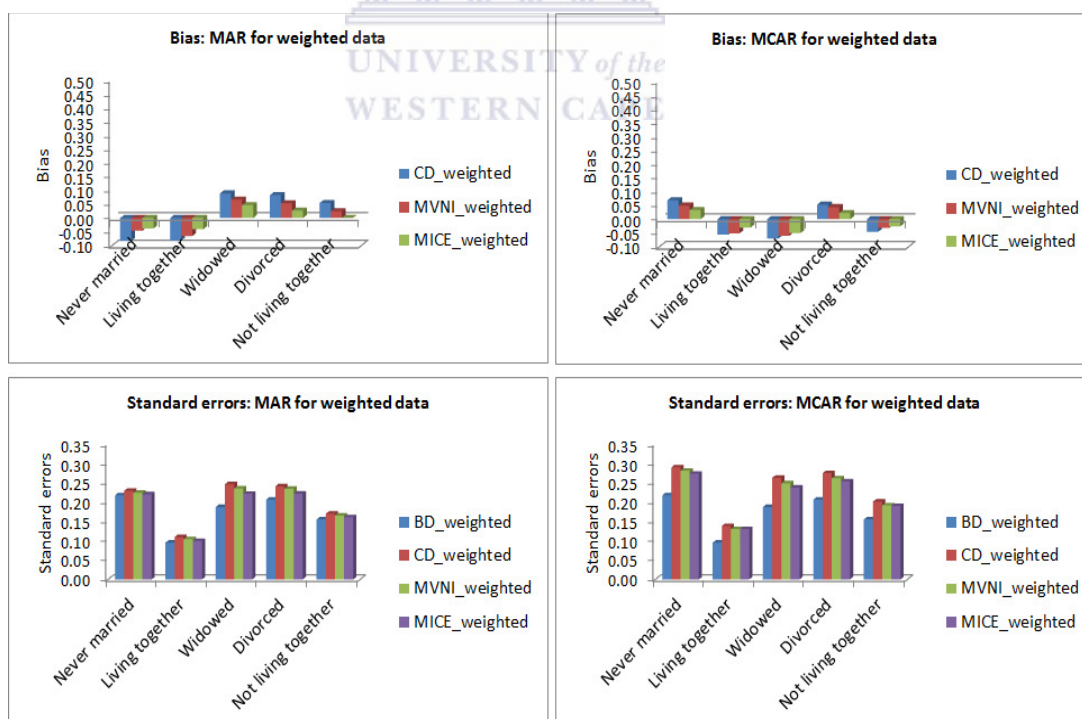


FIGURE 5.18: Model 2.2: Plot of bias and standard errors of the modern method use category when 50% data are MAR and MCAR for weighted data sets.

5.3.2.3 Model diagnostics

The Monte Carlo errors of the statistics involved in the estimation of the multinomial logistic regression models are reported in Appendix C in Tables 6.49, 6.50, 6.51, 6.52, 6.53, 6.54, 6.55 and 6.56. The results show that the Monte Carlo errors on the coefficients are less than 10% of the standard errors for unweighted and weighted data for both MVNI and MICE. The Monte Carlo errors of the p-values are also less than 0.1 when 5% level of significance was used for both MVNI and MICE. The Monte Carlo errors of the t-test statistics are approximately less than 0.1 for all the methods and data sets. Therefore, based on these results, it can be concluded that the imputation values were drawn from the desired distributions under MAR and MCAR assumptions.

The convergence of imputations to the desired distributions was assessed for both unweighted and weighted data sets, under MAR and MCAR assumptions. The estimates of the WLF were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE methods. Under the MAR assumption, the results are shown in Appendix D for unweighted (see Tables 6.121, 6.122, 6.123 and 6.124) and weighted (see Tables 6.125, 6.126, 6.127 and 6.128) data sets. Under the MCAR assumption, similar results were obtained and presented in Figures 6.129, 6.130, 6.131, 6.132, 6.133, 6.134, 6.135 and 6.136. As observed, the plots of the estimates of WLF against the iteration numbers show no visible trend, which is an indication that convergence was attained when 1000 iterations were used. On the other hand, the plots of WLF's estimates against the lag numbers showed that the autocorrelations that died off quickly, which is an indication that even a smaller number of iterations than what was used could be utilised to obtain independent samples.

5.3.3 Scenario 2: Summary of findings

Scenario 2 contains the results of the regression models with missing values on the responses variables. Two types of results are presented. These include the results of Model 2.1, which is a binary logistic regression model with missing values on the unordered outcome variable (contraceptive method use status) that was dichotomous (coded 1 if a woman uses any contraceptive method and 0 otherwise). This variable was imputed using MVNI and MICE, and the results showed that MICE produced more accurate estimates of bias and standard errors than MVNI.

In Model 2.2, the response variable was coded as a categorical variable that has no natural order (1 if a woman has not used any contraceptive method, 2 if she has used a traditional or folkloric method and 3 if she has used a modern contraceptive method). MAR and MCAR data were simulated on this variable and then imputation with MVNI and MICE was performed. The results were presented in terms of bias and standard errors as well. It was found that MICE performed better than MVNI as well, thus suggesting that when missing values are present on the unordered categorical variables that have to be used as outcome variables in the regression models, MICE, which takes into account the distributional form of the variables to be imputed, would be a better imputation technique than MICE.

5.4 Summary of the chapter

This chapter presents the results on the performance (in terms of bias in regression coefficients and standard errors) of the multiple imputation methods of interest, namely MVNI and MICE. The 2007 Democratic Republic of Congo Demographic Health Survey was used for analysis. Using this data set, data sets with missing at random (MAR) or missing completely at random (MCAR) were simulated for imputation purposes.

Two scenarios were considered for this study. The first scenario contained logistic regression models in which missing values were observed on nominal variables with more than three categories or levels that were treated as predictors in the regression models. In this scenario, the performance of the multiple imputation methods of interest was explored on a single and multiple predictors as well as on different rates of missingness. The second scenario contained two regression models in which missing values were observed on the response variables; binary and polytomous. The first model or Model 2.1 consisted of a binary logistic regression model of contraceptive method use status (coded 1 if a woman has used a contraceptive method and 0 otherwise). The second model or Model 2.2 consisted of a multinomial logistic regression model of contraceptive method use status as well, but in this case coded 1 if a woman has not used any contraceptive method, 2 if she has used a traditional folkloric method and 3 if she has used a modern contraceptive method.

For each model, variables with missing values were imputed using MVNI or

MICE. Under MVNI, missing values on the binary variables were imputed as continuous as suggested by Allison (2001). Categorical variables with more than three levels were dichotomised first and then imputed as binary variables. Imputation with MICE was done taking into account the distributional form of the variables with missing data. That is, a binary and multinomial regression models were used to draw imputation values of the binary and polytomous variables respectively.

Baseline models (models from data sets with no missing values) were estimated first. Then models with missing values (using the case deletion or CD method) and models with completed (observed + imputed) data sets were fitted. To ensure that the imputed values came from the desired distributions, the models diagnostics were done for each model and the results indicated that convergence was attained. The estimates of bias in the regression coefficients were computed and reported along with the standard errors for each model and under MCAR and MAR assumptions. The analysis was done assuming that the sample was not weighted first, then the sample weight was taken into account to assess whether the sample design would affect the performance of the multiple imputation methods of interest, namely MVNI and MICE.

As expected, the results showed that for all the models, MVNI and MICE produced less biased smaller standard errors than the case deletion method, which discards items with missing values from the analysis. Furthermore, it was found that when data were missing (MCAR or MAR) on the nominal variable treated as a predictor in the regression model (Scenario 1 Model 1), MVNI produced less bias in the regression coefficients and standard errors, for both unweighted and weighted data sets. However, the results indicated that when data were missing on the response variables, either the binary or polytomous, MICE produced less biases estimates than MVNI, which suggests that the imputation of outcome variables in the regression models should take into account the distributional form of the variables with missing values. Furthermore, it was noted that the sample design (sample weights), the rates of missingness and the missing data mechanisms (MCAR or MAR) did not affect the performance of the multiple imputation methods (MVNI and MICE) that were considered in this study.

Note that the results from Model 1.2.1 (model with missing values on two predictors measured on nominal scales) were summarized into an article that was accepted for publication in the Brazilian Journal of Probability and Statistics under the heading: "Multiple imputation of unordered categorical missing data: A comparison of the multivariate normal imputation and multiple imputation by

chained equations” ([Karangwa et al., 2015](#)).



Chapter 6

Discussion and conclusion

Missing data commonly arise in many fields of empirical research. They lead to a loss of information and more importantly may cause serious bias into the estimates and lead to incorrect inferences, especially when there is a lot of missing data in the data sets subject to analysis. Therefore, there is a need to adequately handle them in order to obtain reliable results.

In recent years, the multiple imputation technique has gained popularity as the best technique to handle missing values (Carrig et al., 2015; Mukhopadhyay, 2015). With this technique, missing values are replaced with random draws from a predictive distribution based on the available data. The imputation is done multiple times, which results in multiple data sets that are analysed individually and the resulting estimates are combined using Rubin (1978) rule to produce a final estimate that is used for imputation. MVNI and MICE have emerged as the best ways of combining these estimates, and as noted by Ware et al. (2012), researchers are increasingly being encouraged to use them. The former is a parametric-based multiple imputation technique in which variables in the imputation model are assumed to follow a normal distribution, whereas the latter is a flexible technique that takes into account the distributional form of the variables with missing data. The two techniques were designed under the Bayesian framework in which, given specific priors, imputation values are drawn from the conditional distribution of missing values given the observed data.

The primary objective of this study was to examine the behaviour of MVNI and MICE, when missing values are observed specifically on unordered categorical variables treated as either predictors or response variables in the regression models.

Other specific objectives were:

1. To review the literature on MVNI and MICE methods and illustrate their performance when data are missing on continuous variables.
2. To show that as expected, multiple imputation methods of interest produce less biased estimates than the case deletion technique, which discard missing values from the analysis.
3. To investigate whether the rates of missing values in the data sets can impact on the performance of the multiple imputation methods of interest, namely MVNI and MICE.
4. To determine whether the sample design can impact on the performance of MVNI and MICE.
5. To draw relevant conclusions on how specifically non-ordered or nominal categorical data containing MCAR or MAR data should be imputed under different circumstances, especially when missing values are present on the outcome or predictor variables in the regression models.

The purpose of doing this research was to provide some guidance on how and when MVNI and MICE should be used in similar cases. [Van Buuren \(2007\)](#), [Lee and Carlin \(2010\)](#) and [Kropko et al. \(2014\)](#) have previously compared these two methods and obtained mixed results: van Buuren and Kropko, Goodrich, Gelman and Hill found that MICE outperformed MVNI, whereas Lee and Carlin found that MVNI performed equally well as MICE.

The scope of these studies is limited. However, the common approach to all these authors is that they compared the ability of MICE and MVNI to return less or more biased estimates of the regression coefficients. [Lee and Carlin \(2010\)](#) for instance considered a regression model of a continuous variable on binary and ordinal variables containing missing values. The research topic of this study that was started in 2011 was formulated based on the recommendations by these two authors, who suggested that further investigations were still needed to look at the performance of MVNI and MICE when data are missing at random on nominal variables. Three years later, [Kropko et al. \(2014\)](#) explored this topic. They considered four models: 1) a regression model with a continuous response variable and a set of covariates in which missing values were only found on the outcome variable (continuous). 2) A regression model in which the outcome variable was binary and contained missing values, and the covariates were as in model 1. 3) A

regression model with an ordered outcome categorical variable, and covariates as in models 1 and 2. 4) A regression model in which the response variable was an unordered categorical variable and predictors were continuous, binary or nominal.

This study considered regression models as well, with missing values on either the response or independent variables measured on the nominal scale. This approach is totally different from that of [Lee and Carlin \(2010\)](#), who considered only regression models with missing values on binary and ordinal variable predictors. The study is partly similar to that of [Kropko et al. \(2014\)](#) in the case where missing values were observed on the outcome variables measured as binary or polytomous. However, the difference was on the fact that the performance of these methods was investigated when missing values were present on the nominal variables alone, with no influence of other types of variables first, then in the presence of other types of variables (ordinal and continuous) as in the case of [Kropko et al. \(2014\)](#).

A common similarity that was observed for all the approaches is that they all focused on the ability of MVNI and MICE to return accurate regression coefficients. This study reported the bias in the regression coefficients and standard errors from the regression models estimated after the CD, MVNI and MICE approaches were used.

In relation to the main objective, two scenarios were considered for analysis. The first scenario contained regression models in which missing values were observed on the covariates that were measured on nominal scales. The second scenario contained regression models as well, but in this case with missing values on the outcome categorical variables that have no natural order. In Scenario 1, two types of regression models were estimated. The first model regressed contraceptive method use status on marital status alone, with missing values on the independent variable (marital status). The second model regressed contraceptive method use status on marital status, controlling for other variables (nominal, continuous and ordinal). This model was split into two models; Models 1.2.1 and 1.2.2. In Model 1.2.1, two unordered categorical covariates (marital status and region) were considered for analysis, whereas in Model 1.2.2 independent variables of different types (continuous, nominal and ordinal) with missing values on only the nominal ones were considered for analysis. These variables were the woman's age, education in completed years (continuous variables), marital status, region (nominal variables) and wealth index (ordinal variable). The second scenario contained two models with missing values on the response variables. The first model

or Model 2.1 regressed contraceptive method use status (a dichotomous variable) on marital status, whereas the second model or Model 2.2 regressed contraceptive method use status (measured as a polytomous variable). Throughout this thesis, the rate of missingness that was considered is 50%. However, to verify whether the rates of missingness could have an effect on the performance of the multiple imputation methods of interest, 50%, 30% and 10% missing data were considered for only Model 1.1.

Missing values on the variables of these models were imputed using the parametric imputation technique (MVNI) and the MICE method which took into account the distributional form of the variables to be imputed. Two main findings were highlighted for this specific objective:

1. MVNI outperformed MICE when data were missing on the unordered categorical variables treated as predictors in the regression models.
2. MICE outperformed MVNI when missing values were observed on the nominal outcome variables (binary or polytomous), treated as response variables in the regression models.

The first finding was neither highlighted by [Lee and Carlin \(2010\)](#) nor by [Kropko et al. \(2014\)](#) and therefore constitutes one of the contributions of this thesis to the missing data research field. This finding was summarised into a research paper that was accepted for publication in *Brazilian Journal of Probability and Statistics* ([Karangwa et al., 2015](#)). The second finding is consistent with the findings by [Kropko et al. \(2014\)](#), who showed that MICE outperformed MVNI when these two techniques were used to impute missing values of the unordered categorical variables with more than two levels that were treated as outcome measures in the regression models. Note that in addition of what [Kropko et al. \(2014\)](#) has done, this study explored this fact in several circumstances such as under different missing mechanisms (MAR and MCAR) and considering or not the sample design or weights. However, this fact has not affected the behaviour of MICE and MVNI. When data were missing on the binary categorical variables that were treated as response variables in the regression model, the results differed. Our findings indicated that MICE outperformed MVNI, whereas it was the opposite in the [Kropko et al. \(2014\)](#) study. This disparity may be due to the fact that under the MVNI technique, the response variable was imputed as a continuous variable and imputed values were rounded to the nearest integer (0 or 1) to keep its dichotomous

nature as it had to be used as binary variable in the logistic regression model. This fact was not explained by [Kropko et al. \(2014\)](#) and if the same approach was not used, the source of difference may lie there. The difference may also be due to the data sets and the number of imputations that were used.

In relation to other specific objectives, objective 1 was addressed in Chapter 3 where the literature review on the MVNI and MICE methods was provided. A practical example using a real data set was used to illustrate the behaviour of these methods when data were missing completely at random on continuous variables that were treated as predictors in the regression model. Similar results as in the literature were found ([Raghunathan et al., 2001](#)). In fact, it was found that the MVNI and MICE produced similar results when missing values were observed on the continuous variables that were used. In addition to this finding and the existing knowledge about the behaviour of MICE and MVNI on continuous data, this study was able to explore the impact of different rates of missing data on the behaviour of these methods. The results revealed that at some stage, neither the imputation methods used nor the CD can help to maintain the relationship that exists between the dependent and independent variables when the analysis is done using the data set with no missing values. This is an indication that at some stage, the missing values techniques may not be successful and therefore the data users may be forced to give up on the data set that was intended to be used or the analysis that needed to be done. This finding was not highlighted before and therefore is a contribution to the existing knowledge about the imputation of continuous variables. These results were summarised into a single article that was published in [Karangwa and Kotze \(2013\)](#).

Concerning the specific objective 2, models with missing values on the variables of interest (using the CD method) were estimated and the results were compared to the results from the models fitted after MVNI and MICE were used. As expected, the multiple imputation methods of interest yielded more accurate estimates than the CD technique. This is generally the case, but we believed that an empirical research would also be needed to indicate that imputations were needed.

In relation to the specific objective 3, the impact of the rates of missing values in the data sets on the behaviour of MVNI and MICE was explored. This was done using Model 1.1 in the first scenario where 50%, 30% and 10% rates of missingness were considered for analysis. The results indicated that whatever rate of missingness used, the behaviour of the two methods did not change. That is, MVNI outperformed MICE at lower and higher rates of missingness. This fact was

not also explored by [Lee and Carlin \(2010\)](#) and [Kropko et al. \(2014\)](#) and therefore could add value to the existing knowledge about the missing values' treatment for unordered categorical data.

Regarding the specific objective 4, the analysis took into account the sample design (sample weights) first and then assumed that there was no sample weights in the data set that was used. In reality, the data set that was used is a complex survey data set that contains the sample weights. Therefore, only the results obtained when the sample design was taken into account are valid. However, as all the data sets do not contain sample weights, the study assumed that the data set had no sample weights (which is not true) and explored the behaviour of MVNI and MICE. This was done to assess whether the sample design may affect the behaviour of these methods. The results indicated that this fact had no impact on the behaviour of the multiple imputation methods of interest. That is, the same conclusion was obtained when the survey design was taken into account and not. This fact was not highlighted by missing data analysts, especially [Lee and Carlin \(2010\)](#) and [Kropko et al. \(2014\)](#), and therefore could also add value to the existing knowledge about missing data handling.

Finally, as always suggested by researchers on this particular topic such as [Lee and Carlin \(2010\)](#) and [Kropko et al. \(2014\)](#) amongst others, it is not easy to draw general conclusions from a single data or simulation study, however we believe that this research has given a good setting for comparing the multiple imputation techniques of interest, namely MICE and MVNI. Further investigations using different data sets are still needed to strengthen the findings from this study, but it is beyond the scope of this thesis.

Bibliography

- Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, 67(4), 1012–1028.
- Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in medicine*, 20(17-18), 2709–2722.
- Agresti, A. (2002). *Categorical data analysis* (Vol. 359). New York, NY: John Wiley & Sons.
- Allison, P. D. (2001). *Missing data* (Vol. 136). Thousand Oaks, CA: Sage.
- Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, 55(1), 193–196.
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of abnormal psychology*, 112(4), 545.
- Allison, P. D. (2005). Imputation of categorical variables with proc mi. In *Sas users group international, 30th meeting (sugi 30)*.
- Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1), 40–49.
- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5–37.
- Batista, G., & Monard, M. C. (2001). A study of k-nearest neighbour as a model-based method to treat missing data. In *Proceedings of the argentine symposium on artificial intelligence* (pp. 1–9).
- Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in medicine*, 26(6), 1368–1382.
- Bethlehem, J. (2009). *Applied survey methods: A statistical perspective* (Vol. 558). New York, NY: John Wiley & Sons.

- Brand, J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Erasmus MC: University Medical Center Rotterdam.
- Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. New York, NY: John Wiley & Sons.
- Carrig, M. M., Manrique-Vallier, D., Ranby, K. W., Reiter, J. P., & Hoyle, R. H. (2015). A nonparametric, multiple imputation-based method for the retrospective integration of data sets. *Multivariate Behavioral Research*, *50*(4), 383–397.
- Catellier, D. J., Hannan, P. J., Murray, D. M., Addy, C. L., Conway, T. L., Yang, S., & Rice, J. C. (2005). Imputation of missing data when measuring physical activity by accelerometry. *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S555.
- Cattle, B. A., Baxter, P. D., Greenwood, D. C., Gale, C. P., & West, R. M. (2011). Multiple imputation for completion of a national clinical audit dataset. *Statistics in medicine*, *30*(22), 2736–2753.
- Chen, H. Y., & Little, R. J. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, *94*(447), 896–908.
- Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, *49*(4), 327–335.
- Cochran, W. (1977). *Sampling techniques* (Vol. 98). New York, NY: John Wiley & Sons.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.
- De Leeuw, E. D., Hox, J. J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *19*(2), 153–176.
- Deltour, I., Richardson, S., & Hesran, J.-Y. L. (1999). Stochastic algorithms for markov models estimation with intermittent missing data. *Biometrics*, *55*(2), 565–573.
- Demirtas, H. (2004). Simulation driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica*, *58*(4), 466–482.
- Demirtas, H. (2005). Multiple imputation under bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, *24*(15), 2345–2363.

- Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1), 69–84.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.
- Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426), 463–475.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science*, 8, 361–378.
- Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete data in sample surveys*, 2(Part IV), 185–207.
- Galati, J. C., & Carlin, J. B. (2009). Inorm: Stata module to perform multiple imputation using schaffer's method. *Statistical Software Components*.
- Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. (1990). Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412), 972–985.
- Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410), 398–409.
- Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., & Meulders, M. (2005). Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*, 61(1), 74–85.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(6), 721–741.
- Gentle, J. E. (2009). *Computational statistics* (Vol. 308). New York, NY: Springer.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60, 549–576.
- Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31(2), 197–218.

- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1), 97–109.
- He, Y. (2010). Missing data analysis using multiple imputation getting to the heart of the matter. *Circulation: Cardiovascular Quality and Outcomes*, 3(1), 98–105.
- He, Y., Zaslavsky, A. M., Landrum, M., Harrington, D., & Catalano, P. (2009). Multiple imputation in a large-scale complex survey: a practical guide. *Statistical methods in medical research*, 1–18.
- Hedderley, D. (1995). A comparison of imputation techniques for internal preference mapping, using monte carlo simulation. *Food quality and preference*, 6(4), 281–297.
- Helenowski, I. (2015). Advantages and advancements of multiple imputation. *Biometrics and Biostatistics International Journal*, 2(5), 00033.
- Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29(3), 122–126.
- Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3), 244–254.
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.
- Jackman, S. (2000). Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo. *American Journal of Political Science*, 44, 375–404.
- Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills: CA: SAGE.
- Karangwa, I., & Kotze, D. (2013). Using the markov chain monte carlo method to make inferences on items of data contaminated by missing values. *American Journal of Theoretical and Applied Statistics*, 2(3), 48–53. Retrieved from <http://article.sciencepublishinggroup.com/pdf/10.11648.j.ajtas.20130203.12.pdf>
- Karangwa, I., Kotze, D., & Blignaut, R. (2015). Multiple imputation of unordered categorical missing data: A comparison of the multivariate normal imputation and multiple imputation by chained equations. *Brazilian Journal of Probability and Statistics*.. Retrieved from http://imstat.org/bjps/future_papers.html
- Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance

- matrices for multivariate incomplete data. *Psychometrika*, 67(4), 609–623.
- Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.
- Korn, E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49(3), 291–295.
- Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis*, 22(4), 497–519.
- Lavine, M. (2005). *Introduction to statistical thought*. Retrieved from <http://www.public.iastate.edu/~pcaragea/S40608/MLavineBook.pdf>
- Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology*, 171(5), 624–632.
- Lee, K. J., Galati, J. C., Simpson, J. A., & Carlin, J. B. (2012). Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. *Statistics in medicine*, 31(30), 4164–4174.
- Lessler, J., & Kalsbeek, W. (1992). *Nonsampling error in surveys*. New York, NY: Wiley.
- Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: methods and applications*. New York, NY: John Wiley & Sons.
- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202.
- Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3), 292–326.
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons.
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York, NY: Guilford Press.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21, 1087.
- Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247), 335–341.
- Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G.

- (2014). *Handbook of missing data methodology*. New York, NY: CRC Press.
- Mukhopadhyay, P. (2015). *Multiple imputation of missing data using sas*. New York, NY: John Wiley & Sons.
- Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, *24*(7), 815–829.
- Norazian, M. N., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*, *34*(3), 341–345.
- Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, *27*(1), 85–96.
- Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, *47*(1), 13–26.
- Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, *32*(2), 143.
- Robert, C., & Casella, G. (2010). *Introducing monte carlo methods with r*. New York, NY: Springer.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.
- Rubin, D. B. (1978). Multiple imputations in sample surveys—a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the american statistical association* (Vol. 1, pp. 20–34).
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*(394), 366–374.
- Sande, G. T. (1983). Replacement for a ten-minute gap. *Madow, WG and Olkin, I., Incomplete Data in Sample Surveys*, *2*, 337–33.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, *33*(4), 545–571.

- Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, 11(2), 437–457.
- Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association*, 101(475), 924–933.
- Seefeld, K., & Linder, E. (2007). *Statistics using r with biological examples*. University of New Hampshire Press.
- Siddique, J., Brown, C. H., Hedeker, D., Duan, N., Gibbons, R. D., Miranda, J., & Lavori, P. W. (2008). Missing data in longitudinal trials-part b, analytic issues. *Psychiatric annals*, 38(12), 793.
- Siddique, J., Harel, O., & Crespi, C. M. (2012). Addressing missing data mechanism uncertainty using multiple-model multiple imputation: Application to a longitudinal clinical trial. *The annals of applied statistics*, 6(4), 1814.
- SPSS, I. (2013). *Spss statistical software*. Armonk, NY: IBM Corporation.
- Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: a case study of the children's mental health initiative. *American journal of epidemiology*, 169(9), 1133–1139.
- Tsikriktsis, N. (2005). A review of techniques for treating missing data in om survey research. *Journal of Operations Management*, 24(1), 53–62.
- Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of clinical epidemiology*, 66(9), 1022–1028.
- Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3), 219–242.
- Van Buuren, S., Boshuizen, H. C., Knook, D. L., et al. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6), 681–694.
- Ware, J. H., Harrington, D., Hunter, D. J., & D'Agostino, R. B. (2012). Missing data. *New England Journal of Medicine*, 367(14), 1353–1354.
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4), 377–399.
- Wißmann, M., Toutenburg, H., et al. (2007). *Role of categorical variables in*

multicollinearity in the linear regression model (technical report number 08).

Department of Statistics, University of Munich.

Yu, Z., & Schaid, D. J. (2007). Methods to impute missing genotypes for population data. *Human genetics*, 122(5), 495–504.



Appendix A

The following program codes were developed by the author and are available on request. The headings of the specific routines are supplied.



R codes

R code to generate samples from a beta (3, 7) using the Metropolis-Hastings technique.

Function that uses the Gibbs sampler to simulate a bivariate normal distribution by iteratively sampling from the conditional distributions of random variables X and Y.

R code to generate missing completely at random (MCAR) data on the variables of interest.

R code to generate missing at random (MAR) data on the variables of interest.

STATA codes



STATA codes to impute missing values on the variable marital status that was treated as an independent variable in the binary logistic regression model.

STATA codes to impute missing values on the variable contraceptive method use status that was treated as an outcome variable in the binary logistic regression model.

STATA codes to impute missing values on the variable contraceptive method use status that was treated as an outcome variable in the multinomial logistic regression model.

Appendix B




```

. ttest V133, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    18002     4.771914     .0295022     3.958351     4.714087     4.829741
          1 |    11546     4.038455     .0346591     3.724196     3.970517     4.106392
-----+-----
combined |    29548     4.485312     .0226009     3.884996     4.441013     4.529611
-----+-----

      diff |           .7334594           .0461249           .6430525           .8238662
-----+-----

      diff = mean(0) - mean(1)                                t = 15.9016
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 1.0000                                Pr(|T| > |t|) = 0.0000                                Pr(T > t) = 0.0000

```

FIGURE 6.2: Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

```

. ttest V012, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    14607    35.1337    .067131     8.113412    35.00212    35.26529
          1 |    14941    35.22535    .0668389    8.169941    35.09434    35.35637
-----+-----
combined |    29548    35.18005    .0473662    8.142037    35.08721    35.27289
-----+-----

      diff |           -.09165    .0947387           -.277342    .0940419
-----+-----

      diff = mean(0) - mean(1)                                t = -0.9674
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.1667                                Pr(|T| > |t|) = 0.3334                                Pr(T > t) = 0.8333

```

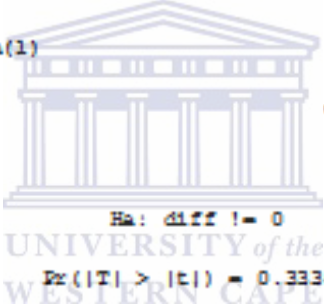


FIGURE 6.3: Model 1.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.

```

. ttest V133, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    14607    4.493873    .032466     3.923819     4.430235     4.55751
          1 |    14941    4.476943    .0314707    3.846775     4.415256     4.538629
-----+-----
combined |    29548    4.485312    .0226009    3.884996     4.441013     4.529611
-----+-----

      diff |           .0169302    .0452054           - .0716745    .1055348
-----+-----

      diff = mean(0) - mean(1)                                t =    0.3745
Ho: diff = 0                                                degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.6460                                Pr(|T| > |t|) = 0.7080                                Pr(T > t) = 0.3540

```




FIGURE 6.4: Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.

Model 1.1 with 30% missing values on the covariate

```
. ttest V012, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |   22617   35.12389   .0537164    8.078379    35.0186    35.22918
          1 |    6931   35.3633    .1002293    8.34435    35.16682    35.55978
-----+-----
combined |   29548   35.18005   .0473662    8.142037    35.08721    35.27289
-----+-----
      diff |          -.2394062   .1117777          -.4584955   -.0203169
-----+-----
      diff = mean(0) - mean(1)                                t = -2.1418
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.0161                                Pr(|T| > |t|) = 0.0322                                Pr(T > t) = 0.9839
```

FIGURE 6.5: Model 1.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

```

. ttest V133, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    22617    4.617986    .0261014    3.925371    4.566826    4.669147
          1 |     6931    4.052373    .0446557    3.717708    3.964834    4.139912
-----+-----
combined |    29548    4.485312    .0226009    3.884996    4.441013    4.529611
-----+-----
      diff |           .5656131    .0532376                .461265    .6699611
-----+-----

      diff = mean(0) - mean(1)                t = 10.6243
Ho: diff = 0                degrees of freedom = 29546

      Ha: diff < 0                Ha: diff != 0                Ha: diff > 0
Pr(T < t) = 1.0000                Pr(|T| > |t|) = 0.0000                Pr(T > t) = 0.0000

```

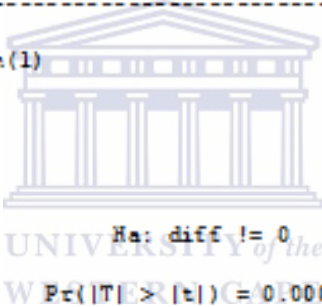


FIGURE 6.6: Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

```

. ttest V012, by(miss1)

Two-sample t test with equal variances
-----+-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |   20611   35.21949   .0564538    8.104811    35.10884    35.33015
          1 |    8937   35.08907   .0870251    8.22698    34.91848    35.25966
-----+-----
combined |   29548   35.18005   .0473662    8.142037    35.08721    35.27289
-----+-----
      diff |           .1304265   .103121           -.0716953   .3325483
-----+-----

      diff = mean(0) - mean(1)                                t =      1.2648
Ho: diff = 0                                                  degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.8970                                Pr(|T| > |t|) = 0.2060                                Pr(T > t) = 0.1030

```

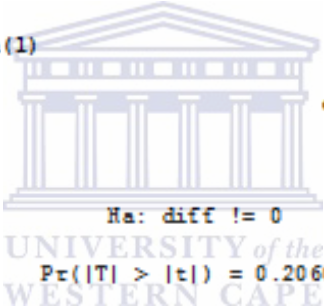


FIGURE 6.7: Model 1.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.


```

. ttest V133, by(miss1)

Two-sample t test with equal variances
-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |   20611   4.49983   .0271748   3.90136   4.446565   4.553095
          1 |    8937   4.451829  .0406936   3.847   4.372061   4.531598
-----+-----
combined |   29548   4.485312  .0226009   3.884996   4.441013   4.529611
-----+-----
      diff |           .0480007   .049205           -.0484433   .1444448
-----+-----

      diff = mean(0) - mean(1)                                t =    0.9755
Ho: diff = 0                                                degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.8353                                Pr(|T| > |t|) = 0.3293                                Pr(T > t) = 0.1647

```

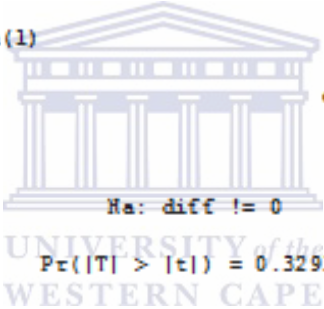


FIGURE 6.8: Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.

Model 1.1 with 10% missing values on the covariate

```
. ttest V012, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |   27227   35.16517   .0492048    8.119082    35.06872    35.26161
          1 |    2321   35.35459   .1744921    8.406462    35.01241    35.69677
-----+-----
combined |   29548   35.18005   .0473662    8.142037    35.08721    35.27289
-----+-----
      diff |           -.1894216   .1760591           -.5345052   .155662
-----+-----
      diff = mean(0) - mean(1)                                t =  -1.0759
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.1410                                Pr(|T| > |t|) = 0.2820                                Pr(T > t) = 0.8590
```

FIGURE 6.9: Model 1.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption under MAR assumption.

```

. ttest V133, by(miss1)

Two-sample t test with equal variances
-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |   27227   4.524406   .0236282   3.898799   4.478093   4.570718
          1 |    2321   4.026713   .0765772   3.68924   3.876546   4.176879
-----+-----
combined |   29548   4.485312   .0226009   3.884996   4.441013   4.529611
-----+-----
      diff |           .4976933   .0839588           .3331303   .6622563
-----+-----

      diff = mean(0) - mean(1)                                t =    5.9278
Ho: diff = 0                                                degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 1.0000                                Pr(|T| > |t|) = 0.0000                                Pr(T > t) = 0.0000

```

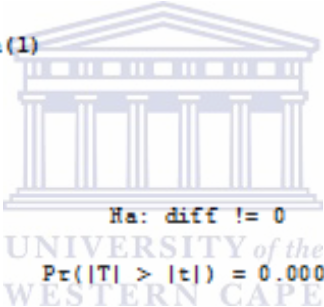


FIGURE 6.10: Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

```

. ttest V012, by(miss1)

Two-sample t test with equal variances
-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |   26581   35.17727   .0498677    8.130276    35.07953    35.27501
          1 |    2967   35.20492   .1514223    8.247998    34.90802    35.50182
-----+-----
combined |   29548   35.18005   .0473662    8.142037    35.08721    35.27289
-----+-----
      diff |           -.0276513   .1576013                -.3365568   .2812542
-----+-----

      diff = mean(0) - mean(1)                                t = -0.1755
Ho: diff = 0                                                  degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.4304                                Pr(|T| > |t|) = 0.8607                                Pr(T > t) = 0.5696

```

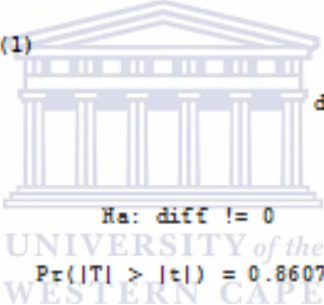


FIGURE 6.11: Model 1.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.

```

. ttest V133, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    26581    4.495692    .0238793    3.893207    4.448888    4.542497
          1 |     2967    4.392315    .069947    3.810026    4.255166    4.529465
-----+-----
combined |    29548    4.485312    .0226009    3.884996    4.441013    4.529611
-----+-----
      diff |          .1033769    .0751975          - .0440136    .2507675
-----+-----

      diff = mean(0) - mean(1)                                t =      1.3747
Ho: diff = 0                                                degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.9154                                Pr(|T| > |t|) = 0.1692                                Pr(T > t) = 0.0846

```

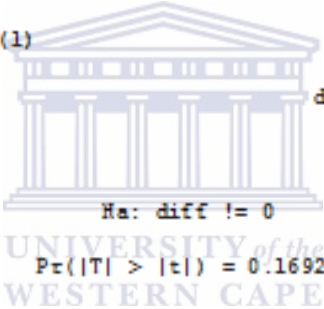


FIGURE 6.12: Model 1.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.


```

. ttest V133, by(miss1)

Two-sample t test with equal variances
-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |    18002    4.771914   .0295022    3.958351    4.714087    4.829741
          1 |    11546    4.038455   .0346591    3.724196    3.970517    4.106392
-----+-----
combined |    29548    4.485312   .0226009    3.884996    4.441013    4.529611
-----+-----
      diff |           .7334594   .0461249                .6430525   .8238662
-----+-----

      diff = mean(0) - mean(1)                                t = 15.9016
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 1.0000                                Pr(|T| > |t|) = 0.0000                                Pr(T > t) = 0.0000

```

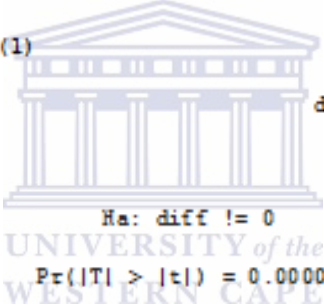


FIGURE 6.14: Model 1.2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.


```

. ttest V012, by(miss1)

Two-sample t test with equal variances
-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
      0 |   14607   35.1337   .067131   8.113412   35.00212   35.26529
      1 |   14941   35.22535   .0668389   8.169941   35.09434   35.35637
-----+-----
combined |   29548   35.18005   .0473662   8.142037   35.08721   35.27289
-----+-----
      diff |           -.09165   .0947387           -.277342   .0940419
-----+-----

      diff = mean(0) - mean(1)                                t = -0.9674
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.1667                                Pr(|T| > |t|) = 0.3334                                Pr(T > t) = 0.8333

```

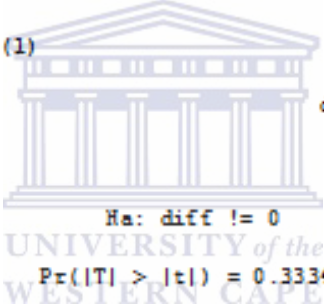


FIGURE 6.15: Model 1.2.1: Independent-samples t-test to compare age (V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.

```

. ttest V133, by(miss1)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |   14607   4.493873   .032466    3.923819   4.430235   4.55751
          1 |   14941   4.476943   .0314707   3.846775   4.415256   4.538629
-----+-----
combined |   29548   4.485312   .0226009   3.884996   4.441013   4.529611
-----+-----

      diff |           .0169302   .0452054           - .0716745   .1055348
-----+-----

      diff = mean(0) - mean(1)                                t =    0.3745
Ho: diff = 0                                                    degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.6460                                Pr(|T| > |t|) = 0.7080                                Pr(T > t) = 0.3540

```

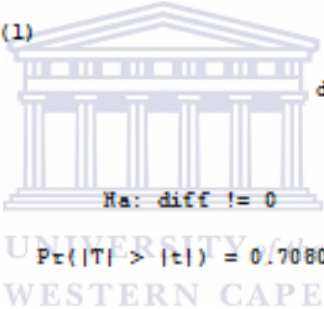


FIGURE 6.16: Model 1.2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MCAR assumption.


```

. ttest V133, by(miss1)

Two-sample t test with equal variances
-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    21604    4.570265    .0263219    3.868878    4.518672    4.621858
          1 |     7944    4.25428    .0439748    3.919439    4.168078    4.340482
-----+-----
combined |    29548    4.485312    .0226009    3.884996    4.441013    4.529611
-----+-----
      diff |           .3159848    .050944           .2161324    .4158372
-----+-----

      diff = mean(0) - mean(1)                                t =    6.2026
Ho: diff = 0                                                degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 1.0000                                Pr(|T| > |t|) = 0.0000                                Pr(T > t) = 0.0000

```

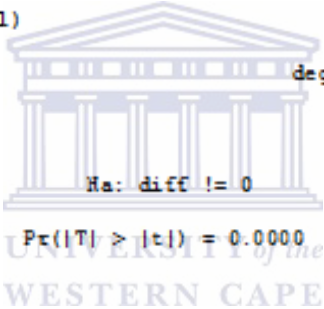


FIGURE 6.18: Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

Model 2.1 with missing completely at random values on the binary outcome variable

```

. ttest V012, by(miss1)

Two-sample t test with equal variances
-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |   14858   35.09497   .0668423   8.147629   34.96395   35.22598
          1 |   14690   35.2661    .0671254   8.135748   35.13453   35.39767
-----+-----
combined |   29548   35.18005   .0473662   8.142037   35.08721   35.27289
-----+-----
      diff |           -.1711337   .0947304                -.3568095   .0145421
-----+-----

      diff = mean(0) - mean(1)                                t = -1.8065
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.0354                                Pr(|T| > |t|) = 0.0708                                Pr(T > t) = 0.9646

```

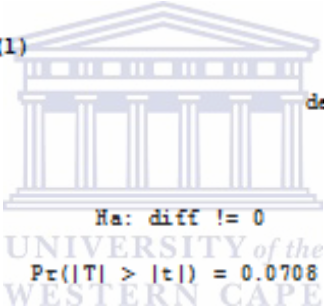


FIGURE 6.19: Model 2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

```

. ttest V013, by(miss1)

Two-sample t test with equal variances
-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    14858    4.607551    .0133678    1.629444    4.581349    4.633754
          1 |    14690    4.64275    .0134227    1.626863    4.61644    4.66906
-----+-----
combined |    29548    4.625051    .0094722    1.628229    4.606485    4.643617
-----+-----
      diff |           -.0351987    .0189439                -.0723297    .0019323
-----+-----

      diff = mean(0) - mean(1)                                t = -1.8580
Ho: diff = 0                                                degrees of freedom = 29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.0316                                Pr(|T| > |t|) = 0.0632                                Pr(T > t) = 0.9684

```

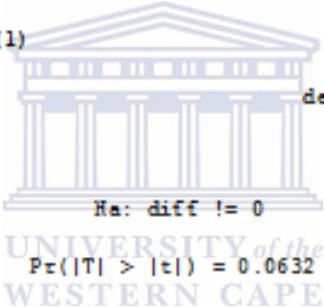


FIGURE 6.20: Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

Model 2.2 with missing at random values on the polytomous variable if a woman is aged at least 35 years

```
. ttest V012, by(miss1)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	21806	32.87682	.0537924	7.943452	32.77139 32.98226
1	7742	41.66727	.0485733	4.273901	41.57205 41.76249
combined	29548	35.18005	.0473662	8.142037	35.08721 35.27289
diff		-8.790447	.0948048		-8.976268 -8.604625

diff = mean(0) - mean(1) t = -92.7216

Ho: diff = 0 degrees of freedom = 29546

UNIVERSITY of the WESTERN CAPE

Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
Pr(T < t) = 0.0000	Pr(T > t) = 0.0000	Pr(T > t) = 1.0000

FIGURE 6.21: Model 2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

```

. ttest V133, by(miss1)

Two-sample t test with equal variances
-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |    21806    4.549803    .0262137    3.870937    4.498422    4.601184
          1 |     7742    4.303668    .0445384    3.918876    4.216361    4.390976
-----+-----
combined |    29548    4.485312    .0226009    3.884996    4.441013    4.529611
-----+-----
      diff |           .2461345    .0513782                .145431    .346838
-----+-----

      diff = mean(0) - mean(1)                                t =      4.7906
Ho: diff = 0                                                degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 1.0000                                Pr(|T| > |t|) = 0.0000                                Pr(T > t) = 0.0000

```

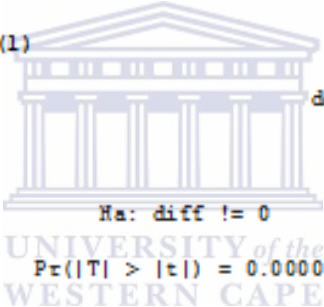


FIGURE 6.22: Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

Model 2.2 with missing completely at random values on the polytomous outcome variable


```

. ttest V012, by(miss2)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |   14833   35.10382   .0670039   8.160465   34.97249   35.23516
          2 |   14715   35.25688   .066963   8.122972   35.12562   35.38814
-----+-----
combined |   29548   35.18005   .0473662   8.142037   35.08721   35.27289
-----+-----
      diff |           -.1530582   .0947307           -.3387345   .0326181
-----+-----

      diff = mean(0) - mean(2)                                t =  -1.6157
Ho: diff = 0                                                    degrees of freedom =  29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.0531                                Pr(|T| > |t|) = 0.1062                                Pr(T > t) = 0.9469

```

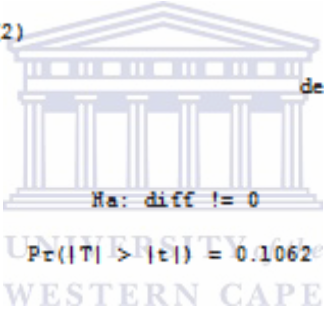


FIGURE 6.23: Model 2.1: Independent-samples t-test to compare age (in V012) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

```

. ttest V133, by(miss2)

Two-sample t test with equal variances

-----+-----
      Group |      Obs      Mean      Std. Err.      Std. Dev.      [95% Conf. Interval]
-----+-----
          0 |   14833    4.52336    .0320332    3.901343    4.460571    4.586149
          2 |   14715    4.446959   .0318881    3.868202    4.384454    4.509464
-----+-----
combined |   29548    4.485312   .0226009    3.884996    4.441013    4.529611
-----+-----
      diff |           .0764012   .0452008                -.0121944    .1649968
-----+-----

      diff = mean(0) - mean(2)                                t =    1.6903
Ho: diff = 0                                                degrees of freedom =    29546

      Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.9545                                Pr(|T| > |t|) = 0.0910                                Pr(T > t) = 0.0455

```

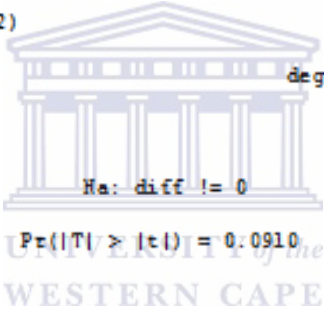


FIGURE 6.24: Model 2.1: Independent-samples t-test to compare education in completed years (V133) for missing (coded 1 in the table) and present or not missing (coded 0) under MAR assumption.

Appendix C



Model 1.1: Estimates of Monte Carlo errors after MVNI and MICE are used

Estimates of Monte Carlo errors after MVNI and MICE are applied to data sets with 50% MAR and MCAR data on the covariate

TABLE 6.1: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.005	0.001	0.050	0.000	0.005	0.005
Living together	0.002	0.000	0.080	0.000	0.002	0.002
Widowed	0.007	0.003	0.260	0.000	0.010	0.008
Divorced	0.005	0.001	0.050	0.004	0.006	0.005
Not living together	0.003	0.001	0.070	0.000	0.004	0.004

TABLE 6.2: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.007	0.002	0.060	0.001	0.009	0.006
Living together	0.003	0.001	0.060	0.000	0.003	0.003
Widowed	0.008	0.004	0.360	0.000	0.011	0.011
Divorced	0.005	0.001	0.030	0.021	0.006	0.005
Not living together	0.005	0.001	0.060	0.000	0.006	0.005

TABLE 6.3: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.007	0.002	0.070	0.000	0.008	0.008
Living together	0.002	0.001	0.140	0.000	0.003	0.003
Widowed	0.003	0.001	0.060	0.000	0.003	0.004
Divorced	0.005	0.001	0.050	0.004	0.005	0.005
Not living together	0.003	0.001	0.060	0.000	0.004	0.004

TABLE 6.4: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.011	0.005	0.090	0.004	0.016	0.013
Living together	0.003	0.001	0.100	0.000	0.004	0.004
Widowed	0.005	0.001	0.070	0.000	0.005	0.005
Divorced	0.007	0.002	0.040	0.027	0.007	0.008
Not living together	0.004	0.001	0.060	0.000	0.004	0.005

TABLE 6.5: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.011	0.005	0.100	0.000	0.015	0.013
Living together	0.004	0.002	0.020	0.000	0.005	0.006
Widowed	0.010	0.005	0.100	0.000	0.013	0.016
Divorced	0.011	0.006	0.020	0.005	0.015	0.016
Not living together	0.006	0.003	0.100	0.016	0.008	0.010

TABLE 6.6: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.016	0.009	0.100	0.016	0.029	0.020
Living together	0.007	0.004	0.060	0.000	0.008	0.012
Widowed	0.014	0.007	0.070	0.000	0.018	0.020
Divorced	0.015	0.007	0.070	0.016	0.024	0.017
Not living together	0.008	0.005	0.080	0.015	0.014	0.010

TABLE 6.7: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.011	0.005	0.100	0.001	0.014	0.016
Living together	0.005	0.002	0.080	0.000	0.007	0.006
Widowed	0.014	0.007	0.030	0.000	0.018	0.020
Divorced	0.011	0.006	0.100	0.004	0.017	0.015
Not living together	0.006	0.003	0.100	0.022	0.008	0.010

TABLE 6.8: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.015	0.010	0.090	0.009	0.020	0.028
Living together	0.005	0.002	0.100	0.000	0.007	0.007
Widowed	0.019	0.011	0.080	0.000	0.030	0.026
Divorced	0.013	0.006	0.070	0.010	0.019	0.015
Not living together	0.009	0.004	0.070	0.014	0.014	0.010

Estimates of Monte Carlo errors after MVNI and MICE are applied to data sets with 30% MAR and MCAR data on the covariate

TABLE 6.9: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.003	0.001	0.040	0.000	0.004	0.003
Living together	0.001	0.000	0.050	0.000	0.001	0.001
Widowed	0.005	0.001	0.020	0.000	0.006	0.005
Divorced	0.004	0.001	0.030	0.000	0.004	0.004
Not living together	0.002	0.000	0.030	0.000	0.002	0.002

TABLE 6.10: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.005	0.001	0.040	0.001	0.006	0.004
Living together	0.002	0.000	0.040	0.000	0.002	0.002
Widowed	0.008	0.003	0.070	0.000	0.010	0.008
Divorced	0.004	0.001	0.020	0.007	0.005	0.003
Not living together	0.003	0.001	0.030	0.000	0.003	0.003

TABLE 6.11: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.004	0.001	0.050	0.000	0.005	0.005
Living together	0.002	0.000	0.080	0.000	0.002	0.002
Widowed	0.002	0.000	0.030	0.000	0.002	0.002
Divorced	0.003	0.001	0.030	0.000	0.003	0.003
Not living together	0.002	0.000	0.040	0.000	0.002	0.002

TABLE 6.12: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 30% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.005	0.001	0.030	0.001	0.005	0.005
Living together	0.002	0.001	0.040	0.000	0.002	0.002
Widowed	0.003	0.000	0.030	0.000	0.003	0.003
Divorced	0.004	0.001	0.020	0.008	0.004	0.004
Not living together	0.003	0.000	0.030	0.000	0.003	0.003

TABLE 6.13: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 30% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.008	0.004	0.100	0.000	0.009	0.012
Living together	0.003	0.001	0.090	0.000	0.003	0.004
Widowed	0.008	0.003	0.070	0.000	0.010	0.010
Divorced	0.007	0.003	0.070	0.004	0.010	0.008
Not living together	0.005	0.002	0.080	0.002	0.007	0.006

TABLE 6.14: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 30% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.010	0.004	0.080	0.000	0.011	0.013
Living together	0.004	0.002	0.110	0.000	0.005	0.006
Widowed	0.015	0.006	0.030	0.000	0.018	0.021
Divorced	0.009	0.003	0.060	0.024	0.010	0.012
Not living together	0.007	0.003	0.080	0.007	0.008	0.009

TABLE 6.15: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 30% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.007	0.003	0.020	0.000	0.011	0.008
Living together	0.003	0.001	0.090	0.000	0.004	0.004
Widowed	0.009	0.004	0.040	0.000	0.011	0.012
Divorced	0.007	0.003	0.070	0.003	0.010	0.009
Not living together	0.004	0.001	0.070	0.001	0.005	0.005

TABLE 6.16: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 30% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.010	0.004	0.080	0.000	0.013	0.013
Living together	0.004	0.002	0.030	0.000	0.006	0.004
Widowed	0.012	0.006	0.070	0.000	0.018	0.016
Divorced	0.009	0.004	0.040	0.021	0.014	0.008
Not living together	0.007	0.003	0.060	0.006	0.010	0.007

Estimates of Monte Carlo errors after MVNI and MICE are applied to data sets with 10% MAR and MCAR data on the covariate

TABLE 6.17: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.001	0.000	0.010	0.000	0.002	0.001
Living together	0.001	0.000	0.030	0.000	0.001	0.001
Widowed	0.003	0.001	0.060	0.000	0.004	0.003
Divorced	0.002	0.000	0.020	0.000	0.002	0.002
Not living together	0.001	0.000	0.010	0.000	0.001	0.001

TABLE 6.18: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.002	0.000	0.010	0.000	0.002	0.002
Living together	0.001	0.000	0.020	0.000	0.001	0.001
Widowed	0.006	0.002	0.060	0.000	0.007	0.006
Divorced	0.002	0.000	0.010	0.004	0.003	0.002
Not living together	0.002	0.000	0.020	0.000	0.002	0.002

TABLE 6.19: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.002	0.000	0.020	0.000	0.002	0.003
Living together	0.001	0.000	0.030	0.000	0.001	0.001
Widowed	0.001	0.000	0.010	0.000	0.001	0.001
Divorced	0.002	0.000	0.010	0.000	0.002	0.002
Not living together	0.001	0.000	0.010	0.000	0.001	0.001

TABLE 6.20: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 10% MAR data on marital status if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.003	0.000	0.020	0.000	0.003	0.003
Living together	0.001	0.000	0.020	0.000	0.001	0.001
Widowed	0.002	0.000	0.010	0.000	0.002	0.002
Divorced	0.003	0.000	0.010	0.005	0.003	0.002
Not living together	0.001	0.000	0.010	0.000	0.001	0.001

TABLE 6.21: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 10% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.004	0.001	0.050	0.000	0.005	0.004
Living together	0.001	0.000	0.060	0.000	0.002	0.001
Widowed	0.006	0.002	0.100	0.000	0.008	0.006
Divorced	0.003	0.001	0.030	0.000	0.004	0.003
Not living together	0.002	0.001	0.040	0.000	0.002	0.002

TABLE 6.22: Model 1.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 10% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.005	0.001	0.040	0.000	0.006	0.005
Living together	0.002	0.001	0.060	0.000	0.003	0.002
Widowed	0.010	0.004	0.030	0.000	0.012	0.012
Divorced	0.004	0.001	0.020	0.009	0.006	0.004
Not living together	0.004	0.001	0.040	0.000	0.006	0.004

TABLE 6.23: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 10% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.004	0.001	0.040	0.000	0.005	0.004
Living together	0.001	0.000	0.060	0.000	0.001	0.002
Widowed	0.005	0.001	0.080	0.000	0.005	0.005
Divorced	0.004	0.001	0.040	0.001	0.004	0.004
Not living together	0.002	0.001	0.040	0.000	0.003	0.003

TABLE 6.24: Model 1.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 10% MCAR data on marital status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.005	0.001	0.030	0.000	0.005	0.007
Living together	0.003	0.001	0.060	0.000	0.003	0.003
Widowed	0.010	0.005	0.080	0.000	0.009	0.016
Divorced	0.005	0.001	0.030	0.013	0.006	0.005
Not living together	0.003	0.001	0.050	0.000	0.003	0.004

Model 1.2: Estimates of Monte Carlo errors after MVNI and MICE are used

Model 1.2.1: Estimates of Monte Carlo errors after MVNI and MICE are applied to data sets with 50% MAR and MCAR data on the covariate

TABLE 6.25: Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.001	0.040	0.027	0.007	0.006	
Living together	0.000	0.050	0.000	0.002	0.002	
Widowed	0.002	0.200	0.000	0.006	0.008	
Divorced	0.002	0.070	0.000	0.008	0.007	
Not living together	0.001	0.080	0.000	0.004	0.004	
Bas Kongo	0.001	0.040	0.001	0.003	0.003	
Bandundu	0.000	0.070	0.000	0.002	0.002	
Equateur	0.001	0.060	0.000	0.003	0.003	
Orientale	0.001	0.060	0.000	0.004	0.004	
Nord Kivu	0.001	0.080	0.000	0.003	0.003	
Maniema	0.001	0.080	0.000	0.003	0.003	
Sud Kivu	0.001	0.060	0.000	0.003	0.004	
Katanga	0.001	0.080	0.000	0.003	0.003	
Kasai Occidental	0.001	0.030	0.000	0.004	0.005	
Kasai Oriental	0.001	0.050	0.000	0.004	0.004	

TABLE 6.26: Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.007	0.002	0.040	0.022	0.009	0.007
Living together	0.003	0.001	0.040	0.006	0.003	0.003
Widowed	0.008	0.003	0.300	0.000	0.009	0.010
Divorced	0.007	0.002	0.040	0.011	0.008	0.007
Not living together	0.004	0.001	0.050	0.000	0.005	0.004
Bas Kongo	0.004	0.002	0.190	0.000	0.006	0.006
Bandundu	0.002	0.000	0.030	0.000	0.002	0.002
Equateur	0.003	0.001	0.070	0.000	0.003	0.003
Orientale	0.003	0.001	0.080	0.000	0.004	0.003
Nord Kivu	0.005	0.002	0.070	0.000	0.007	0.006
Maniema	0.006	0.002	0.050	0.000	0.007	0.008
Sud Kivu	0.005	0.001	0.080	0.000	0.006	0.006
Katanga	0.003	0.001	0.010	0.000	0.003	0.003
Kasai Occidental	0.004	0.001	0.010	0.000	0.004	0.005
Kasai Oriental	0.003	0.001	0.020	0.000	0.003	0.003

TABLE 6.27: Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.012	0.005	0.060	0.032	0.011	0.019
Living together	0.003	0.001	0.100	0.000	0.003	0.003
Widowed	0.003	0.001	0.050	0.000	0.004	0.004
Divorced	0.005	0.001	0.050	0.000	0.006	0.005
Not living together	0.003	0.001	0.090	0.000	0.003	0.004
Bas Kongo	0.004	0.001	0.070	0.001	0.004	0.004
Bandundu	0.003	0.001	0.110	0.000	0.003	0.003
Equateur	0.003	0.001	0.010	0.000	0.003	0.004
Orientale	0.003	0.001	0.030	0.000	0.003	0.003
Nord Kivu	0.003	0.001	0.020	0.000	0.004	0.004
Maniema	0.003	0.001	0.090	0.000	0.003	0.003
Sud Kivu	0.003	0.001	0.040	0.000	0.003	0.003
Katanga	0.003	0.001	0.090	0.000	0.003	0.003
Kasai Occidental	0.004	0.001	0.030	0.000	0.005	0.004
Kasai Oriental	0.003	0.001	0.030	0.000	0.004	0.003

TABLE 6.28: Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on marital status and region if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.010	0.003	0.050	0.025	0.010	0.014
Living together	0.003	0.001	0.040	0.007	0.003	0.004
Widowed	0.004	0.001	0.070	0.000	0.004	0.005
Divorced	0.005	0.001	0.030	0.009	0.006	0.005
Not living together	0.004	0.001	0.060	0.000	0.004	0.005
Bas Kongo	0.005	0.002	0.150	0.000	0.007	0.006
Bandundu	0.003	0.001	0.050	0.000	0.003	0.003
Equateur	0.003	0.001	0.100	0.000	0.003	0.003
Orientale	0.003	0.001	0.070	0.000	0.003	0.003
Nord Kivu	0.004	0.001	0.080	0.000	0.005	0.005
Maniema	0.004	0.002	0.090	0.000	0.005	0.005
Sud Kivu	0.004	0.001	0.120	0.000	0.004	0.004
Katanga	0.004	0.001	0.090	0.000	0.004	0.005
Kasai Occidental	0.004	0.001	0.090	0.000	0.005	0.004
Kasai Oriental	0.004	0.001	0.090	0.000	0.004	0.004

TABLE 6.29: Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.013	0.006	0.070	0.053	0.018	0.017
Living together	0.005	0.003	0.080	0.000	0.007	0.008
Widowed	0.010	0.006	0.040	0.000	0.017	0.015
Divorced	0.011	0.007	0.030	0.003	0.018	0.017
Not living together	0.008	0.004	0.030	0.001	0.011	0.010
Bas Kongo	0.005	0.002	0.100	0.004	0.006	0.007
Bandundu	0.005	0.002	0.020	0.000	0.007	0.006
Equateur	0.005	0.002	0.080	0.000	0.007	0.006
Orientale	0.006	0.003	0.040	0.000	0.009	0.008
Nord Kivu	0.006	0.002	0.050	0.000	0.007	0.006
Maniema	0.006	0.003	0.070	0.000	0.008	0.009
Sud Kivu	0.008	0.004	0.063	0.000	0.009	0.011
Katanga	0.006	0.003	0.054	0.000	0.008	0.008
Kasai Occidental	0.006	0.003	0.060	0.000	0.009	0.009
Kasai Oriental	0.006	0.003	0.060	0.000	0.009	0.008

TABLE 6.30: Model 1.2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.016	0.008	0.070	0.052	0.021	0.023
Living together	0.006	0.003	0.120	0.001	0.009	0.008
Widowed	0.012	0.006	0.270	0.000	0.016	0.017
Divorced	0.014	0.007	0.080	0.018	0.020	0.018
Not living together	0.009	0.004	0.110	0.002	0.013	0.012
Bas Kongo	0.007	0.003	0.080	0.018	0.009	0.010
Bandundu	0.005	0.002	0.180	0.000	0.006	0.007
Equateur	0.006	0.002	0.240	0.000	0.008	0.006
Orientale	0.007	0.003	0.370	0.000	0.009	0.010
Nord Kivu	0.008	0.004	0.320	0.000	0.012	0.011
Maniema	0.010	0.004	0.330	0.000	0.013	0.014
Sud Kivu	0.011	0.004	0.330	0.000	0.013	0.014
Katanga	0.008	0.004	0.390	0.000	0.011	0.011
Kasai Occidental	0.008	0.004	0.440	0.000	0.010	0.011
Kasai Oriental	0.007	0.003	0.400	0.000	0.008	0.010

TABLE 6.31: Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.017	0.009	0.080	0.061	0.021	0.028
Living together	0.005	0.003	0.090	0.000	0.007	0.009
Widowed	0.013	0.007	0.031	0.000	0.020	0.018
Divorced	0.012	0.007	0.050	0.004	0.017	0.019
Not living together	0.008	0.005	0.040	0.001	0.014	0.010
Bas Kongo	0.005	0.002	0.010	0.006	0.006	0.008
Bandundu	0.005	0.002	0.030	0.000	0.007	0.006
Equateur	0.005	0.002	0.090	0.000	0.007	0.007
Orientale	0.007	0.003	0.040	0.000	0.009	0.009
Nord Kivu	0.007	0.003	0.050	0.000	0.011	0.008
Maniema	0.006	0.002	0.060	0.000	0.009	0.007
Sud Kivu	0.007	0.003	0.030	0.000	0.010	0.009
Katanga	0.006	0.004	0.030	0.000	0.010	0.009
Kasai Occidental	0.006	0.004	0.030	0.000	0.010	0.009
Kasai Oriental	0.007	0.004	0.050	0.000	0.012	0.010

TABLE 6.32: Model 1.2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.018	0.010	0.070	0.056	0.022	0.032
Living together	0.006	0.003	0.120	0.001	0.009	0.008
Widowed	0.019	0.009	0.250	0.000	0.024	0.026
Divorced	0.014	0.008	0.070	0.015	0.025	0.017
Not living together	0.010	0.005	0.120	0.003	0.013	0.015
Bas Kongo	0.008	0.005	0.090	0.024	0.014	0.012
Bandundu	0.006	0.003	0.200	0.000	0.009	0.007
Equateur	0.006	0.003	0.250	0.000	0.009	0.007
Orientale	0.007	0.004	0.420	0.000	0.009	0.012
Nord Kivu	0.010	0.005	0.340	0.000	0.013	0.015
Maniema	0.009	0.004	0.330	0.000	0.012	0.013
Sud Kivu	0.010	0.004	0.320	0.000	0.012	0.013
Katanga	0.006	0.003	0.350	0.000	0.007	0.008
Kasai Occidental	0.006	0.002	0.290	0.000	0.007	0.007
Kasai Oriental	0.007	0.004	0.400	0.000	0.010	0.010

Model 1.2.2: Estimates of Monte Carlo errors after MVNI and MICE are applied to data sets with 50% data missing at random or completely at random on the covariate

TABLE 6.33: Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% of data missing at random on marital status and region if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.006	0.002	0.050	0.005	0.007	0.007
Living together	0.002	0.001	0.060	0.000	0.002	0.002
Widowed	0.006	0.002	0.080	0.000	0.007	0.008
Divorced	0.006	0.002	0.050	0.004	0.009	0.007
Not living together	0.003	0.001	0.070	0.000	0.004	0.004
Bas Kongo	0.003	0.001	0.110	0.000	0.004	0.003
Bandundu	0.003	0.001	0.040	0.000	0.003	0.003
Equateur	0.003	0.001	0.040	0.013	0.004	0.004
Orientale	0.004	0.001	0.070	0.000	0.004	0.004
Nord Kivu	0.003	0.001	0.050	0.001	0.004	0.004
Maniema	0.003	0.001	0.050	0.001	0.004	0.004
Sud Kivu	0.003	0.001	0.020	0.000	0.003	0.004
Katanga	0.003	0.001	0.070	0.000	0.003	0.003
Kasai Occidental	0.004	0.001	0.020	0.000	0.004	0.005
Kasai Oriental	0.004	0.001	0.120	0.000	0.004	0.004
Age	0.000	0.000	0.030	0.000	0.000	0.000
Education	0.000	0.000	0.040	0.000	0.000	0.000
Poorer	0.001	0.000	0.020	0.000	0.001	0.001
Middle	0.001	0.000	0.020	0.000	0.001	0.001
Richer	0.001	0.000	0.030	0.000	0.001	0.001
Richest	0.002	0.000	0.050	0.000	0.002	0.002

TABLE 6.34: Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% of data missing at random on marital status and region if a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.007	0.002	0.040	0.024	0.009	0.007
Living together	0.003	0.001	0.040	0.005	0.003	0.003
Widowed	0.008	0.003	0.300	0.000	0.010	0.010
Divorced	0.007	0.002	0.030	0.024	0.008	0.006
Not living together	0.004	0.001	0.050	0.000	0.005	0.004
Bas Kongo	0.005	0.002	0.070	0.008	0.006	0.006
Bandundu	0.003	0.001	0.050	0.000	0.003	0.003
Equateur	0.004	0.001	0.030	0.026	0.004	0.004
Orientale	0.004	0.001	0.040	0.000	0.004	0.004
Nord Kivu	0.006	0.002	0.100	0.000	0.007	0.006
Maniema	0.006	0.002	0.090	0.000	0.008	0.008
Sud Kivu	0.005	0.001	0.100	0.000	0.005	0.005
Katanga	0.003	0.001	0.050	0.000	0.004	0.003
Kasai Occidental	0.004	0.001	0.100	0.000	0.005	0.005
Kasai Oriental	0.003	0.001	0.070	0.000	0.004	0.004
Age	0.000	0.000	0.020	0.000	0.000	0.000
Education	0.000	0.000	0.020	0.000	0.000	0.000
Poorer	0.001	0.000	0.020	0.005	0.001	0.001
Middle	0.001	0.000	0.020	0.001	0.001	0.001
Richer	0.001	0.000	0.020	0.000	0.001	0.002
Richest	0.002	0.000	0.040	0.000	0.002	0.002

TABLE 6.35: Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on marital status and region is a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.010	0.004	0.080	0.025	0.011	0.015
Living together	0.003	0.001	0.010	0.000	0.003	0.004
Widowed	0.004	0.001	0.050	0.000	0.004	0.004
Divorced	0.005	0.001	0.040	0.002	0.006	0.005
Not living together	0.003	0.001	0.090	0.000	0.003	0.004
Bas Kongo	0.004	0.001	0.030	0.000	0.005	0.005
Bandundu	0.004	0.001	0.080	0.000	0.004	0.004
Equateur	0.004	0.001	0.050	0.016	0.005	0.005
Orientale	0.003	0.001	0.070	0.000	0.004	0.004
Nord Kivu	0.004	0.001	0.060	0.001	0.004	0.004
Maniema	0.004	0.001	0.060	0.001	0.005	0.005
Sud Kivu	0.004	0.001	0.110	0.000	0.004	0.004
Katanga	0.003	0.001	0.020	0.000	0.004	0.004
Kasai Occidental	0.005	0.002	0.060	0.000	0.006	0.005
Kasai Oriental	0.004	0.001	0.100	0.000	0.005	0.004
Age	0.000	0.000	0.020	0.000	0.000	0.000
Education	0.000	0.000	0.050	0.000	0.000	0.000
Poorer	0.001	0.000	0.020	0.000	0.001	0.001
Middle	0.001	0.000	0.020	0.000	0.001	0.001
Richer	0.001	0.000	0.020	0.000	0.001	0.001
Richest	0.002	0.000	0.070	0.000	0.002	0.002

TABLE 6.36: Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on marital status and region is a woman is not using any contraceptive method.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.010	0.003	0.050	0.037	0.009	0.013
Living together	0.003	0.001	0.040	0.006	0.003	0.004
Widowed	0.004	0.001	0.070	0.000	0.004	0.005
Divorced	0.006	0.001	0.030	0.021	0.006	0.006
Not living together	0.004	0.001	0.060	0.000	0.004	0.005
Bas Kongo	0.006	0.002	0.070	0.015	0.008	0.007
Bandundu	0.004	0.001	0.080	0.000	0.005	0.005
Equateur	0.004	0.001	0.040	0.026	0.005	0.005
Orientale	0.004	0.001	0.040	0.000	0.004	0.004
Nord Kivu	0.005	0.002	0.090	0.000	0.006	0.006
Maniema	0.005	0.002	0.080	0.000	0.006	0.006
Sud Kivu	0.004	0.001	0.080	0.000	0.004	0.005
Katanga	0.004	0.001	0.080	0.000	0.004	0.005
Kasai Occidental	0.004	0.002	0.110	0.000	0.006	0.004
Kasai Oriental	0.004	0.001	0.100	0.000	0.005	0.005
Age	0.000	0.000	0.020	0.000	0.000	0.000
Education	0.000	0.000	0.020	0.000	0.000	0.000
Poorer	0.001	0.000	0.020	0.005	0.001	0.001
Middle	0.001	0.000	0.020	0.002	0.001	0.001
Richer	0.002	0.000	0.030	0.000	0.002	0.002
Richest	0.002	0.000	0.060	0.000	0.002	0.002

TABLE 6.37: Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.012	0.006	0.090	0.008	0.019	0.016
Living together	0.005	0.003	0.200	0.000	0.007	0.008
Widowed	0.011	0.006	0.050	0.000	0.018	0.015
Divorced	0.012	0.007	0.100	0.023	0.019	0.017
Not living together	0.007	0.004	0.030	0.001	0.011	0.010
Bas Kongo	0.007	0.004	0.022	0.000	0.009	0.010
Bandundu	0.007	0.003	0.090	0.024	0.010	0.009
Equateur	0.007	0.003	0.090	0.003	0.010	0.008
Orientale	0.008	0.004	0.090	0.000	0.010	0.011
Nord Kivu	0.007	0.003	0.050	0.000	0.010	0.009
Maniema	0.008	0.004	0.070	0.000	0.011	0.011
Sud Kivu	0.009	0.005	0.030	0.000	0.013	0.013
Katanga	0.006	0.003	0.050	0.000	0.008	0.010
Kasai Occidental	0.008	0.004	0.030	0.000	0.012	0.011
Kasai Oriental	0.008	0.004	0.090	0.000	0.011	0.010
Age	0.000	0.000	0.050	0.000	0.000	0.000
Education	0.000	0.000	0.100	0.000	0.000	0.000
Poorer	0.001	0.000	0.020	0.001	0.001	0.002
Middle	0.001	0.000	0.030	0.000	0.001	0.001
Richer	0.001	0.000	0.040	0.000	0.001	0.001
Richest	0.003	0.001	0.030	0.000	0.004	0.004

TABLE 6.38: Model 1.2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.016	0.009	0.080	0.033	0.022	0.025
Living together	0.006	0.003	0.030	0.000	0.009	0.008
Widowed	0.012	0.006	0.050	0.000	0.017	0.017
Divorced	0.014	0.007	0.070	0.038	0.021	0.018
Not living together	0.009	0.004	0.010	0.002	0.013	0.012
Bas Kongo	0.009	0.004	0.020	0.000	0.011	0.012
Bandundu	0.008	0.004	0.070	0.037	0.011	0.012
Equateur	0.008	0.004	0.090	0.002	0.011	0.010
Orientale	0.010	0.005	0.021	0.000	0.013	0.016
Nord Kivu	0.010	0.005	0.030	0.000	0.015	0.014
Maniema	0.013	0.005	0.040	0.000	0.016	0.017
Sud Kivu	0.012	0.005	0.010	0.000	0.016	0.016
Katanga	0.009	0.004	0.021	0.000	0.012	0.013
Kasai Occidental	0.010	0.005	0.021	0.000	0.014	0.015
Kasai Oriental	0.008	0.004	0.021	0.000	0.011	0.011
Age	0.000	0.000	0.030	0.000	0.000	0.000
Education	0.000	0.000	0.040	0.000	0.000	0.000
Poorer	0.002	0.000	0.020	0.010	0.002	0.002
Middle	0.002	0.000	0.020	0.004	0.002	0.002
Richer	0.002	0.000	0.040	0.000	0.002	0.002
Richest	0.005	0.001	0.100	0.000	0.005	0.005

TABLE 6.39: Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.014	0.007	0.100	0.019	0.019	0.019
Living together	0.005	0.003	0.200	0.000	0.006	0.008
Widowed	0.014	0.007	0.090	0.000	0.019	0.019
Divorced	0.012	0.007	0.100	0.022	0.017	0.018
Not living together	0.008	0.004	0.030	0.001	0.013	0.010
Bas Kongo	0.007	0.004	0.021	0.000	0.009	0.011
Bandundu	0.008	0.004	0.010	0.011	0.012	0.011
Equateur	0.008	0.004	0.090	0.028	0.010	0.011
Orientale	0.009	0.004	0.070	0.000	0.011	0.012
Nord Kivu	0.008	0.005	0.030	0.001	0.014	0.011
Maniema	0.009	0.004	0.020	0.001	0.013	0.011
Sud Kivu	0.009	0.004	0.090	0.000	0.012	0.012
Katanga	0.008	0.005	0.050	0.000	0.012	0.012
Kasai Occidental	0.008	0.005	0.031	0.000	0.013	0.012
Kasai Oriental	0.009	0.006	0.040	0.000	0.015	0.015
Age	0.000	0.000	0.040	0.000	0.000	0.000
Education	0.000	0.000	0.100	0.000	0.000	0.000
Poorer	0.001	0.000	0.030	0.001	0.001	0.001
Middle	0.001	0.000	0.030	0.000	0.002	0.001
Richer	0.002	0.000	0.060	0.000	0.002	0.002
Richest	0.004	0.001	0.090	0.000	0.005	0.005

TABLE 6.40: Model 1.2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on marital status and region.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.017	0.009	0.090	0.049	0.021	0.028
Living together	0.006	0.003	0.030	0.001	0.009	0.008
Widowed	0.019	0.009	0.040	0.000	0.025	0.027
Divorced	0.013	0.007	0.060	0.030	0.023	0.016
Not living together	0.010	0.005	0.020	0.003	0.013	0.014
Bas Kongo	0.009	0.006	0.021	0.000	0.017	0.015
Bandundu	0.008	0.005	0.100	0.028	0.014	0.011
Equateur	0.008	0.004	0.080	0.010	0.011	0.011
Orientale	0.009	0.004	0.060	0.000	0.011	0.013
Nord Kivu	0.011	0.006	0.040	0.002	0.014	0.017
Maniema	0.011	0.005	0.030	0.001	0.015	0.016
Sud Kivu	0.011	0.005	0.020	0.000	0.014	0.015
Katanga	0.007	0.003	0.090	0.000	0.008	0.010
Kasai Occidental	0.007	0.003	0.070	0.000	0.009	0.008
Kasai Oriental	0.009	0.005	0.020	0.000	0.014	0.013
Age	0.000	0.000	0.020	0.000	0.000	0.000
Education	0.000	0.000	0.030	0.000	0.000	0.000
Poorer	0.002	0.000	0.020	0.009	0.002	0.002
Middle	0.002	0.000	0.020	0.004	0.002	0.002
Richer	0.002	0.000	0.040	0.000	0.002	0.002
Richest	0.004	0.002	0.050	0.000	0.006	0.005

Model 2: Estimates of Monte Carlo errors after MVNI and MICE are used

Model 2.1: Estimates of Monte Carlo errors after MVNI and MICE are applied to data set with missing values on the binary outcome variable

TABLE 6.41: Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.003	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.003	0.000	0.000	0.000	0.000	0.000

TABLE 6.42: Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.003	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.003	0.000	0.000	0.000	0.000	0.000

TABLE 6.43: Model 2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.002	0.000	0.010	0.000	0.002	0.002
Living together	0.002	0.000	0.090	0.000	0.002	0.002
Widowed	0.002	0.000	0.020	0.000	0.002	0.002
Divorced	0.002	0.000	0.010	0.000	0.002	0.002
Not notliving together	0.002	0.000	0.030	0.000	0.002	0.002

TABLE 6.44: Model 2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.002	0.000	0.010	0.000	0.002	0.002
Living together	0.002	0.001	0.050	0.000	0.002	0.002
Widowed	0.002	0.000	0.020	0.000	0.002	0.002
Divorced	0.002	0.000	0.010	0.004	0.002	0.002
Not notliving together	0.002	0.000	0.020	0.000	0.002	0.002

TABLE 6.45: Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on contraceptive method use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.011	0.005	0.100	0.000	0.015	0.013
Living together	0.004	0.002	0.020	0.000	0.005	0.006
Widowed	0.010	0.005	0.100	0.000	0.013	0.016
Divorced	0.011	0.006	0.020	0.005	0.015	0.016
Not living together	0.006	0.003	0.100	0.016	0.008	0.010

TABLE 6.46: Model 2.1: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on contraceptive method use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.016	0.009	0.100	0.016	0.029	0.020
Living together	0.007	0.004	0.060	0.000	0.008	0.012
Widowed	0.014	0.007	0.070	0.000	0.018	0.020
Divorced	0.015	0.007	0.070	0.016	0.024	0.017
Not living together	0.008	0.005	0.080	0.015	0.014	0.010

TABLE 6.47: Model 2.1: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on contraceptive method use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.011	0.005	0.100	0.001	0.014	0.016
Living together	0.005	0.002	0.080	0.000	0.007	0.006
Widowed	0.014	0.007	0.030	0.000	0.018	0.020
Divorced	0.011	0.006	0.100	0.004	0.017	0.015
Not living together	0.006	0.003	0.100	0.022	0.008	0.010

TABLE 6.48: Model 2.1: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR on contraceptive method use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Never married	0.015	0.010	0.090	0.009	0.020	0.028
Living together	0.005	0.002	0.100	0.000	0.007	0.007
Widowed	0.019	0.011	0.080	0.000	0.030	0.026
Divorced	0.013	0.006	0.070	0.010	0.019	0.015
Not living together	0.009	0.004	0.070	0.014	0.014	0.010

Model 2.2: Estimates of Monte Carlo errors after MVNI and MICE are applied to outcome variable with more than two levels or categories

TABLE 6.49: Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.000	0.000	0.000	0.000	0.000	0.000
Modern method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 6.50: Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.003	0.000	0.000	0.000	0.000	0.000
Modern method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.003	0.000	0.000	0.000	0.000	0.000

TABLE 6.51: Model 2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.004	0.001	0.030	0.008	0.005	0.004
Living together	0.002	0.001	0.080	0.000	0.003	0.003
Widowed	0.018	0.009	0.230	0.000	0.029	0.019
Divorced	0.009	0.005	0.110	0.005	0.011	0.016
Not living together	0.005	0.002	0.070	0.000	0.006	0.005
Modern method						
Never married	0.005	0.001	0.040	0.000	0.005	0.005
Living together	0.004	0.002	0.140	0.000	0.005	0.005
Widowed	0.015	0.008	0.120	0.003	0.021	0.022
Divorced	0.010	0.004	0.060	0.027	0.012	0.013
Not living together	0.006	0.002	0.050	0.037	0.007	0.008

TABLE 6.52: Model 2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MAR data on contraceptive method use status if a woman is aged at least 35 years.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.006	0.030	0.030	0.018	0.006	0.006
Living together	0.003	0.050	0.050	0.003	0.004	0.004
Widowed	0.021	0.240	0.240	0.000	0.031	0.027
Divorced	0.011	0.050	0.050	0.011	0.013	0.012
Not living together	0.007	0.110	0.110	0.000	0.008	0.010
Modern method						
Never married	0.006	0.030	0.030	0.000	0.006	0.007
Living together	0.005	0.090	0.090	0.000	0.005	0.006
Widowed	0.016	0.110	0.110	0.001	0.019	0.022
Divorced	0.012	0.050	0.050	0.012	0.012	0.020
Not living together	0.007	0.002	0.040	0.019	0.008	0.009

TABLE 6.53: Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to unweighted data set with approximately 50% MCAR data on contraceptive method use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.000	0.000	0.000	0.000	0.000	0.000
Modern method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 6.54: Model 2.2: Estimates of Monte Carlo errors after MVNI is applied to weighted data set with approximately 50% MCAR data on contraceptive method use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.003	0.000	0.000	0.000	0.000	0.000
Modern method						
Never married	0.000	0.000	0.000	0.000	0.000	0.000
Living together	0.000	0.000	0.000	0.000	0.000	0.000
Widowed	0.000	0.000	0.000	0.000	0.000	0.000
Divorced	0.000	0.000	0.000	0.000	0.000	0.000
Not living together	0.003	0.000	0.000	0.000	0.000	0.000

TABLE 6.55: Model 2.2: Estimates of Monte Carlo errors after MICE is applied to unweighted data set with approximately 50% MCAR data on contraceptive methods use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.015	0.008	0.070	0.052	0.022	0.022
Living together	0.005	0.002	0.100	0.000	0.006	0.006
Widowed	0.021	0.011	0.090	0.000	0.032	0.028
Divorced	0.014	0.006	0.120	0.000	0.019	0.018
Not living together	0.007	0.004	0.100	0.008	0.010	0.011
Modern method						
Never married	0.016	0.008	0.070	0.000	0.022	0.023
Living together	0.006	0.003	0.090	0.000	0.009	0.008
Widowed	0.017	0.007	0.110	0.002	0.023	0.021
Divorced	0.015	0.007	0.070	0.046	0.020	0.020
Not living together	0.011	0.005	0.080	0.058	0.018	0.011

TABLE 6.56: Model 2.2: Estimates of Monte Carlo errors after MICE is applied to weighted data set with approximately 50% MCAR data on contraceptive methods use status.

	Statistics					
	Coef.	Std. Error	t	P > t	Lower limit of the 95% C.I	Upper limit of the 95% C.I
Traditional method						
Never married	0.023	0.011	0.070	0.059	0.030	0.032
Living together	0.006	0.002	0.100	0.002	0.008	0.007
Widowed	0.031	0.019	0.070	0.000	0.052	0.045
Divorced	0.013	0.004	0.050	0.004	0.017	0.013
Not living together	0.009	0.006	0.110	0.004	0.018	0.013
Modern method						
Never married	0.016	0.007	0.110	0.000	0.020	0.022
Living together	0.009	0.005	0.090	0.000	0.013	0.012
Widowed	0.023	0.013	0.100	0.003	0.039	0.029
Divorced	0.015	0.007	0.060	0.039	0.019	0.021
Not living together	0.012	0.006	0.060	0.050	0.015	0.017

Appendix D



Model 1.1: Imputation models diagnostics

Imputation models's diagnostics when 50% of data are missing at random or completely at random on the covariate

Unweighted data sets

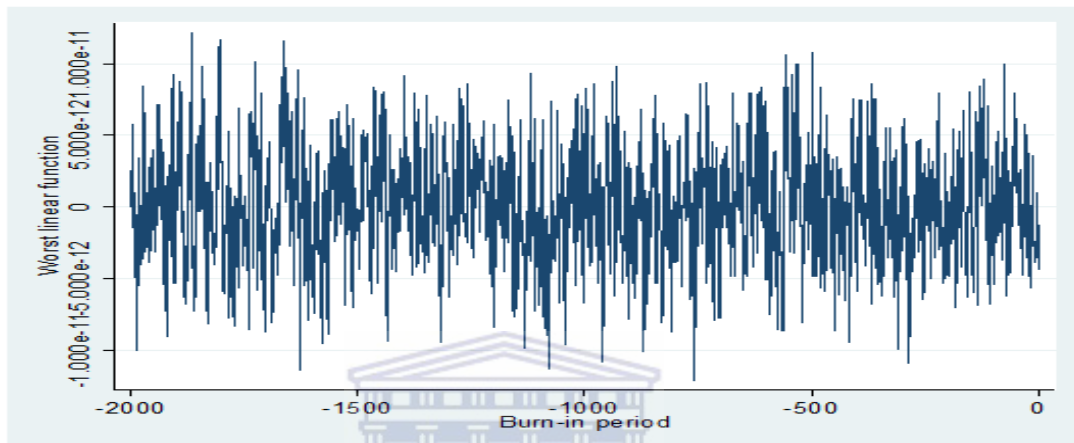


FIGURE 6.25: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

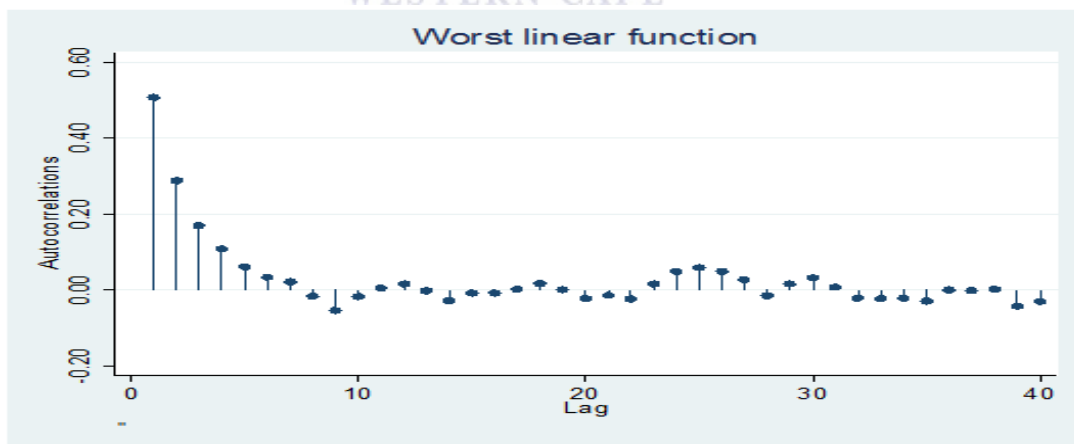


FIGURE 6.26: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

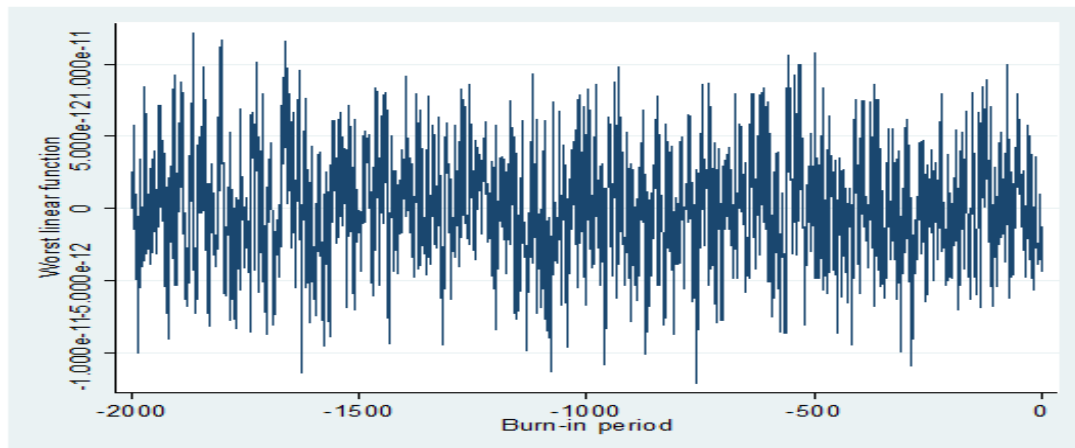


FIGURE 6.27: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

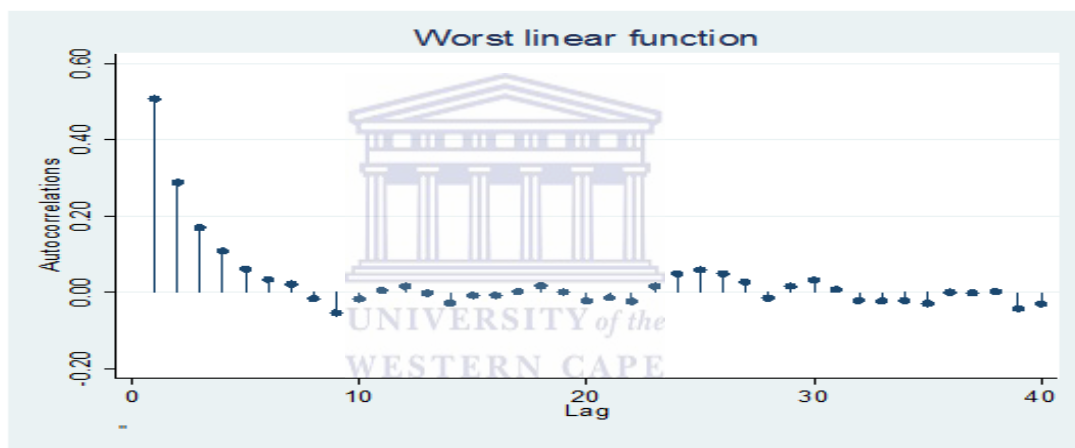


FIGURE 6.28: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

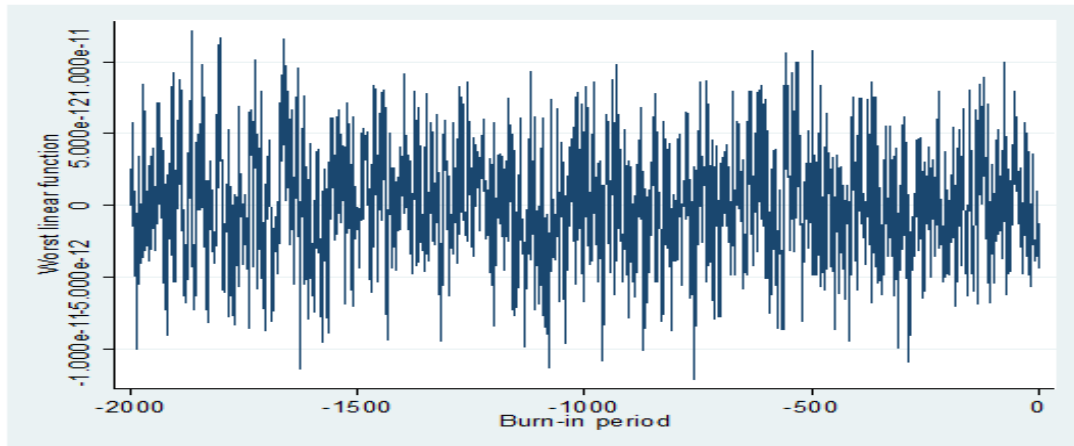


FIGURE 6.29: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

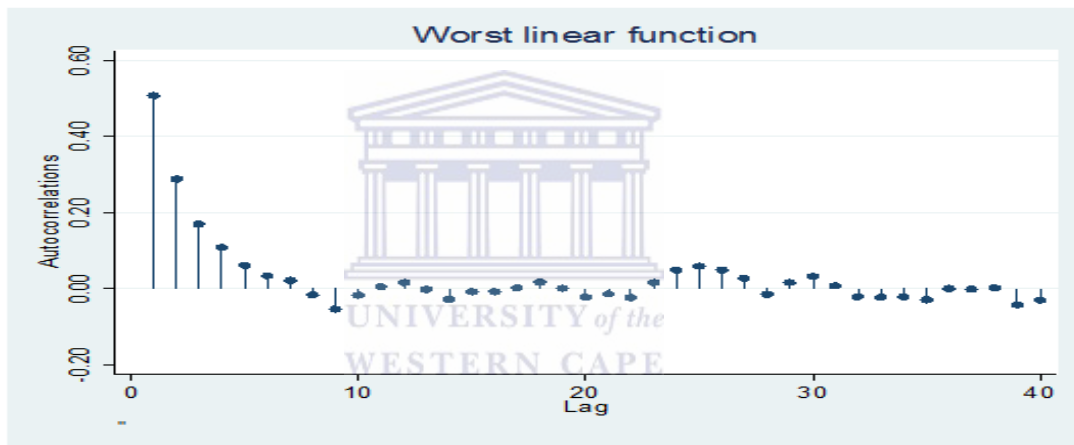


FIGURE 6.30: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

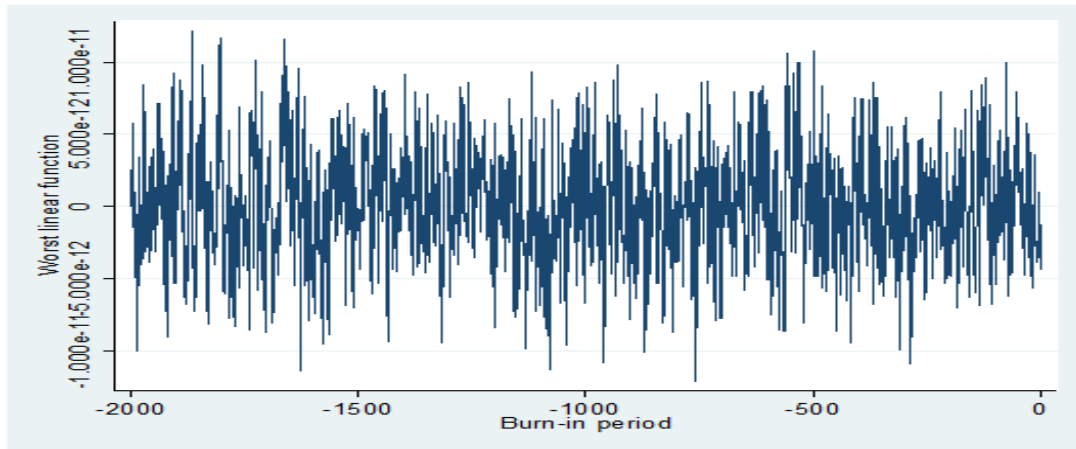


FIGURE 6.31: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

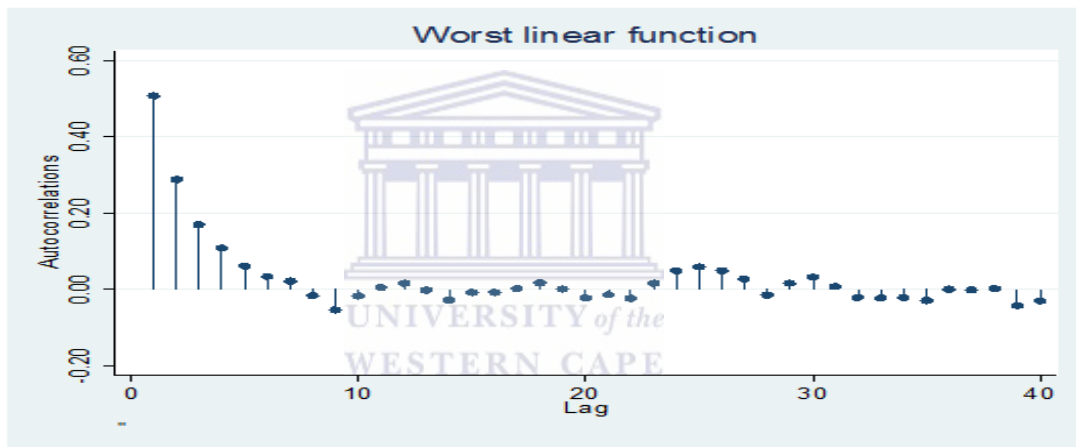


FIGURE 6.32: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

Weighted data sets

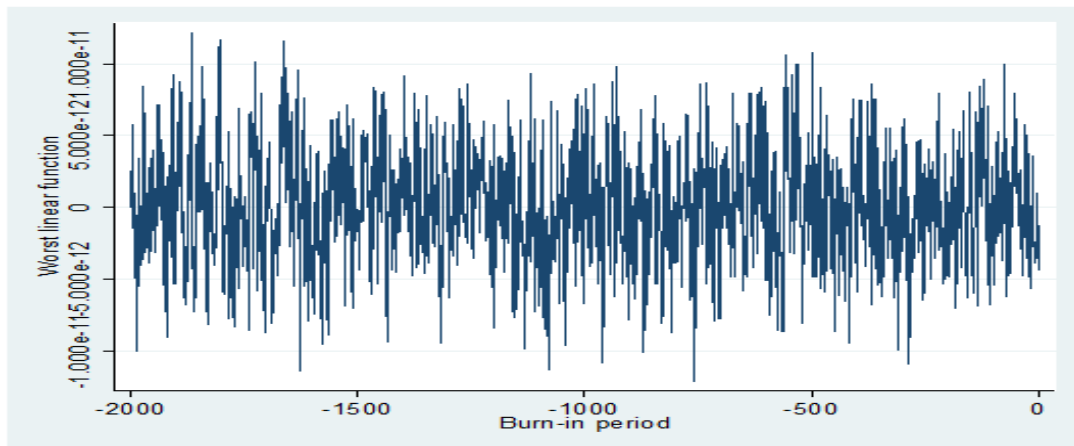


FIGURE 6.33: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

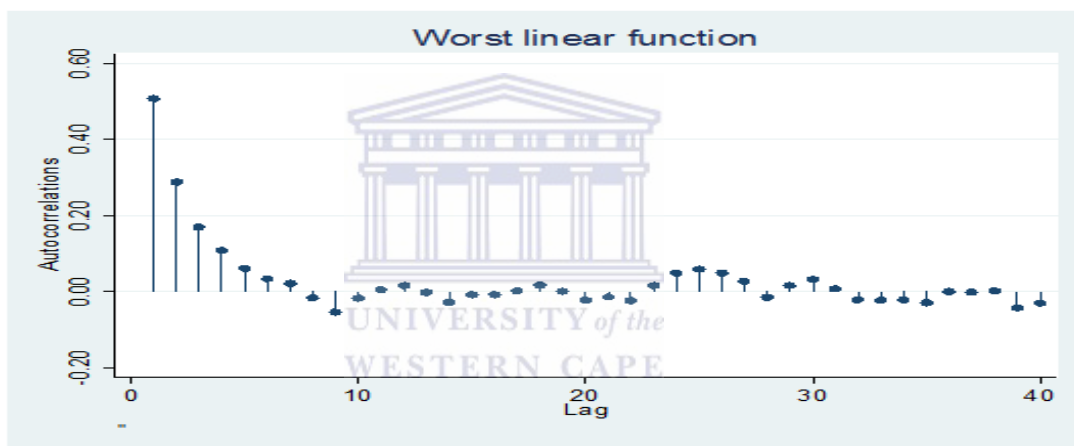


FIGURE 6.34: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

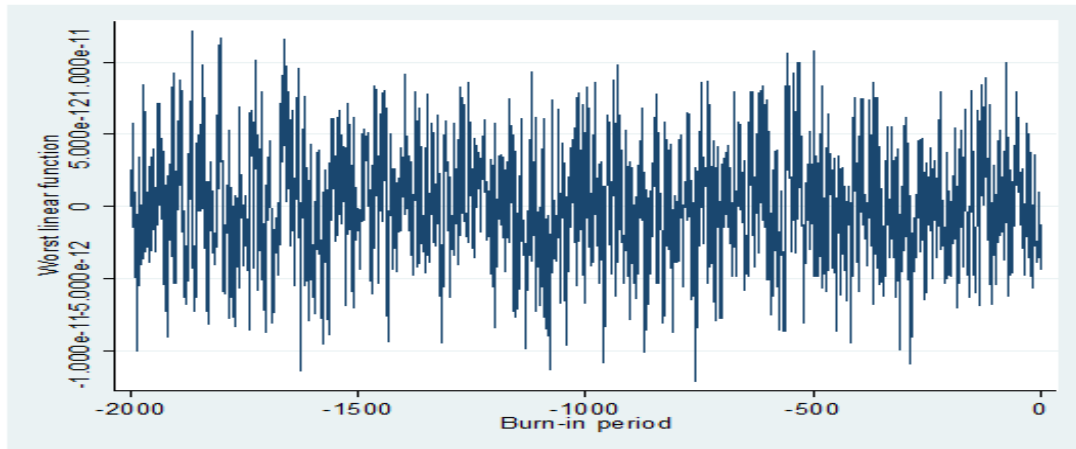


FIGURE 6.35: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

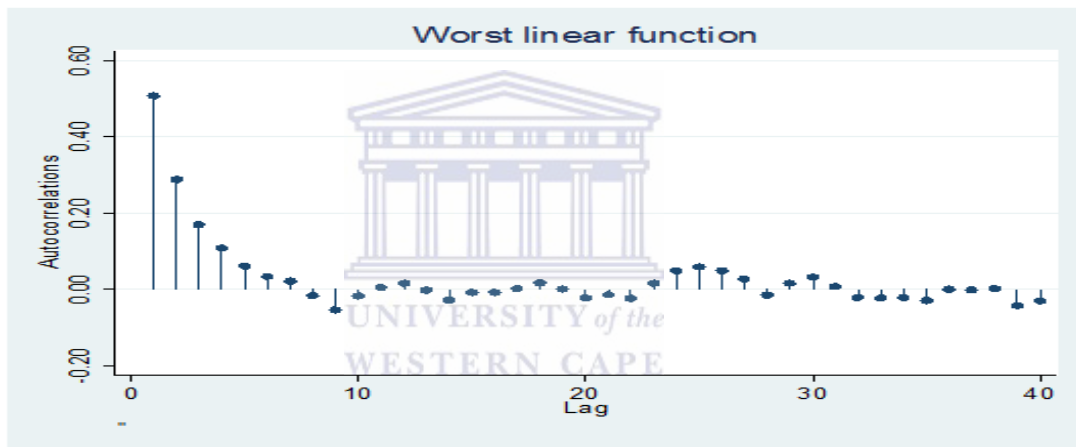


FIGURE 6.36: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

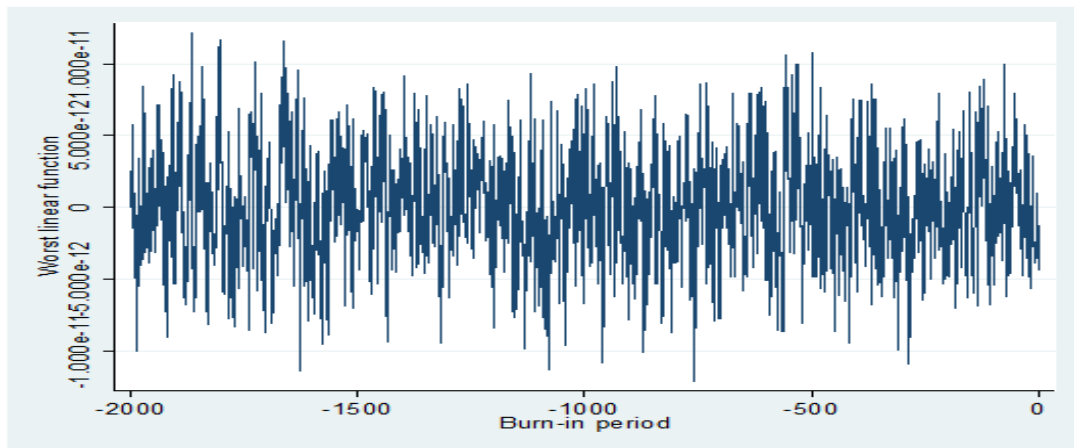


FIGURE 6.37: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

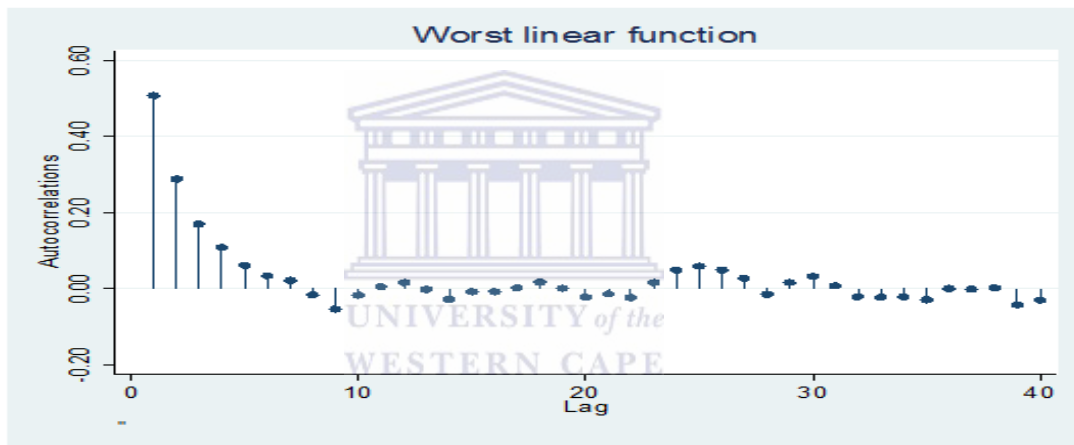


FIGURE 6.38: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

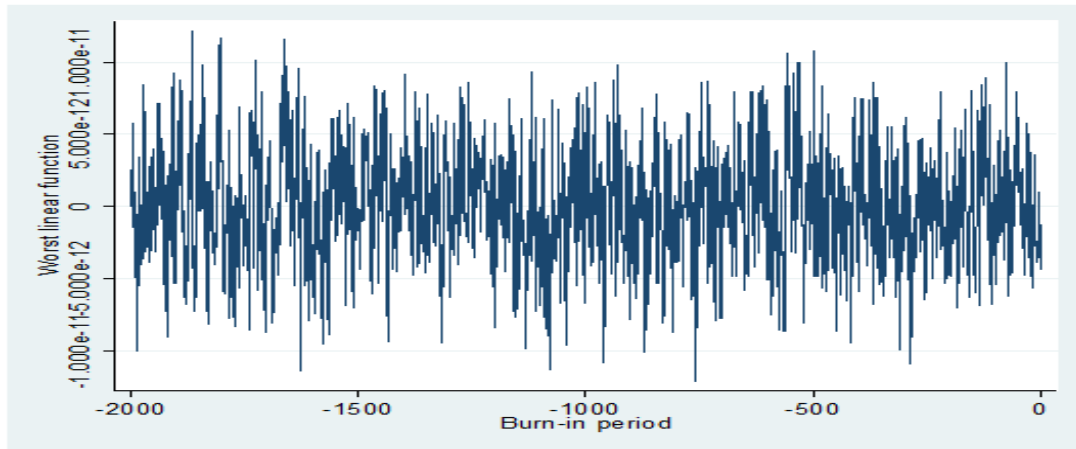


FIGURE 6.39: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

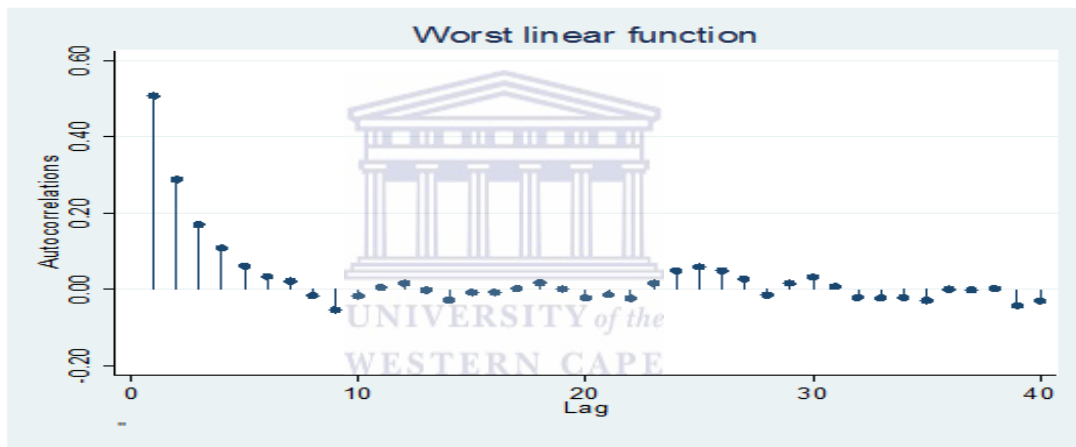


FIGURE 6.40: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

Imputation models's diagnostics when 30% of data are missing at random or completely at random on the covariate

Unweighted data sets

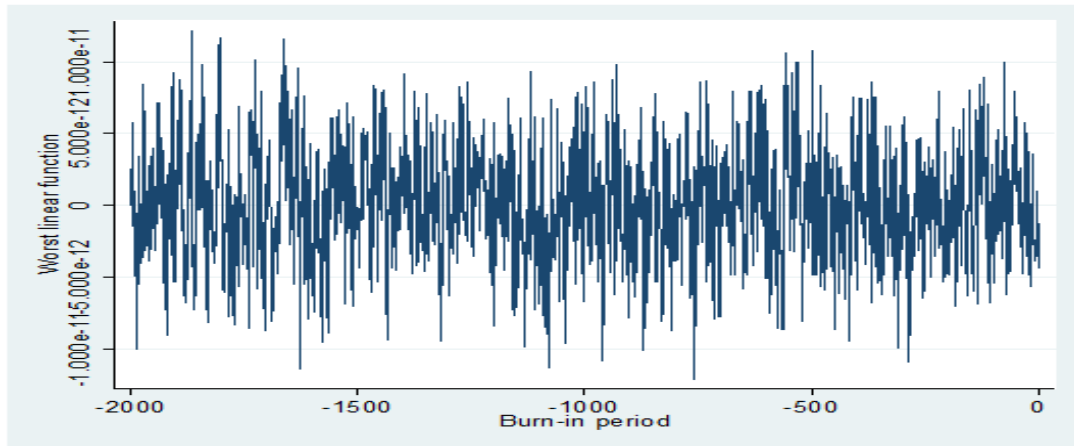


FIGURE 6.41: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

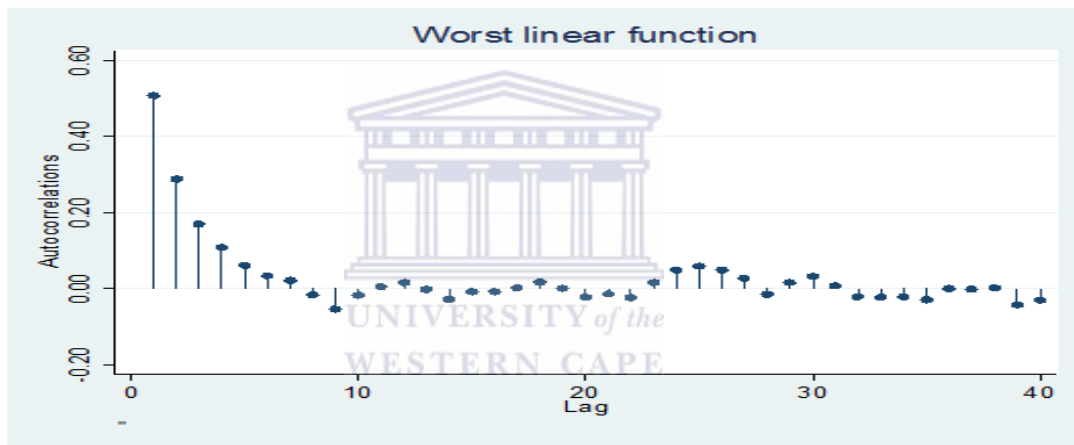


FIGURE 6.42: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set

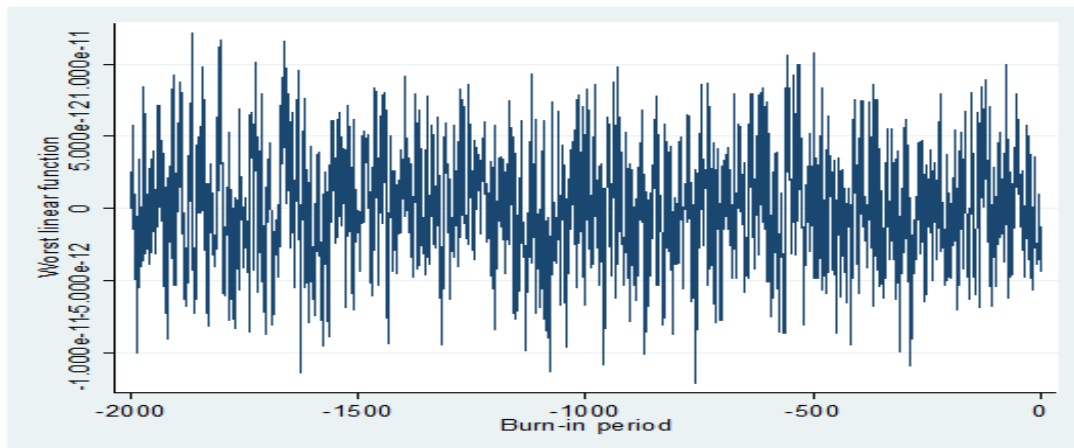


FIGURE 6.43: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set

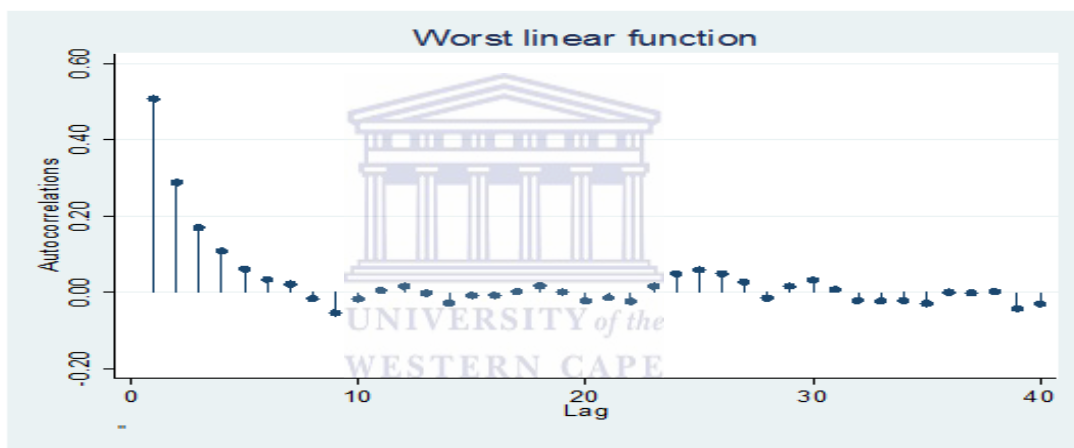


FIGURE 6.44: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set

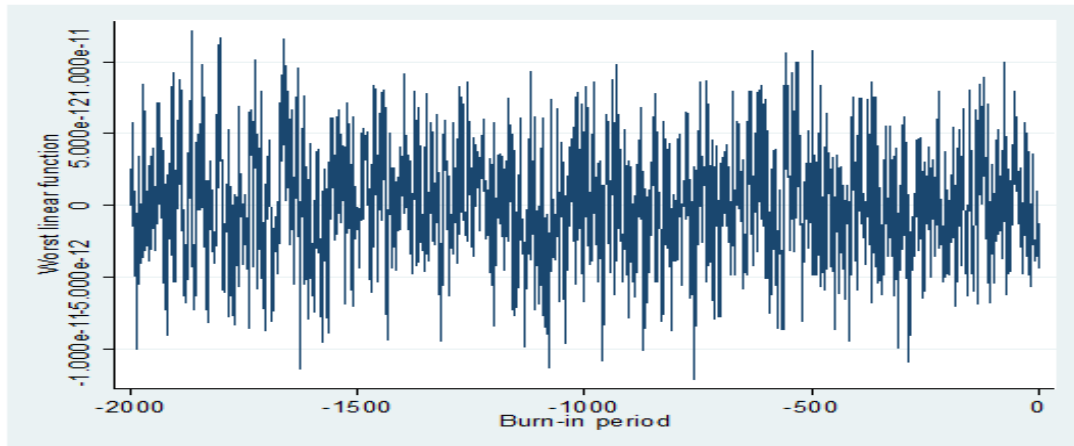


FIGURE 6.45: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

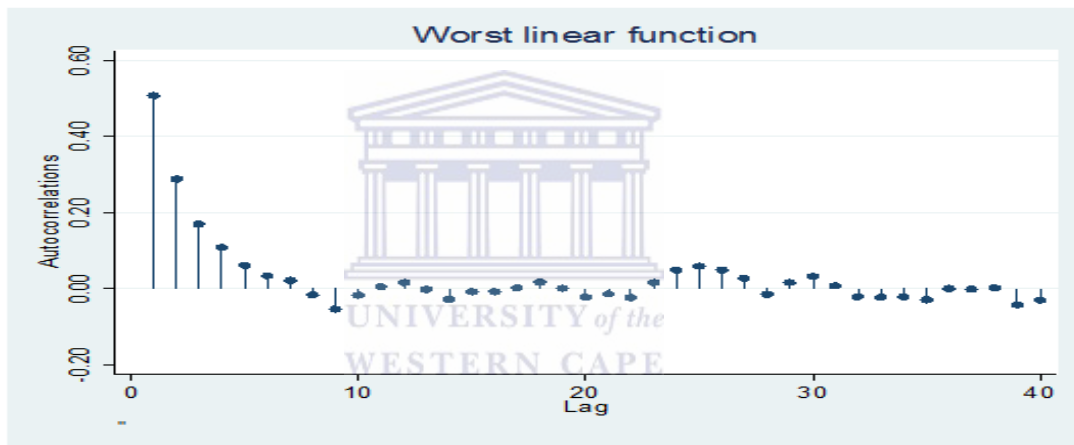


FIGURE 6.46: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set

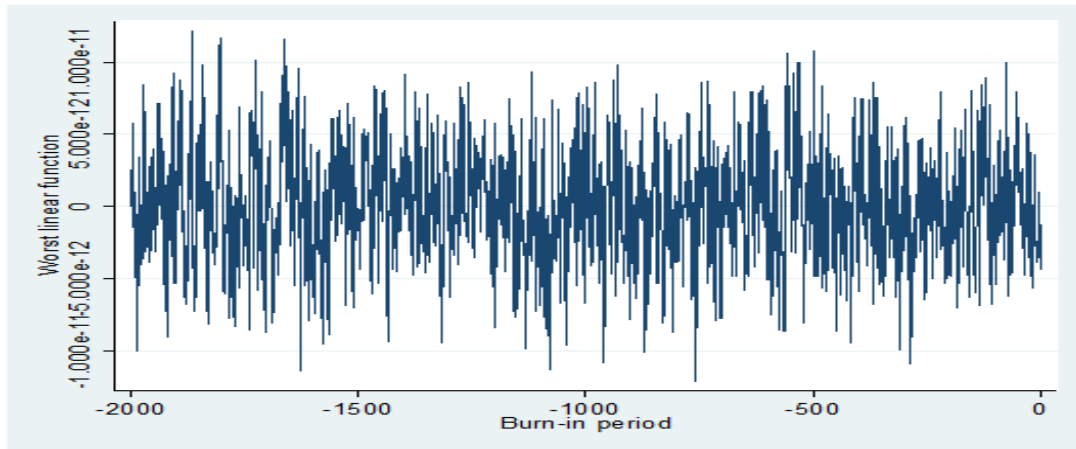


FIGURE 6.47: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set

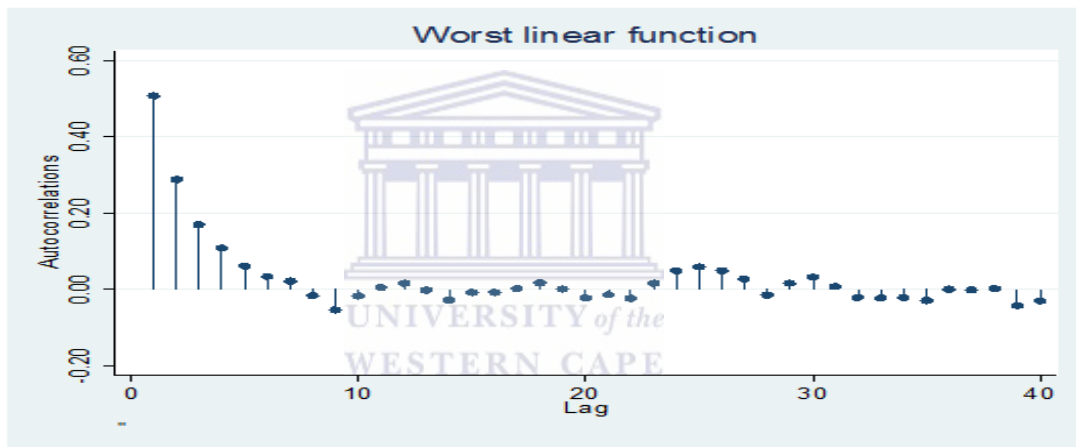


FIGURE 6.48: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set

Weighted data sets

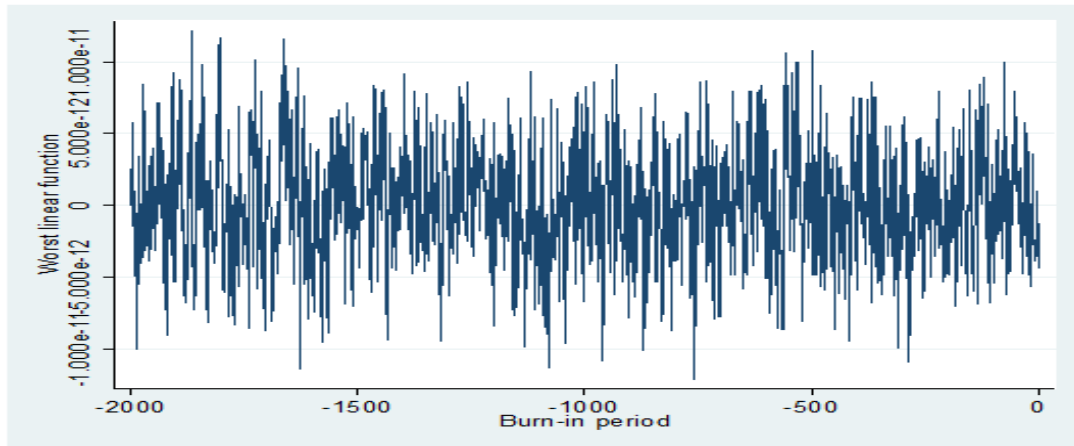


FIGURE 6.49: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

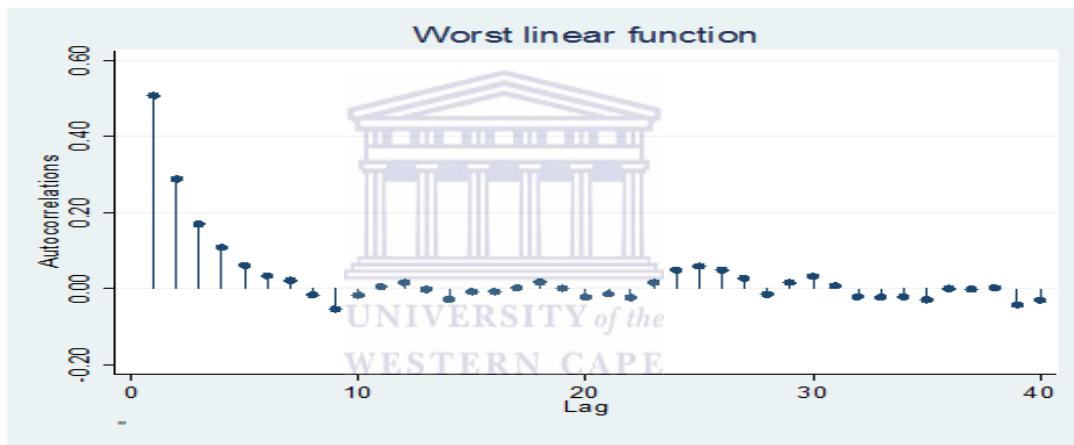


FIGURE 6.50: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

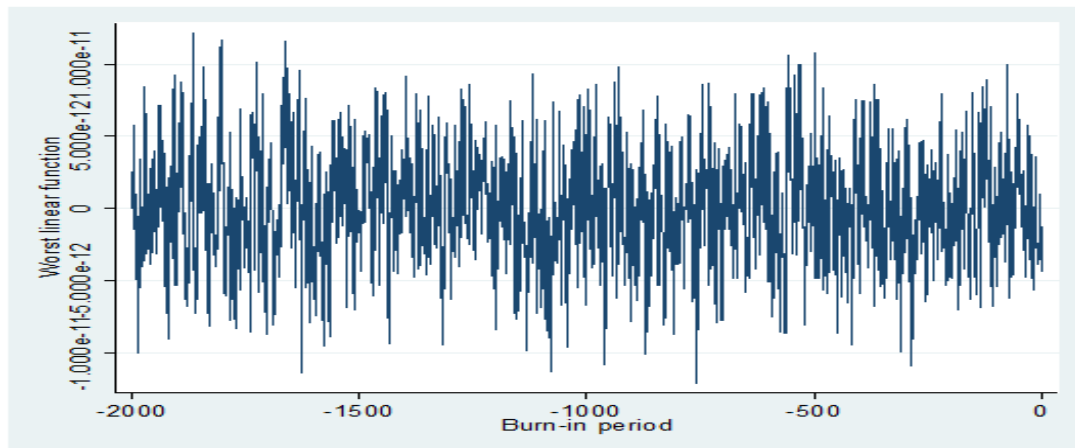


FIGURE 6.51: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

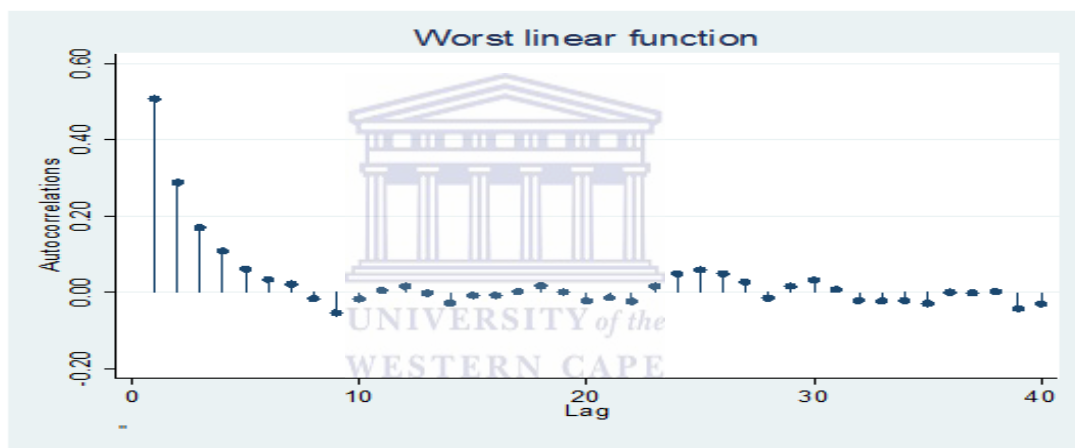


FIGURE 6.52: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set

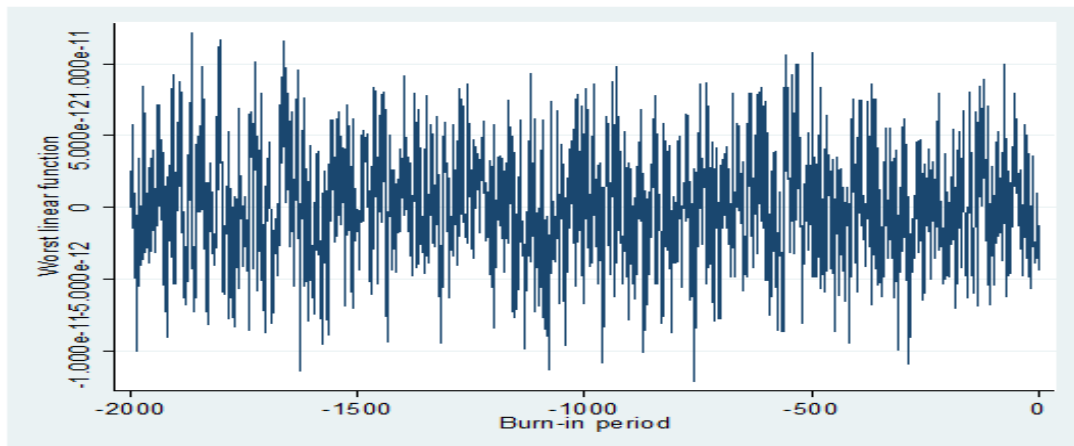


FIGURE 6.53: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

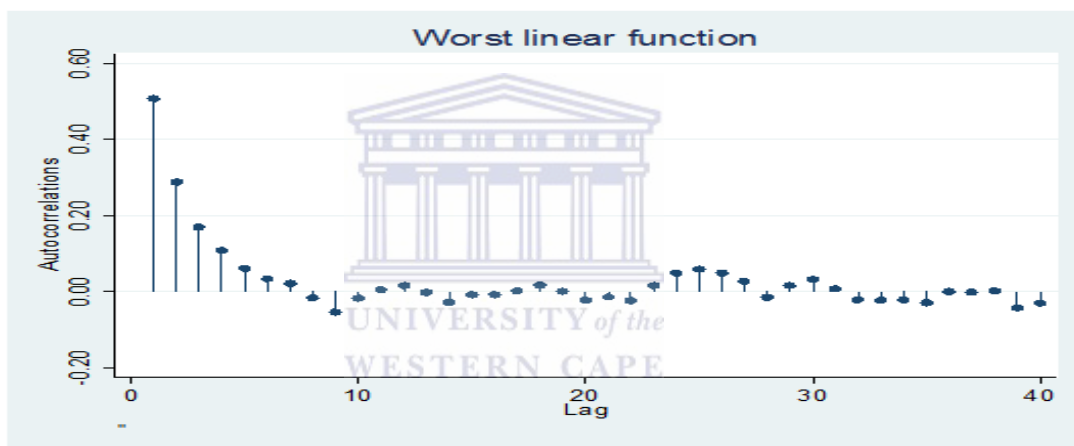


FIGURE 6.54: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

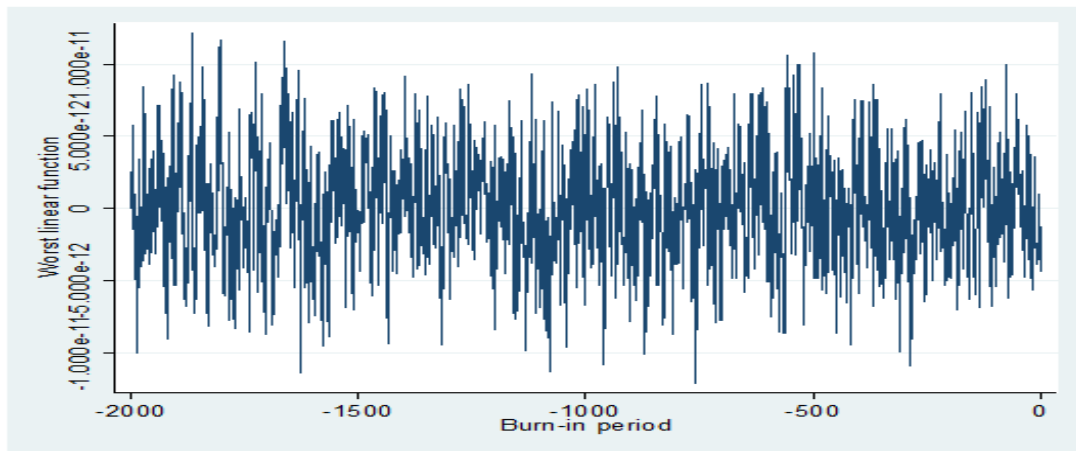


FIGURE 6.55: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

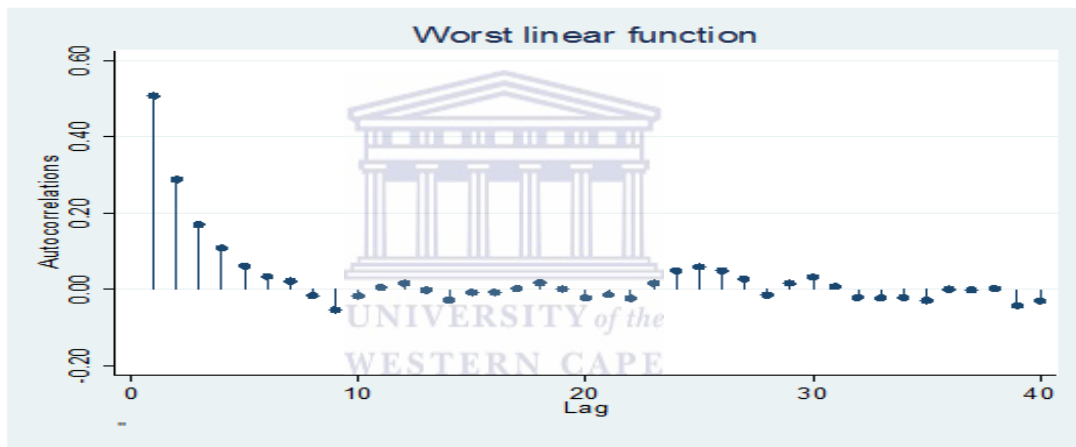


FIGURE 6.56: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set

Imputation models's diagnostics when 10% of data are missing at random or completely at random on the covariate

Unweighted data sets

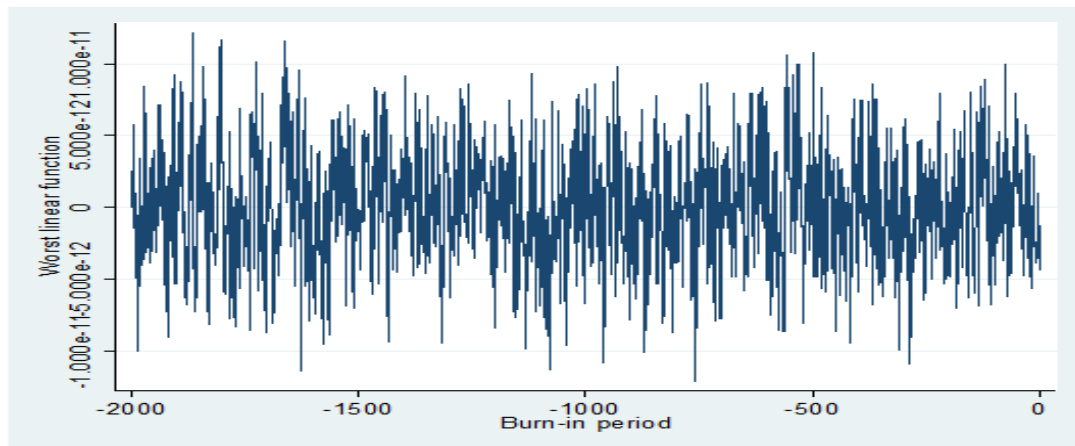


FIGURE 6.57: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

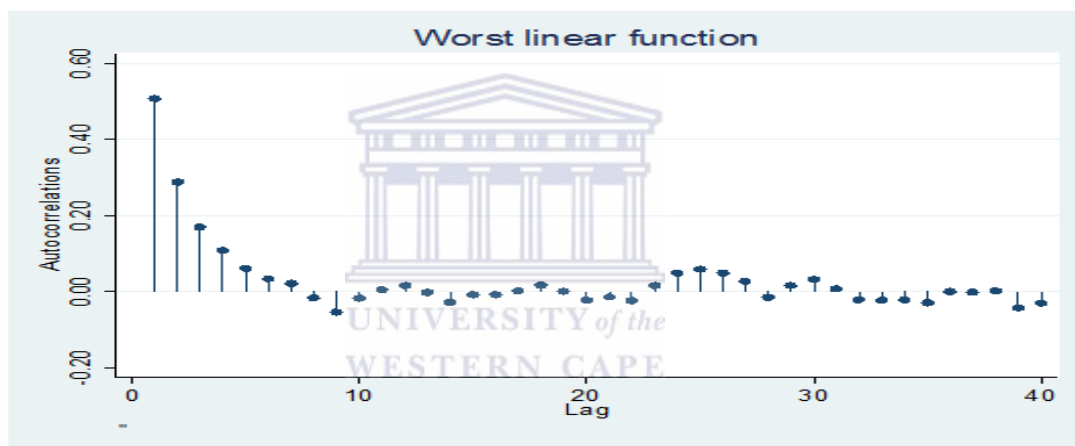


FIGURE 6.58: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

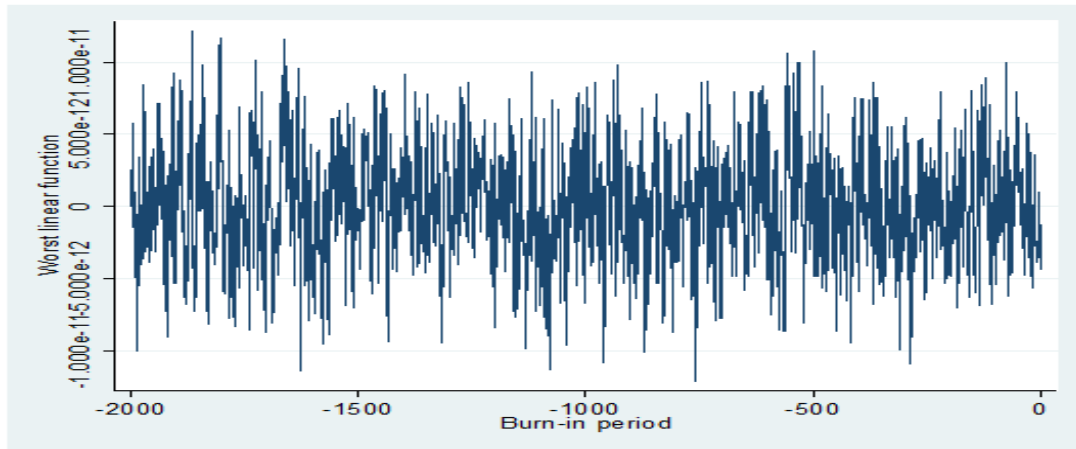


FIGURE 6.59: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

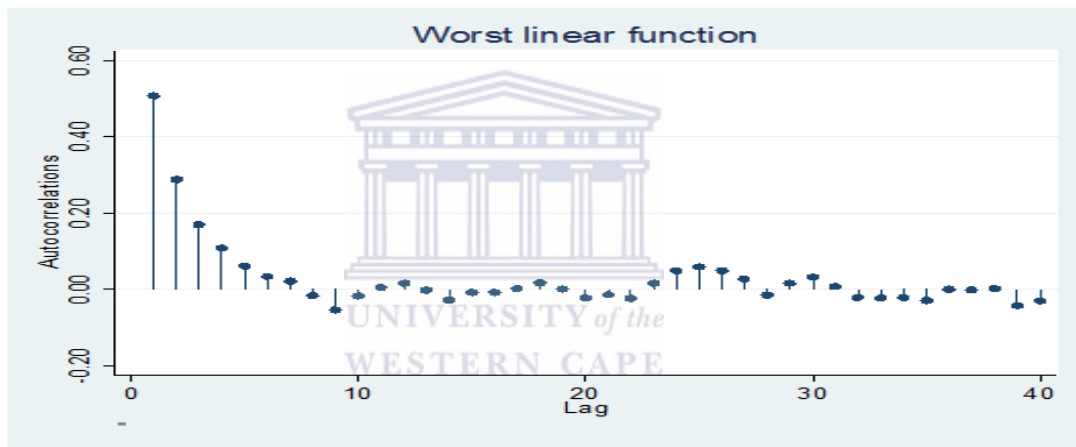


FIGURE 6.60: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

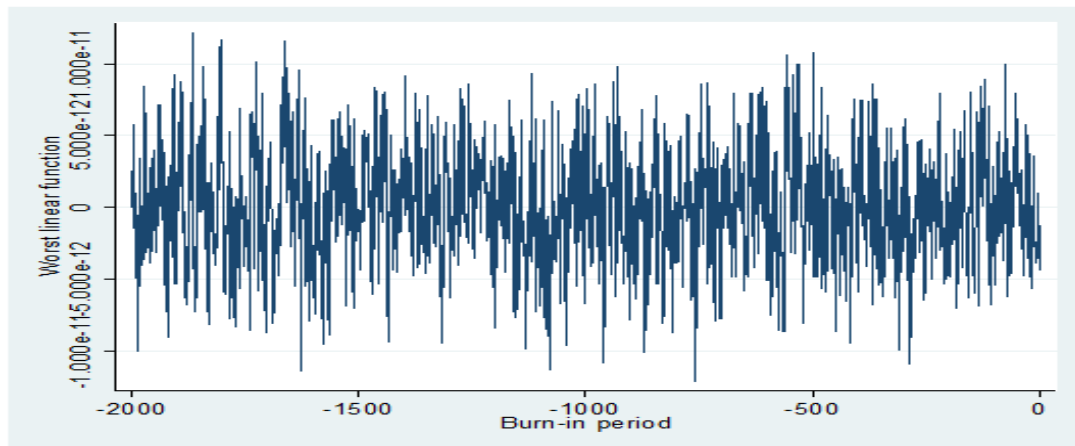


FIGURE 6.61: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

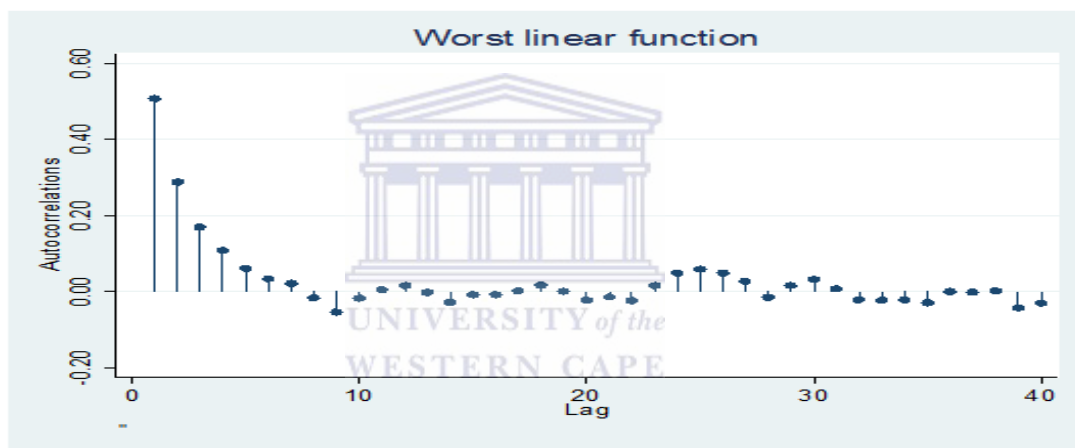


FIGURE 6.62: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

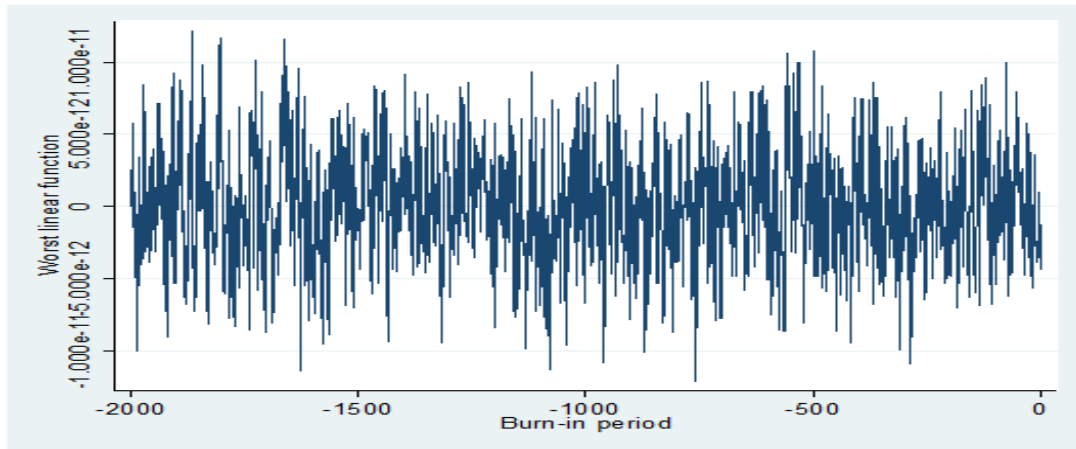


FIGURE 6.63: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

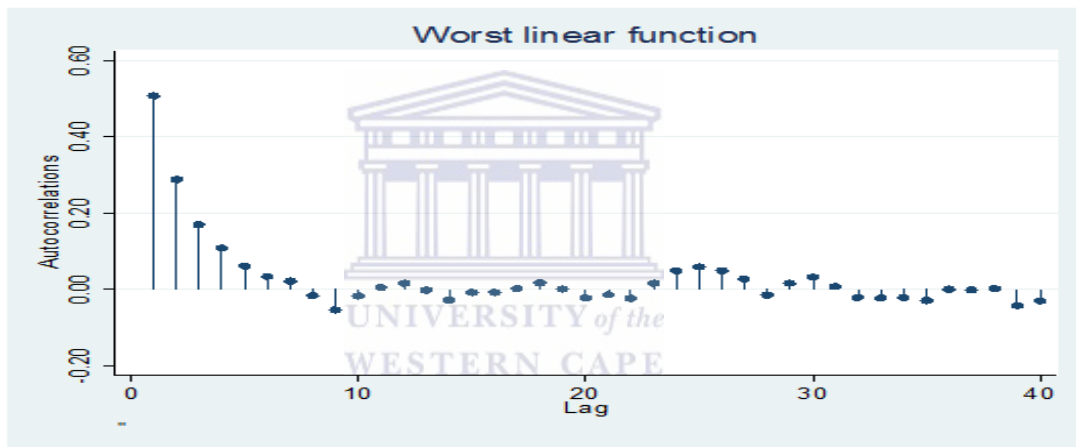


FIGURE 6.64: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

Weighted data sets

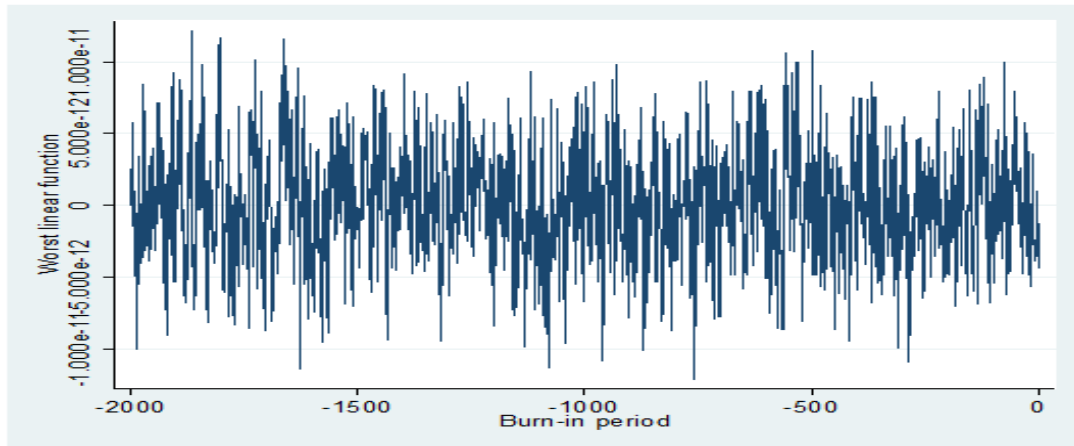


FIGURE 6.65: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

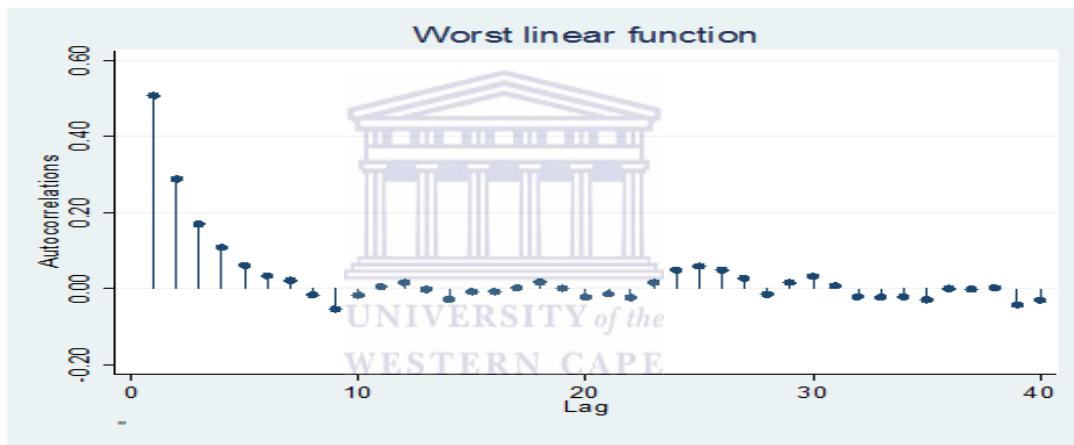


FIGURE 6.66: Model 1.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

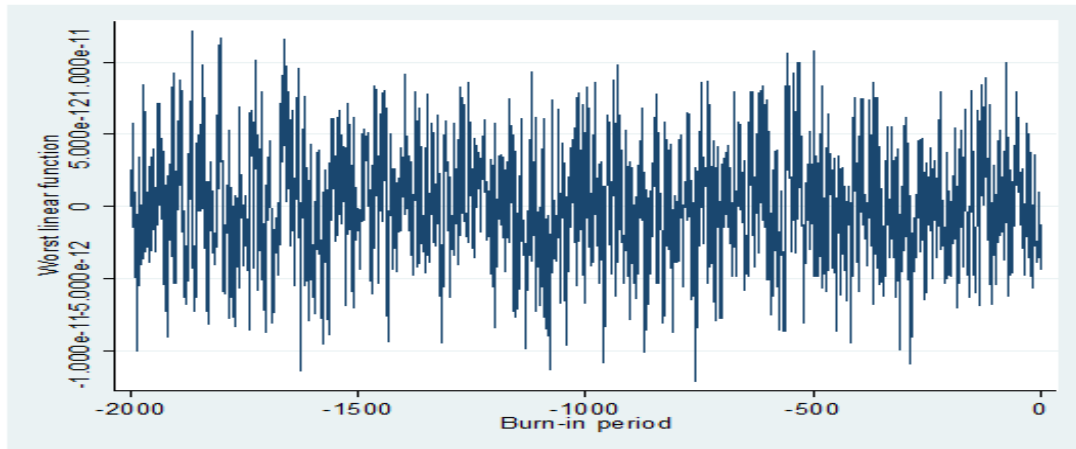


FIGURE 6.67: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

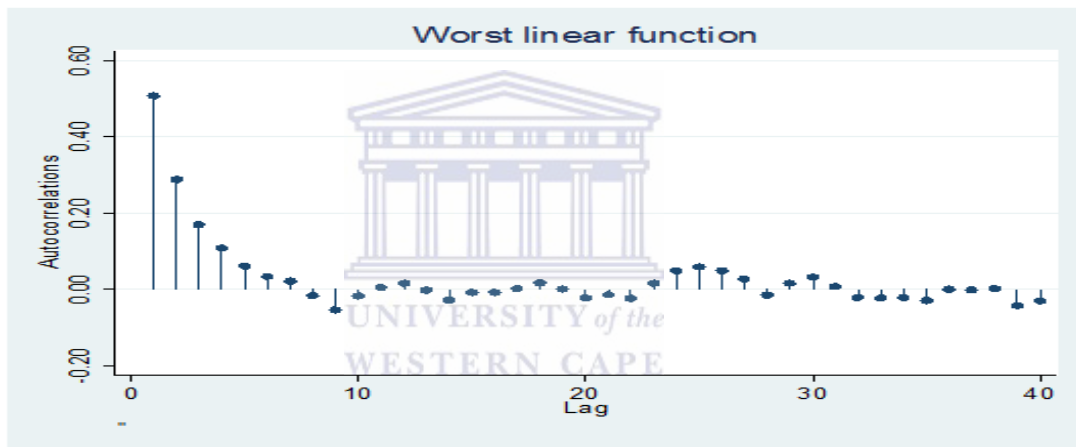


FIGURE 6.68: Model 1.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

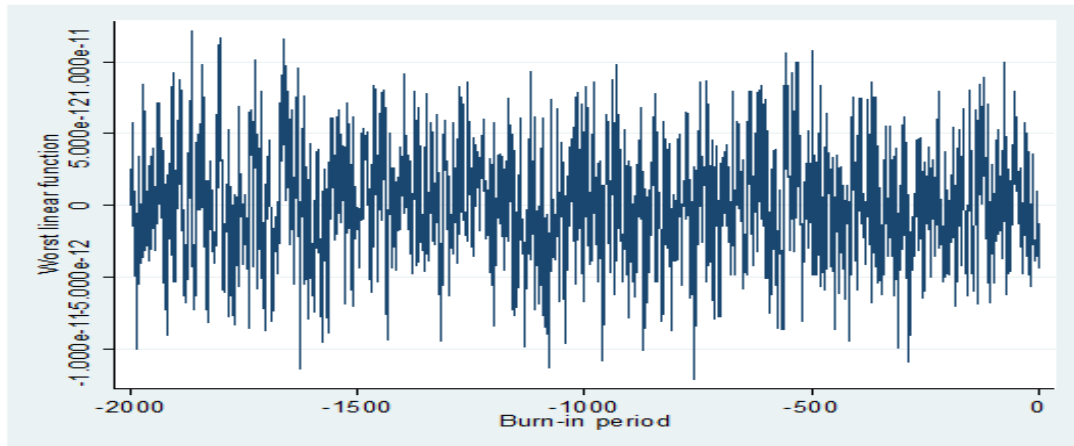


FIGURE 6.69: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

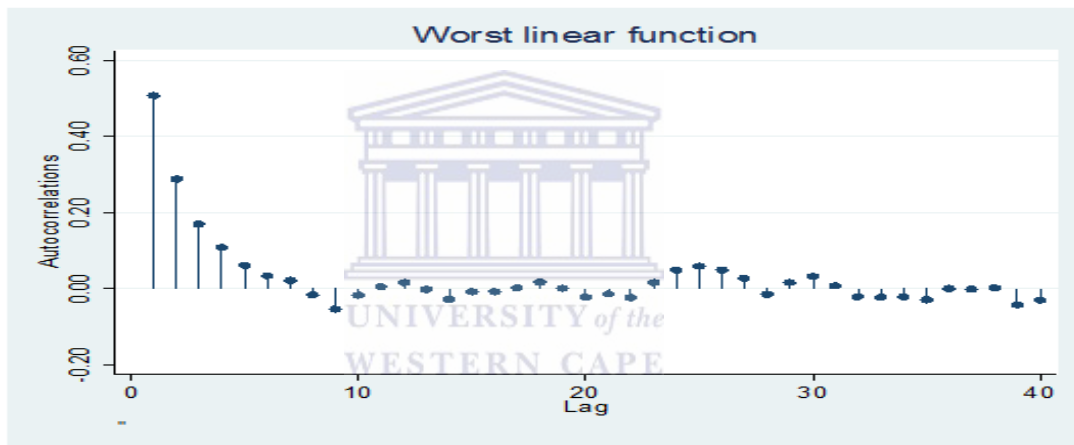


FIGURE 6.70: Model 1.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

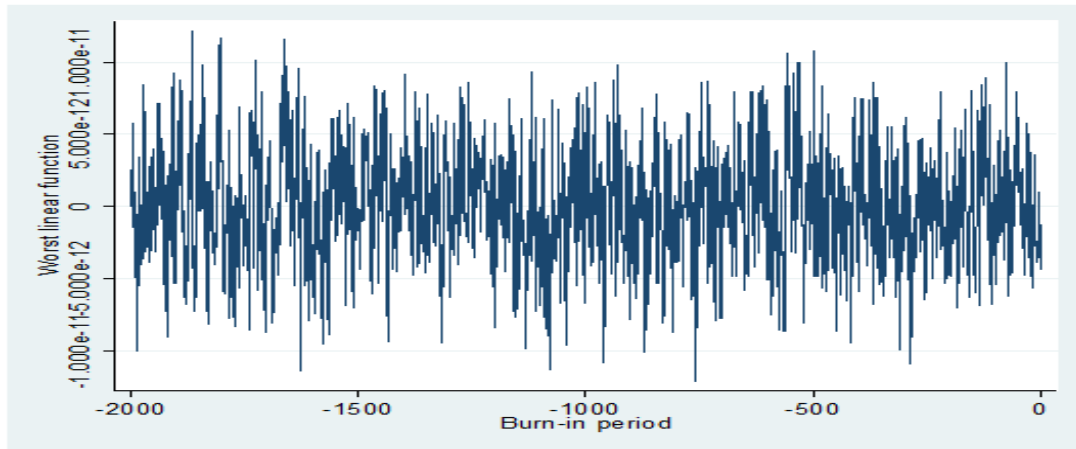


FIGURE 6.71: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

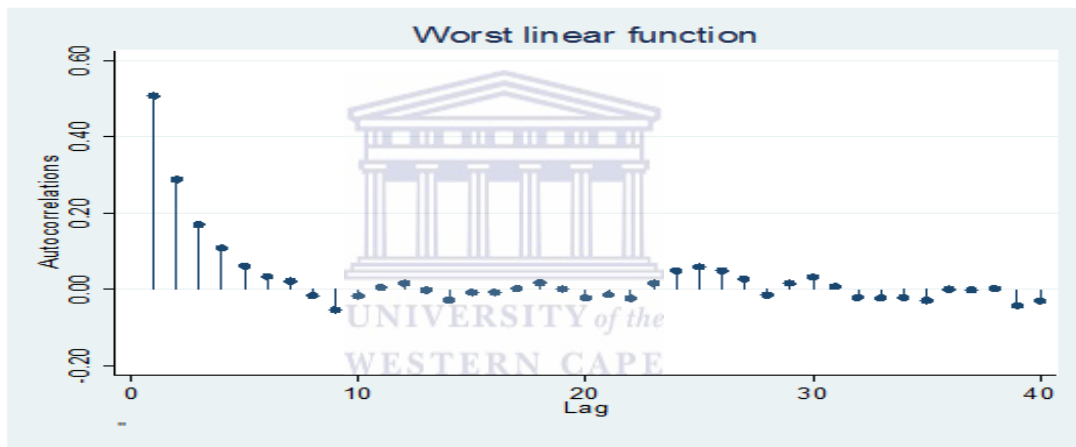


FIGURE 6.72: Model 1.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

Model 1.2: Imputation models diagnostics

Model 1.2.1: Imputation models diagnostics

Unweighted data sets

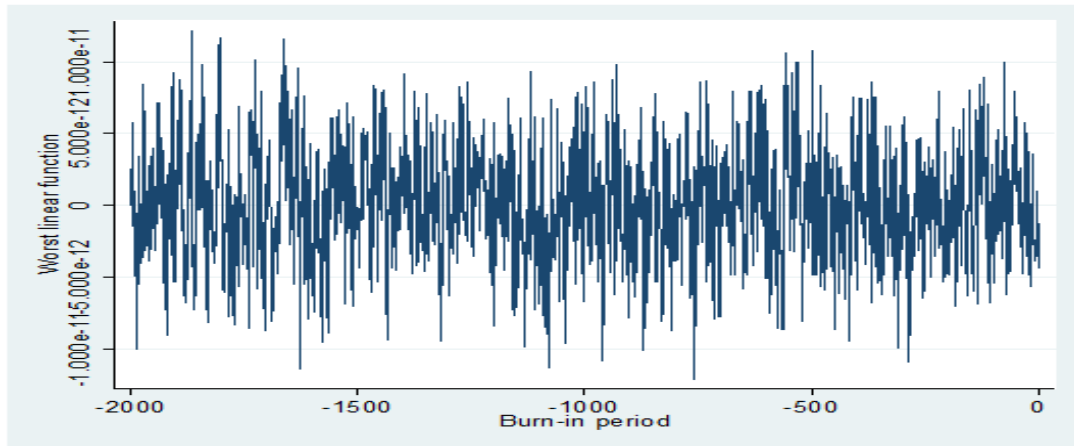


FIGURE 6.73: Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

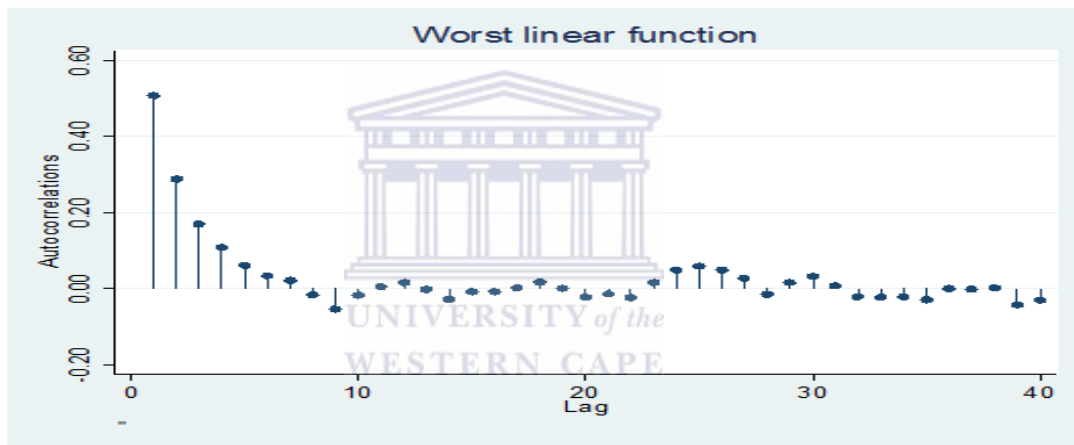


FIGURE 6.74: Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

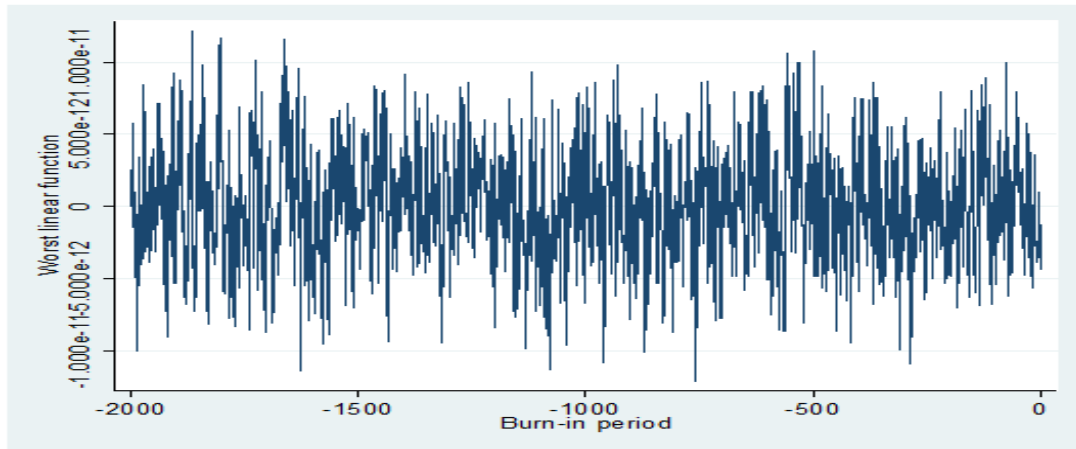


FIGURE 6.75: Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

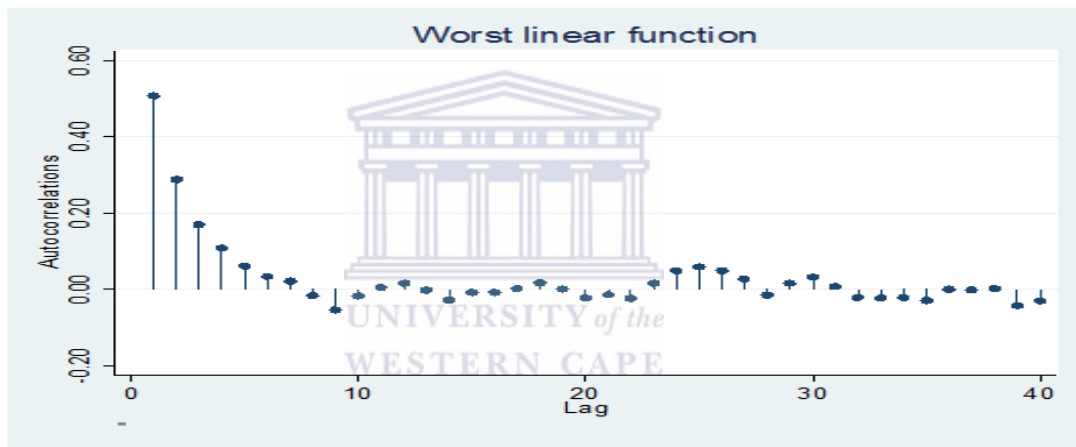


FIGURE 6.76: Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

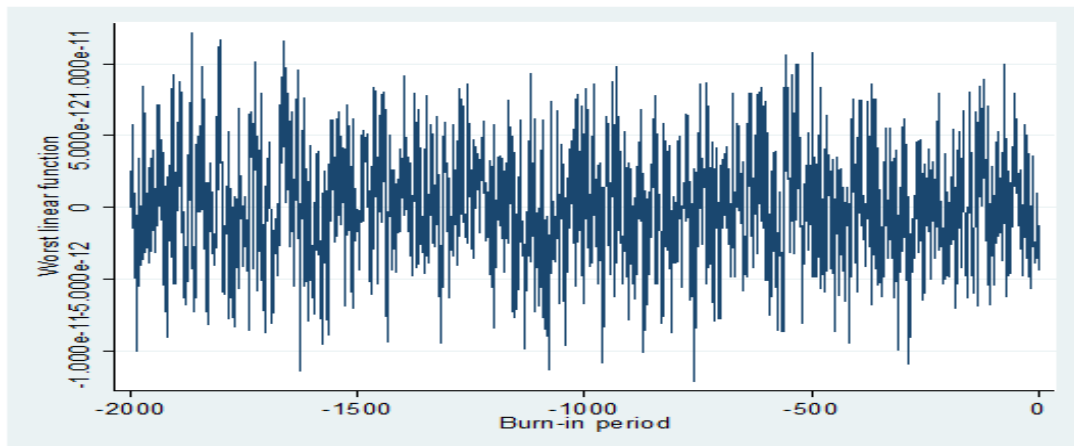


FIGURE 6.77: Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

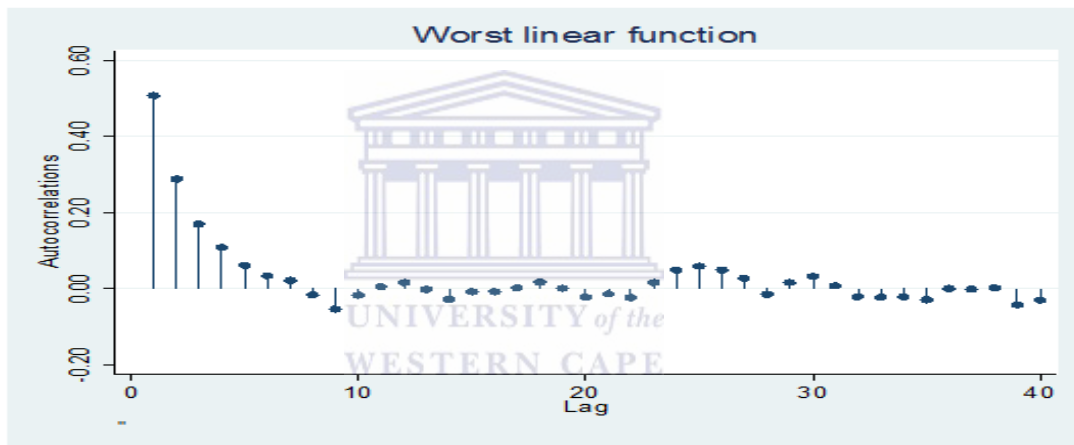


FIGURE 6.78: Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

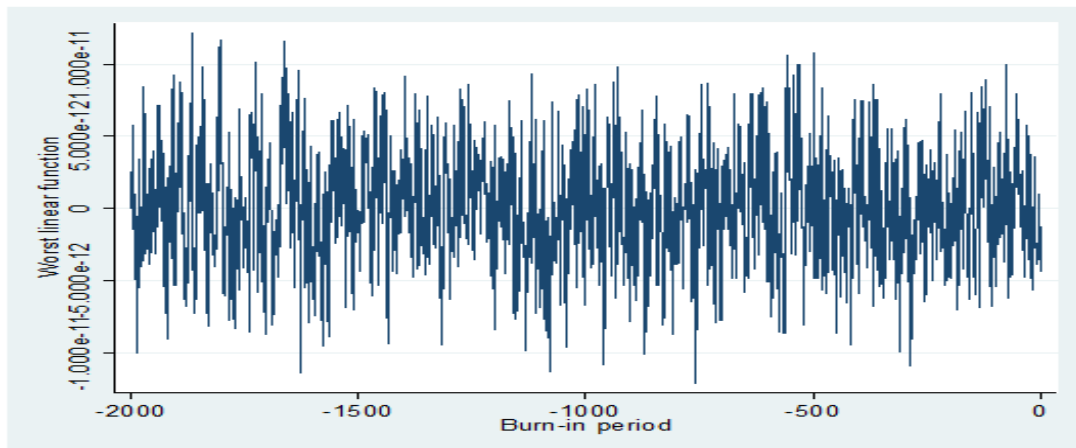


FIGURE 6.79: Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

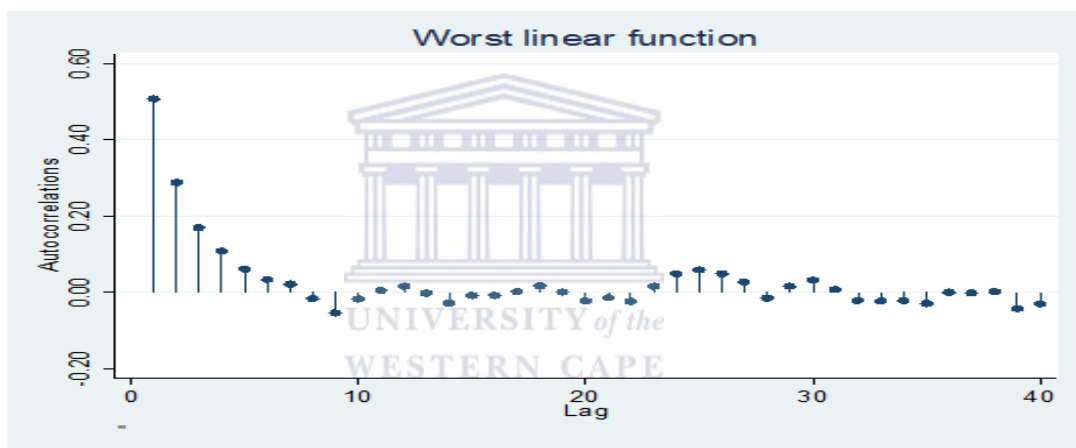


FIGURE 6.80: Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

Weighted data sets

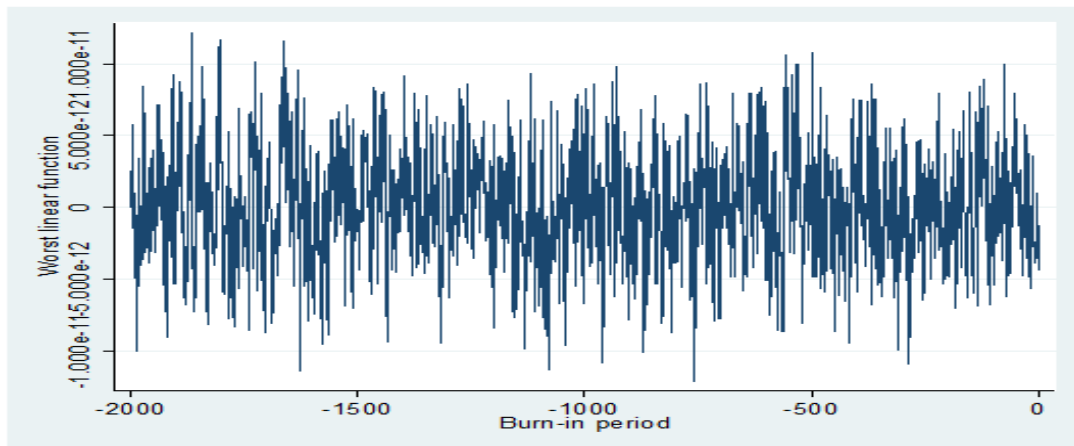


FIGURE 6.81: Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

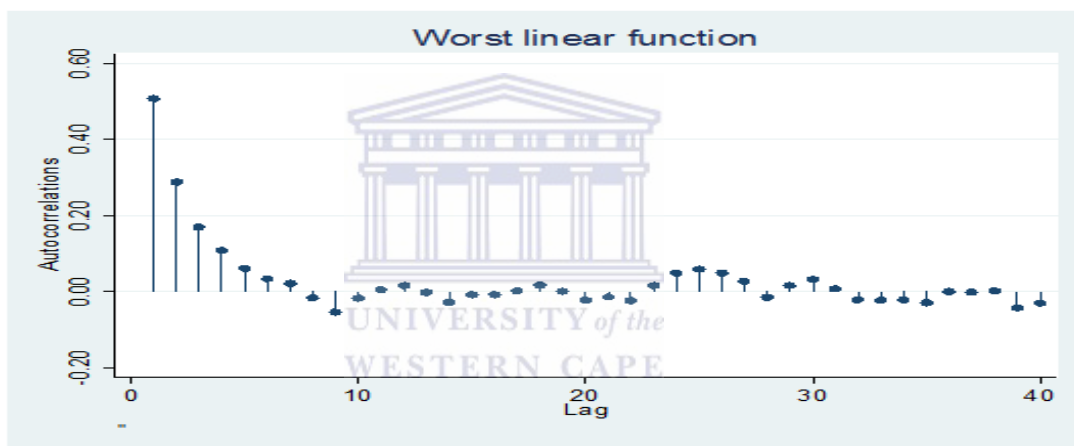


FIGURE 6.82: Model 1.2.1: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

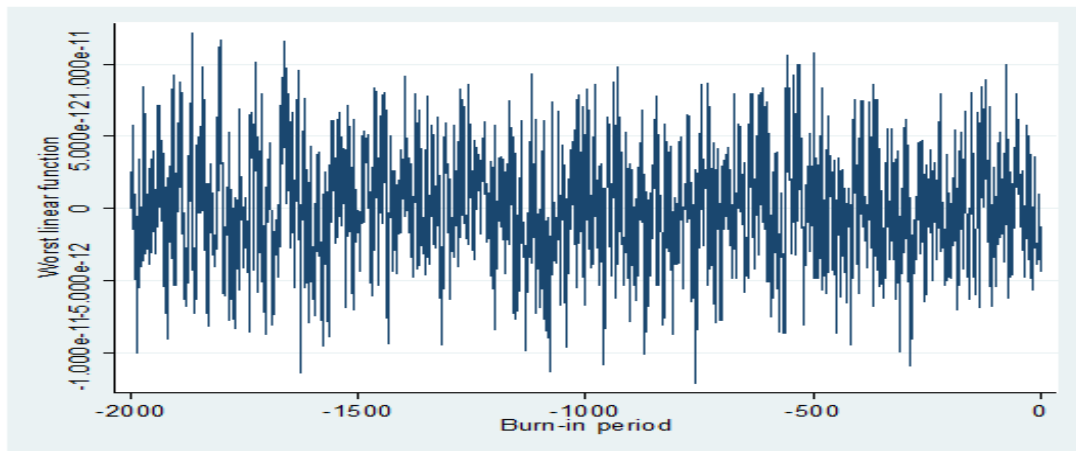


FIGURE 6.83: Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

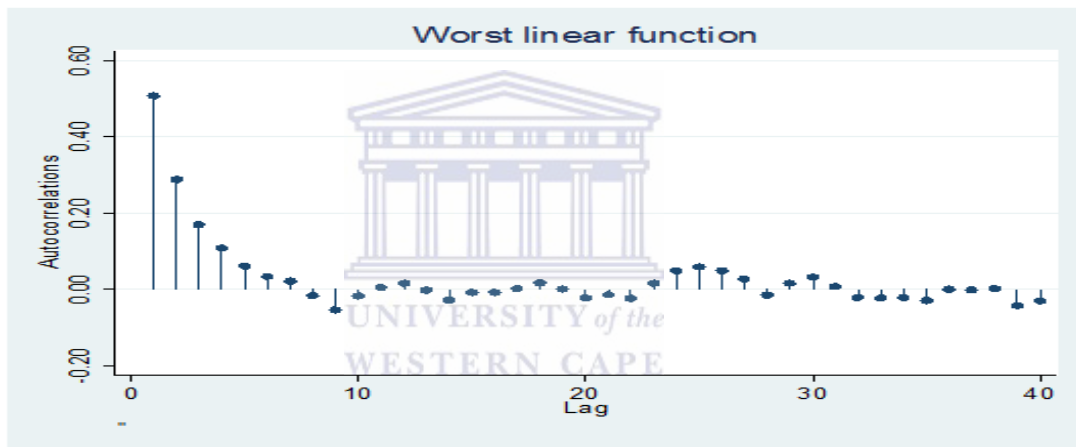


FIGURE 6.84: Model 1.2.1: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

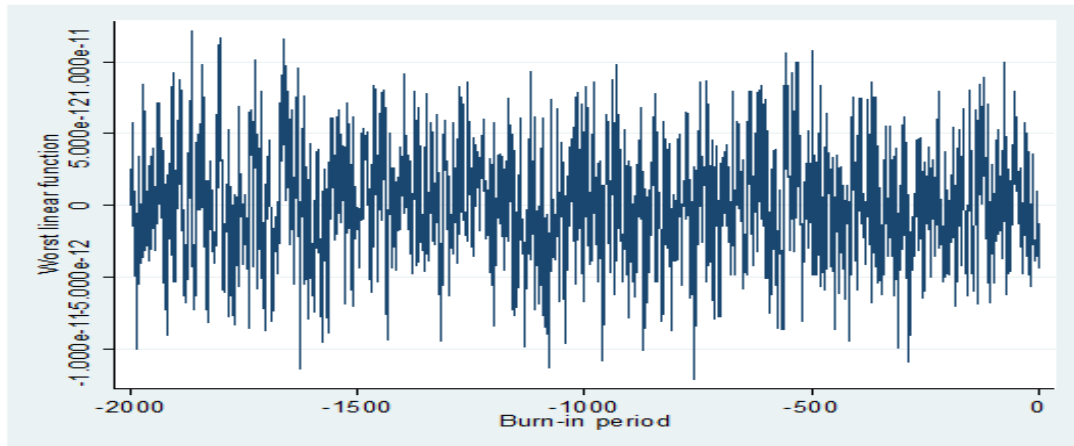


FIGURE 6.85: Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

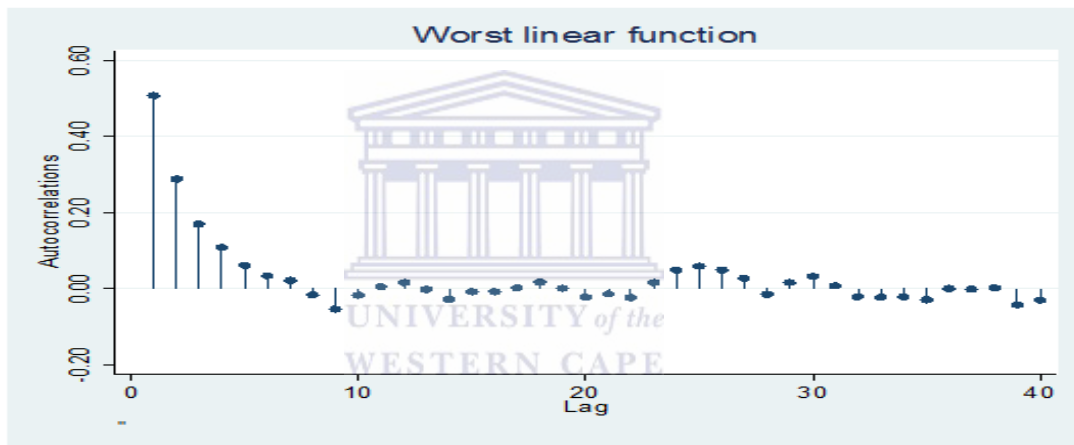


FIGURE 6.86: Model 1.2.1: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

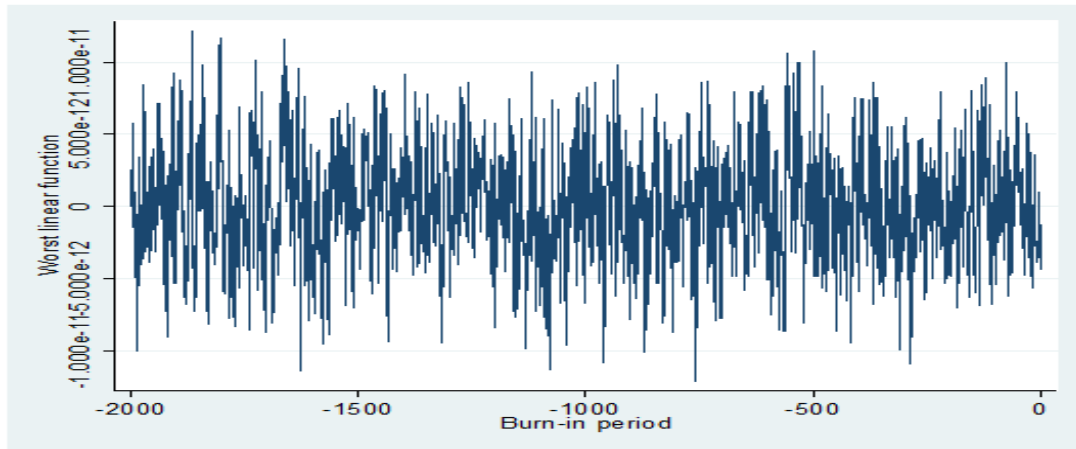


FIGURE 6.87: Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

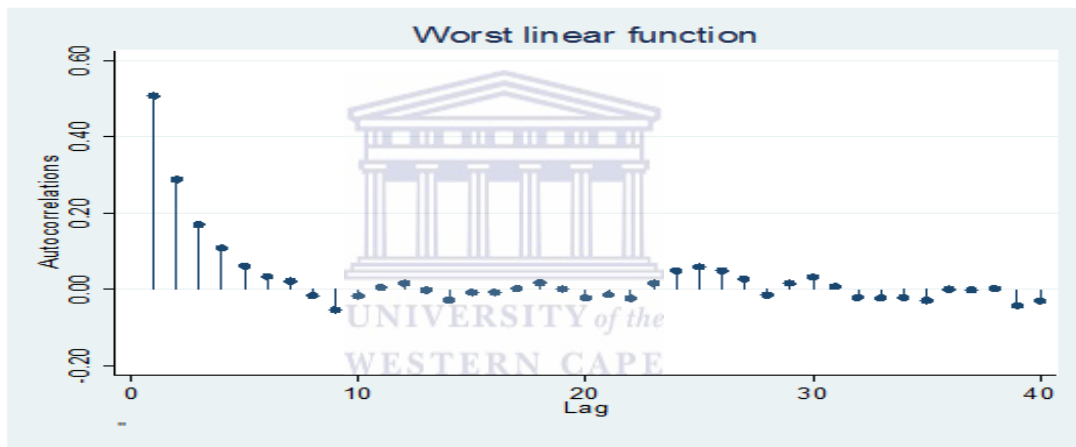


FIGURE 6.88: Model 1.2.1: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

Model 1.2.2: Imputation models diagnostics

Unweighted data sets

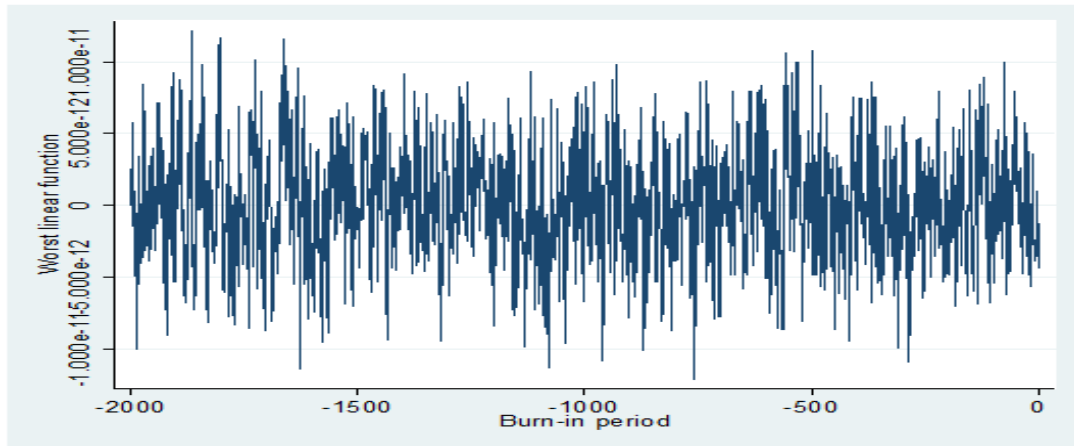


FIGURE 6.89: Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

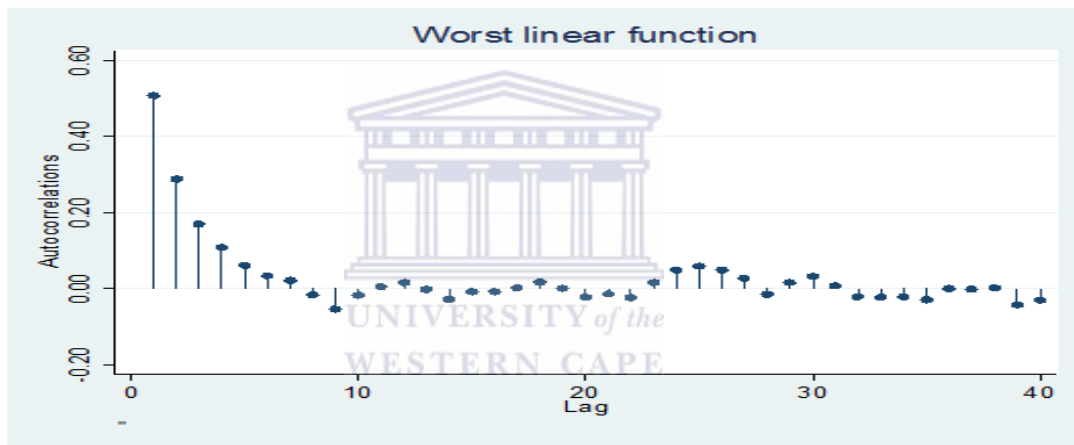


FIGURE 6.90: Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

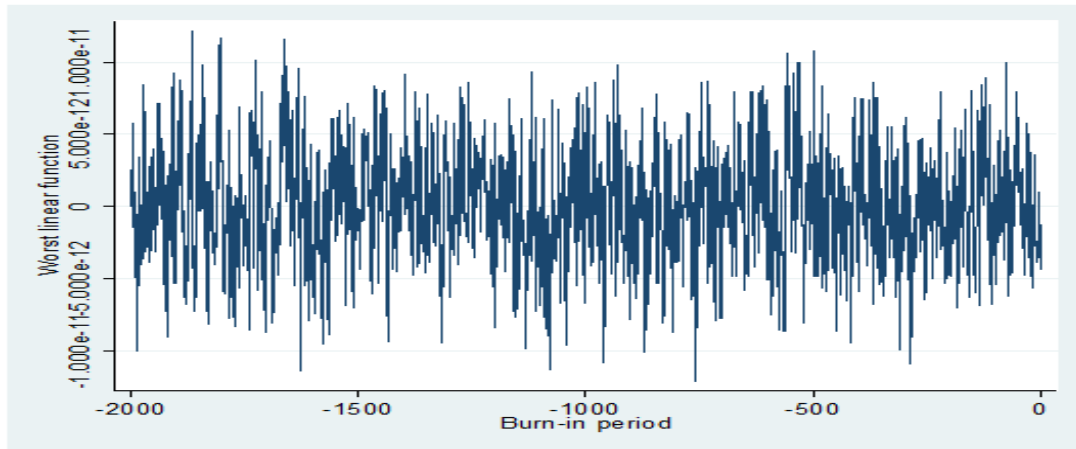


FIGURE 6.91: Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

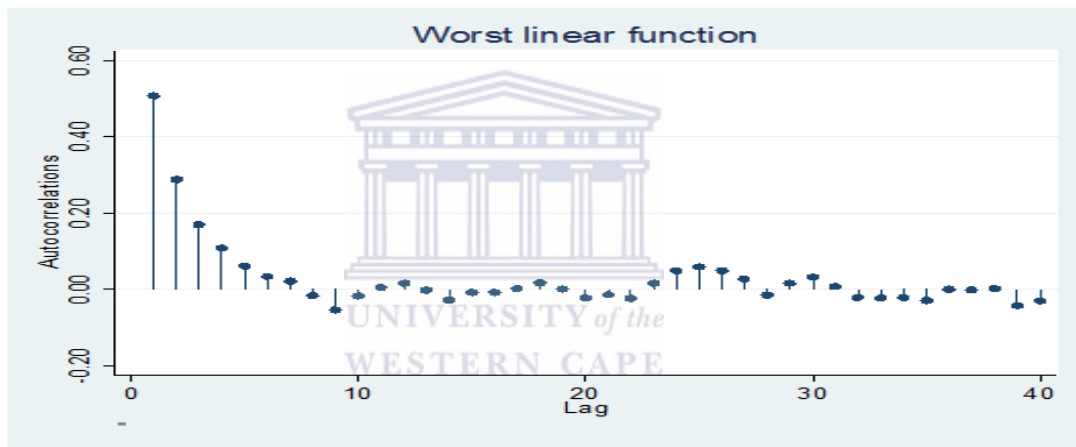


FIGURE 6.92: Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

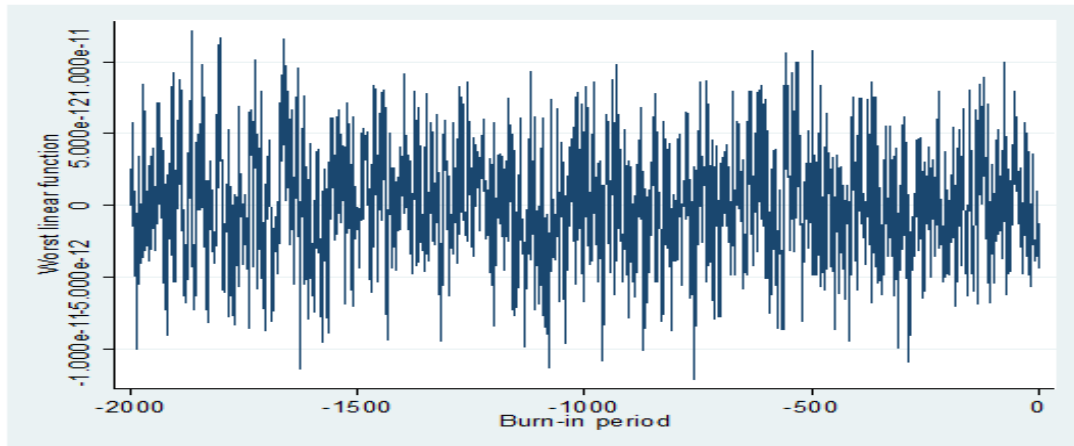


FIGURE 6.93: Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

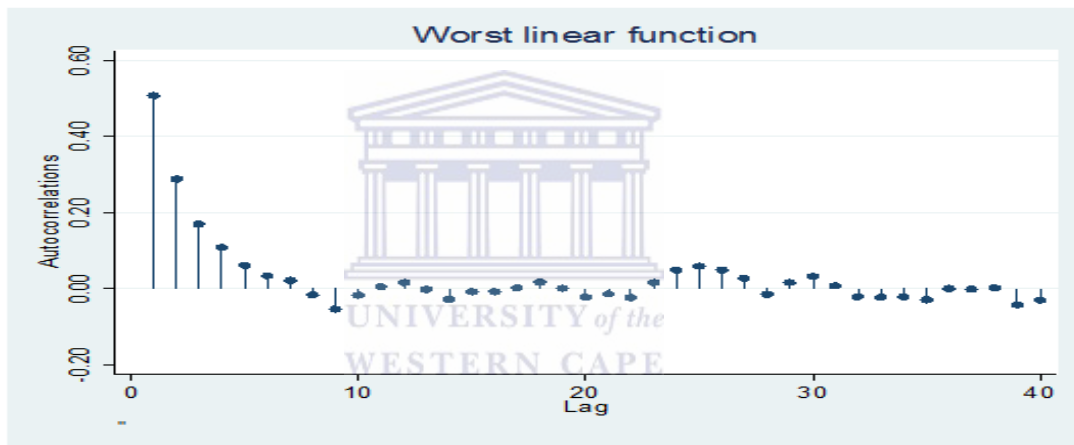


FIGURE 6.94: Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

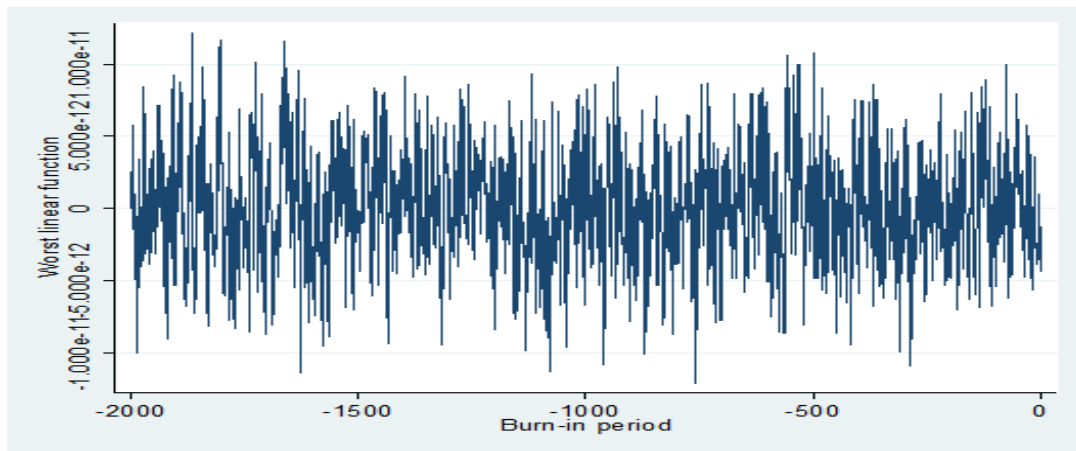


FIGURE 6.95: Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for unweighted data set.

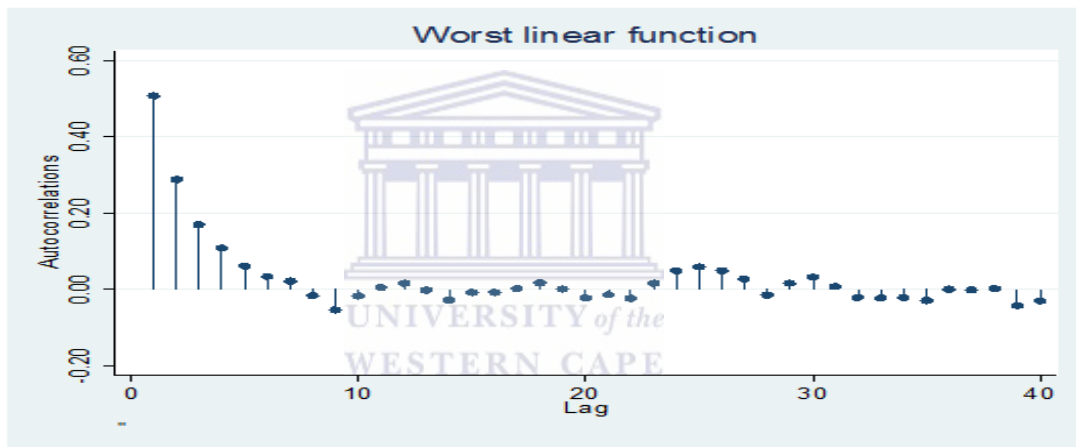


FIGURE 6.96: Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for unweighted data set.

Weighted data sets

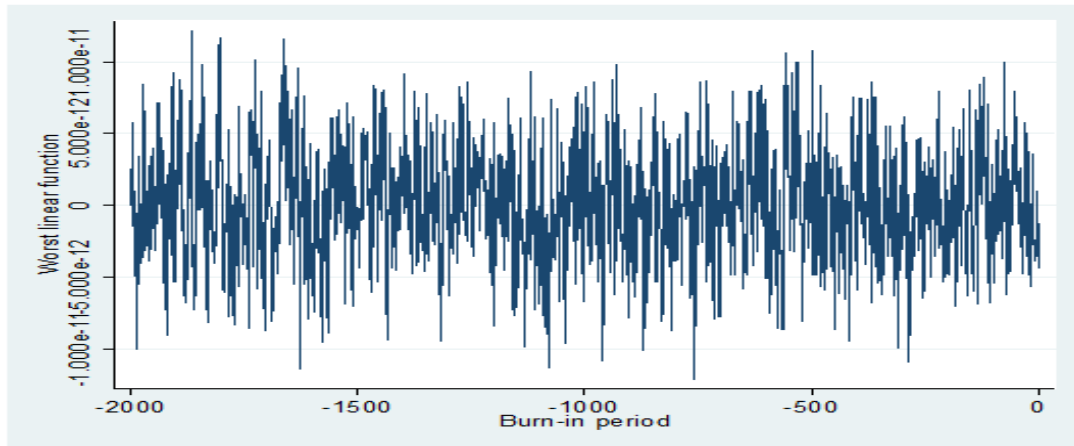


FIGURE 6.97: Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set

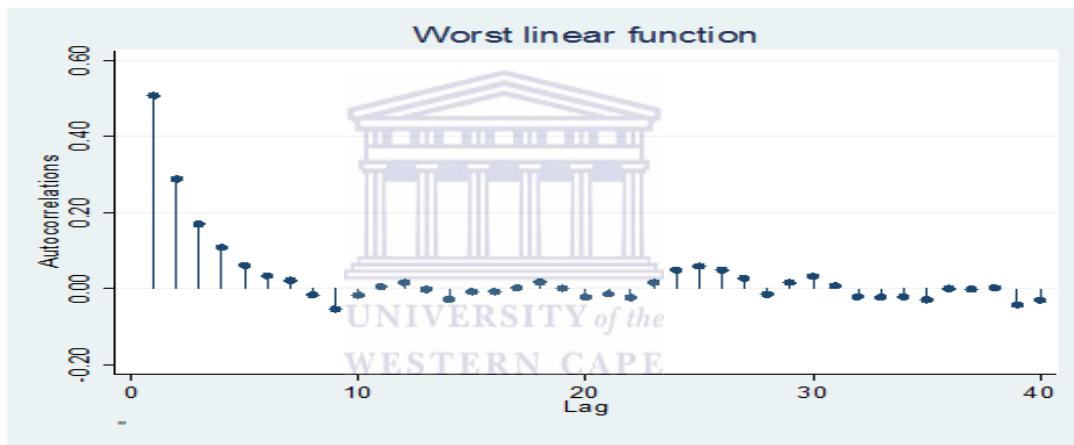


FIGURE 6.98: Model 1.2.2: Convergence of MCMC after MVNI under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

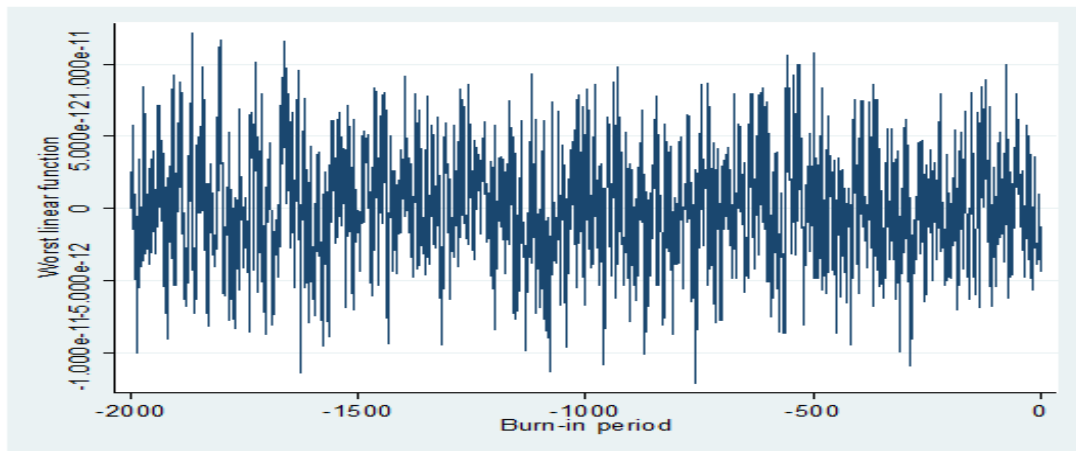


FIGURE 6.99: Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

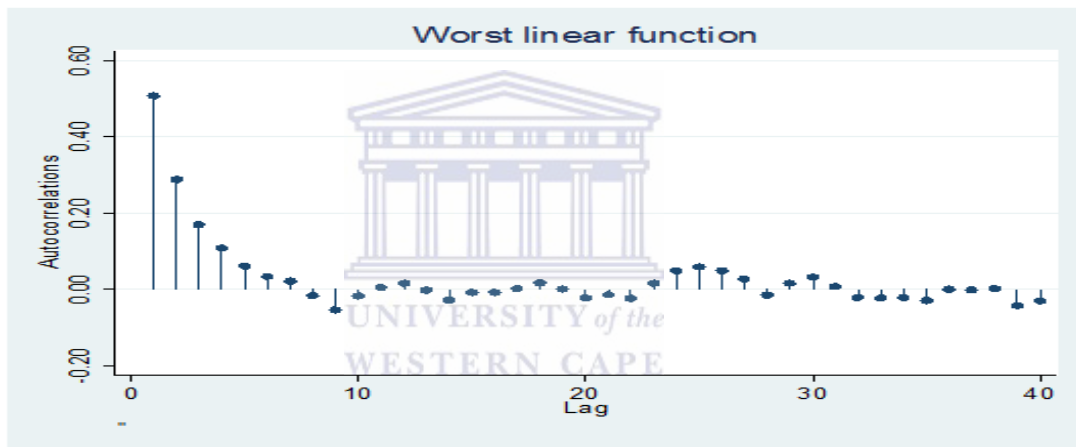


FIGURE 6.100: Model 1.2.2: Convergence of MCMC after MICE under MAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

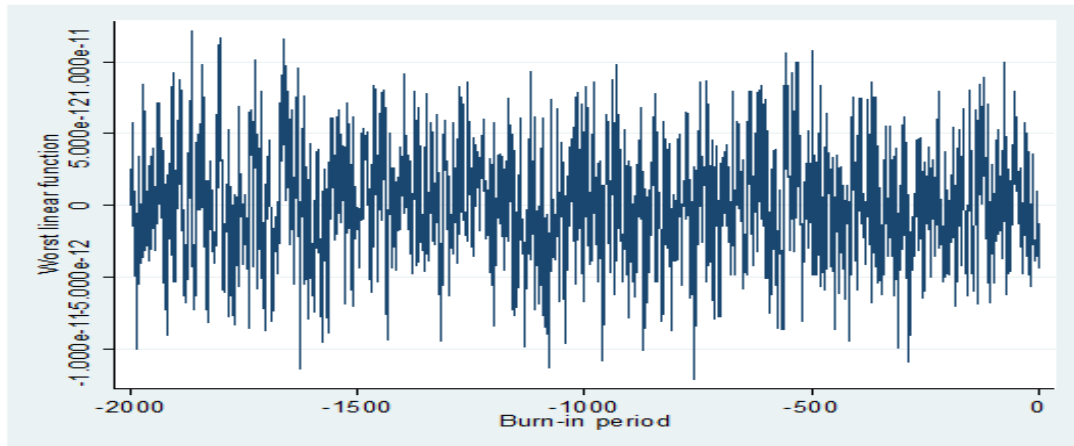


FIGURE 6.101: Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set

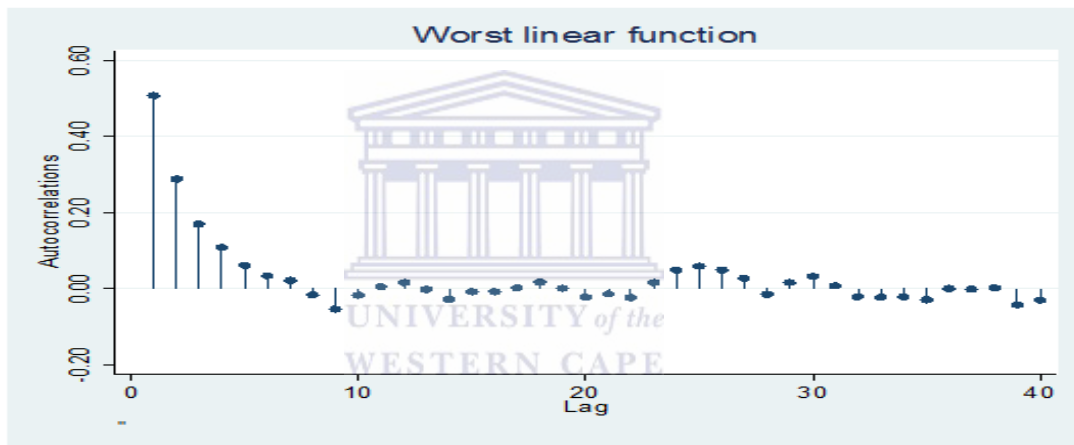


FIGURE 6.102: Model 1.2.2: Convergence of MCMC after MVNI under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

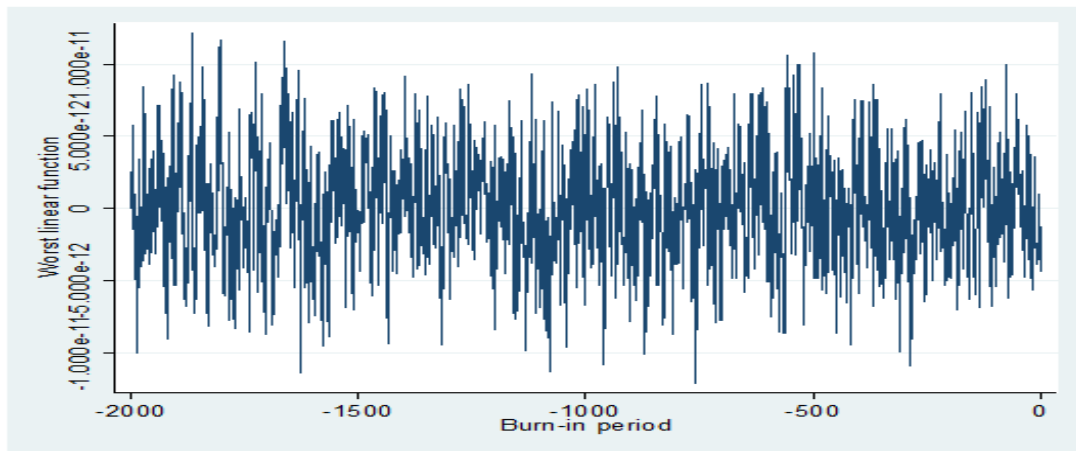


FIGURE 6.103: Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF against the iteration numbers for weighted data set.

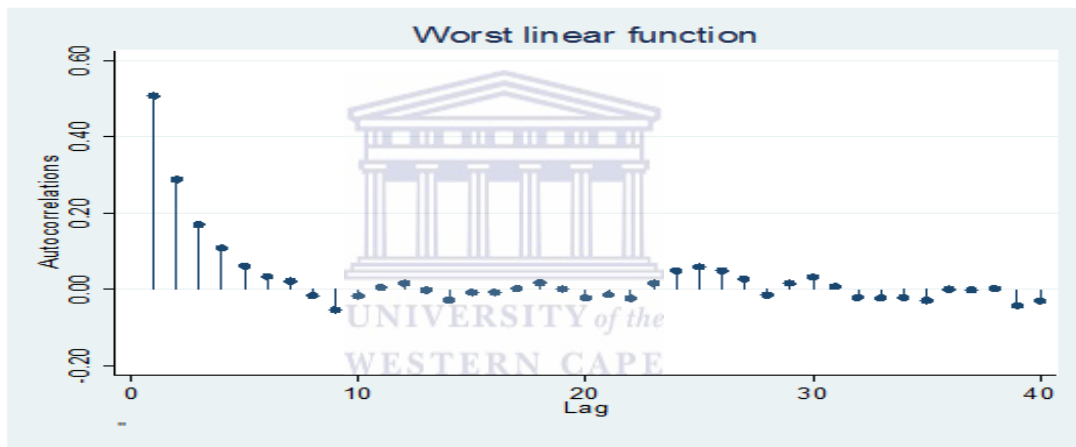


FIGURE 6.104: Model 1.2.2: Convergence of MCMC after MICE under MCAR assumption on marital status: plot of the estimates of WLF versus the lag numbers for weighted data set.

Model 2.1: Imputation models diagnostics

Unweighted data sets

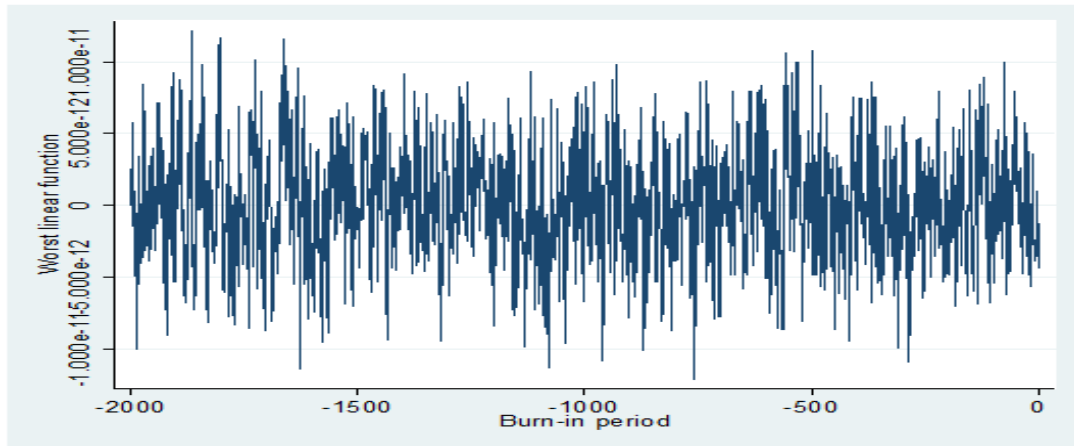


FIGURE 6.105: Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

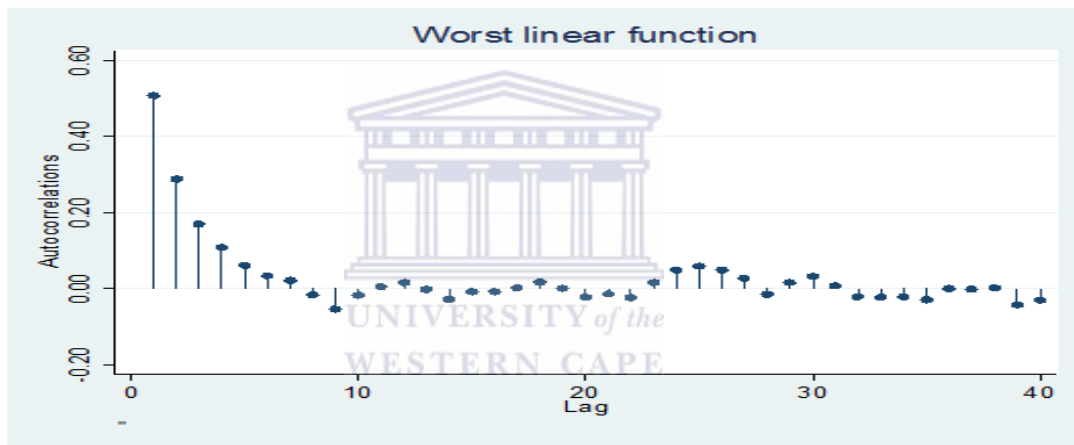


FIGURE 6.106: Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

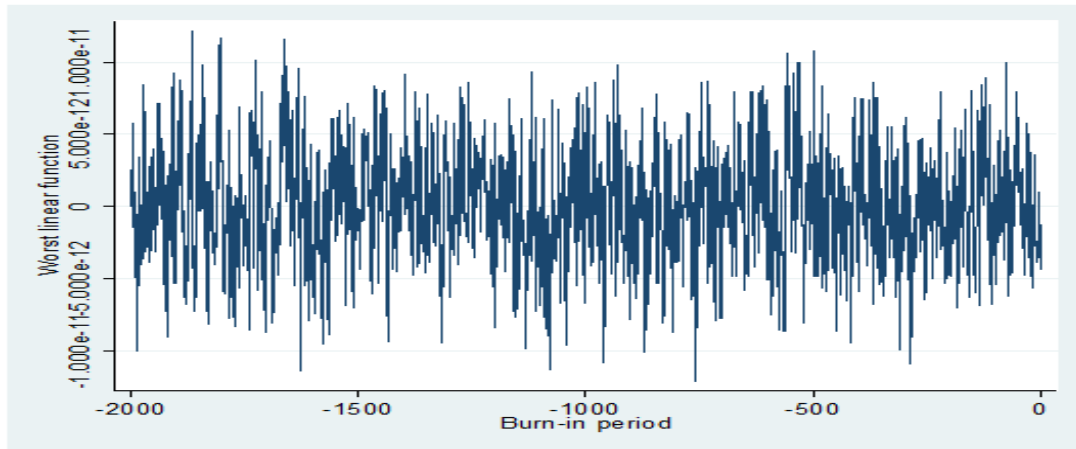


FIGURE 6.107: Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

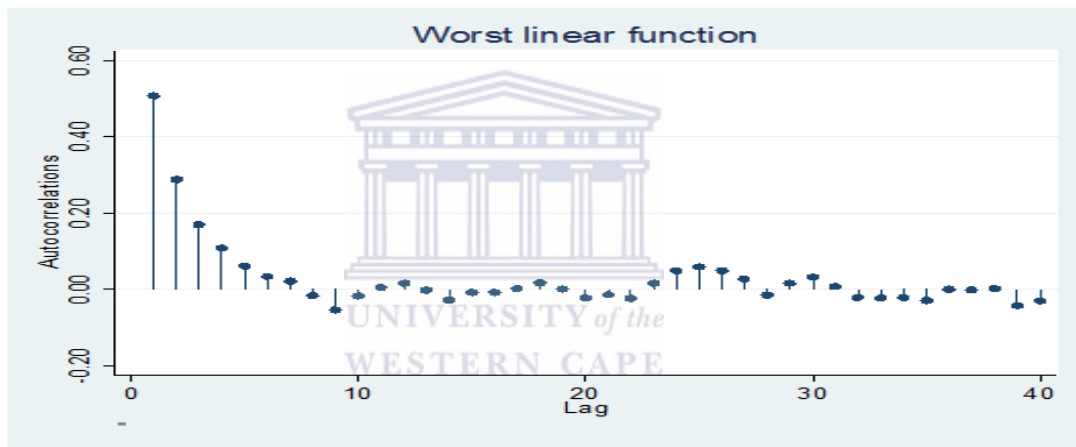


FIGURE 6.108: Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

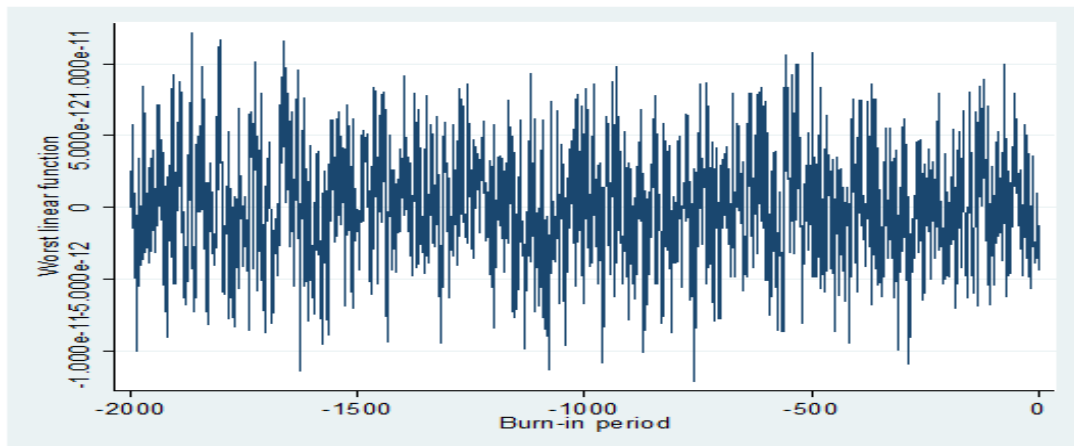


FIGURE 6.109: Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

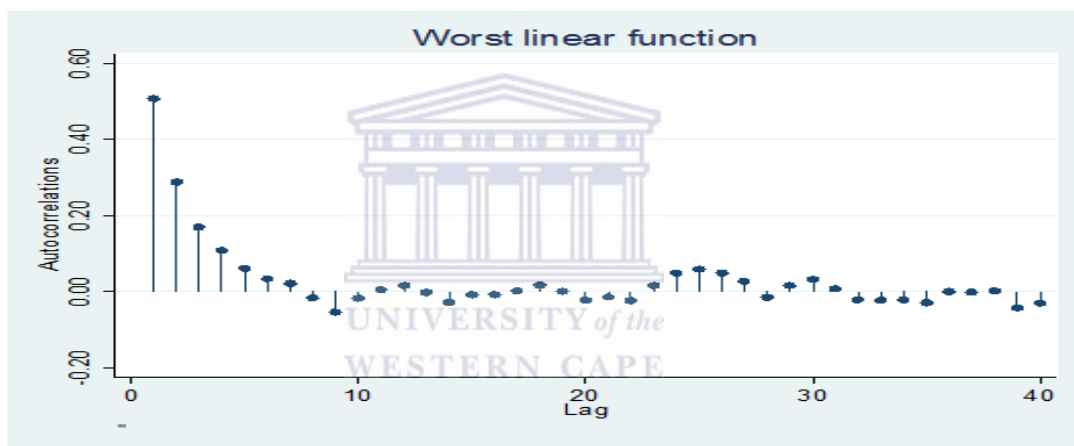


FIGURE 6.110: Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

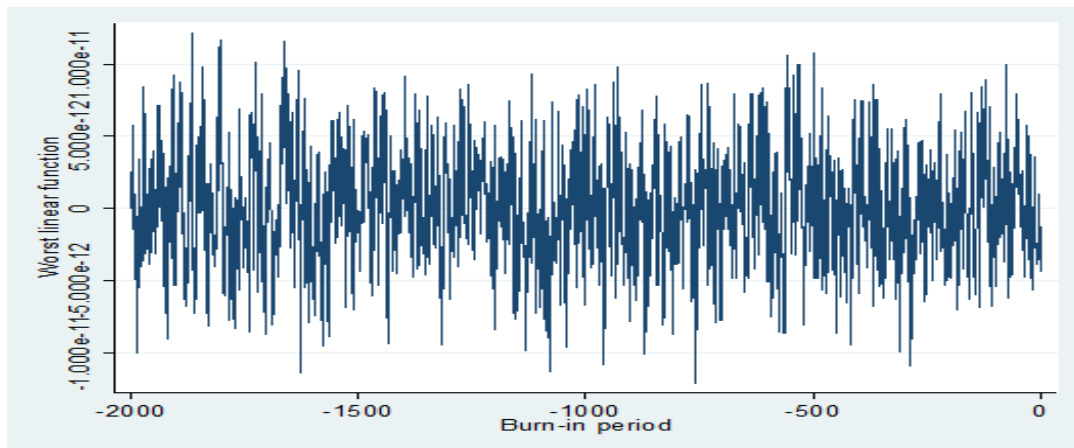


FIGURE 6.111: Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

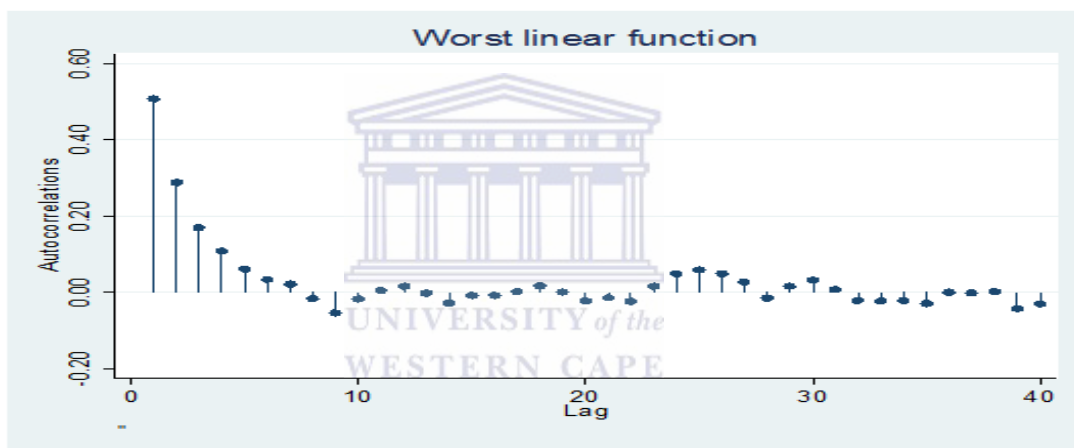


FIGURE 6.112: Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

Weighted data sets

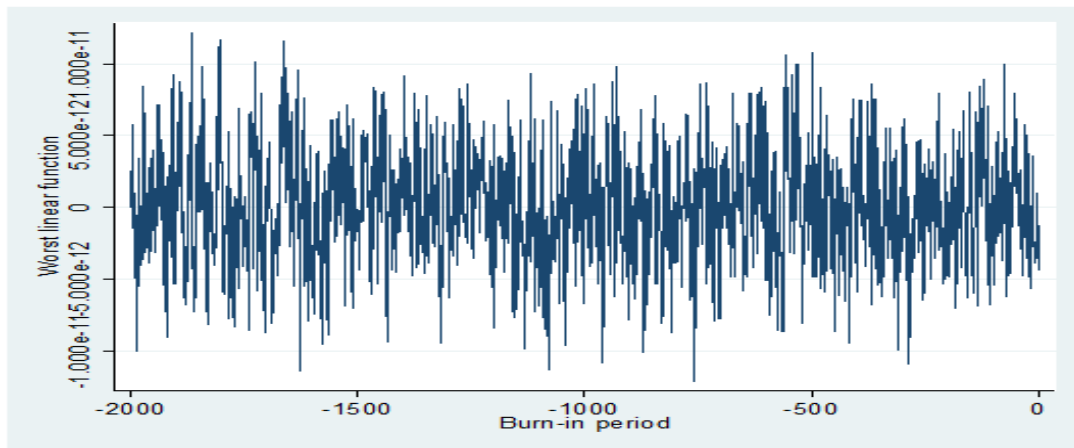


FIGURE 6.113: Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

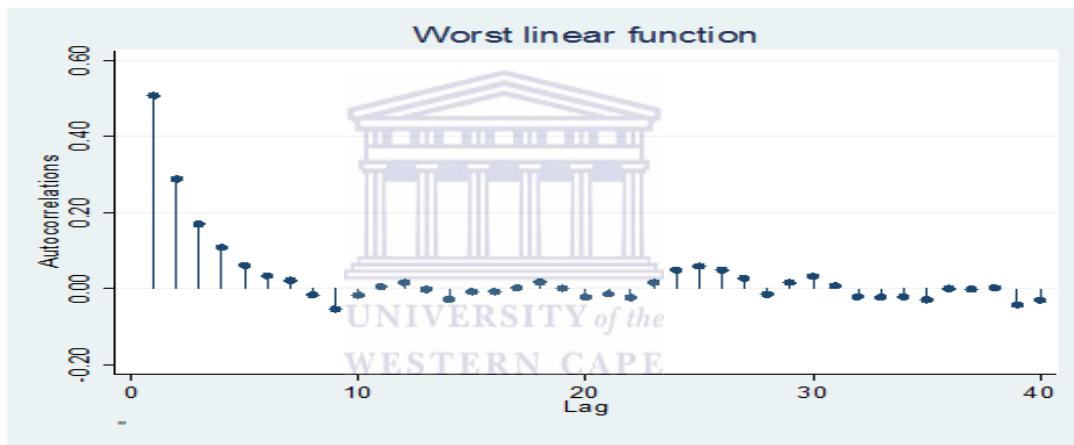


FIGURE 6.114: Model 2.1: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.

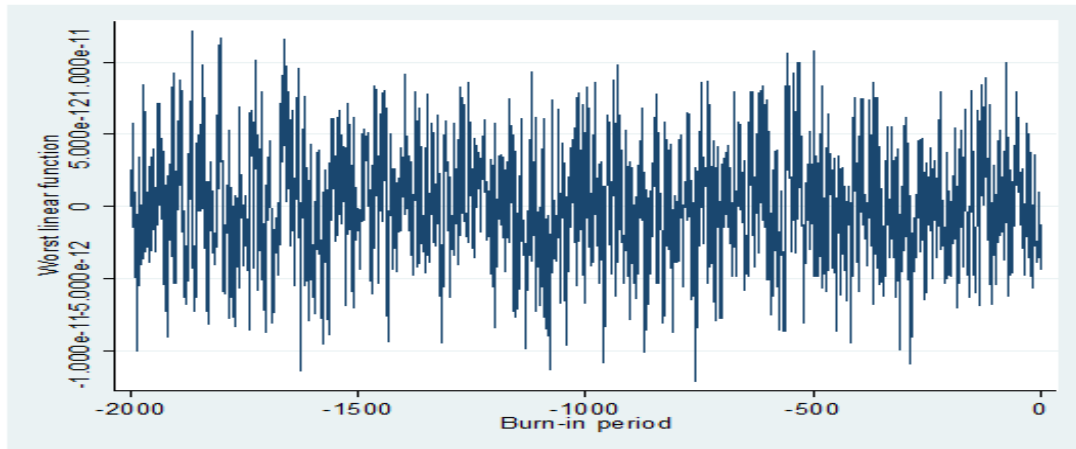


FIGURE 6.115: Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

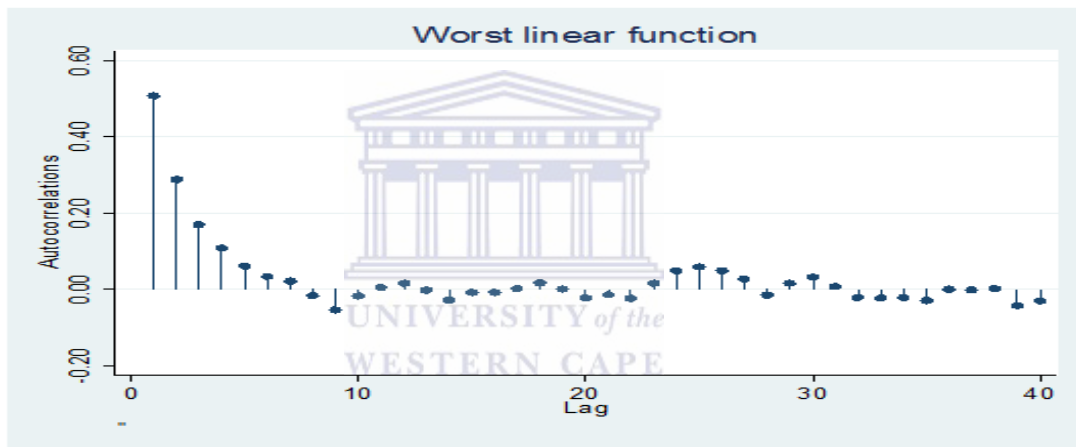


FIGURE 6.116: Model 2.1: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.

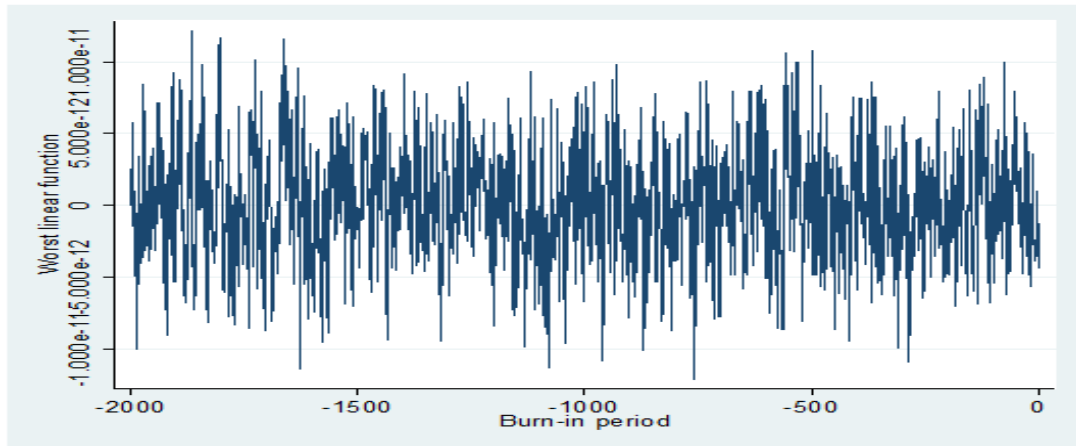


FIGURE 6.117: Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

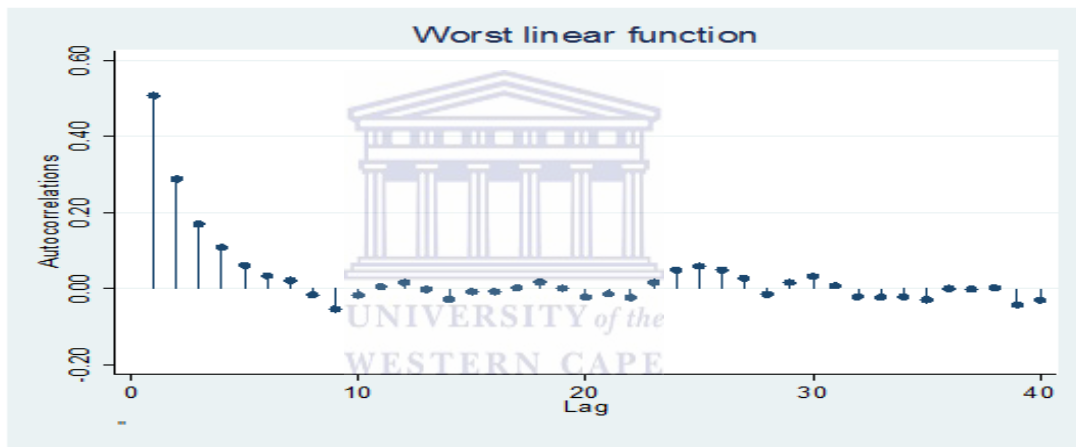


FIGURE 6.118: Model 2.1: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.

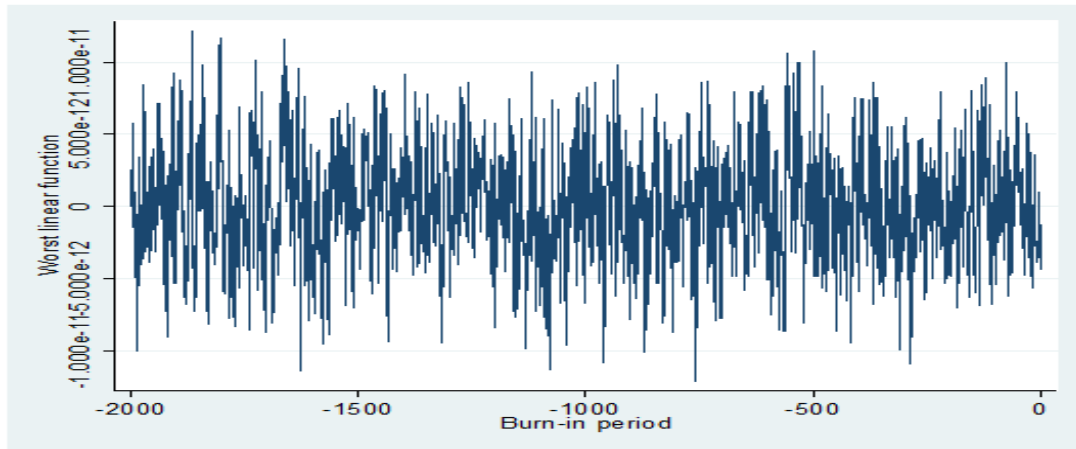


FIGURE 6.119: Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

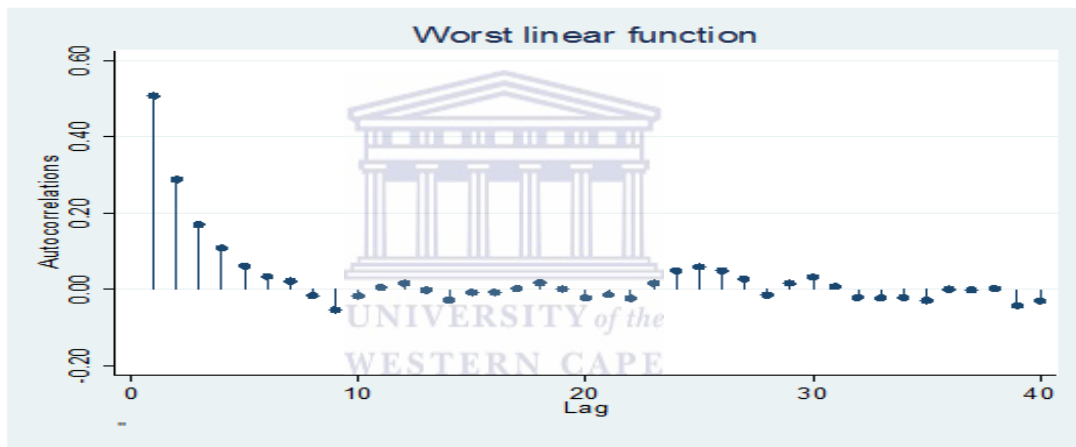


FIGURE 6.120: Model 2.1: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.

Model 2.2: Imputation models diagnostics

Unweighted data sets

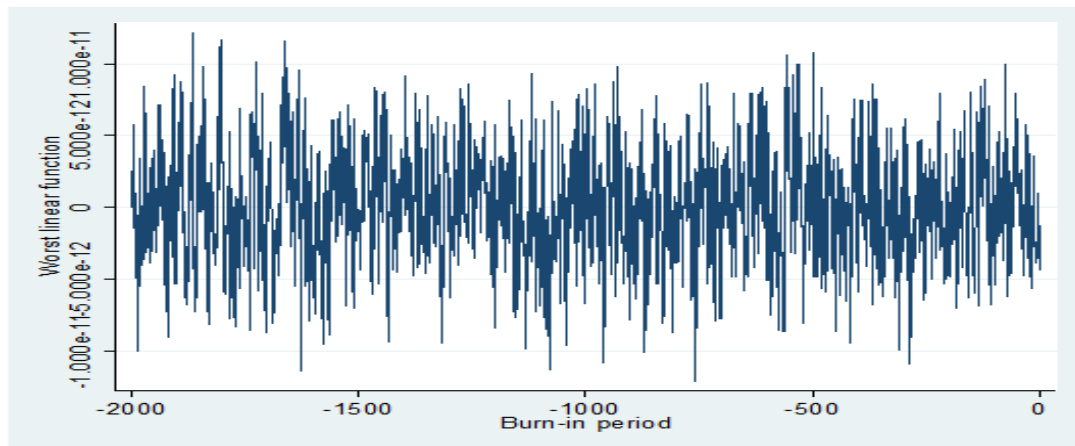


FIGURE 6.121: Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

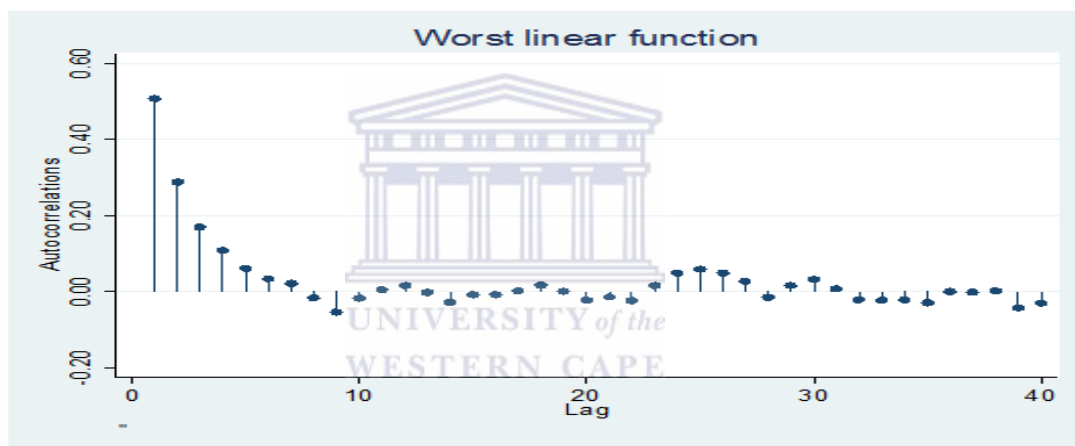


FIGURE 6.122: Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

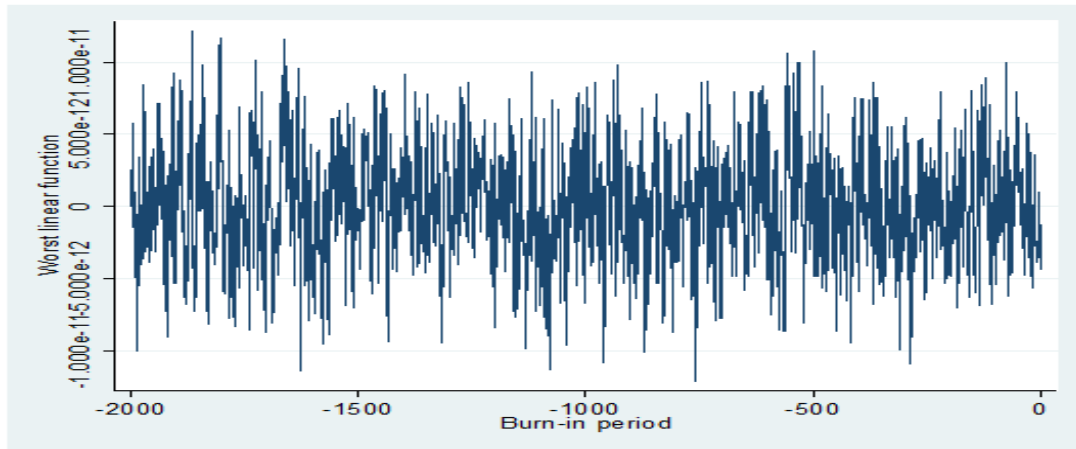


FIGURE 6.123: Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

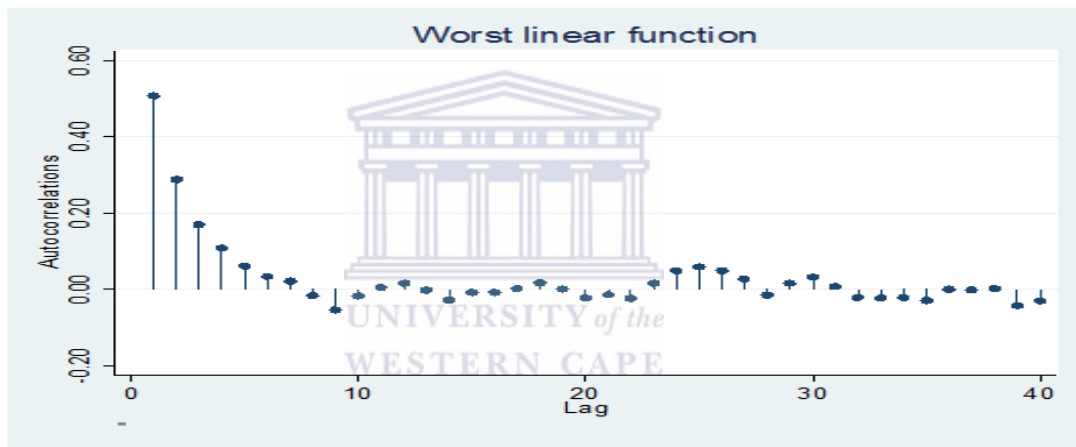


FIGURE 6.124: Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

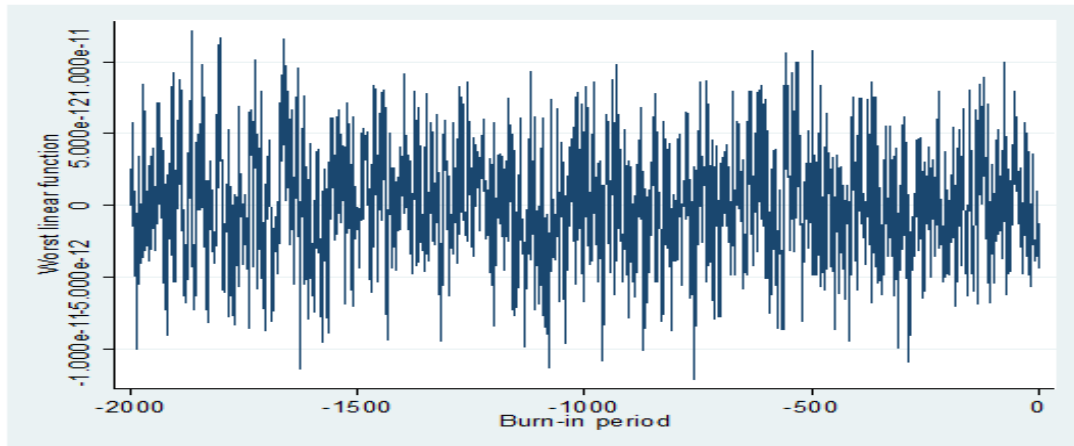


FIGURE 6.125: Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

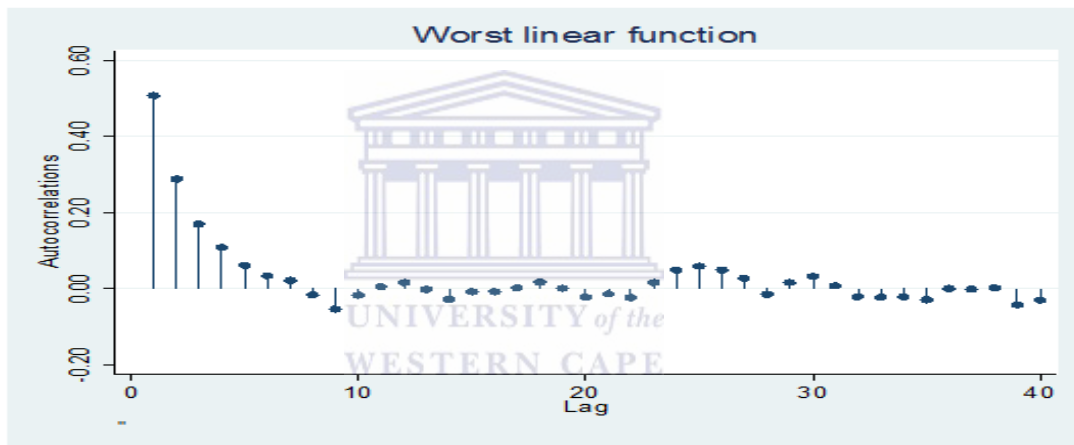


FIGURE 6.126: Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

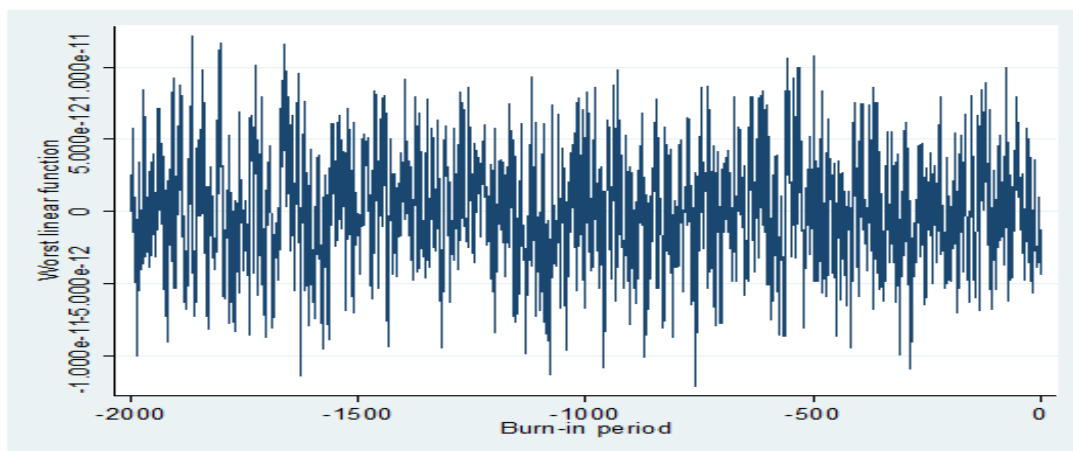


FIGURE 6.127: Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for unweighted data set.

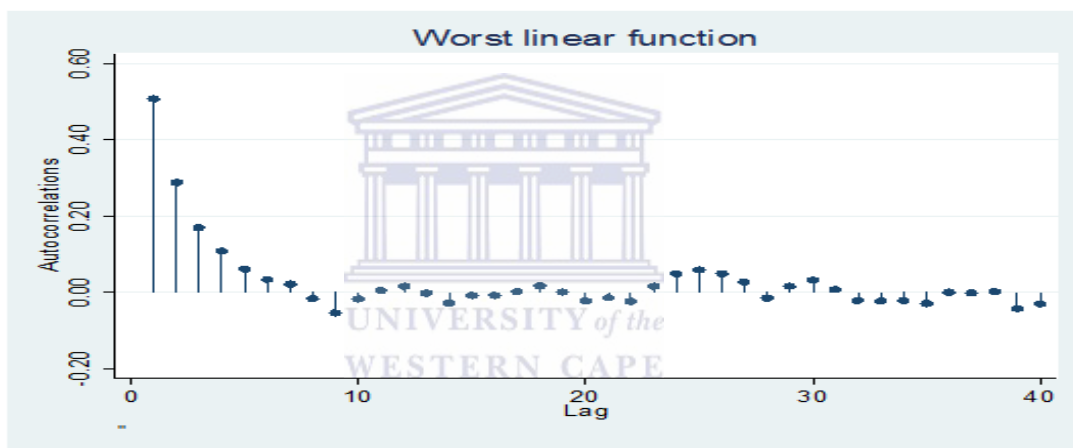


FIGURE 6.128: Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for unweighted data set.

Weighted data sets

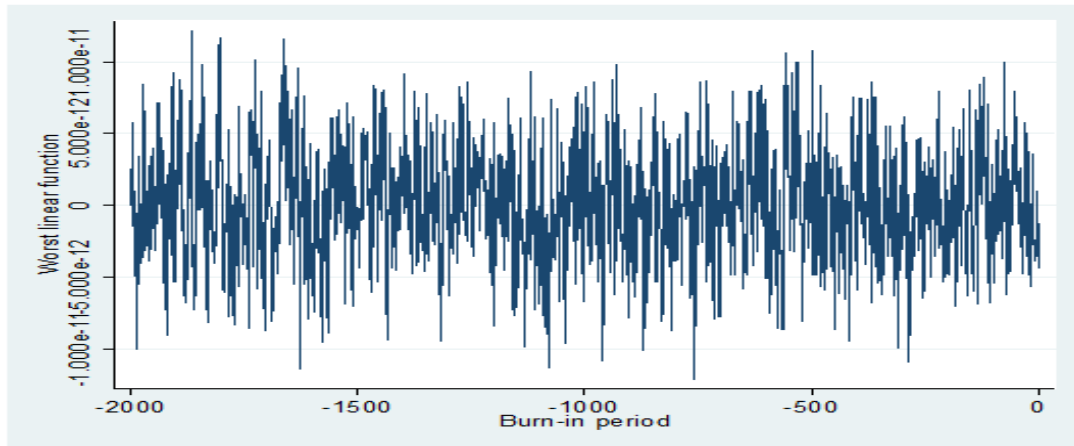


FIGURE 6.129: Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

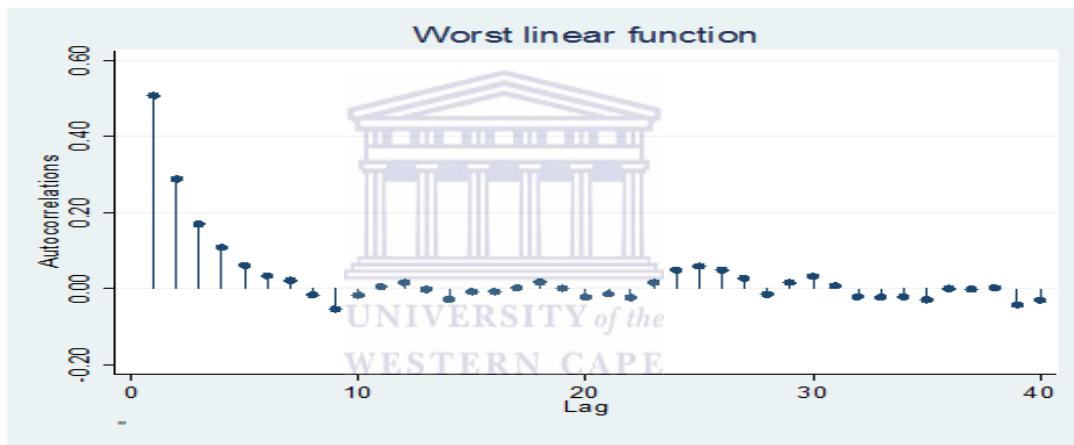


FIGURE 6.130: Model 2.2: Convergence of MCMC after MVNI under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.

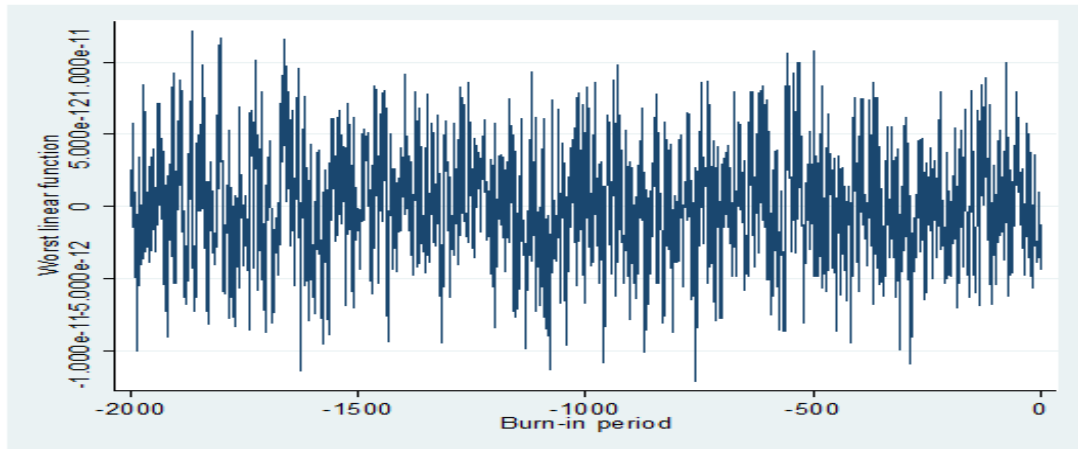


FIGURE 6.131: Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

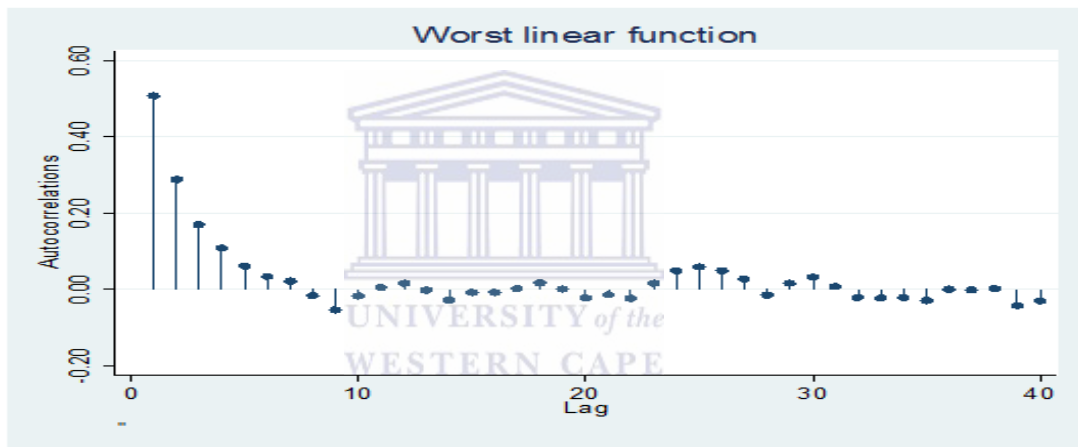


FIGURE 6.132: Model 2.2: Convergence of MCMC after MICE under MAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.

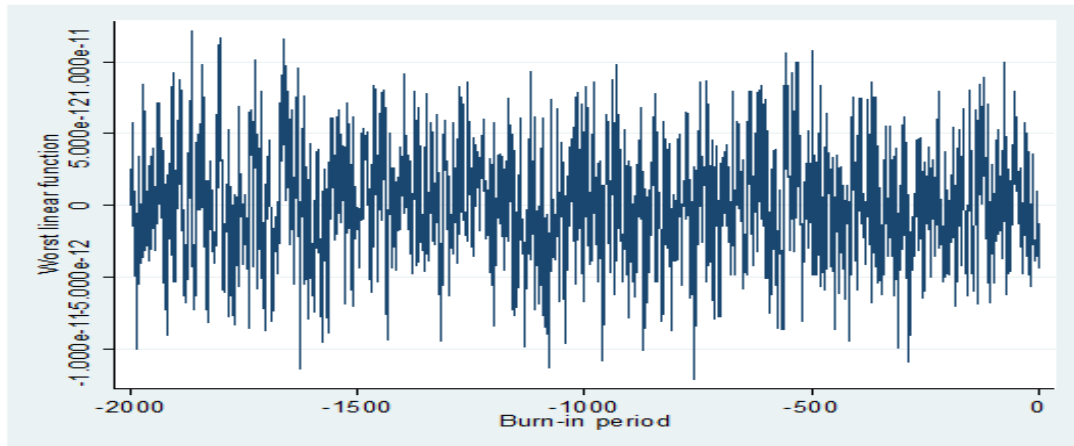


FIGURE 6.133: Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

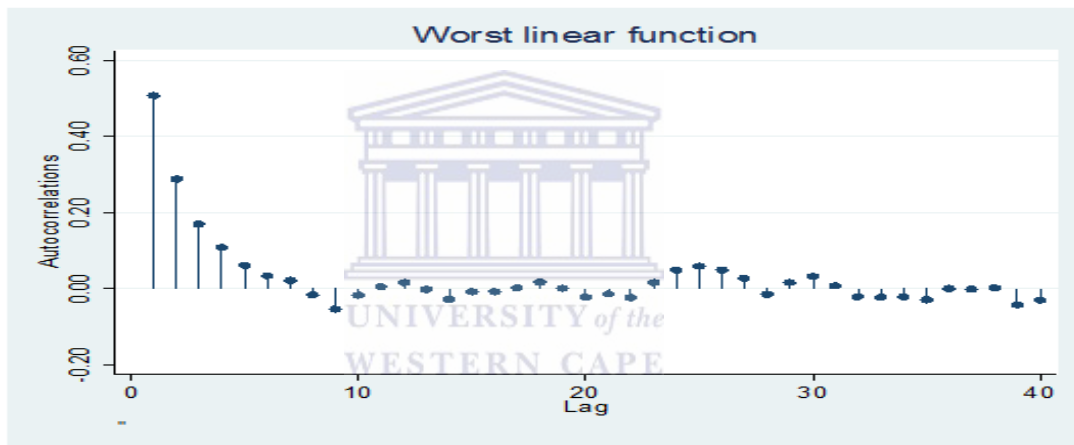


FIGURE 6.134: Model 2.2: Convergence of MCMC after MVNI under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.

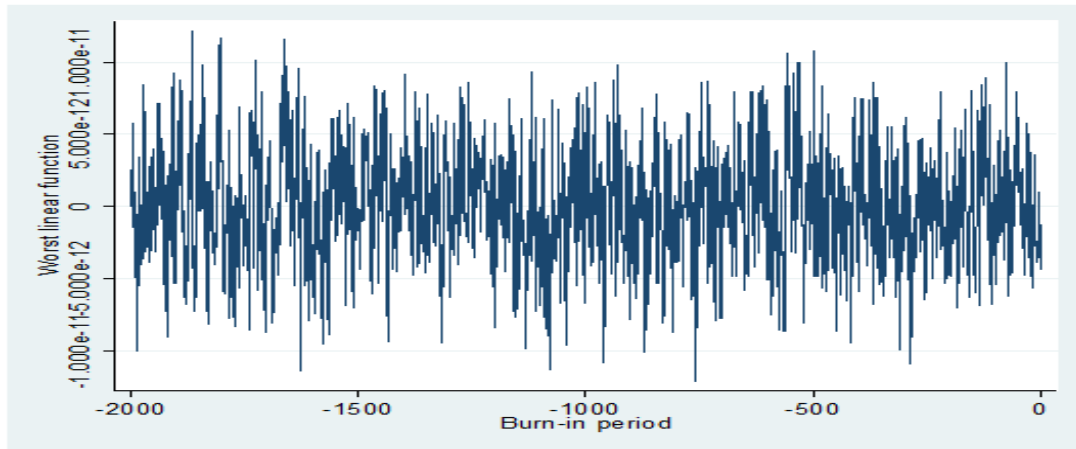


FIGURE 6.135: Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF against the iteration numbers for weighted data set.

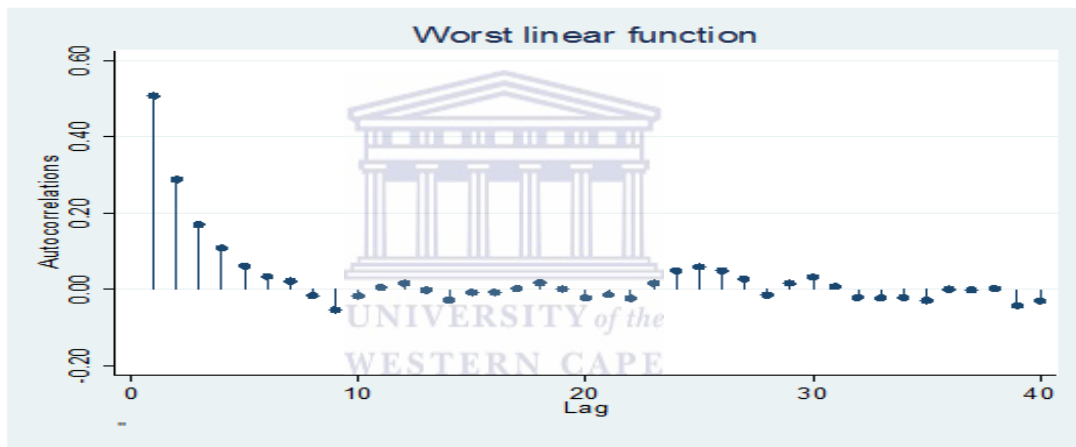


FIGURE 6.136: Model 2.2: Convergence of MCMC after MICE under MCAR assumption on contraceptive method use status (dichotomous variable): plot of the estimates of WLF versus the lag numbers for weighted data set.