

# **Imputation on the Food, Nutrition and Environment Surveys 2007 and 2009 data**

Author: Maoxin Luo  
Supervisor: Dr. Richard Arnold

A thesis submitted  
to Victoria University of Wellington  
in fulfillment of the requirements for  
the Degree of Master of Science in Statistics

School of Mathematics, Statistics and Operations Research  
Victoria University of Wellington  
PO Box 600  
Wellington  
New Zealand

Victoria University of Wellington  
2013

## **Abstract**

The Food Nutrition Environment Survey (FNES) is a survey of New Zealand early childhood centres and schools and the food and nutritional services that they provide for their pupils. The 2007 and 2009 FNES surveys were managed by the Ministry of Health. Like all the other social surveys, the FNES has the common problem of unit and item non-responses. In other words, the FNES has missing data. In this thesis, we have surveyed a wide variety of missing data handling techniques and applied most of them to the FNES datasets.

This thesis can be roughly divided into two parts. In the first part, we have studied and investigated the different nature of missing data (i.e. missing data mechanisms), and all the common and popular imputation methods, using the Synthetic Unit Record File (SURF) which has been developed by the Statistics New Zealand for educational purposes. By comparing all those different imputation methods, Bayesian Multiple Imputation (MI) method is the preferred option to impute missing data in terms of reducing non-response bias and properly propagating imputation uncertainty.

Due to the overlaps in the samples selected for the 2007 and 2009 FNES surveys, we have discovered that the Bayesian MI can be improved by incorporating the matched dataset. Hence, we have proposed a couple of new approaches to utilize the extra information from the matched dataset. We believe that adapting the Bayesian MI to use the extra information from the matched dataset is a preferable imputation strategy for imputing the FNES missing data. This is because the use of the matched dataset provides more prediction power to the imputation model.

# Acknowledgement

Writing up this thesis is not an easy task. It would be extremely harder without the guidance and help from my supervisor Dr. Richard Arnold and support from people around me.

To my supervisor, Richard, thanks for your teaching, encouragement and understanding. You have been an excellent mentor. There are no words can express my gratitude for your help!

To my wife, Echo, thanks for your love and support. In order to let me concentrate on writing up this thesis, you have shared a huge proportion of housework and looked after our baby almost all by yourself. I am grateful for your sacrifice!

To my parents, Yongsho Luo and Yongxiu Mao, thank you for believing in me all the time. I hope I have made you proud!

To my employer, Statistics New Zealand, thanks heaps for supporting my postgraduate study and giving me time to write this thesis.

Finally, to my thirteen-month-old daughter Susana, thank you very much for coming to this world. You are the new reason I want to complete this thesis because I wish I could inspire you when you are older. “Infinity” is the word to describe how much I love you!

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Introducing Missing Data</b>	<b>4</b>
2.1	Missing data . . . . .	4
2.2	Creating Missing data in a Synthetic Unit Record File . . . . .	7
2.2.1	Description of the SURF dataset . . . . .	7
2.2.2	Creating the missing SURF income values . . . . .	8
2.3	Conclusion . . . . .	16
<b>3</b>	<b>Dealing with missing data</b>	<b>17</b>
3.1	Introduction . . . . .	17
3.2	Dealing with missing data I: Data Deletion Methods . . . . .	17
3.2.1	Listwise Deletion . . . . .	18
3.2.2	Pairwise Deletion . . . . .	19
3.2.3	Reweighting . . . . .	20
3.2.4	Averaging the Available Items . . . . .	21
3.3	Dealing with missing data II: Imputation . . . . .	21
3.3.1	Single Imputation Methods . . . . .	21
3.3.2	Likelihood Based Approaches . . . . .	24
3.3.3	Bayes Theory and Simulation methods . . . . .	25
3.3.4	Multiple Imputation . . . . .	27
3.4	Distinguishing Non-response bias and Imputation uncertainty . . . . .	27
3.5	Conclusion . . . . .	29
<b>4</b>	<b>Applying single imputation methods to the SURF data</b>	<b>30</b>
4.1	Introduction . . . . .	30
4.2	Explicit modelling methods . . . . .	30
4.2.1	Unconditional mean imputation . . . . .	30
4.2.2	Conditional mean imputation (Cell mean) . . . . .	34
4.2.3	Regression imputation . . . . .	37
4.2.4	Stochastic regression imputation . . . . .	39
4.3	Implicit modelling methods . . . . .	40
4.3.1	Hot deck imputation - random hot deck imputation with replacement	42
4.3.2	Hot deck imputation - random hot deck imputation without replacement	46
4.3.3	Hot deck imputation - sequential hot deck . . . . .	50
4.3.4	Hot deck imputation - Hot deck within adjustment cells . . . . .	52
4.3.5	Hot deck imputation - Nearest-Neighbour Hot deck Imputation . . . .	54
4.4	Conclusion . . . . .	56

<b>5</b>	<b>Using Non-parametric Resampling Methods to Incorporate Imputation Uncertainty</b>	<b>57</b>
5.1	Introduction . . . . .	57
5.2	Relationship between Resampling Methods and Imputation Uncertainty . . .	57
5.3	The Simple Bootstrap for Complete Data . . . . .	59
5.4	The Simple Bootstrap Applied to Imputed Incomplete Data . . . . .	60
5.5	The Simple Jackknife for Complete Data . . . . .	62
5.6	The Simple Jackknife Applied to Imputed Incomplete Data . . . . .	63
5.7	Conclusion . . . . .	68
<b>6</b>	<b>Likelihood based imputation methods</b>	<b>69</b>
6.1	Introduction . . . . .	69
6.2	Applying EM Algorithm to the Exponential Family . . . . .	70
6.3	Example one: applying EM Algorithm to the Univariate Normal Data with Missing Values . . . . .	72
6.4	Applying EM Algorithm to Bivariate Normal data with Missing Data on Both Variables . . . . .	74
6.5	Convergence of EM algorithm . . . . .	79
6.6	Conclusion . . . . .	80
<b>7</b>	<b>Bayesian Multiple Imputation</b>	<b>81</b>
7.1	Introduction . . . . .	81
7.2	Bayesian Iterative Simulation Methods - Markov Chain Monte-Carlo (MCMC)	82
7.2.1	Gibbs sampling algorithm . . . . .	82
7.2.2	Metropolis-Hastings (MH) algorithm . . . . .	83
7.2.3	Relationship between Gibbs and MH sampling . . . . .	84
7.2.4	Block Updating . . . . .	85
7.3	Applying MCMC methods to Normal data with Ignorable Non-response . . .	85
7.3.1	Applying the Gibbs sampler to Univariate Normal data . . . . .	85
7.3.2	Applying the MH algorithm to Univariate Normal data . . . . .	88
7.3.3	Applying Gibbs sampler to Bivariate Normal data . . . . .	90
7.4	Convergence Diagnostics . . . . .	92
7.4.1	The Theory of Convergence . . . . .	92
7.4.2	Pre-convergence: the burn-in period . . . . .	93
7.4.3	Some of the popular Methods of Convergence Diagnostics . . . . .	93
7.5	Applying Gibbs sampler to multiple regression with missing data . . . . .	96
7.6	Conclusion . . . . .	98
<b>8</b>	<b>Multiple Imputation</b>	<b>100</b>
8.1	Introduction . . . . .	100
8.2	Analysis of multiply-imputed data . . . . .	101
8.3	The MI Process . . . . .	102
8.4	Proper and Improper Multiple Imputation . . . . .	109
8.5	Conclusion . . . . .	114
<b>9</b>	<b>Imputation methods for categorical variables</b>	<b>115</b>
9.1	Introduction . . . . .	115
9.2	Types of categorical variable . . . . .	115
9.3	Single imputation methods for categorical data . . . . .	115
9.3.1	Mode imputation . . . . .	115
9.3.2	Logistic regression imputation . . . . .	120

9.3.3	The Nearest-Neighbour hot deck imputation methods . . . . .	121
9.4	Likelihood based and Bayesian iterative simulation imputation methods for categorical data . . . . .	122
9.4.1	EM algorithm for categorical variable . . . . .	122
9.4.2	Bayesian iterative simulation methods for categorical variables with missing data . . . . .	125
9.5	Conclusion . . . . .	131
<b>10</b>	<b>Introduction to the Food Nutrition Environment Survey (FNES)</b>	<b>132</b>
10.1	Purpose . . . . .	132
10.2	Survey background . . . . .	132
10.3	Periods . . . . .	132
10.4	Target Population . . . . .	132
10.5	Survey Population . . . . .	133
10.6	Sample Frame . . . . .	133
10.7	Sample Size . . . . .	134
10.8	Matching process . . . . .	135
10.9	Sample Weights . . . . .	137
10.10	Sample Design . . . . .	139
10.11	Questionnaire . . . . .	139
<b>11</b>	<b>Imputation of FNES missing data</b>	<b>140</b>
11.1	Exploratory Data Analysis (EDA) . . . . .	140
11.2	Investigating the Missing Data Mechanism . . . . .	145
11.2.1	Univariate comparisons . . . . .	146
11.2.2	Logistic regression assessment method . . . . .	147
11.3	Applying Imputation Methods to the FNES . . . . .	152
11.3.1	Preparing for Imputation . . . . .	152
11.3.2	Incorporating sample weights in the imputation models . . . . .	154
11.3.3	Imputing the 2007 and the 2009 ECE FNES missing data . . . . .	155
11.3.4	Imputing the School FNES data . . . . .	165
11.3.5	Imputation using the matched 2007 and 2009 ECE FNES sample . . . . .	171
11.4	Discussion . . . . .	184
<b>12</b>	<b>Some final thoughts</b>	<b>186</b>
12.1	Summary of previous chapters . . . . .	186
12.2	Future work . . . . .	187
	<b>Appendices</b>	<b>195</b>
<b>A</b>	<b>R code for chapter 5</b>	<b>196</b>
A.1	The simple bootstrap . . . . .	196
A.2	The simple jackknife . . . . .	197
<b>B</b>	<b>R code for chapter 6</b>	<b>199</b>
B.1	EM algorithm - Univariate Normal Data . . . . .	199
B.2	Recipe: EM algorithm - Bivariate Normal Sample with Missing Data on both Variables . . . . .	200
<b>C</b>	<b>R code for chapter 7</b>	<b>203</b>
C.1	Applying MH algorithm to Univariate Normal data . . . . .	203
C.2	Applying Gibbs sampling algorithm to Bivariate Normal data . . . . .	204

<b>D</b>	<b>R code for Chapter 8</b>	<b>206</b>
D.1	The MI Process . . . . .	206
D.2	Proper and Improper Multiple Imputation . . . . .	207
<b>E</b>	<b>R code for Chapter 9</b>	<b>209</b>
E.1	Single imputation methods for categorical data . . . . .	209
E.1.1	Mode imputation . . . . .	209
E.1.2	Logistic regression imputation . . . . .	209
E.2	Likelihood based and Bayesian iterative simulation imputation methods for categorical data . . . . .	210
E.2.1	EM algorithm for categorical variable . . . . .	210
<b>F</b>	<b>R code for Chapter 11</b>	<b>212</b>
F.1	Impute the 2007 and 2009 ECE FNES missing data . . . . .	212
F.2	Imputation using the matched 2007 and 2009 ECE FNES sample . . . . .	216
F.2.1	The simple approach . . . . .	216
F.2.2	The complex approach . . . . .	220

# List of Tables

2.1	Frequency table of SURF's categorical variables . . . . .	7
2.2	Estimates for complete SURF . . . . .	9
8.1	Mean Estimates (total mean and means for each qualification) from Each of the Five MI Data Sets . . . . .	105
8.2	Overall Estimates for means . . . . .	106
8.3	Within-Imputations Variance for means . . . . .	106
8.4	Between Imputation Variance (or <i>B</i> ) for means . . . . .	107
8.5	Tests for Parameter Estimates Produced Using MI . . . . .	107
8.6	Tests for Parameter Estimates Produced Using MI . . . . .	108
8.7	Rate of Missing Information . . . . .	109
10.1	Sample frame for both 2007 and 2009 . . . . .	133
10.2	Selected sample size for both 2007 and 2009 . . . . .	134
10.3	Final sample size for both 2007 and 2009 . . . . .	134
10.4	Selected sample size of ECEs in 2007, 2009 and size of ECEs in both the 2007 and 2009 FNES . . . . .	135
10.5	Table of selected sample size of schools in 2007 and 2009 and size of schools in both 2007 and 2009 FNES . . . . .	136
10.6	Responding sample size of ECEs in 2007, 2009 and size of ECEs that responded to both the 2007 and 2009 FNES . . . . .	136
10.7	Responding sample size of schools in 2007, 2009 and size of schools that responded to both the 2007 and 2009 FNES . . . . .	137
10.8	Sample selection weights for both 2007 and 2009 . . . . .	138
10.9	Matched Sample selection weights for both 2007 and 2009 . . . . .	138
10.10	Responding Sample weights for both 2007 and 2009 . . . . .	138
10.11	Matched responding Sample weights for both 2007 and 2009 . . . . .	139
10.12	The five categories of ECE service types . . . . .	139
11.1	The sample frame, sample size, responded sample size, and response rate of the FNES 2007 . . . . .	140
11.2	The sample frame, sample size, responded sample size, and response rate of the FNES 2009 . . . . .	140
11.3	Design and Sample variables for 2007 and 2009 . . . . .	141
11.4	Collected Sample variables for 2007 and 2009 . . . . .	141
11.5	Example of Q151 . . . . .	142
11.6	Composite of low response (< 50%) questions . . . . .	142
11.7	The selected variables . . . . .	145
11.8	Sample response rate of the five missing data scenarios . . . . .	146
11.9	Split the "Authority 2009" based on the missingness of "Q3a 2009" . . . . .	147
11.10	Subset variables form the ECE 2009 data . . . . .	148
11.11	Responding sample size breaks down by explanatory variables . . . . .	149



11.12	Responding sample size breaks down by regrouped variables . . . . .	150
11.13	Cross tabulation between Stratum and Authority . . . . .	151
11.14	Investigate missing mechanism: logistic modelling results . . . . .	152
11.15	Investigate missing mechanism: Analysis of Deviance Table . . . . .	152
11.16	Responding sample size for 2009 ECE FNES Q3a breaks down by answer categories . . . . .	153
11.17	Multiway table “Region” and “Stratum” . . . . .	153
11.18	Multiway table “Super region” and “Stratum” . . . . .	153
11.19	Breakdown of incomplete Q3a of 2009 ECE FNES by Stratum and Super region	160
11.20	Imputation results for 2009 ECE FNES Q3a . . . . .	160
11.21	Imputation results for 2009 ECE FNES Q7b . . . . .	161
11.22	Imputation results for 2009 ECE FNES Q7c . . . . .	162
11.23	Imputation results for 2007 ECE FNES Q3a . . . . .	162
11.24	Imputation results for 2007 ECE FNES Q7b . . . . .	163
11.25	Imputation results for 2007 ECE FNES Q7c . . . . .	164
11.26	Scenarios where the imputation based on logical rules can be applied . . . . .	165
11.27	Cross tabulation of Q5a and Q5b for 2009 School FNES . . . . .	165
11.28	Cross tabulation of Q5a and Q5b for 2007 School FNES . . . . .	166
11.29	Combining Q5a and Q5b into Q5 . . . . .	166
11.30	Imputation results for 2009 School FNES Q5a . . . . .	168
11.31	Imputation results for 2009 School FNES Q5b . . . . .	169
11.32	Imputation results for 2007 School FNES Q5a . . . . .	170
11.33	Imputation results for 2007 School FNES Q5b . . . . .	170
11.34	Update the matched 2009 ECE FNES data with the matched 2007 data . . . . .	171
11.35	Impute 2009 ECE FNES with the matched 2007 data by using MI . . . . .	180
11.36	Impute 2007 ECE FNES with the matched 2009 data by using MI . . . . .	180
11.37	Impute 2009 ECE FNES with the matched 2007 data by using MI on both response and explanatory variables . . . . .	183
11.38	Impute 2007 ECE FNES with the matched 2009 data by using MI on both response and explanatory variables . . . . .	183

# List of Figures

2.1	Missing data Patterns . . . . .	4
2.2	SURF's age distribution . . . . .	8
2.3	SURF's weekly working hours distribution . . . . .	8
2.4	SURF's weekly income distribution . . . . .	8
2.5	Missing SURF Income (MCAR) . . . . .	10
2.6	Mean of Income by sex (MCAR) . . . . .	11
2.7	Variance of Income by sex (MCAR) . . . . .	11
2.8	Missing SURF Income (MAR) . . . . .	13
2.9	Mean of Income by sex (MAR) . . . . .	13
2.10	Variance of Income by sex (MAR) . . . . .	13
2.11	Missing SURF Income (NMAR) . . . . .	15
2.12	Mean of Income by sex (NMAR) . . . . .	15
2.13	Variance of Income by sex (NMAR) . . . . .	15
3.1	Listwise Deletion . . . . .	18
3.2	Pairwise Deletion . . . . .	19
3.3	Mean and Confidence interval for imputed data . . . . .	29
4.1	Unconditional Mean Imputed MAR SURF income . . . . .	32
4.2	Unconditional Mean Imputed MCAR SURF income . . . . .	33
4.3	Conditional Mean Imputed MAR SURF Income . . . . .	35
4.4	Regression Imputed MAR SURF Income . . . . .	38
4.5	Stochastic regression Imputed MAR SURF Income . . . . .	40
4.6	Random hot deck Imputed MAR SURF Income with replacement . . . . .	45
4.7	Random hot deck Imputed MAR SURF Income without replacement . . . . .	49
4.8	Sequential hot deck Imputed MAR SURF Income . . . . .	51
4.9	Hot deck within adjustment cells Imputed MAR SURF Income . . . . .	54
4.10	Nearest-Neighbor hot deck Imputed MAR SURF Income . . . . .	56
5.1	Comparing the Adjustment cell hot deck income means and variances with the bootstrap income means and variances for the simulated 100 incomplete SURF data . . . . .	62
5.2	Comparing the Adjustment cell hot deck income means and variances with the jackknife income means and variances for the simulated 100 incomplete SURF data . . . . .	65
5.3	Comparing the Adjustment cell hot deck income means and variances with the naive jackknife income means and variances for the simulated 100 incomplete SURF data . . . . .	66
5.4	Comparing the distributions of the pseudo-value of jackknife mean $\tilde{\mu}_{j,Jack}$ , the naive pseudo-value of jackknife mean $\tilde{\mu}_{j,naive}$ , and the SURF Income $Y_{Income}$ . . . . .	67

6.1	The distributions of means and variances of the 1000 replicate SURF data's income variables imputed by the EM algorithm . . . . .	75
6.2	The distributions of the means and variances of the 1000 replicate SURF data's income and hours variables imputed by the EM algorithm . . . . .	79
7.1	Applying the Gibbs sampler to the SURF data with 50 income values missing completely at random (MCAR) . . . . .	88
7.2	Applying the MH algorithm to the SURF data with 50 MCAR income values	89
7.3	Applying the Gibbs sampler to the SURF data with 50 MCAR values for each income and hours variables . . . . .	92
7.4	Time series plots for the simulated SURF's Income mean . . . . .	94
7.5	Applying the Gibbs sampler to the multiple regression with missing data . . .	98
8.1	Matrix of multivariate data with missing values and multiple imputation . . .	101
8.2	Time Series convergence diagnostics . . . . .	104
8.3	Comparison of the total variance of improper MI and proper MI . . . . .	113
9.1	The proportions of the four qualification categories - unconditional mode imputation . . . . .	119
9.2	The proportions of the four qualification levels - conditional mode imputation	119
9.3	The proportion of Male and Female - logistic regression imputation . . . . .	121
9.4	The ratio of counts of females over males - EM imputation . . . . .	125
9.5	Applying Data Augmentation(DA) to impute missing values for the SURF Gender variables by using the MH algorithm . . . . .	128
10.1	Selected sample size of ECEs in 2007, 2009 and size of ECEs in both the 2007 and 2009 FNES . . . . .	135
10.2	Selected sample size of schools in 2007, 2009 and size of schools in both 2007 and 2009 FNES . . . . .	135
10.3	Responding sample size of ECEs in 2007, 2009 and size of ECEs that responded to both the 2007 and 2009 FNES . . . . .	136
10.4	Responding sample size of schools in 2007, 2009 and size of schools that responded to both the 2007 and 2009 FNES . . . . .	137
11.1	Response rate of responding sample for 2007 and 2009 FNES questions . . .	143
11.2	Response rate of responding sample less than 50% for 2007 and 2009 FNES questions . . . . .	143
11.3	Response rate of responding sample between 50% and 90% for 2007 and 2009 FNES questions . . . . .	144
11.4	Response rate of Q3a breaks down by the four explanatory variables . . . . .	150
11.5	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q3a 2009 ECE .	161
11.6	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q7b 2009 ECE .	161
11.7	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q7c 2009 ECE .	162
11.8	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q3a 2007 ECE .	163
11.9	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q7b 2007 ECE .	163
11.10	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q7c 2007 ECE .	164
11.11	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q5a 2009 School	169
11.12	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q5b 2009 School	169
11.13	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q5a 2007 School	170
11.14	The 95% confidence interval of proportion $\hat{P}$ for the imputed Q5b 2007 School	171
11.15	Utilizing information from the matched data . . . . .	173
11.16	Distributions of $\alpha$ . . . . .	176

11.17Distributions of $\beta$ . . . . .	177
11.18Time series plots of $\hat{P}$ and $se$ for $m = 5$ Bayesian chains . . . . .	178
11.19The simple approach: plot of Gelman-Rubin PSRF by iteration for $\hat{P}$ and standard error of $\hat{P}$ . . . . .	179
11.20Utilizing information from the matched and unmatched data. . . . .	181
11.21The complex approach: plot of Gelman-Rubin PSRF by iteration for $\hat{P}$ and standard error of $\hat{P}$ . . . . .	183

# Chapter 1

## Introduction

The first purpose of this project is to provide a review of standard imputation methods for continuous and categorical variables in survey data, including multiple imputation methods. The second purpose is to apply these methods to missing data in the Food Nutrition Environment Survey (FNES), which was conducted by the Ministry of Health in 2007 and 2009, and to select the best imputation method. The outline of the thesis is as follows.

Chapter 2 introduces the general forms of missing data patterns; the three missing data mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR). This chapter also explores what might happen to our sample estimates if we ignore the missing data, under the three missing data mechanisms. The investigation was carried out by creating missing data under the three missing data mechanisms on the Synthetic Unit Record File (SURF) which is based on the Statistics New Zealand income survey (Statistics New Zealand 2011).

Chapter 3 explains the two most common methods of dealing with missing data: the first approach is to ignore the missing data by deleting records with missing data from datasets; the second approach is the imputation of the missing data.

Chapter 4 discusses standard single imputation methods. We have also applied them to the replicate SURF datasets with missing data.

Chapter 5 discusses and explores the use of resampling methods to incorporate **imputation uncertainty**, and shows why the resampling and Multiple Imputation (MI) methods are a means of including the imputation uncertainty in survey estimates. We have also discovered that the jackknife resampling method does not work well in imputation procedures.

Chapter 6 discusses and explores likelihood based imputation methods by using the EM algorithm. The EM algorithm is normally used to set up the initial estimates of the parameters before the use of Multiple Imputation.

Chapter 7 distinguishes standard Multiple Imputation and Bayesian Multiple Imputation (MI). Then, the chapter focuses on how to apply Bayesian iterative simulation methods (the Gibbs sampler and the Metropolis-Hastings algorithm), which is the Bayesian part of the Bayesian MI, to missing data, and how to measure the convergence of Bayesian simulation chain.

Chapter 8 demonstrates how Bayesian MI is applied, the exact MI process, and how the sample estimates from multiple imputed datasets are combined to get the final MI estimates. Then, the chapter discusses the difference between proper and improper MI.

Chapter 9 applies the imputation methods that have been introduced in previous chapters to a particular missing categorical data problem. Chapters 4 to 8 only present imputation methods for continuous data. Hence, the purpose of Chapter 9 is to present analogous methods for categorical data.

Chapter 10 describes the 2007 and 2009 FNES surveys, detailing the sample design used, and the collected sample data.

Chapter 11 displays the missing data pattern of the 2007 and 2009 FNES data, describes the methods of investigating the missing data mechanism, and applies previously introduced imputation methods to a few categorical variables with missing data. The chapter also proposes the use of Bayesian MI to incorporate the extra information we get from matching the 2007 and 2009 datasets to gain precision. This is our new development of the Bayesian MI for the case of partially matched datasets. Then, after comparing the results from different imputation methods, the chapter concludes that the use of Bayesian MI is the best imputation option for the FNES data.

Chapter 12 summarizes the previous chapters and proposes future improvements.

# Chapter 2

## Introducing Missing Data

### 2.1 Missing data

Missing data mean there are “holes” in our datasets<sup>1</sup>. In the real world, it is very common to have a dataset with missing data. For example, an interviewer may fail to ask a question; a respondent may refuse to answer the question or cannot provide the information; a data processor entering the data may skip the value (Lohr 1999). If missing data haven’t been dealt with properly, statistical estimates can be seriously distorted. In other words, missing data or poorly handled missing data introduce bias. There will be detailed discussion of how exactly missing data lead to bias, but for now, an intuitive explanation is to imagine that missing data hold some unique information of our target population. If we cannot retrieve the missing data, the inferences of the target population based on the observed data would be wrong.

Hence, in order to deal with missing data properly, we need to investigate the patterns, the causes, and the characteristics of missing data.

The first thing for dealing with missing data is to investigate its missing pattern. Little & Rubin (2002) and Enders (2010) classify missing data patterns into six prototypical patterns. Figure 2.1 displays the six missing data patterns, with the shaded areas representing the observed values in the data set.

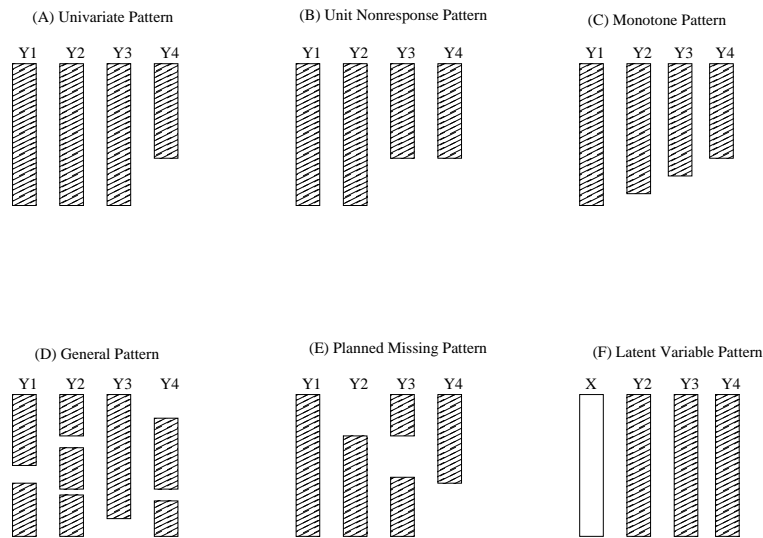


Figure 2.1: Missing data Patterns

<sup>1</sup>In this thesis, the data we have investigated are survey data

The **univariate pattern** in panel A describes the situation that has only one variable with missing data for some observations. This type of missing pattern is not common in survey data, but it is an ideal example for researchers to explore missing data handling theories and techniques. The **unit non-response pattern** in panel B, also called “Multivariate Two Patterns” by Little & Rubin (2002), has more than one variable with missing data. We usually encounter this situation in survey data. For example, there might be some variables which are available for everyone from a sample frame, but some respondents may refuse to answer some of the survey questions. The **monotone missing data pattern** in panel C, often refers to attrition<sup>2</sup> for a longitudinal study. The **general pattern** in panel D is perhaps the most common missing data pattern data users encounter. It is not surprising to see missing values dispersed throughout the data matrix in a random fashion. The **planned missing data pattern (or structural missingness pattern)** in panel E is another common pattern in survey data. For example, if a respondent gives values for variables  $Y_2$  and  $Y_4$ , then she/he does not need to provide information for variable  $Y_3$ . This can reduce respondent burden if we are collecting a large number of questionnaire items. Strictly speaking, the **latent variable pattern** in panel F, is not a missing data pattern. This is because there is no missing data in our collected data set. However, we derive another variable, called a “latent” variable<sup>3</sup> based on our collected variable values. From a missing data perspective, we consider this “latent” variable to be completely missing for the entire sample.

As an aside, researchers often describe missing data as **item non-response** and **unit non-response**. Item non-response means that survey participants answer some of the survey questions, but refuse to provide information for all the asked questions. Unit non-response means that survey participants refuse to answer any of the survey questions. For example, panel A, D, E, and F can be considered as item non-response, and panel B and C are unit non-response.

Missing data arise from the design of data collection stage to the data processing stage. For instance, during a design stage, a survey or experimental design does not include all the variables which might contribute to the data user’s research at the data analysis stage. These unobserved “latent” variables are completely missing and were regarded them as missing data (Little & Rubin 2002). However, the majority of missing data come from the data collection stage: people may refuse to respond to all or a part of a survey; there might be no outcome for an experiment; drop outs may occur when we repeat our survey after a certain time in a longitudinal study. These are missing data due to non-response. Missing data are also caused by data collectors and data processors: the data may not be properly collected, or mistakes happen at the data entry stage (Ader & Mellenbergh 2008).

Understanding the causes of missing data can help us to solve some of the missingness problems. For example, we can investigate whether some of observed variables are related to the “latent” variables and build up a model based on their relationship to predict “latent” variables’ values; we can re-contact the survey non-respondents to get their data; we can double check our edited data to make sure there is no typographic error.

However, not all the missing data can be fixed by using the above methods. There will be missing data in the raw data sets whether we like it or not.

---

<sup>2</sup>In a longitudinal study, we collect information from the same respondents repetitively during different time periods. Some of our participants may drop out and never return to the study. These drop outs are referred to as attrition.

<sup>3</sup>latent variables are variables which haven’t been surveyed or collected.



The question is what we can do about the irretrievable missing data. Little & Rubin (2002) think that there are possible relationships between observed variables and the probability of missing data, so they have identified three mechanisms which lead to missing data. They are: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little & Rubin 2002).

Let  $Y = (y_{ij})$  denote an  $(n \times K)$  rectangular complete data set, with  $i$ th row  $y_i = (y_{i1}, \dots, y_{iK})$  where  $y_{ij}$  is the value of variable  $Y_j$  for observation  $i$ . If there are missing data, let us define the missing data indicator matrix  $R = (r_{ij})$ , such that:

$$r_{ij} = \begin{cases} 0 & \text{if } y_{ij} \text{ is missing} \\ 1 & \text{if } y_{ij} \text{ is not missing} \end{cases}$$

**Missing completely at random (MCAR):** If the conditional distribution of  $R$  given  $Y$  does not depend on the values of the data  $Y$ , missing or observed, the data are called missing completely at random (MCAR):

$$f(R|Y, \phi) = f(R|\phi) \text{ for all } Y, \phi, \quad (2.1)$$

where  $f$  is a generic symbol for a probability distribution,  $\phi$  is a parameter (or a set of parameters).

One easy way to understand this concept is to imagine that we draw a simple random sample from the population. If we randomly delete a few observations' values of one complete variable from our sample, the remaining sample is still considered a simple random sample. Theoretically, our sample is still the same as the sample without missing data, except it has a smaller sample size for that variable with missing data. This means that the estimates of our incomplete sample data are not biased against complete sample data, although they are less efficient due to reduced sample size.

**Missing at random (MAR):** Let  $Y_{obs}$  denote the observed components or entries of  $Y$ , and  $Y_{mis}$  the missing components. If missingness depends only on the components  $Y_{obs}$  of  $Y$  that are observed, and not on the components that are missing, then the missing data are called missing at random (MAR):

$$f(R|Y, \phi) = f(R|Y_{obs}, \phi) \text{ for all } Y_{mis}, \phi, \quad (2.2)$$

Now, suppose we stratify our sample data into several groups according to one or a few of complete variables. Then, we randomly delete some observations' values of another complete variable from one<sup>4</sup> of the groups. Again, we can consider these deleted observations as our missing data. Clearly, our missing data is MCAR within that particular group. However, for the whole sample data, there is only one group of data that has missing data. This means missingness is somehow related to the groups. If the computation of the sample estimates does not incorporate the fact that missingness depends on the variables that forms the groups, the result could be biased.

**Not Missing at random (NMAR):** The mechanism is called not missing at random (NMAR) if the distribution of  $R$  depends on the missing values in the data matrix  $Y$ :

$$f(R|Y, \phi) = f(R|Y_{mis}, Y_{obs}, \phi) \text{ for all } Y, \phi, \quad (2.3)$$

---

<sup>4</sup>could be more than one group. We used one here only for demonstration purposes

Let us still use the above example. This time we first divide the sample into several groups according to the characteristics of the complete variable before deleting a few observations' values from it. Then, we randomly delete a few observations' values of that variable within one of the several groups. However, the information about the groups is not included in the final dataset. Hence, we have no idea which group has the missing values by looking at the final dataset. This is NMAR. Equivalently, for the MAR example, if we delete the information of the groups and the variables that have been used to stratify our sample data, then MAR becomes NMAR.

## 2.2 Creating Missing data in a Synthetic Unit Record File

In this section, incomplete data are generated based on complete synthetic unit record data. The generations are designed to create cases of MCAR, MAR, and NMAR. Because we have the complete data set, it is possible for us to compare the estimates from the incomplete datasets which are generated from different missing data mechanisms with the estimates from the complete data set. The purpose of this comparison is to assess the impact of different kinds of missing data on the sample estimates.

### 2.2.1 Description of the SURF dataset

The SURF dataset we use is a synthetic unit record file based on the Statistics New Zealand income survey(Statistics New Zealand 2011). The SURF is a complete dataset. It has 200 observations and 8 variables: PersonID, Gender, Highest qualification, Age, Weekly working hours, Weekly Income, Marital status, and Ethnicity. PersonID is the identification variable. Age, Weekly working hours and Weekly Income are numeric variables. The rest are categorical variables. Table 2.1 describes the frequency counts of SURF's categorical variables, and Figure 2.2 to Figure 2.4 show the distribution of the SURF's numerical variables.

Table 2.1: Frequency table of SURF's categorical variables

Gender	Male	Female		
	93	107		
Highest qualification	None	School	Vocational	Degree
	39	66	67	28
Marital status	Never	Married	Previously	Other
	88	70	21	21
Ethnicity	Maori	Pacific	European	Other
	24	7	156	13

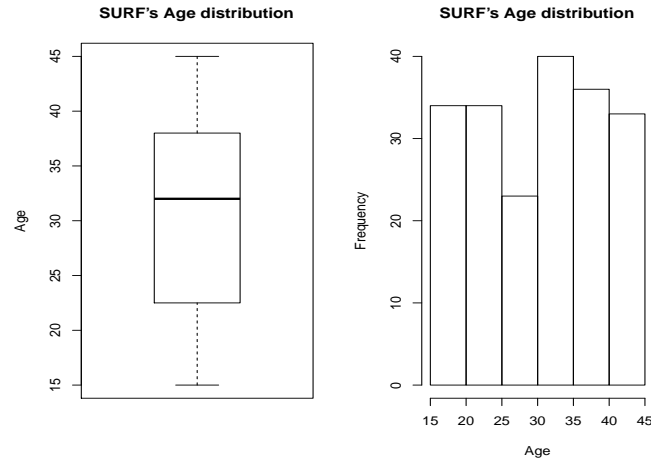


Figure 2.2: SURF's age distribution

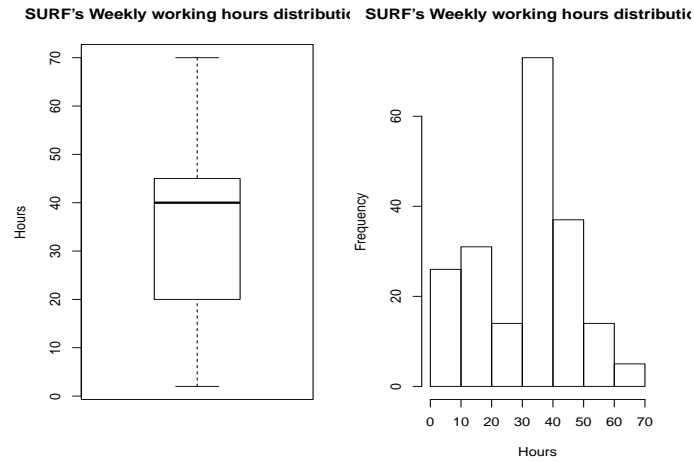


Figure 2.3: SURF's weekly working hours distribution

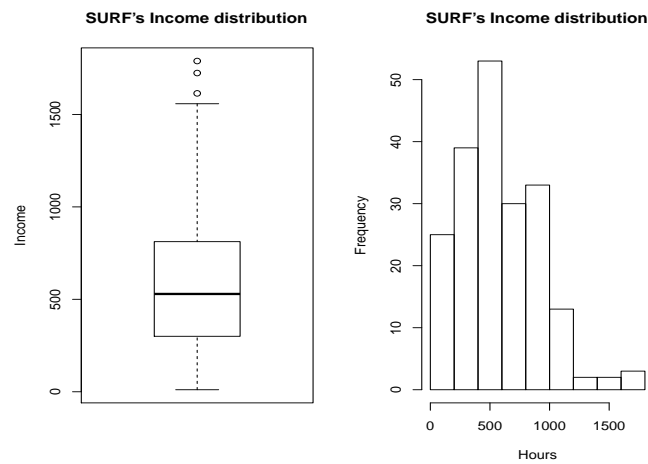


Figure 2.4: SURF's weekly income distribution

## 2.2.2 Creating the missing SURF income values

For simplicity and demonstration purposes, we only create missing data for the SURF's income variable. This makes it a univariate missing data pattern.

Table 2.2 shows estimates of weekly Income variable from the complete<sup>5</sup> SURF data. The estimates are: means, variances and confidence intervals (95%) for the means. The estimates have also been broken down by gender. We refer these estimates as “true” estimates<sup>6</sup> in our latter sections and chapters. We also consider the SURF as a simple random sample (SRS) from the population aged 16 – 45.

Table 2.2: Estimates for complete SURF

	Mean	Variance	Confidence Interval (95%)
All	575.36	120137.60	527.32–623.40
Female	438.50	73568.97	400.91–476.10
Male	732.82	128253.30	683.18–782.45

### Missing completely at random (MCAR)

**Method:** We can indicate the missing data with an indicator variable for each unit, so missing data can be created by drawing a Bernoulli random variable  $r_i$  with probability  $p$ . This method creates MCAR missing data. Mathematically, suppose we have  $n$  observations, for each observation  $i, i = 1, \dots, n$ , we draw an indicator number  $r_i$  from the Bernoulli distribution with a probability  $p$ , if  $r_i = 0$ , a missing value is assigned to the unit  $i$  for the “Income” variable.

$$r_i \sim \text{Bernoulli}(p) \quad (2.4)$$

One thousand replicate SURF datasets with missing data for the income variable were generated by using this method. Estimates have been calculated for each replicate dataset.

The means and variances are calculated as follows:

$$\hat{Y} = \frac{\sum_{i=1}^n r_i y_i}{\sum_{i=1}^n r_i} \quad (2.5)$$

$$\text{var}(\hat{Y}) = \frac{\sum_{i=1}^n r_i (y_i - \hat{Y})^2}{\sum_{i=1}^n r_i - 1} \quad (2.6)$$

where

$$r_i = \begin{cases} 0 & \text{if } y_i \text{ is missing} \\ 1 & \text{if } y_i \text{ is not missing} \end{cases}$$

and  $\hat{Y}$  is the estimated mean.

Histograms and density plots (Figure 2.5) show the set of one thousand means and variances. The red vertical dashed lines are the true means and variances.

<sup>5</sup>Throughout this thesis, I use “complete” to describe the original complete dataset which has no missing data. This is different from the imputed complete dataset which has been used in other literature.

<sup>6</sup>Technically, these estimates should be the sample estimates. “True” estimates are often used to refer the population estimates. The term “true” estimate is used here because we want to compare the complete data estimates with the imputed data estimates.

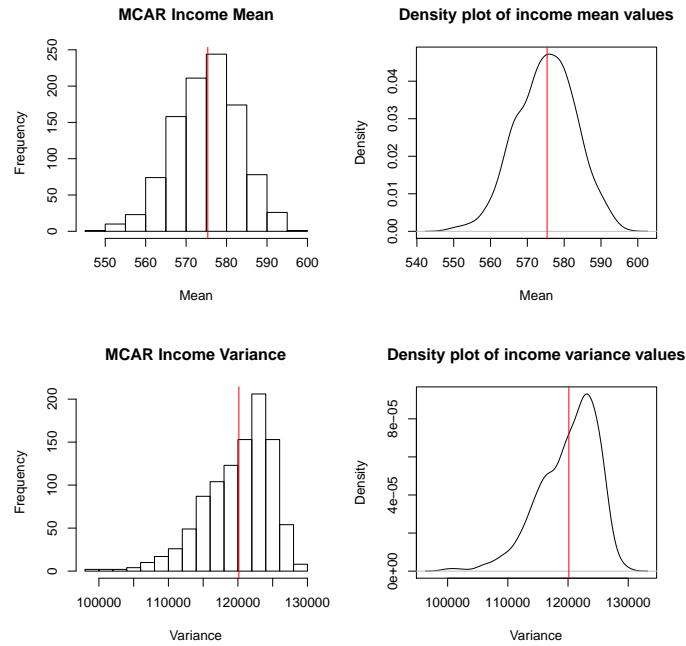


Figure 2.5: Missing SURF Income (MCAR)

Here below is the R program which can be used to create MCAR missing data for any variables. This program uses the R function “sample” which implements the process of drawing a Bernoulli random variable. We have used it to create 50 missing income values, e.g. MCAR(SURF, 50, “Income”).

```
#MCAR
#Apply missing completely at random mechanism to create missing data
#this program can produce different numbers of MCAR data for more than
#one variables
#dat-input dataset, nmissing-number of missing data, variables-variables with
#missing data, same-create missing data for the same observations of more than
#one variables
MCAR=function(dat,nmissing,variables,same=FALSE){
  n=nrow(dat)
  nvar=length(variables)
  if (length(nmissing)<nvar & length(nmissing)==1){
    nmissing_temp=rep(nmissing,nvar)
  }else{
    nmissing_temp=nmissing
  }
  if (same==F){
    for (i in 1:nvar){
      idx=sort(sample(n,nmissing_temp[i],replace=F))
      dat[idx,variables[i]]=NA
    }
  }else{
    idx=sort(sample(n,nmissing,replace=F))
    dat[idx,variables]=NA
  }dat }
}
```

As seen in Figure 2.5, the means and variances of the simulated missing data for the income variable are distributed around the true mean and variance which are shown by the red vertical dashed lines. These results confirm that the estimates of MCAR incomplete sample data are not biased.

Figure 2.5 also shows that the distribution of income variances of the replicate SURF datasets is skewed. This is related to the distribution of the Income variable. Figure 2.4 tells us that there are a few very high income observations (outliers) in the data, but there are no extremely low income observations. Hence, removing those very high income observations greatly reduces the variance, on the other hand, the removal of some of the lowest income values has little impact on the variance. Because we randomly sample observations as our missing data with equal probability, those few extremely high income observations have little chance to be chosen as missing, comparing to the majority of observations with normal range income values. This is why we have many high variances, but few low variances.

However, the distribution of simulated means is symmetric and bell shaped. This is because the outliers have smaller impact on the mean than on the variance. Equation (2.5) shows that the mean of incomplete income variable has a linear relationship with  $y_i$ , but the variance of incomplete income variable has a quadratic relationship with  $y_i$ , according to equation (2.6).

We have broken down each of our simulated datasets by sex, and shown the separate mean and variance estimates for male and female in Figure 2.6 and 2.7. These plots show the results of means and variances of 1000 simulated replicate datasets. Again, the vertical lines represent the complete data means and variances for male and female groups. These results provide further evidence that MCAR sample data does not deviate from the complete data estimates. The reason is that MCAR does not introduce bias to the sample. Hence, we see that the estimates of replicate datasets are centred around the true estimates.

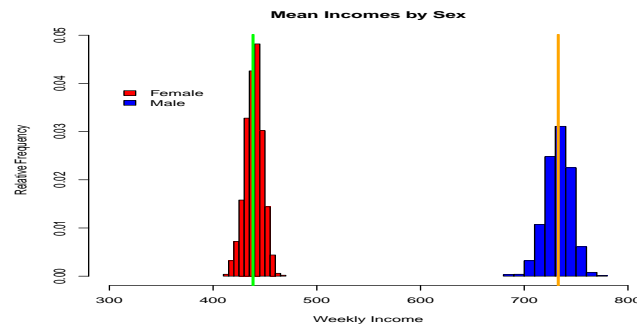


Figure 2.6: Mean of Income by sex (MCAR)

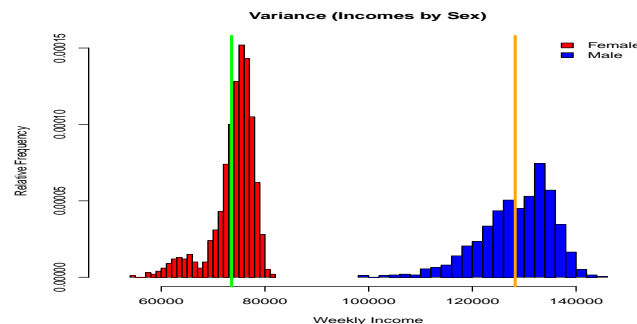


Figure 2.7: Variance of Income by sex (MCAR)

## Missing at random (MAR)

We assume the missingness of Income is related to other variables. For example: suppose that male respondents are less likely to give their income details in a survey. Then, the missing data for the income variable are no longer MCAR, because male respondents are less likely to respond than female respondents. However, the missing pattern is still random within the male group. This is called Missing at random (MAR)

Let's assume male respondents have higher probability of missing "Income" than female respondents. We created missing Income data by this probability. One thousand simulated replicate datasets were also created. Mathematically, suppose we have  $n = 200$  units, for each gender group, we draw a number of indicator  $r_i$  from the Bernoulli distribution with a probability  $p_i$  (the probabilities are different for different gender). Units in each gender group with indicator  $r_i = 0$  are treated as having missing Income values. For example:

$$r_i \sim \text{Bernoulli}(p_i) \quad (2.7)$$

where

$$r_i = \begin{cases} 0 & \text{if } y_i \text{ is missing} \\ 1 & \text{if } y_i \text{ is not missing} \end{cases}$$

and

$$p_i = \begin{cases} 0.5 & \text{if Male} \\ 0.2 & \text{if Female} \end{cases}$$

Here below is the R program which can be used to create the MAR missing data for any variables. We have used it to create MAR missing data for income variable, e.g. MAR(SURF,"Gender", "Income", c(0.5,0.2))

```
#MAR -- Apply missing at random mechanism to create missing data
#Only can be used for single variable
#group-divide data into groups by its own categorical variable
#pmiss-probability of being missing for different groups
MAR=function(dat,group,variable,pmiss){
  group_temp=levels(dat[,group]) #assign probability of missing to each group
  for (i in 1:length(group_temp)){ pmiss[group_temp[i]]=pmiss[i]}
  #get rid of extra columns which are the products of above for loop
  pmiss=pmiss[-(1:length(group_temp))]
  dat= by(dat,dat[,group],
    function(msub){
      g=as.character(msub[,group][1])
      idx=(rbinom(nrow(msub),1,pmiss[g])==1)
      msub[idx,variable]=NA
      return(msub)
    })
  d=dat[[group_temp[1]]]
  for (i in 2:length(group_temp)){
    d= rbind(d,dat[[group_temp[i]]]) }
  d=cbind(d, randid=runif(nrow(d)))
  dat=d[order(d$randid),]
  dat
}
```

Figure 2.8 clearly indicates that most of means and variances of the 1000 simulated MAR datasets are very different from the true mean and variance. These estimates are biased because the male group has more missing data than the female group. We also noticed that most of the incomplete data estimates are less than the true estimates. This is because males earn more income than females. If more male respondents were deleted than female respondents, then the estimates of incomplete data would tend to be smaller than the true estimates. Hence, we have learnt that treating MAR missingness as if it were MCAR missingness leads to bias

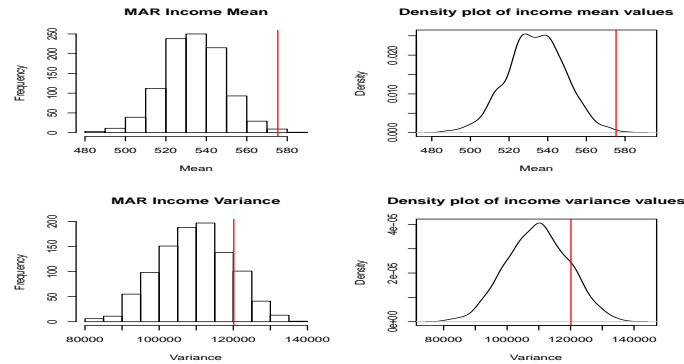


Figure 2.8: Missing SURF Income (MAR)

Figure 2.9 and Figure 2.10 show us that the estimates of male and female of MAR datasets do not deviate systematically from the complete data estimates. This is because missing data is still MCAR within each gender group. This means the within gender estimates are unbiased. We can therefore fully adjust for the missingness and conduct unbiased population estimates by incorporating the variables that are related to the missingness.

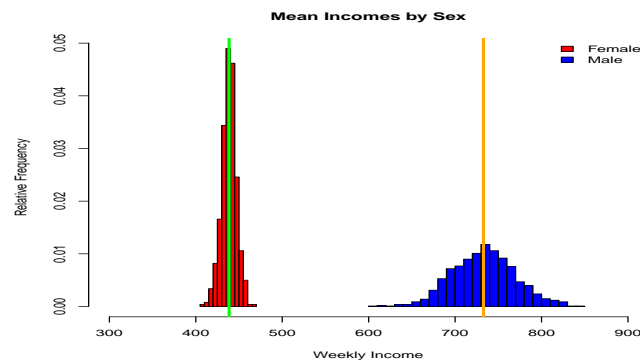


Figure 2.9: Mean of Income by sex (MAR)

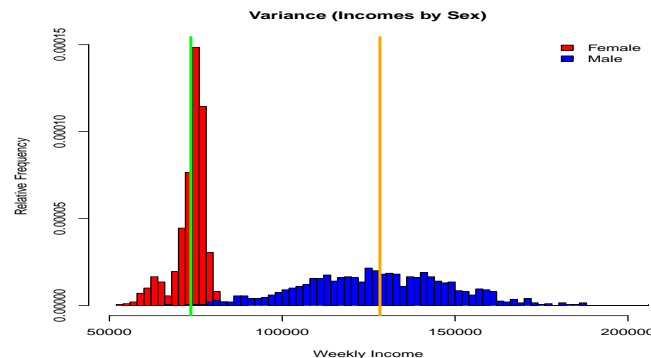


Figure 2.10: Variance of Income by sex (MAR)



### Not missing at random (NMAR)

Now, if we assume the missingness of Income variable is related to income itself, then this is called not missing at random. For example: high income male respondents are less likely to give their income detail in a survey.

We create missing Income values probabilistically. We set high income male respondents to have the most missing Income values. One thousand simulated replicate SURF datasets were created as previously. Mathematically, suppose we have  $n = 200$  observations, for each observation  $i$ , we draw an indicator  $r_i$  from the Bernoulli distribution with a probability  $p_i$ . If  $r_i = 0$ , the income value is missing for the  $i$ th unit. The  $p_i^0$  for male group is higher than the female group (0.6 : 0.1), where  $p_i^0$  is part of  $p_i$ . The  $p_i$  is the combination of  $p_i^0$ , and the fraction  $\frac{0.4y_i}{2000}$ . The fraction makes  $p_i$  higher for the high income observations than low income observations.

$$r_i \sim \text{Bernoulli}(p_i) \quad (2.8)$$

where

$$p_i^0 = \begin{cases} 0.6 & \text{if Male} \\ 0.1 & \text{if Female} \end{cases}$$
$$p_i = p_i^0 + \frac{0.4y_i}{2000} \quad (p_i \leq 1)$$

Here is the R program which creates the NMAR missing data for any variables. We have used it to create NMAR missing data for income variable, e.g NMAR(SURF, "Gender", "Income", c(0.6, 0.1)).

```
#NMAR
NMAR=function(dat,group,variable,pmiss,ph=0.4,nh=2000){
  pmiss=pmiss[dat$group]
  # make this higher for higher incomes
  pmiss = pmiss + ph*(dat$variable/nh)
  pmiss=pmin(1,pmax(0,pmiss))
  # now toss the coin for each row
  idx <- as.logical(sapply(pmiss, function(p) rbinom(1,size=1,prob=p)))
  # anything with idx==1 will be missing
  dat[,variable][idx] <- NA
  dat
}
```

Figure 2.11 shows that the simple estimates of means and variances of 1000 replicate NMAR SURF datasets are concentrated far away from the true mean and variance. It also shows that most of the means and variances are much less than the true mean and variance. The reason for this is that after losing more male and high income observations, the estimates become much smaller than the estimates of the complete data.

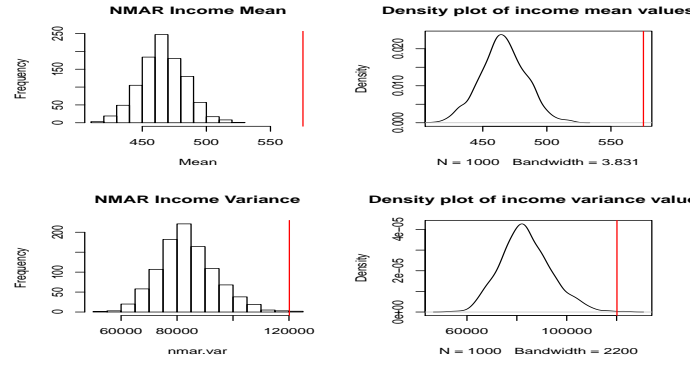


Figure 2.11: Missing SURF Income (NMAR)

Figure 2.12 and Figure 2.13 also show that the means and variances of 1000 simulated NMAR SURF datasets breakdown by sex deviate from the true mean and variance. The means and variances of male group are higher than the means and variances of female group. In addition, they are both less than the true estimates. This further confirms that high income observations are more than likely to be male, and after losing some high income male observations, the estimates becomes much less than the complete data estimates.

Since missingness is related to the outcome variable, we cannot create unbiased estimates in either “Gender” subgroups or the total population. NMAR missingness is much more problematic than MCAR or MAR missingness.

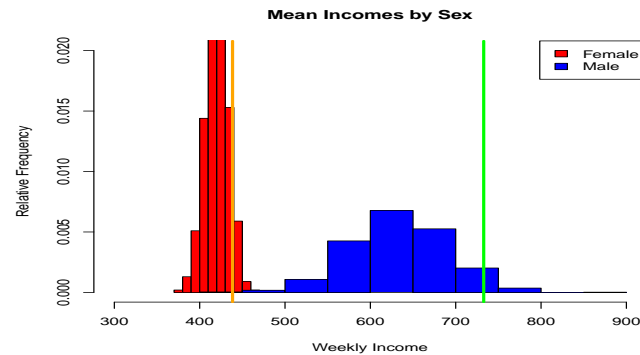


Figure 2.12: Mean of Income by sex (NMAR)

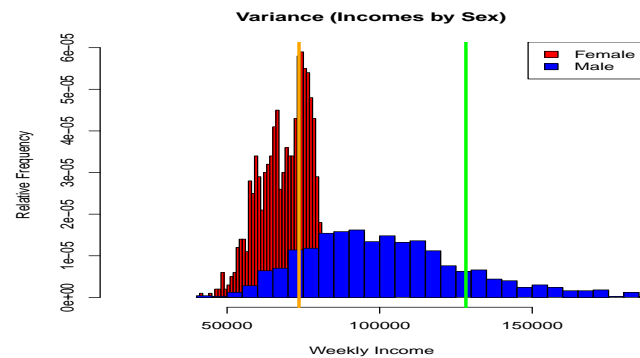


Figure 2.13: Variance of Income by sex (NMAR)

## 2.3 Conclusion

In this Chapter, we have introduced the concept of missing data for survey samples, and the missing mechanisms which lead to missing data. If the missing mechanism is MCAR, we have seen that our incomplete data estimates are unbiased. If the missing mechanism is MAR, we have seen that the overall incomplete data estimates are biased. This is because one or some groups of the observations drive the estimates apart from the complete data estimates. However, we found that the incomplete data estimates which incorporate the variables which determine the pattern of missingness are not biased against the complete data estimates. This is because for each individual groups the missing mechanism is actually MCAR. If the missing mechanism is NMAR, we have shown that both the overall estimates and the breakdown estimates by each groups are biased against the complete data. This is because some groups with particular characteristics are missing and we cannot find any groups with similar characteristics in the observed data.

This thesis does not discuss the impact of the proportion  $p$  of missing data on the bias of estimates. Scheffer (2002) has demonstrated that the impact on estimates is small even when the proportion of missing data is up to 50%, if the missing data mechanism is not NMAR. However, according to Scheffer (2002) simulations, if the missing data mechanism is NMAR, the impact on the estimates can be substantial, if there is a large proportion of missing data. Hence, this thesis focuses on investigating missing data handling methods which can reduce the impact on the bias of estimates from the missing data mechanisms, assuming MAR missingness.

# Chapter 3

## Dealing with missing data

### 3.1 Introduction

When people encounter missing data in their datasets, the easiest approach is to get rid of them. The easiest and quickest method we can think of is to remove the missing data directly by discarding all data for any unit that has missing information. However, it has been shown that missing data cause bias for statistical estimates if the missingness is not MCAR (Chapter 2). This bias is called non-response bias. We would like to deal with this non-response bias before conducting any further statistical analysis. Apparently, deletion is not a good method of reducing non-response bias, although some deletion methods which are associated with a reweighting method can reduce bias when dealing with unit non-responses. If we do not delete missing data and leave them in the dataset, then some analysis cannot be performed because they depend on the data being complete. Furthermore, in most cases, missingness is not MCAR. This means missingness causes bias.

If you cannot delete or ignore the missing data, then it is reasonable to think a way of filling in the missing data with some “guessed” values. This is called imputation. The apparent advantage of imputation is that you get a balanced data set which can feed into any statistical software. If your fill in values are close to the true missing values, imputation can also reduce the non-response bias. However, if you only fill in the missing data once and treat them as if they were true, then you have introduced another source of estimation error - the uncertainty introduced by non-response. This uncertainty is referred as “imputation uncertainty”. This may mean your parameter estimates are biased and it also means that standard errors are underestimated due to apparently larger sample size (the complete data sample size) than the actual sample size (the responding sample size). Imputation methods which only impute missing data once are classified as “single imputation” methods in most statistical literature. To cope with the drawbacks of single imputation, Rubin (1977) developed a method called “multiple imputation”. The simplest and most intuitive explanation of multiple imputation is that missing data are filled in multiple times, so researchers can use the multiple imputed missing values to estimate the imputation uncertainty. In the following sections of this Chapter, we briefly introduce some of the common methods of data deletion and imputation. The following chapters will give more detailed discussion.

### 3.2 Dealing with missing data I: Data Deletion Methods

One of the simplest and most commonly used methods for dealing with missing data is data deletion procedure (McKnight, M.McKnight, Sidani & Figueredo 2007). Schafer & Graham

(2002) identify four main data deletion methods. They are: (1) listwise deletion, (2) pairwise deletion, (3) reweighting, and (4) averaging the available items. Data deletion is often not the ideal solution for handling missing data, because it may introduce large bias as shown in Chapter 2. However, it can still provide suitable parameter estimates when its methods have been used appropriately. The advantage of data deletion methods is that they are very easy to be implemented. To be clear, the deletion methods can only be used appropriately under the assumption that we have MCAR data, or the variables upon which missingness depends are known and observed fairly.

### 3.2.1 Listwise Deletion

The most popular data deletion method is to discard units which have missing data. In other words, drop all units with any missing data for any of their variables. This method is called listwise deletion (LD) or complete-case analysis. LD is a default routine for most statistical software. According to Little & Rubin (2002), Schafer & Graham (2002), listwise deletion is appropriate if data are MCAR. As we have demonstrated in the previous Chapter, if the missing mechanism is MCAR, the incomplete data estimates are not biased against the complete data estimates. Hence, LD can be performed if data are MCAR. Figure 3.1 demonstrate how listwise deletion works. The double question marks stand for the missing data, the two horizontal lines which strike through the centre of the cells mean that the whole records for those units are deleted. Figure 3.1 also clearly displayed the downside of LD even

Gender	Income	Ethnicity	Qualification
	??		
??		??	

Missing data

Figure 3.1: Listwise Deletion: units with missing data are deleted

if data are MCAR. As we can see, the LD method tends to delete more data than we want. That is, if an unit has even only one variable with missing data, the unit is deleted from the data. Obviously, this process deletes observed values of other variables of that unit. Hence, LD reduces the efficiency of estimates from survey data.

We express LD mathematically. Suppose we have  $n$  units with  $K$  variables, so  $y_{ij}$ , where  $i = (1, \dots, n)$  and  $j = (1, \dots, K)$ , represent the value of cell  $ij$ , and we use indicator  $R_i$  to indicate rows with complete responses. Hence:

$$R_i = \begin{cases} 0 & \text{if any } y_{ij} \text{ is missing, for } j = 1, \dots, K \\ 1 & \text{if no } y_{ij} \text{ is missing, for } j = 1, \dots, K \end{cases}$$

For the mean of  $Y_j$ ,

$$\bar{Y}_j = \frac{\sum_{i=1}^n R_i y_{ij}}{\sum_{i=1}^n R_i}$$

For the covariance of  $Y_j$  and  $Y_k$ ,

$$\hat{Cov}(Y_j, Y_k) = \frac{\sum_{i=1}^n R_i (y_{ij} - \bar{Y}_j)(y_{ik} - \bar{Y}_k)}{\sum_{i=1}^n R_i}$$

### 3.2.2 Pairwise Deletion

Pairwise deletion (PD) or available-case analysis method is in contrast to LD. Instead of deleting units with missing data, it chooses to discard data only at the level of the variable (McKnight, M.McKnight, Sidani & Figueredo 2007). Obviously, PD preserves all units. For example, we may use every observed value of  $Y_j$  to estimate the mean on  $Y_j$ , and every observed pair of values  $(Y_j, Y_k)$  to estimate the covariance of  $Y_j$  and  $Y_k$ . However, we might end up with a different number of units for each variable. The differing numbers of units affects the stability of the estimates while also affecting the characteristics of the correlation matrix (McKnight, M.McKnight, Sidani & Figueredo 2007). Figure 3.2 demonstrate how pairwise deletion works. The double question marks are the missing data, and the crosses in some cells mean that those cells have been deleted.

Gender	Income	Ethnicity	Qualification
	??		
??		??	

Missing data

Figure 3.2: Pairwise Deletion: Missing data are deleted and not shown in the final dataset

We express PD mathematically. Suppose we have  $n$  units with  $K$  variables, and we use indicator  $r_{ij}$  where  $i = (1, \dots, n)$  and  $j = (1, \dots, K)$  to indicate cells with observed data. Hence:

$$r_{ij} = \begin{cases} 0 & y_{ij} \text{ is missing} \\ 1 & y_{ij} \text{ is not missing} \end{cases}$$

For the mean of  $Y_j$ ,

$$\bar{Y}_j = \frac{\sum_{i=1}^n r_{ij} y_{ij}}{\sum_{i=1}^n r_{ij}}$$

For the covariance of  $Y_j$  and  $Y_k$ ,

$$\hat{Cov}(Y_j, Y_k) = \frac{\sum_{i=1}^n r_{ij} r_{ik} (y_{ij} - \bar{Y}_j)(y_{ik} - \bar{Y}_k)}{\sum_{i=1}^n r_{ij} r_{ik}}$$

### 3.2.3 Reweighting

As discussed, LD and PD methods introduce bias in non-MCAR cases, but it is possible to reduce the bias by applying a weighting class adjustment method. This method is called reweighting in Schafer & Graham (2002). The basic idea is to re-weight the observed units in order to represent the full sample or population after removing units with missing data. Please note that reweighting is more suitable for unit non-response than item non-response in practice. This is because a dataset normally only has one set of weights instead of creating a set of weights for each variable. Also, the reweighting method cannot handle a function of more than one variable unless both are missing together.

Sample can be considered to be a miniature version of the population of interest. The weights can be used to scale up the sample estimates to the population estimates. This is because a weight  $w_i$  for a sample unit  $i$  represents that there are  $w_i$  similar units in the population. For example, if the sample size is  $n$ , then an estimate of the population total is  $\sum_{i=1}^n w_i y_i$ , where  $y_i$  is the sampled value for unit  $i$ . Now, there are only  $r$  respondents in our sample out of the sample size  $n$ . This means that our actual sample size is  $r$  instead of  $n$ . Hence, it is reasonable to think that if the weights  $w_i$  can be adjusted somehow according to the “new” sample size  $r$ , we effectively adjust weights for the non-response.

If the probabilities of selecting a unit from a target population is  $\pi_i$ , and the probabilities of response for each responding unit  $i$  is  $\phi_i$ , then:

$$P(\text{unit } i \text{ selected in sample and responds}) = P(\text{selection}) \times P(\text{response}|\text{selection}) = \pi_i \phi_i$$

Hence, the final adjusted weights for non-response is  $\tilde{w}_i = 1/(\pi_i \phi_i)$ .

So far, we have only considered the general case of non-response weighting adjustment. If our missing data is MAR, we can apply the weighting class adjustment methods to adjust for non-response. We know that MAR means that our missing data is related to some other variables. Therefore, we can divide the variable with missing data into several classes according to the other related variables. These classes are called “weighting classes”(Lohr 1999). It is assumed that the probability of response is to be the same within each weighting class. The weight for a respondent in weighting class  $c$  is  $1/(\pi_i \hat{\phi}_c)$ .

To estimate the population total using weighting-class adjustments, let  $k_{ci} = 1$  if unit  $i$  is in class  $c$ , and 0 otherwise. Then the new weight for respondent  $i$  is:

$$\tilde{w}_i = \sum_c \frac{w_i k_{ci}}{\hat{\phi}_c}$$

where  $w_i = 1/\pi_i$ .

Little & Rubin (2002) classify “Reweighting” method as data deletion methods. This is because it seems that the reweighting method discards missing data completely as the listwise deletion and pairwise deletion methods, and the reweighting method also shares the same dilemmas as those two deletion methods, such as losing observed values for some variables or creating unbalanced datasets. However, we tend to think the reweighting method as some kind of imputation method. As stated previously, a weight  $w_i$  stands for  $w_i$  units which are similar to the sample unit  $i$  in the population, reweighting increases  $w_i$  to count the number of non-responses, so the new weight  $\tilde{w}_i$  represents the non-respondents as well. This can be considered as replacing the missing data with the observed data which fundamentally is the same concept as imputation which we talk about in section 3.3.

### 3.2.4 Averaging the Available Items

Many variables are not able to be measured directly by survey questions, such as self-esteem, intelligence, depression, and anxiety. These variables are normally derived variables. They can only be approximated by combining different survey questions. Usually, we tend to create a scale to find derived variables by averaging the items. If there are missing items, we average the remaining items. This is equivalent to replacing the missing items with the mean of observed items (Schafer & Graham 2002). However, we need to point out that this method only works if we have sufficient item responses, and equivalent items.

Mathematically, suppose we have  $n$  units with  $K$  variables,  $r$  responses. The derived variable  $Y = (y_i)$ , where  $i = (1, \dots, n)$ , can be estimated by a vector of variables  $X = (x_{ij})$ , where  $i = (1, \dots, n)$  indicates the units, and  $j = 1, \dots, K$  indicates the variables. Then:

$$y_i = \frac{\sum_{j=1}^K R_{ij} x_{ij}}{\sum_{j=1}^K R_{ij}} \quad (3.1)$$

where the  $R_{ij}$  is the indicator of missing data:

$$R_{ij} = \begin{cases} 1 & x_{ij} \text{ is missing} \\ 0 & x_{ij} \text{ is not missing} \end{cases}$$

## 3.3 Dealing with missing data II: Imputation

Missing data often reduce the usability of data. Imputation methods are designed to handle missing data problems. Imputation can reduce the non-response bias and produce datasets without “holes” (missing data). The basic idea is to impute (fill in) the values of items that are missing by replacing values from other respondents in the survey who are similar to the item non-responses on other variables (Lohr 1999). There are a variety of imputation methods that deal with missing data. These imputation methods can be used to impute one value for each missing item (single imputation) or, in some cases, to impute more than one value in order to reflect imputation uncertainty (multiple imputation) (Little & Rubin 2002).

### 3.3.1 Single Imputation Methods

Little & Rubin (2002) classify single imputation into two generic approaches: Explicit modelling and Implicit modelling.

#### Explicit modelling

The imputed missing values are generated from formal statistical models. Hence, the results are explicit. Here are some explicit modelling examples:

#### (1) Mean imputation

One of the simplest imputation methods is to replace all the missing values with the mean of the observed values for the numerical variable. There are two mean imputation methods. One is called “unconditional mean imputation”. The other one is called “conditional mean imputation”



**Unconditional mean imputation:** This method simply replaces missing values with the mean of all the observed values for the variable. It is very easy but it also is very likely to distort the distribution for the variable. Normally, this method is not recommended. Let us define  $\bar{Y}$  as the mean of our variable  $Y$  with missing values. For  $n$  units, we have:

$$\bar{Y}_k = \frac{1}{C_k} \sum_{j=1}^n R_{jk} Y_{jk} \quad (3.2)$$

where

$$C_k = \sum_{j=1}^n R_{jk}$$

$$R_{jk} = \begin{cases} 1 & \text{if item } k \text{ is not missing for unit } j \\ 0 & \text{Otherwise} \end{cases}$$

**Conditional mean imputation:** This is an improvement of unconditional mean imputation. It simply divides data into several groups or strata based on fully observed variables or auxiliary data. Then, means of variables with missing data will be calculated for each stratum. The missing values can be replaced by the means of their stratum. Compared to unconditional mean imputation, this method can preserve the distribution of the variable. Let us define  $\bar{Y}_g$  as the mean of our variable  $Y$  with missing values for group  $g$ . For  $n$  units, we have item  $k$ :

$$\bar{Y}_{gk} = \frac{1}{C_{gk}} \sum_{j=1}^n R_{jk} I_{jg} Y_{jk} \quad (3.3)$$

where

$$C_{gk} = \sum_{j=1}^n R_{jk} I_{jg}$$

$$R_{jk} = \begin{cases} 1 & \text{if item } k \text{ is not missing for unit } j \\ 0 & \text{Otherwise} \end{cases}$$

$$I_{jg} = \begin{cases} 1 & \text{if unit } j \text{ is in group } g \\ 0 & \text{Otherwise} \end{cases}$$

Then

$$\bar{Y}_k = \sum_g W_g \bar{Y}_{gk}$$

where

$$W_g = \frac{n_g}{n}$$

$$\sum_g W_g = 1$$

## (2) Regression imputation

A statistical model is established by using observed data. Then the model can be used to predict the missing values. Conditional mean imputation can be considered a special case of regression imputation. Suppose a random variable  $Y$  has density  $f(Y|X, \theta)$  for random variable  $X$ , given the  $X$  is observed. We can estimate  $\theta$  from complete data  $Y_{obs}$  and  $X_{obs}$ . Then, each missing  $Y$  can be imputed independently as

$$Y_{mis,i} = E[Y_i | X_i; \hat{\theta}] \quad (3.4)$$

where  $i = 1, \dots, n$ ,  $n$  is the sample size, and  $\hat{\theta} = \hat{\theta}(Y_{obs}, X_{obs})$ .

### (3) Stochastic regression imputation

This is like the regression imputation method. The only difference is that we introduce uncertainty to the predicted missing values. The missing  $Y$  ( $Y_{mis}$ ) is randomly drawn from the distribution  $f(Y|X; \hat{\theta})$ :

$$Y_{mis,i} \sim f(Y_i|X_i, \hat{\theta}) \quad (3.5)$$

### Implicit modelling

In implicit modelling, the imputation is based on an algorithm. There are no formal statistical models. Here are some implicit modelling examples:

#### (1) Simple random imputation

This is the simplest imputation method. It simply imputes missing values of variable  $y$  from a random draw of the observed data from all observed records for this variable.

#### (2) Hot deck imputation

The basic idea of this method is to replace individual missing values drawn from the observed values of “similar” responding units. In other words, for each unit with a missing  $Y$ , find a unit with similar values of  $X$  in the observed data and take its  $Y$  value. The hot deck method can be very complex. Here is some examples of hot deck imputation.

- **Sequential hot deck Imputation:** There are hot deck imputation procedures that impute the value in the same subgroup that was last read by the computer in a single scan of the data
- **Random hot deck Imputation:** A donor is randomly chosen from the respondents with information on all missing items. It is just like the “Simple random imputation”
- **Adjustment cell hot deck imputation:** Adjustment cells are formed from the joint levels of categorical variables which have observed values for variables with missing data. Then, a donor is randomly chosen from the respondents within each adjustment cell to replace the missing data with that cell. This is a similar idea to conditional mean imputation.
- **Nearest-Neighbor hot deck Imputation:** Define a distance measure between units, and impute the value of a respondent who is “closest” to the person with the missing item, where closeness is defined using the distance function, such as the Mahalanobis distance (Andridge & Little 2010):

$$d(i, j) = (x_i - x_j)^T \widehat{Var}(x_i)^{-1} (x_i - x_j)$$

where  $\widehat{Var}(x_i)$  is an estimate of the covariance matrix of  $x_i$ .

### (3) Substitution

This method is basically to replace non responding units with alternative units not selected into the sample. The method can only be applied during the data collection stage. This method is not recommended. This is because the substituted responding units may differ systematically from non-respondents. Although we may end up with a complete dataset, the non-respondents' information is actually missing (Little & Rubin 2002).

### (4) Cold deck imputation

This method imputes missing values from historical records of a particular unit. For example, we might be able to find a value of an unit from the same survey of previous period to replace the missing value of current survey.

### (5) Imputation based on logical rules (or deductive imputation)

Sometimes we can impute using logical rules: for example, suppose a survey has questions such as “whether drinking water is free of charge to students during the school day” and “Whether drinking water is available to students through drinking water fountains”. If the answer for free drinking water is “no” and for availability of drinking water fountains is missing, it is reasonable for us to conclude that there is no drinking water fountains in the school. This is also called the deductive imputation method (Kalton 1983).

## 3.3.2 Likelihood Based Approaches

### Maximum Likelihood Estimator

Suppose a simple random sample with size  $n$  is designed to collect data for variable  $Y$ . Assuming its probability density function (pdf) is  $f_y(Y; \theta)$ ,  $-\infty < y < \infty$ . In other words, this means that the pdf of observing  $Y_1$  is  $f_y(Y_1; \theta)$ , if  $\theta$  is known. The same logic can be applied for  $Y_2, Y_3, \dots, Y_n$ . Hence, the probability of observing sample  $Y = (Y_1, Y_2, \dots, Y_n)$  altogether is the product of  $f_y(Y; \theta)$ ,  $y \in (Y_1, Y_2, \dots, Y_n)$ , or it can be expressed as below:

$$f(Y|\theta) \equiv L(\theta|y) = \prod_{i=1}^n f_y(Y_i; \theta)$$

We call the above function the likelihood function (Azzalini 1996).

The purpose of maximum likelihood estimation is to find  $\hat{\theta}$  which produces the highest probability of observing sample  $Y = (Y_1, Y_2, \dots, Y_n)$ . The  $\hat{\theta}$  is called the Maximum Likelihood Estimator. In order to make the calculation of the maximum likelihood estimator easier, we take the logarithm of the likelihood function, hence, the likelihood function is transformed into log-likelihood function. Mathematically:

$$\text{Log-Likelihood} \equiv \ell(\theta|Y) = \log(f(Y|\theta)) \quad (3.6)$$

Now, if we repeat our survey over and over again, each time we draw a sample size of  $n$ , and find the maximum likelihood estimator  $\hat{\theta}$ . Then, we can get the sampling distribution of our maximum likelihood estimator  $\hat{\theta}$ . In other words,  $\hat{\theta}$  is just an estimate and is not guaranteed to equal to the population parameter  $\theta$  in any particular sample. Hence, there is uncertainty in estimating  $\theta$  with  $\hat{\theta}$  from a sample. Fortunately, we can use the inverse of the negative expected value of the second derivative of the log-likelihood function to estimate the variance of the sampling distribution of  $\hat{\theta}$ . Mathematically:

$$Var(\hat{\theta}) = I(\theta)^{-1} = \left( -E \left( \frac{d^2 LL(\theta|Y)}{d\theta^2} \right) \right)^{-1} \quad (3.7)$$

Given the expected  $\theta$  and the variance of  $\theta$ , we can find the confidence interval for  $\theta$ .

$I(\theta)$  is called the “information matrix” (Scott 2007). We use the inverse of information  $I(\theta)$  to define the means although we do not know the exact true  $\theta$ , we can find an interval which  $\theta$  lies in to some specific confidence level.

### Expectation Maximization (EM) Algorithm

We know that ML can be used to estimate distribution parameters. Given the parameters, we will be able to draw samples from the probability distribution. This could be very useful for imputing missing values. Actually, the EM algorithm is one of the methods based on ML to solve incomplete data problems. The EM algorithm is a very general iterative algorithm which formalizes an intuitive idea for obtaining parameter estimates when some of the data are missing. This *ad hoc* idea has four steps: (1) Replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values using the new parameter estimates, (4) re-estimate parameters, and so forth, iterating until convergence (Healy & Westmacott 1956). Roughly speaking the EM algorithm follows the very same idea, but instead of filling in any missing values and iterating, EM computes missing values by using the conditional expectation of the “missing data”, or in other words, the functions of “missing data”. Hence, its E step finds the conditional expectation of the “missing data” given the observed data and current estimated parameters, and then fills in the missing value; the M step is just simply performing ML estimation of  $\theta$  (Little & Rubin 2002).

Mathematically, let  $\theta^{(t)}$  be the current estimate of the parameter  $\theta$ . The E step of EM finds the expected complete-data loglikelihood if  $\theta$  were  $\theta^{(t)}$ :

$$Q(\theta|\theta^{(t)}) = \int \ell(\theta|y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis} \quad (3.8)$$

The M step of EM determines  $\theta^{(t+1)}$  by maximizing this expected complete-data loglikelihood:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \text{ for all } \theta \quad (3.9)$$

### 3.3.3 Bayes Theory and Simulation methods

The fundamental idea of Bayes’ Theorem is to provide us with the reversed conditional probability. This is extremely useful. For example, it might be easy to calculate the probability of later event, given earlier event ( $P(A=\text{later event}|B=\text{earlier event})$ ), but it could be very hard to find the probability of earlier event, given later event ( $P(B=\text{earlier event}|A=\text{later event})$ ), because it is not possible to have information of the later event. If  $P(A=\text{later event}|B=\text{earlier event})$

can be reversed, we can easily work out  $P(B=\text{earlier event}|A=\text{later event})$ . The Bayesian approach offers us this kind of possibility. Bayes' theorem states:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (3.10)$$

The theorem says that a conditional probability for event  $B$  given event  $A$  is equal to the conditional probability of event  $A$  given event  $B$ , multiplied by the marginal probability for event  $B$  and divided by the marginal probability for event  $A$  (Scott 2007).

Bayes' theorem can be applied to probability distributions. Bayes' Theorem, expressed in terms of probability distributions, appear as:

$$f(\theta|data) = \frac{f(data|\theta)P(\theta)}{f(data)} \quad (3.11)$$

where  $f(\theta|data)$  is the posterior distribution for the parameter  $\theta$ ,  $f(data|\theta)$  is the sampling density for the data which is proportional to the likelihood function,  $f(data)$  is the marginal probability of the data, and  $P(\theta)$  is the prior distribution. The  $P(\theta)$  is not usually known, but we may use a fairly flat distribution. The posterior is not very sensitive to the prior. It can be computed as:

$$f(data) = \int f(data|\theta)P(\theta)d\theta \quad (3.12)$$

In Eq (3.11), the prior knowledge of  $\theta$ , summarized by  $P(\theta)$ , is updated to the posterior distribution gives the addition of new data. Equation 3.11 is a density function of  $\theta$ , and treats the data as a parameter. As such it can be simplified as:

$$Posterior \propto Likelihood \times Prior \quad (3.13)$$

In terms of imputing missing values, Bayes' theorem offers us the opportunity to find the joint posterior distribution of  $(\theta, Y_{mis})$  given  $Y_{obs}$ . This is very powerful, because we can draw samples from the posterior distribution to replace our missing values and create updated estimates of the parameter  $\theta$ . We talk about this in detail in Chapter 7.

### A Simple Iterative Simulation method: Data Augmentation

Data Augmentation is a Bayesian iterative simulation method which can be used to solve missing data problems. It simulates the posterior distribution of  $\theta$  (Tanner & Wong 1987). We can consider data augmentation as a refinement of the EM algorithm using simulation, with the imputation (or I) step corresponding to the E step and the posterior (or P) step corresponding to the M step (Little & Rubin 2002). Basically, it includes multiple imputation to enhance the EM algorithm. Mathematically, given a starting value  $\theta^{(t)}$  of  $\theta$  drawn at iteration  $t$ :

- I Step (Imputation Step): Generate a sample  $Y_{mis}^{(t+1)}$  from the predictive density  $p(Y_{mis}|Y_{obs}, \theta^{(t)})$
- P Step (Posterior Step): Draw  $\theta^{(t+1)}$  with density  $p(\theta|Y_{obs}, Y_{mis}^{(t+1)})$

Data augmentation can be easily understood if we compare it with EM algorithm. For the EM algorithm we substituted a predicted value on the basis of the variables that were available. In data augmentation we will substitute random data (Howell 2009).

We talk about more generic Bayesian iterative simulation methods in Chapter 7.

### 3.3.4 Multiple Imputation

#### The purpose of multiple imputation

Multiple imputation (MI) (Durrant 2005) is a procedure which replaces missing values multiple times. Hence, as a result, MI produces several datasets with different imputed missing values. According to Little & Rubin (2002), normally 2 to 10 data sets are sufficient. However, Enders (2010) argues that a minimum of 20 data sets might be better for most situations. This is discussed in later chapters. However, the essential question is why we impute missing values multiple times?

The reason to impute missing values multiple times is to reflect sampling variability and uncertainty for those missing values. The idea is based on the resampling methodology. Instead of having only a single imputed value which obviously cannot reflect the variability and uncertainty of the missing or unknown, a sample of missing values has the ability to capture sampling variability and uncertainty.

The performance of multiple imputation in a variety of missing data situations has been well-studied and it has shown good results. The results show that, under MCAR and MAR assumptions, MI produces unbiased parameter estimates which reflect the uncertainty associated with estimating missing data. Furthermore, MI has been shown to be robust to departures from normality assumptions and provides adequate results in the presence of small sample size or a high proportion of missing data (Schafer & Graham 2002, Graham & Schafer 1999).

#### The multiple imputation process

The MI procedure is intuitive, (Wayman 2003) but it is also complex. For a proper MI <sup>1</sup>, Bayesian theory and Bayesian simulation methods are heavily involved. This will be explored in the later chapters. For now, a simple description of MI is that it has two main stages. Stage I: impute missing data  $D$  times. Hence, we end up creating  $D$  datasets; Stage II: derive the final set of estimates by averaging over the estimates of our datasets following a set of rules provided by Rubin (Little & Rubin 2002). Again, the formulas are given in Chapter 8.

## 3.4 Distinguishing Non-response bias and Imputation uncertainty

Throughout this chapter, we have mentioned the concepts of non-response bias and imputation uncertainty several times. In this section, let's spend some time to discuss what exactly they are, their relationship, and their differences, because these two terms can be confused, and are the foundation for understanding the concept of multiple imputation.

First of all, the name “non-response bias” is too specific. It misleads us to think bias is only caused by non-response in survey data, but the term actually can be applied to any types of missing data in any types of data, not just non-response in survey data. I personally think “missing data bias” might be more appropriate. Anyway, as we have demonstrated in Chapter

---

<sup>1</sup>There are proper and improper multiple imputation. Their formal definitions and differences are discussed in Chapter 8.

2, if missing data are not MCAR, they are likely to bias our estimates. It is hard, if not impossible, to figure out the exact size of the bias of the estimates. However, it is possible to show how the biases happen. Let's consider simple univariate survey data with missing values. We can divide the respondents and non-respondents into two groups. These two groups can be treated as strata. The respondent stratum has  $N_R$  units, and the non-respondent stratum has  $N_M$  (M for missing) units. In total, there are  $N$  units. Let  $\bar{y}_R$  be the mean of the respondent stratum, and  $\bar{y}_M$  be the mean of the non-respondent stratum, then, the overall mean  $\bar{y}$  is:

$$\bar{y} = \frac{N_R}{N} \bar{y}_R + \frac{N_M}{N} \bar{y}_M,$$

According to Lohr (1999), the bias of mean is approximately

$$bias = E[\bar{y}_R] - \bar{y} \approx \frac{N_M}{N} (\bar{y}_R - \bar{y}_M) \quad (3.14)$$

Hence, we see that the bias happens when there is a difference between the mean of respondent stratum and the mean of non-respondent stratum. The bias can be ignored if the difference between the mean of respondents and the mean of non-respondents is close to zero, or the fraction  $\frac{N_M}{N}$  is very small, which means little non-responses.

Of course, the best way to eliminate bias is to have no missing data, but we usually do not have the luxury of getting non-missing survey data since it may require too much time and resources to go back and collect data from non-respondents. Imputation offers us a cheaper and more efficient way to deal with missing data in the hope of reducing non-response bias. We have to make it clear that if the missingness is MCAR, then there is no non-response bias. The non-response bias happens when the missingness is not MCAR, and we can only reduce the bias if the missingness is MAR.

However, all imputation methods make the assumption that missing data have similar values to the entire or some particular groups of the observed data, if missing data is MCAR or MAR. If the missing data is NMAR, then we cannot even make the assumption that missing data is similar to observed data. Even if we are sure that the missing data is MCAR or MAR, the assumption of missing data's similarity with observed data is still an assumption that can not be verified. Hence, we cannot be sure that our imputed values are the true values. This "lack of sureness" is the root of "imputation uncertainty". This is somewhat like sampling error in survey statistics. The sampling error results from taking only a sample instead of the entire population. If a different sample were taken from the same population, the results, or estimates might be different from the estimates of the previous sample. Similarly, the values we impute for the missing data might be different from the true values if there were no missing data. Loosely speaking, imputation is like sampling from the population of missing data.

We know that our survey data estimates are already subject to sampling error. A sampling error is usually quantified by the standard error of a particular estimates (mean, regression parameters, etc.). The standard error can be used to calculate confidence intervals within which the true population estimates may fall. Imputation uncertainty may be thought of as the imputed values' "sampling error". There are likewise intervals within which the true values of missing data may fall. This means imputation adds extra variance to the sampling error of a survey data.

Figure 3.3 demonstrates what the imputation uncertainty means for the confidence interval of mean estimation. The graph shows what the mean and confidence would be like if the imputation method has propagated the imputation uncertainty, comparing to imputation methods which do not. The plots are not based on any data, they are just examples to help us get a tangible idea of imputation uncertainty.

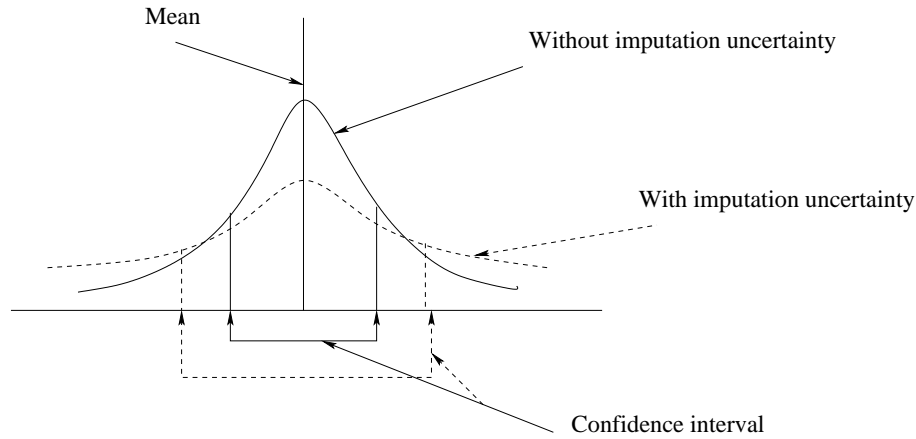


Figure 3.3: Mean and Confidence interval for imputed data: a simple demonstration of the added variance by imputation uncertainty

### 3.5 Conclusion

This Chapter simply exhibits two ways to handle missing data: data deletion and imputation. Data deletion methods are very easy to implement. It is suitable to use data deletion methods if our missing data is MCAR. If missing data is not MCAR, data deletion methods simply introduce bias and reduce the usability of the data. All imputation methods aim to increase the usability of data with missing values and reduce the non-response bias, although some of them are not easy to use, compared to data deletion methods.

Then we discussed the concepts of non-response bias and imputation uncertainty. Generally, if missing data is not MCAR, the estimates are likely to be biased. Imputation is designed to tackle the non-response bias issue. However, as we have described, imputation, in some sense, resembles sampling. As sampling error usually associates with sample survey data, imputation produces its own “sampling error”, which we may call it “imputation uncertainty”. If we say imputation uncertainty is like sampling error, does this mean we can measure imputation uncertainty as we do for sampling error? Indeed, statisticians have developed similar methodologies to measure imputation uncertainty as the measure of sampling error. In Chapter 5, we introduce resampling methods which are widely used for measuring sampling error to measure imputation uncertainty. Based on the same concept of resampling, Rubin (1977) developed multiple imputation which, comparing to single imputation methods, has a major advantage in dealing with imputation uncertainty.



# Chapter 4

## Applying single imputation methods to the SURF data

### 4.1 Introduction

In the previous chapter, we discussed several common single imputation methods. In this chapter, we apply these single imputation methods to the SURF data. The purpose of doing these applications is to enhance our understanding of these single imputation techniques and compare their advantages and disadvantages. Some of these single imputation methods are used on the large and complex FNES data in later chapters.

As described in the previous chapters, incomplete data estimates are not biased against the complete data estimates if the missing data are MCAR. Hence, imputation offers little benefit to incomplete data if missing data are MCAR. However, according to Acock (2005) and Buddhavarapu (2007), MCAR is very rare in social science research. MAR is likely to be more common than MCAR. MAR assumption is established on the knowledge or assumption that we have complete information about the variable or variables that the missingness depends on. In order to impute the proper missing values, we need to incorporate variables that the missingness depends on into our imputation methods. If the missing data are NMAR, it means we lost some groups of unit with distinct characteristics. Because all of our imputation methods are based on inference from the observed data, there is no remedy for us to retrieve the missing groups' characteristics. In the situation of NMAR, imputation methods again offers little benefit. The only case which makes imputation methods quite useful is when the missing data are MAR. Again, we have discussed in the previous chapter that the non-response bias can be reduced if the missingness is MAR. In this chapter, we demonstrate how imputation methods can help to reduce the non-response bias if the missing data are MAR.

### 4.2 Explicit modelling methods

As we have introduced before, these methods impute missing values by using formal statistical models, such as mean, regression models, etc.

#### 4.2.1 Unconditional mean imputation

Let us make some of the income variable values of the SURF data missing at random. We do this by making male income missing with a probability of 50% and female income missing with a probability of 20%. Then, throughout this chapter, we apply different imputation

methods to the SURF data with the same missing probabilities. In addition, please note that males have a higher mean income than females.

Eq.(3.2) can be re-expressed as:

$$\bar{Y}_{Income} = \frac{1}{C_{Income}} \sum_{j=1}^n R_{Income(j)} Y_{Income(j)} \quad (4.1)$$

where

$$C_{Income} = \sum_{j=1}^n R_{Income(j)}$$

$$R_{Income(j)} = \begin{cases} 1 & \text{if Income is not missing for unit } j \\ 0 & \text{Otherwise} \end{cases}$$

The following steps demonstrate how to implement Eq.(4.1). The unconditional mean imputation process has also been applied 1000 times to 1000 replicate incomplete SURF data which have the same missing probabilities. This simulation helps us to find the distribution of the estimates of the imputed data.

#### Recipe: Unconditional mean imputation

- Step 1:** Create missing income with the missing probability 50% for male and 20% for female
- Step 2:** Calculate the mean of “Income” variable without missing ”Income” values
- Step 3:** Replace missing “Income” values with the mean of “Income” variable from Step 2
- Step 4:** Estimate mean and variance of imputed “Income” variable
- Step 5:** Repeat Step 1 to Step 4 for 1000 times and record the means and variances for each imputed “Income” variable

The following is the R program which applied the unconditional mean imputation to the simulated 1000 incomplete SURF data.

```
# R program
#(a)1 -- Unconditional mean imputation - impute MAR data
Imp.mean.mar=c()
Imp.var.mar=c()
for (i in (1:1000)){
  #Create missing Income (Missing at random)
  mar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
  #unconditional mean imputation
  Mean.mar.surf=mean(mar.surf$Income,na.rm=T)
  mar.surf$Income[is.na(mar.surf$Income)]=Mean.mar.surf
  #Compute means and variances
  Imp.mean.mar[i]=mean(mar.surf$Income)
  Imp.var.mar[i]=var(mar.surf$Income)
}
```

Figure 4.1 shows the distribution of the means and variances for the 1000 imputed SURF datasets, the red vertical lines represent the true mean and variance. Comparing to Figure 2.8 which displays the means and variances for the 1000 incomplete SURF datasets, we find that the shapes of the distribution of the means of the two figures are about the same, but the variances for the unconditional mean imputed datasets tends to be less than for the simulated 1000 MAR data variances.

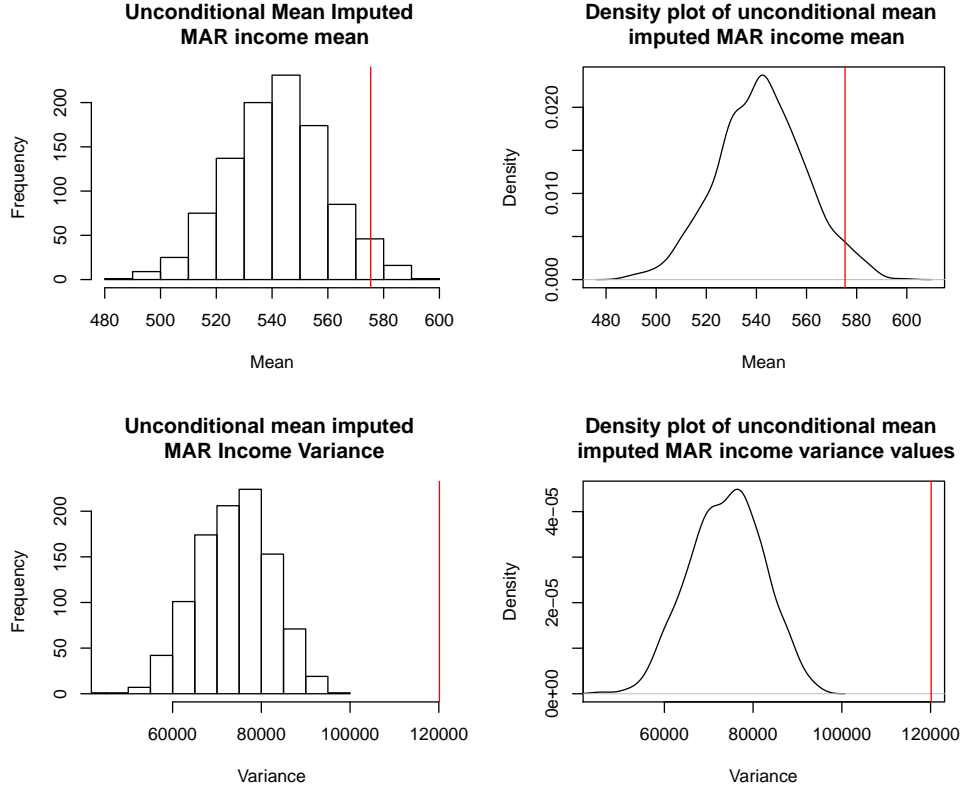


Figure 4.1: Unconditional Mean Imputed MAR SURF income

This is because imputed means has no effect on the original incomplete data mean, but they caused the imputed data variance to be smaller than the incomplete data variance. Intuitively speaking, since missing values are all replaced by a common mean, we can consider these missing values as one group and clearly its variance is zero. This group with zero variance reduces the overall variance. We can prove this mathematically. Suppose  $y_i$  are iid  $N(\mu, \sigma^2)$  where  $y_i$  for  $i = 1, \dots, r$  are observed, and  $y_i$  for  $i = r + 1, \dots, n$  are missing. The  $n - r$  missing values are replaced by one common mean  $\hat{\mu}_{mis} = \frac{\sum_{i=1}^r y_i}{r}$ , then the overall mean is:

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^r y_i + (n - r)\hat{\mu}_{mis}}{n} \\ &= \frac{\sum_{i=1}^r y_i + (n - r)\frac{\sum_{i=1}^r y_i}{r}}{n} \\ &= \frac{\sum_{i=1}^r y_i}{r} = \hat{\mu}_{mis}\end{aligned}$$

Hence, we see that the shapes of the distributions of means of Figure 2.8 and Figure 4.1 are about the same. The slight differences are cause by random creation of MAR data. However,

the variances are different. The overall variance of unconditional mean imputed data is:

$$\begin{aligned}
\hat{V}_{imputed} &= \frac{\sum_{i=1}^n (y_i - \hat{\mu})^2}{n-1} \\
&= \frac{\sum_{i=1}^r (y_i - \hat{\mu})^2 + \sum_{i=r+1}^n (y_i - \hat{\mu})^2}{n-1} \\
&= \frac{\sum_{i=1}^r (y_i - \hat{\mu})^2 + \sum_{i=r+1}^n (\hat{\mu}_{mis} - \hat{\mu})^2}{n-1} \\
&= \frac{\sum_{i=1}^r (y_i - \hat{\mu})^2 + 0}{n-1} \\
&= \frac{\sum_{i=1}^r (y_i - \hat{\mu})^2}{n-1} < \frac{\sum_{i=1}^r (y_i - \hat{\mu})^2}{r-1} = \hat{V}_{MAR}
\end{aligned}$$

where  $\hat{V}_{MAR}$  is the variance of data with missing data that are MAR. The  $\hat{V}_{imputed}$  is also less than the complete data variance  $\hat{V}_{complete}$ , where

$$\hat{V}_{complete} = \frac{\sum_{i=1}^n (y_i - \hat{\mu})^2}{n-1}$$

This is even true if the missing data are MCAR. Figure 4.2 shows the results for the simulated 1000 MCAR SURF data.

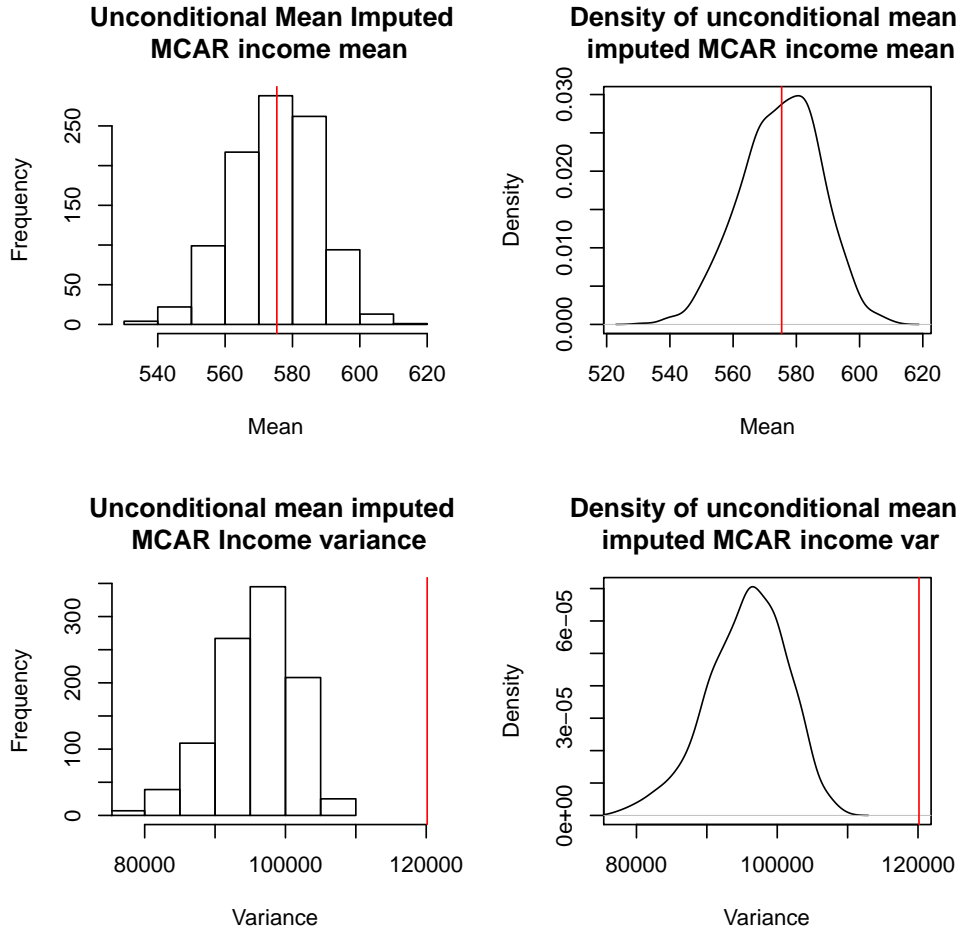


Figure 4.2: Unconditional Mean Imputed MCAR SURF income

#### 4.2.2 Conditional mean imputation (Cell mean)

Figure 4.1 also shows that the imputed means and variances are biased against the true mean and variance. This means the unconditional mean imputation cannot reduce the non-response bias. In our case, the mean of the MAR income variable is already biased against the true mean; using that mean to replace the missing income values produces the same overall mean which is biased. Now, we know that the income missingness depends on the gender variable, then we can apply conditional mean imputation to impute the missingness in the hope of reducing the bias. To be clear, conditional mean imputation reduces non-response bias if, and only if it is conditioned on the missingness depended variables. But, why do we think conditional mean imputation can reduce bias? This is because the missing values are replaced by the group means instead of the overall mean. The group means, if grouped by the missingness depended variables, is unbiased against the true group means, as Figure 2.9 shows that the income means for each gender are unbiased even if the overall income mean is biased (Figure 2.8).

For our case of missing SURF income, conditional mean imputation Eq.(3.3) can be re-expressed as:

$$\bar{Y}_{(g)Income} = \frac{1}{C_{g(Income)}} \sum_{j=1}^n R_{j(Income)} I_{jg} Y_{j(Income)} \quad (4.2)$$

where

$g = \text{Gender}$

$$C_{g(Income)} = \sum_{j=1}^n R_{j(Income)} I_{jg}$$

$$R_{j(Income)} = \begin{cases} 1 & \text{if Income is not missing for unit } j \\ 0 & \text{Otherwise} \end{cases}$$

$$I_{jg} = \begin{cases} 1 & \text{if unit } j \text{ is in group } g \\ 0 & \text{Otherwise} \end{cases}$$

The following steps demonstrate how to implement Eq.(4.2), and also simulated the conditional mean imputation process 1000 times in order to find the distribution of the estimates of the imputed data.

##### Recipe: Conditional mean imputation

- Step 1:** Create missing income with the missing probability 50% for male and 20% for female
- Step 2:** Divide observations into subgroups by variable “Gender”.
- Step 3:** Calculate the mean of “Income” variable without missing “Income” values for each subgroup
- Step 4:** Replace each subgroup’s missing “Income” values with the mean of “Income” variable for each subgroup from “Step 3”
- Step 5:** Estimate mean and variance of imputed “Income” variable
- Step 6:** Repeat “Step 1” to “Step 5” for 1000 times and record the means and variances for each imputed “Income” variable

```

# R program
#(a)1b -- Conditional mean imputation
# Condition on Gender.....
Imp.mean.mar.Con=c()
Imp.var.mar.Con=c()
Condition=c("Gender")
add.condition=function(x){
  if (is.null(ncol(x))) x
  else do.call("paste",c(x,sep=""))
}

for (i in (1:1000)){
  #Create missing at random Income values
  mar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
  #Compute mean for each group
  qual.mean=tapply(mar.surf[which(mar.surf$Income!="NA"),"Income"],
    add.condition(mar.surf[which(mar.surf$Income!="NA"),Condition]), mean)
  #Replace missing Income with computed mean for each group
  mar.surf$Income[is.na(mar.surf$Income)]
    =qual.mean[add.condition(mar.surf[is.na(mar.surf$Income),Condition)]]
  #Store means and variances
  Imp.mean.mar.Con[i]=mean(mar.surf$Income)
  Imp.var.mar.Con[i]=var(mar.surf$Income)
}

```

Figure 4.2 shows the distribution of the means and variances for the imputed 1000 simulated SURF datasets. Comparing the distribution of estimates of unconditional mean imputed data sets (Figure 4.1) and the distribution of estimates of the MAR data sets (Figure 2.8), we see an improvement in the distribution of means, although the variances are still biased against the complete data variances.

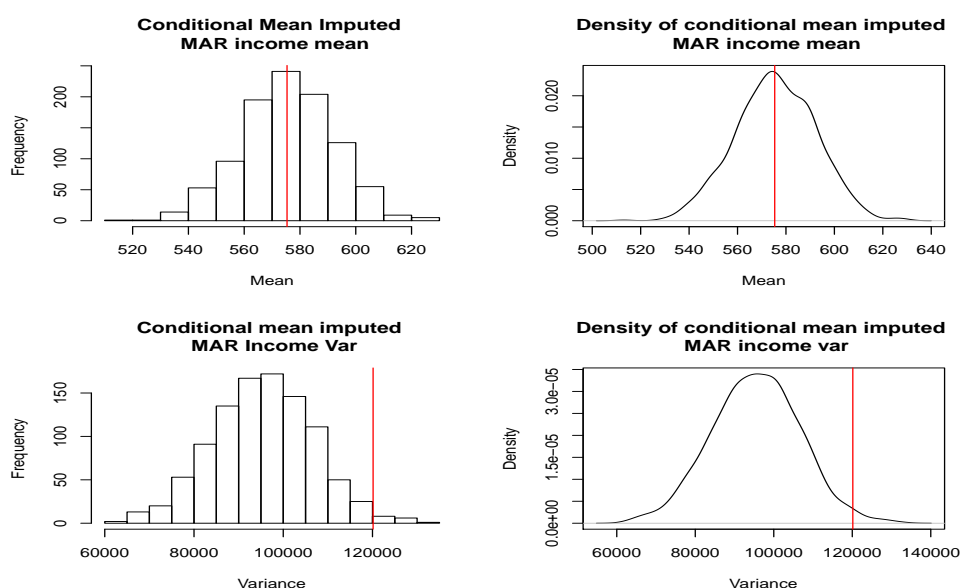


Figure 4.3: Conditional Mean Imputed MAR SURF Income

We can explain this phenomenon by using the following proofs. Again, suppose  $y_i$  are iid  $N(\mu, \sigma^2)$  where  $y_i$  is classified into  $g$  groups where  $g = 1, \dots, G$  by other variables. For each  $g$  group,  $y_i$  for  $i = 1, \dots, r_g$  are observed,  $y_i$  for  $i = r_g + 1, \dots, n_g$  are missing.  $n$  is  $\sum_{g=1}^G n_g$ . Then, we replace each missing value by the mean  $\hat{\mu}_{mis,g}$  of its group. Hence, the overall mean  $\hat{\mu}$  is:

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{g=1}^G (\sum_{i=1}^{r_g} y_i + (n_g - r_g) \hat{\mu}_{mis,g})}{n} \\ &= \frac{\sum_{g=1}^G (\sum_{i=1}^{r_g} y_i + (n_g - r_g) \frac{\sum_{i=1}^{r_g} y_i}{r_g})}{n} \\ &= \frac{\sum_{g=1}^G (\frac{n_g}{r_g} \sum_{i=1}^{r_g} y_i)}{n}\end{aligned}$$

where  $n_g/r_g$  is actually the weight  $w_g = n_g/r_g$ , so we have:

$$\hat{\mu} = \frac{\sum_{g=1}^G (\sum_{i=1}^{r_g} w_g y_i)}{n}$$

If we have no missing data, the overall mean can be expressed as:

$$\begin{aligned}\hat{\mu}_{complete} &= \frac{\sum_{i=1}^n y_i}{n} \\ &= \frac{\sum_{g=1}^G (\sum_{i=1}^{n_g} y_i)}{n}\end{aligned}$$

Lohr (1999) tells us that  $\sum_{i=1}^{r_g} w_g y_i \approx \sum_{i=1}^{n_g} y_i$ . Hence, we can conclude that:

$$\hat{\mu} \approx \hat{\mu}_{complete}$$

However, the variance is different. For the conditional mean imputed incomplete data variance:

$$\begin{aligned}\hat{V} &= \frac{\sum_{g=1}^G (\sum_{i=1}^{r_g} (y_i - \hat{\mu})^2 + \sum_{i=r_g+1}^{n_g-r_g} (y_i - \hat{\mu})^2)}{n-1} \\ &= \frac{\sum_{g=1}^G (\sum_{i=1}^{r_g} (y_i - \hat{\mu})^2 + (n_g - r_g)(\hat{\mu}_g - \hat{\mu})^2)}{n-1}\end{aligned}$$

The variance for the complete data is:

$$\begin{aligned}\hat{V}_{complete} &= \frac{\sum_{i=1}^n (y_i - \hat{\mu}_{complete})^2}{n-1} \\ &= \frac{\sum_{g=1}^G (\sum_{i=1}^{n_g} (y_i - \hat{\mu}_{complete})^2)}{n-1} \\ &= \frac{\sum_{g=1}^G (\sum_{i=1}^{r_g} (y_i - \hat{\mu}_{complete})^2 + \sum_{i=r_g+1}^{n_g} (y_i - \hat{\mu}_{complete})^2)}{n-1}\end{aligned}$$

Now, the reason for having biased variances is clear. Given  $\hat{\mu} \approx \hat{\mu}_{complete}$ ,  $\sum_{i=r_g+1}^{n_g} (y_i - \hat{\mu}_{complete})^2$  tends to be bigger than  $(n_g - r_g)(\hat{\mu}_g - \hat{\mu})^2$  in most cases. Hence, we have:

$$\hat{V} \leq \hat{V}_{complete} \quad (4.3)$$

The question is then how to explain the upper tail of the distribution of variances in Figure 4.3. According to equation (4.3), we shouldn't have variances from the conditional mean imputed data sets being bigger than the variance of the complete data. The explanation lies within the outliers or very big values. Because our data are MAR, there is always a chance that a large number of the lower values will be chosen to become the missing data, but all the very high values will be kept. If this happens, we will see big means, and  $(n_g - r_g)(\hat{\mu}_g - \hat{\mu})^2$  will be bigger than  $\sum_{i=r_g+1}^{n_g} (y_i - \hat{\mu}_{complete})^2$  for some group  $g$ .

### 4.2.3 Regression imputation

Harvey (2001, p.7) points out that the conditional mean imputation is actually a special case of regression imputation. The variables which conditional mean imputation conditions on can be considered as the explanatory variables of the regression model. The advantage of regression imputation, compared to conditional mean imputation, are: (1) it is possible to incorporate as many explanatory variables as we wish, without worrying about the cell size. However, we do not recommend fitting too many variables in the model due to over-modelling problems. On the contrary, if conditional mean imputation has too many conditional variables, the cell sizes could be one or less. If the size one cell happens to have a missing value, then there will be no value available to replace that missing value; (2) regression imputation can have any types of explanatory variables, but the conditional mean imputation can only condition on categorical variables.

For demonstration purposes, we used all the SURF variables to construct a regression model, although the income missingness only depends on the gender variable. Eq.(3.4) can be re-expressed as:

$$Y_{mis}(\text{Income}) \sim f(Y_{\text{Income}} | X_{\text{Gender}}, X_{\text{Qualification}}, X_{\text{Age}}, X_{\text{Hours}}, X_{\text{Marital}}, X_{\text{Ethnicity}}; \hat{\theta}) \quad (4.4)$$

The purpose of this regression imputation is only to show how this imputation method works. Hence, the regression model we used is not the optimal model. Suppose the sample size is  $n$ , there are  $r$  complete observations, and  $n - r$  observations with missing “Income”, and  $K$  is the number of explanatory variables. The regression model which is used to impute missing “Income” can be expressed as:

$$\hat{Y}_{\text{Income}(i)} = \tilde{\beta}_0 + \sum_{j=1}^K \beta_{Kj} X_{ij} \quad (4.5)$$

where  $\hat{Y}_{\text{Income}(i)}$  is the imputed missing “Income” for unit  $i$ ,  $i = (n - r, \dots, n)$ ,  $\tilde{\beta}_0$  is the intercept and  $\beta_{Kj}$  is the coefficient of  $X_j$  in the regression of  $Y_{\text{Income}}$  on  $X_1, \dots, X_K$  based on the  $r$  complete observations.

The following steps demonstrate how to implement Eq.(4.4), and also to simulate the linear regression imputation process 1000 times in order to find out the distribution of the estimates of the imputed data.

#### Recipe: Regression imputation

- Step 1:** Create missing income with the missing probability 50% for male and 20% for female
- Step 2:** Fit a linear regression model to the observed data, using “Income” as response variable.
- Step 3:** Predict “Income” values by using the linear regression model from Step 2
- Step 4:** Replace missing “Income” values with the predicted “Income” values
- Step 5:** Estimate mean and variance of imputed “Income” variable
- Step 6:** Repeat “Step 1” to “Step 5” for 1000 times and record the means and variances for each imputed “Income” variable



```

#R program
RegImp.mean.mar=c()
RegImp.var.mar=c()
for (i in (1:1000)){
  #Create missing Income values -- MAR
  mar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
  #fit regression to income
  lm.imp=lm(Income~Gender+Qualification+Age+Hours+Marital+Ethnicity,
            data=mar.surf[!is.na(mar.surf$Income),])
  pred=predict (lm.imp, mar.surf)
  mar.surf[,"Income"]=impute(mar.surf$Income, pred)
  RegImp.mean.mar[i]=mean(mar.surf$Income)
  RegImp.var.mar[i]=var(mar.surf$Income) }

```

Figure 4.4 shows the distribution of the simulated 1000 means and variances. As we can see, the shapes of the distributions of the means and the variances are almost identical to the distributions of Figure 4.3. This is expected as the real contributor in the regression model's explanatory variables is the gender variable which is the variable the income missingness depends on.

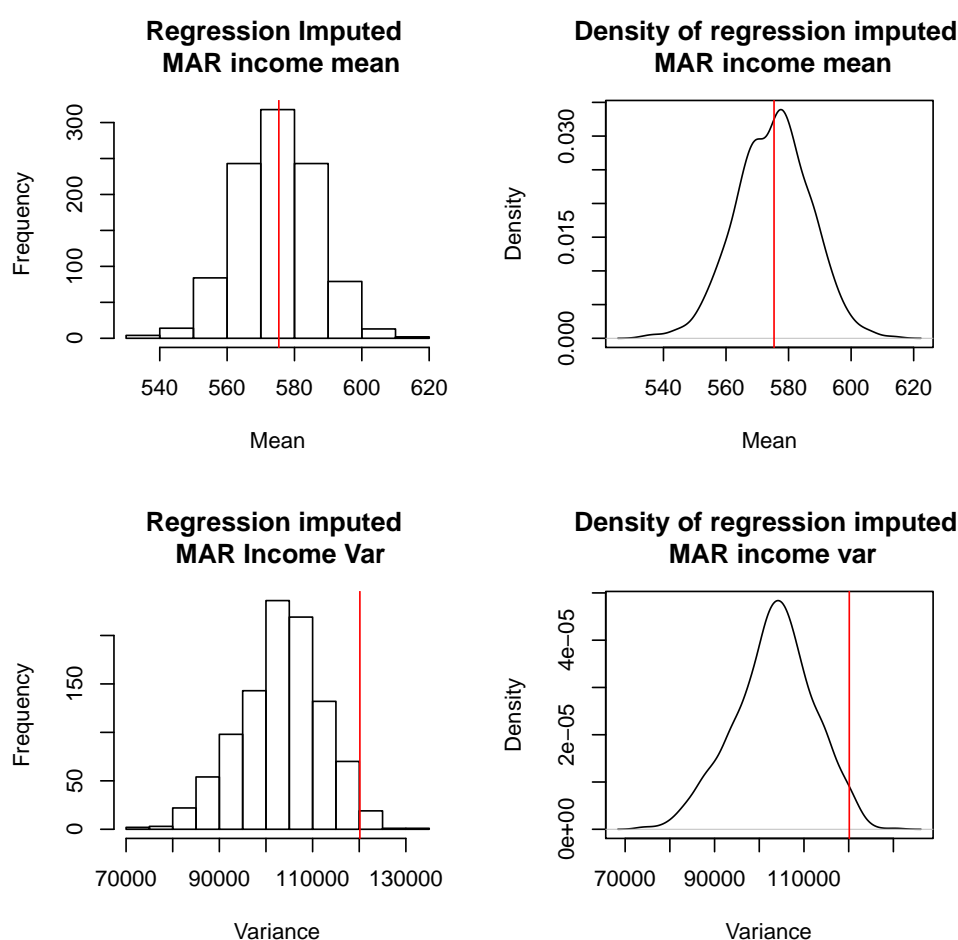


Figure 4.4: Regression Imputed MAR SURF Income

#### 4.2.4 Stochastic regression imputation

As described in the previous chapter, stochastic regression imputation replaces missing values by a model predicted value plus an uncertainty. A simple normal linear regression model was used to impute/predict missing “Income” values. Then, the uncertainties which are random draws from the normal distribution with zero means and residual variance in the regression added to the imputed values. Similarly to Eq (4.5), we can add an uncertainty term to the regression model:

$$\hat{Y}_{Income(i)} = \tilde{\beta}_0 + \sum_{j=1}^K \beta_{Kj} X_{ij} + z_{ik}, \quad (4.6)$$

where  $z_{ik}$  is a random normal deviate with mean 0 and variance  $\tilde{\sigma}_K$ , the residual variance from the regression of  $Y_{Income}$  on  $X_1, \dots, X_K$  based on the  $r$  complete observations.

The following steps demonstrate how to implement the stochastic regression method, and also to simulate the imputation process 1000 times in order to find out the distribution of the estimates of the imputed data.

##### Recipe: Stochastic regression imputation

- Step 1:** Create missing income with the missing probability 50% for male and 20% for female
- Step 2:** Fit a linear regression model to the observed data, using “Income” as response variable.
- Step 3:** Find the standard error of the regression “sigma” which is just the standard deviation of the residuals
- Step 4:** Predict “Income” values by using the linear regression model from Step 2
- Step 5:** Draw stochastic predicted “Income” values from a normal distribution with mean equal to the predicted “Income values” and standard deviation equal to “sigma”
- Step 6:** Replace missing “Income” values with the stochastic predicted “Income” values
- Step 7:** Estimate mean and variance of imputed “Income” variable
- Step 8:** Repeat Step 1 to Step 7 1000 times and record the means and variances for each imputed “Income” variable

```
#R program
#Stochastic regression MAR
Sto.RegImp.mean.mar=c()
Sto.RegImp.var.mar=c()
for (i in (1:1000)){
  #Create missing Income values -- MAR
  #mcar.surf=MCAR(SURF,50,"Income")
  mar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
  #fit regression to income
  lm.imp=lm(Income~Gender+Qualification+Age+Hours+Marital+Ethnicity,
            data=mar.surf[!is.na(mar.surf$Income),])
```

```

sigma <- summary(lm.imp)$sigma
pred.Stochastic=rnorm(nrow(mar.surf),predict(lm.imp,mar.surf),sigma)
mar.surf[, "Income"]=impute(mar.surf$Income, pred.Stochastic)
Sto.RegImp.mean.mar[i]=mean(mar.surf$Income)
Sto.RegImp.var.mar[i]=var(mar.surf$Income)
}

```

Figure 4.5 shows the distribution of the simulated 1000 means and variances. We have seen the improvement on both means and variances of the stochastic imputed data, compared to previous imputation methods. The increase of the variances is due to the addition of a random error ( $z_{ik}$ ) to the regression prediction obtained by using regression imputation. As Little & Schenker (1995) point out, Stochastic regression imputation compensates for the underestimation of the variance of variables with missing data that is associated with regression imputation.

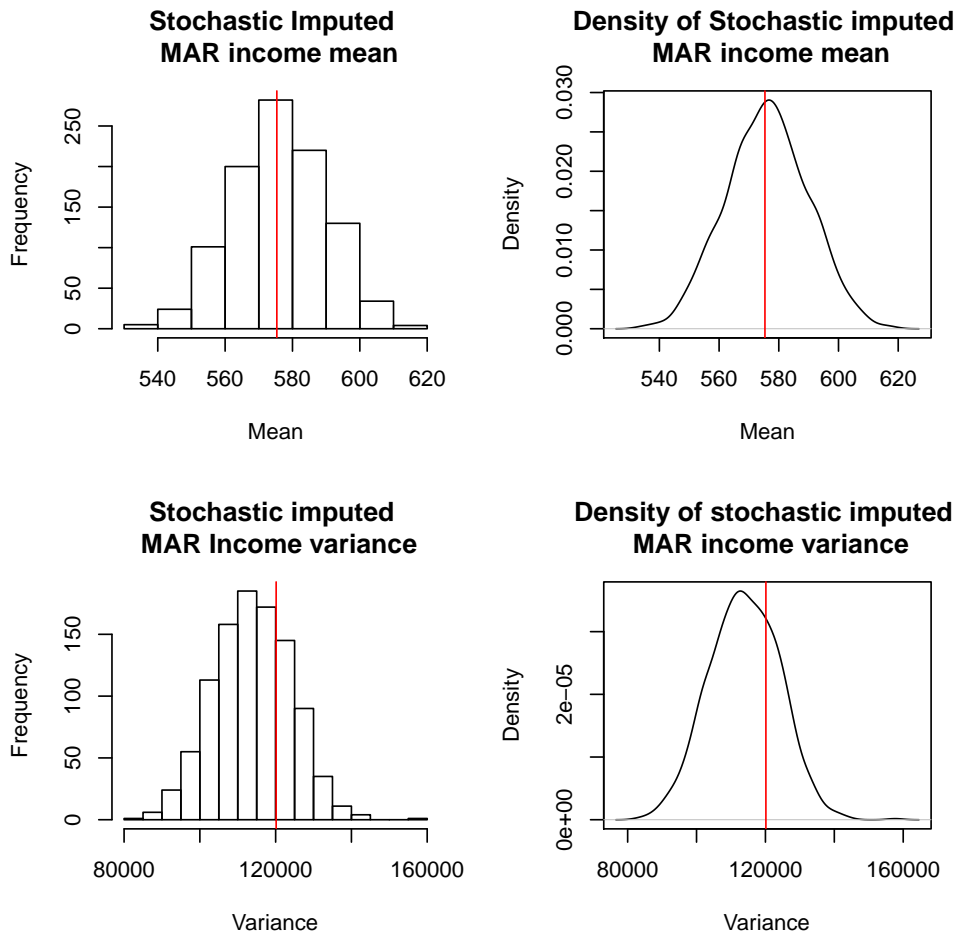


Figure 4.5: Stochastic regression Imputed MAR SURF Income

### 4.3 Implicit modelling methods

As briefly described in Chapter 3, the implicit modelling methods are based on algorithms. Hence, the hot deck method is the best example to illustrate implicit modelling methods. The hot deck basically means replacing missing values by values sampled from similar responding units in the sample. Suppose we have a sample size of  $n$  out of a population of size  $N$ ,  $r$  units

for variable  $Y$  are observed, and  $n - r$  units are missing. If the sampling scheme is simple random sampling without replacement (SRSWOR) and the  $n - r$  units with missing  $Y$  values are imputed, then the mean  $Y$  can be estimated as the mean of the responding and the imputed units. This can be written as:

$$\bar{y}_{HD} = \{r\bar{y}_R + (n - r)\bar{y}_{NR}\}/n \quad (4.7)$$

where  $\bar{y}_R$  is the mean of the respondent units, and the mean of non-responding units  $\bar{y}_{NR}$  is:

$$\bar{y}_{NR} = \sum_{i=1}^r \frac{H_i y_i}{n - r} \quad (4.8)$$

where  $H_i$  is the number of times  $y_i$  is used as a replacement for a missing  $Y$  values, with  $\sum_{i=1}^r H_i = n - r$ , the number of missing units. If the imputed values can be regarded as selected from the values for the responding units by a probability sampling design, then the distribution of  $H_1, \dots, H_r$  in repeated applications of the hot deck method is known. The mean and variance of  $\bar{y}_{HD}$  can be expressed as:

$$E(\bar{y}_{HD}) = E[E(\bar{y}_{HD}|Y_{obs})] \quad (4.9)$$

*Proof:*

$$\begin{aligned} E[E(\bar{y}_{HD}|Y_{obs})] &= \sum_{y_{obs}} E(\bar{y}_{HD}|y_{obs})P(y_{obs}) \\ &= \sum_{y_{obs}} \left( \sum_{\bar{y}_{HD}} \bar{y}_{HD} P(\bar{y}_{HD}|y_{obs}) \right) P(y_{obs}) \\ &= \sum_{y_{obs}} \sum_{\bar{y}_{HD}} \bar{y}_{HD} P(\bar{y}|y_{obs}) P(y_{obs}) \\ &= \sum_{y_{obs}} \sum_{\bar{y}_{HD}} \bar{y}_{HD} P(y_{obs}|\bar{y}_{HD}) P(\bar{y}_{HD}) \\ &= \sum_{\bar{y}_{HD}} \bar{y}_{HD} P(\bar{y}_{HD}) \left( \sum_{y_{obs}} P(y_{obs}|\bar{y}_{HD}) \right) \\ &= \sum_{\bar{y}_{HD}} \bar{y}_{HD} P(\bar{y}_{HD}) \\ &= E(\bar{y}_{HD}) \end{aligned}$$

$$Var(\bar{y}_{HD}) = Var[E(\bar{y}_{HD}|Y_{obs})] + E[Var(\bar{y}_{HD}|Y_{obs})] \quad (4.10)$$

*Proof:*

$$\begin{aligned} Var(\bar{y}_{HD}) &= E(\bar{y}_{HD}^2) - E(\bar{y}_{HD})^2 \\ &= E[E(\bar{y}_{HD}^2|y_{obs})] - E[E(\bar{y}_{HD}|y_{obs})]^2 \\ &= E[Var(\bar{y}_{HD}|y_{obs}) + E(\bar{y}_{HD}|y_{obs})^2] - E[E(\bar{y}_{HD}|y_{obs})]^2 \\ &= E[Var(\bar{y}|y_{obs})] + (E[E(\bar{y}|y_{obs})^2] - E[E(\bar{y}_{HD}|y_{obs})]^2) \\ &= E[Var(\bar{y}_{HD}|y_{obs})] + Var[E(\bar{y}_{HD}|y_{obs})] \end{aligned}$$

where the inner expectations and variances are over the distribution of  $\{H_1, \dots, H_r\}$  given the observed data  $Y_{obs}$ , and the outer expectations and variances are over the model distribution of  $Y$ . The second term in Eq (4.10) represents the additional variance from the imputation procedure (Little & Rubin 2002).

### 4.3.1 Hot deck imputation - random hot deck imputation with replacement

Let  $\bar{y}_{HD}$  denote the hot deck estimator when the  $\{H_i\}$  are obtained by random sampling with replacement from the recorded values of  $Y$ . This means any  $H_i$  can be  $0, 1, 2, \dots, n-r$ . Conditioning on the sampled and recorded values, the distribution of  $H_1, \dots, H_r$  in repetitions of the hot deck is multinomial with sample size  $n-r$  and probabilities  $(1/r, \dots, 1/r)$  (Cochran 1977). Therefore, the moments of the distribution of  $\{H_1, \dots, H_r\}$  given the observed data  $Y$  are:

$$\begin{aligned} E(H_i|Y_{obs}) &= \frac{n-r}{r} \\ \text{Var}(H_i|Y_{obs}) &= \frac{(n-r)(1-1/r)}{r} \\ \text{Cov}(H_i, H_j|Y_{obs}) &= -\frac{(n-r)}{r^2} \end{aligned}$$

Hence, we can express the imputed mean and variance as:

$$E(\bar{y}_{HD}|Y_{obs}) = \bar{y}_R \quad (4.11)$$

*Proof:*

$$\begin{aligned} E[\bar{y}_{HD}|y_{obs}] &= E\left[\frac{r}{n}\bar{y}_R + \frac{n-r}{n} \sum_{i=1}^r \frac{H_i y_i}{n-r}\right] \\ &= \frac{r}{n}\bar{y}_R + \frac{n-r}{n} E\left[\sum_{i=1}^r \frac{H_i y_i}{n-r}\right] \\ &= \frac{r}{n}\bar{y}_R + \frac{n-r}{n} \sum_{i=1}^r \frac{E[H_i y_i]}{n-r}, \quad \text{since } H_i \perp y_i \\ &= \frac{r}{n}\bar{y}_R + \frac{n-r}{n} \frac{n-r}{r} \frac{1}{n-r} \sum_{i=1}^r E[y_i] \\ &= \frac{r}{n}\bar{y}_R + \frac{n-r}{n} \frac{n-r}{r} \frac{1}{n-r} \sum_{i=1}^r \bar{y}_R \\ &= \bar{y}_R \end{aligned}$$

and

$$\text{Var}(\bar{y}_{HD}|Y_{obs}) = (1-r^{-1})(1-r/n)s_{y_R}^2/n \quad (4.12)$$

*Proof:*

$$\begin{aligned} \text{Var}(\bar{y}_{HD}|y_{obs}) &= \text{Var}\left[\frac{r}{n}\bar{y}_R + \frac{n-r}{n} \sum_{i=1}^r \frac{H_i y_i}{n-r}\right] \\ &= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^r H_i y_i\right) \end{aligned}$$

The variance  $Var(\sum_{i=1}^r H_i y_i)$  follows:

$$\begin{aligned}
Var(\sum_{i=1}^r H_i y_i) &= E[\sum_{i=1}^r H_i y_i \sum_{j=1}^r H_j y_j] - E[\sum_{i=1}^r H_i y_i] E[\sum_{j=1}^r H_j y_j] \\
&= \sum_{i=1}^r \sum_{j=1}^r E[H_i H_j] y_i y_j - \sum_{i=1}^r \sum_{j=1}^r E[H_i] E[H_j] y_i y_j \\
&= \sum_{i=1}^r \sum_{j=1}^r Cov[H_i H_j] y_i y_j \\
&= \sum_{i=1}^r y_i^2 Cov[H_i, H_i] + \sum_{i=1}^r \sum_{j=1, j \neq i}^r y_i y_j Cov[H_i, H_j] \\
&= \sum_{i=1}^r y_i^2 \frac{(n-r)(1-1/r)}{r} - \sum_{i=1}^r \sum_{j=1, j \neq i}^r y_i y_j \frac{(n-r)}{r^2} \\
&= \frac{n-r}{r^2} \left( \sum_{i=1}^r y_i^2 (r-1) - \sum_{i=1}^r \sum_{j=1, j \neq i}^r y_i y_j \right) \\
&= \frac{n-r}{r^2} \left( r \sum_{i=1}^r y_i^2 - \sum_{i=1}^r \sum_{j=1}^r y_i y_j \right) \\
&= \frac{n-r}{r^2} \left( r \sum_{i=1}^r y_i^2 - r^2 \bar{y}_R \right) \\
&= (n-r)(r-1) \frac{s_{yR}^2}{r}
\end{aligned}$$

Then, we have  $Var(\bar{y}_{HD} | Y_{obs}) = \frac{1}{n^2} Var(\sum_{i=1}^r H_i y_i) = (1-r^{-1})(1-r/n)s_{yR}^2/n$ .

If we assume the missing data are MCAR, then Eq (4.9) and (4.10) yield

$$\begin{aligned}
E(\bar{y}_{HD}) &= \bar{y} \\
Var(\bar{y}_{HD}) &= (r^{-1} - N^{-1})S_y^2 + (1-r^{-1})(1-r/n)s_y^2/n
\end{aligned}$$

where the first component of the variance is the simple random sample variance of  $\bar{y}_R$ , and the second component represents the increase in variance from the hot deck procedure.

The following steps demonstrate how to implement the simple random hot deck imputation with Replacement method, and also simulate the imputation process 1000 times in order to find out the distribution of the estimates of the imputed data.

### Recipe: Random hot deck imputation with Replacement

- Step 1:** Create missing income with the missing probability 50% for male and 20% for female
- Step 2:** Randomly choose  $n$  “Income” values from the non-missing “Income” values with replacement
- Step 3:** Replace missing “Income” values with the chosen “Income” values from Step 2
- Step 4:** Estimate the mean and variance of imputed “Income” variable
- Step 5:** Repeat Step 1 to Step 4 for 1000 times and record the means and variances for each imputed “Income” variable

```
#R program
#Random hot deck imputation -- with replacement
hotImp.mean.mar=c()
hotImp.var.mar=c()
for (i in (1:1000)){
  #Create missing value
  #mcar.surf=MCAR(SURF,50,"Income")
  mar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
  #Count the number of missing
  nmissing=nrow(mar.surf[is.na(mar.surf$Income),])
  #Hot deck imputation
  mar.surf[is.na(mar.surf$Income),"Income"]
    =sample(mar.surf[!is.na(mar.surf$Income),"Income"],
            size=nmissing,replace=T)
  hotImp.mean.mar[i]=mean(mar.surf$Income)
  hotImp.var.mar[i]=var(mar.surf$Income)
}
```

Figure 4.6 shows the distribution of the simulated 1000 means and variances. Comparing to Figure 2.8, we see that the distributions of the means and variances are about the same. This is because imputed values are randomly selected from the observed values with an equal selection probability. This simple hot deck with replacement method does not change the distribution of male and female income values.

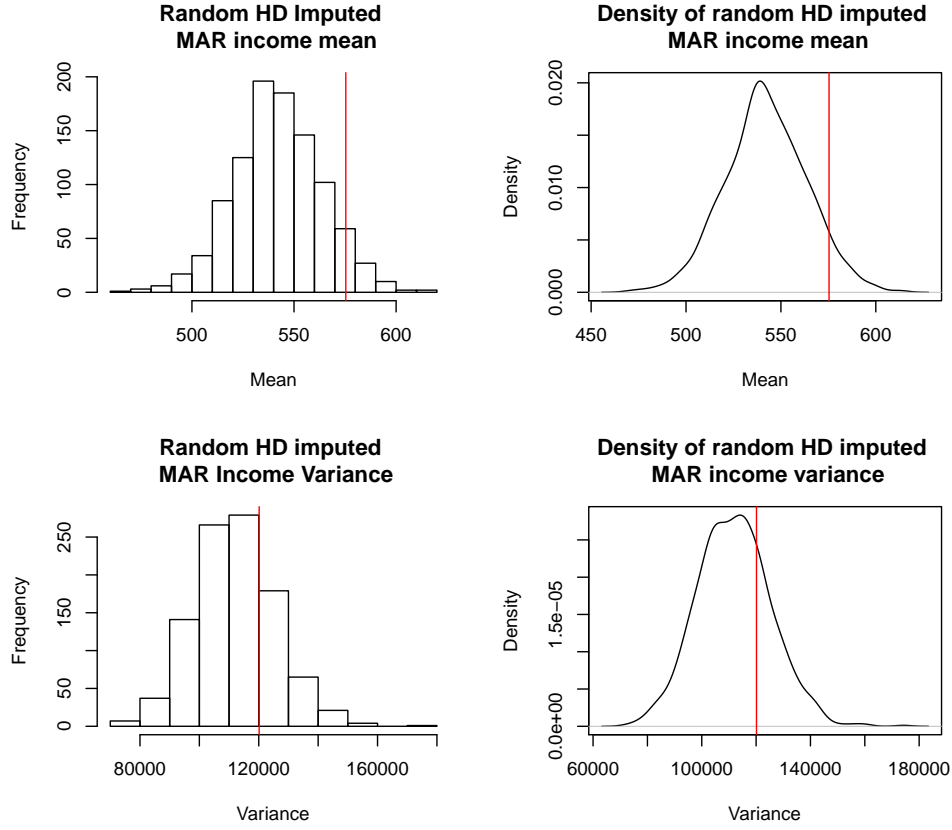


Figure 4.6: Random hot deck Imputed MAR SURF Income with replacement

However, as we have pointed out in Eq. (4.10), the variance of the mean  $\bar{y}_{HD}$  becomes larger than the non imputed data. In fact, Little & Rubin (2002) has worked out the maximum proportionate variance increase of  $\bar{y}_{HD}$  over  $\bar{y}_R$  is 0.25 when  $r/n = 0.5$ .

*Proof:*

$$\begin{aligned}
 \frac{Var(\bar{y}_{HD}|y_{obs})}{Var(\bar{y}_R)} &= \frac{(1-r^{-1})(1-r/n)S_y^2/n}{(r^{-1}-N^{-1})S_y^2} \\
 &= \frac{(1-r^{-1})(1-r/n)(1/n)}{(r^{-1}-N^{-1})} \\
 &= (r-1)\left(1-\frac{r}{n}\right)\frac{1}{n} \\
 &= (r-1)(n-r)\frac{1}{n^2} \\
 &= \frac{r}{n} - \frac{r^2}{n^2} - \frac{1}{n} + \frac{r}{n^2}
 \end{aligned}$$

Now, assume  $x = \frac{r}{n}$ , and  $n$  as constant. We have:  $f(x) = x - x^2 - \frac{1}{n} + \frac{x}{n}$ . Then differentiate  $f(x)$ , we get:

$$\begin{aligned}
 \frac{df(x)}{dx} &= 1 - 2x + \frac{1}{n} \stackrel{set}{=} 0 \\
 \Rightarrow x &= \frac{1}{2} + \frac{1}{2n} \\
 &\stackrel{n \rightarrow \infty}{=} \frac{1}{2}
 \end{aligned}$$



Hence, we have  $\frac{r}{n} = \frac{1}{2}$ . Substituting  $n = 2r$  to Equation (4.3.1), we get  $\frac{1}{4} - \frac{1}{4r}$ . If we assume  $r$  is large, which means  $\frac{1}{4r} \rightarrow 0$ , then the maximum value is  $\frac{1}{4}$ .

### 4.3.2 Hot deck imputation - random hot deck imputation without replacement

This hot deck imputation method is very similar to the previous one, except that imputations are by random sampling of observed values without replacement. In addition, this hot deck method is also slightly different under two conditions. This first condition is when there are more observed units than missing units; the second condition is when there are fewer observed units than missing units. The first condition can be written as  $n - r = t$ , where  $0 < t < r$ . The hot deck without replacement selects  $t$  units randomly without replacement to produce the  $n - r$  values required for the missing data. The second condition can be written as  $n - r = kr + t$ , where  $k$  is a positive integer and  $0 < t < r$ . Then, the hot deck without replacement selects all the recorded units  $k$  times, and then selects  $t$  additional units randomly without replacement to yield the  $n - r$  values required for the missing data.

Hence

$$\bar{y}_{NR} = \begin{cases} t\bar{y}_t / (n - r) & \text{if } n - r < r \\ (kr\bar{y}_R + t\bar{y}_t) / (n - r) & \text{if } n - r > r \end{cases} \quad (4.13)$$

where  $\bar{y}_t$  is the mean of the  $t$  supplementary values of  $Y$ .

Let's express the mean  $Y$  for hot deck imputation without replacement as:

$$\bar{y}_{HD2} = \frac{r\bar{y}_R + (n - r)\bar{y}_{NR}}{n}$$

Given Eq. (4.13), the above equation can be re-expressed as:

$$\bar{y}_{HD2} = \frac{(k + 1)r\bar{y}_R + t\bar{y}_t}{n} \quad (4.14)$$

$\bar{y}_t$  can also be re-expressed as:

$$\bar{y}_t = \frac{\sum_{i=1}^r I_i y_i}{t} \quad (4.15)$$

where the indicator variable  $I_i$  is a random variable which takes the value one if a unit has been selected as the hot deck imputed value, otherwise, the value is zero. Hence:

$$\begin{aligned} E(I_i | Y_{obs}) &= \frac{t}{r} \\ \text{Var}(I_i | Y_{obs}) &= \frac{t}{r} \left(1 - \frac{t}{r}\right) \\ \text{Cov}(I_i, I_j | Y_{obs}) &= -\frac{t}{r} \left(1 - \frac{t}{r}\right) \frac{1}{r - 1} \end{aligned}$$

Then these yield the mean and variance of  $\bar{y}_{HD2}$  given  $Y_{obs}$ :

$$E(\bar{y}_{HD2} | Y_{obs}) = \bar{y}_R \quad (4.16)$$

*Proof:*

$$\begin{aligned} E(\bar{y}_{HD2} | Y_{obs}) &= \frac{(k + 1)r\bar{y}_R}{n} + \frac{tE(\bar{y}_t)}{n} \\ &= \frac{(k + 1)r\bar{y}_R}{n} + \frac{t}{n} \frac{1}{t} r E(H_i) E(y_i) \\ &= \bar{y}_R \end{aligned}$$

and

$$\text{Var}(\bar{y}_{HD2}|Y_{obs}) = (t/n)(1-t/r)s_{yR}^2/n \quad (4.17)$$

*Proof:*

$$\begin{aligned} \text{Var}(\bar{y}_{HD2}|Y_{obs}) &= \frac{t^2}{n^2} \text{Var}(\bar{y}_t) \\ &= \frac{t^2}{n^2} \text{Var}\left(\frac{\sum_{i=1}^r I_i y_i}{t}\right) \\ &= \frac{t^2}{n^2} \frac{1}{t^2} \text{Var}\left(\sum_{i=1}^r I_i y_i\right) \end{aligned}$$

Again,  $\text{Var}(\sum_{i=1}^r I_i y_i)$  follows:

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^r I_i y_i\right) &= E\left[\sum_{i=1}^r I_i y_i \sum_{j=1}^r I_j y_j\right] - E\left[\sum_{i=1}^r I_i y_i\right] E\left[\sum_{j=1}^r I_j y_j\right] \\ &= \sum_{i=1}^r \sum_{j=1}^r E[I_i I_j] y_i y_j - \sum_{i=1}^r \sum_{j=1}^r E[I_i] E[I_j] y_i y_j \\ &= \sum_{i=1}^r \sum_{j=1}^r \text{Cov}[I_i I_j] y_i y_j \\ &= \sum_{i=1}^r y_i^2 \text{Cov}[I_i, I_i] + \sum_{i=1}^r \sum_{j=1, j \neq i}^r y_i y_j \text{Cov}[I_i, I_j] \\ &= \frac{t}{r} \left(1 - \frac{t}{r}\right) \sum_{i=1}^r y_i^2 - \frac{t}{r} \left(1 - \frac{t}{r}\right) \frac{1}{r-1} \sum_{i=1}^r \sum_{j=1, j \neq i}^r y_i y_j \\ &= \frac{t}{r} \left(1 - \frac{t}{r}\right) \frac{1}{r-1} \left( (r-1) \sum_{i=1}^r y_i^2 - \sum_{i=1}^r \sum_{j=1, j \neq i}^r y_i y_j \right) \\ &= \frac{t}{r} \left(1 - \frac{t}{r}\right) \frac{1}{r-1} \left( r \sum_{i=1}^r y_i^2 - \sum_{i=1}^r \sum_{j=1}^r y_i y_j \right) \\ &= \frac{t}{r} \left(1 - \frac{t}{r}\right) \frac{1}{r-1} \left( r \sum_{i=1}^r y_i^2 - r^2 \bar{y}_r \right) \\ &= t \left(1 - \frac{t}{r}\right) \frac{1}{r-1} \frac{\sum_{i=1}^r y_i^2 - r \bar{y}_r}{r-1} (r-1) \\ &= t \left(1 - \frac{t}{r}\right) s_{yR}^2 \end{aligned}$$

Hence,  $\text{Var}(\bar{y}_{HD2}|Y_{obs}) = \frac{t^2}{n^2} \frac{1}{t^2} \text{Var}\left(\sum_{i=1}^r I_i y_i\right) = \frac{t}{n} \left(1 - \frac{t}{r}\right) \frac{s_{yR}^2}{n}.$

The following steps demonstrate how to implement the simple random hot deck imputation without Replacement method, and also simulate the imputation process 1000 times in order to find the distribution of the estimates of the imputed data.

### Recipe: Random hot deck imputation without Replacement

- Step 1:** Create missing income with the missing probability 50% for male and 20% for female
- Step 2:** If  $n - r \leq r$ , then randomly choose  $n$  “Income” values from non-missing “Income” values without replacement. If  $n - r > r$ , then select all the recorded units ( $r$ )  $k$  times, and select  $t$  additional units randomly without replacement to yield the  $n - r$  values.
- Step 3:** Replace missing “Income” values with the chosen “Income” values from Step 1
- Step 4:** Estimate the mean and variance of imputed “Income” variable
- Step 5:** Repeat Step 1 to Step 4 1000 times and record the means and variances for each imputed “Income” variable

```
#R program
#Random hot deck imputation -- without replacement
hotImp.mean.mar=c()
hotImp.var.mar=c()
for (i in (1:1000)){
  #Create missing value
  #mcar.surf=MCAR(SURF,50,"Income")
  mar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
  #Count the number of missing
  nmissing=nrow(mar.surf[is.na(mar.surf$Income),])
  #Count the number of not missing
  not_missing=nrow(mar.surf[!is.na(mar.surf$Income),])
  #Hot deck imputation -- two cases: (1) n-r<r; (2) n-r>r
  if (not_missing>=nmissing){
    mar.surf[is.na(mar.surf$Income),"Income"]
      =sample(mar.surf[!is.na(mar.surf$Income),"Income"],
              size=nmissing,replace=F)
  }else{
    ty_size=round(((nmissing/not_missing)-
                    as.integer(nmissing/not_missing-1))*not_missing)
    kry= rep(mar.surf[!is.na(mar.surf$Income),"Income"],
              as.integer(nmissing/not_missing-1))
    ty= sample(mar.surf[!is.na(mar.surf$Income),"Income"],
               size=ty_size,replace=F)
    mar.surf[is.na(mar.surf$Income),"Income"]= c(kry,ty)
  }
  hotImp.mean.mar[i]=mean(mar.surf$Income)
  hotImp.var.mar[i]=var(mar.surf$Income)
}
```

Figure 4.7 shows the distribution of the simulated 1000 means and variances. Comparing Figure 2.8 and Figure 4.6, we found that the distribution of means and variances are very similar. This is still due to this hot deck imputation method (hot deck without replacement)

randomly selecting observed values as imputed values. This does not change the distribution of observed male and female income.

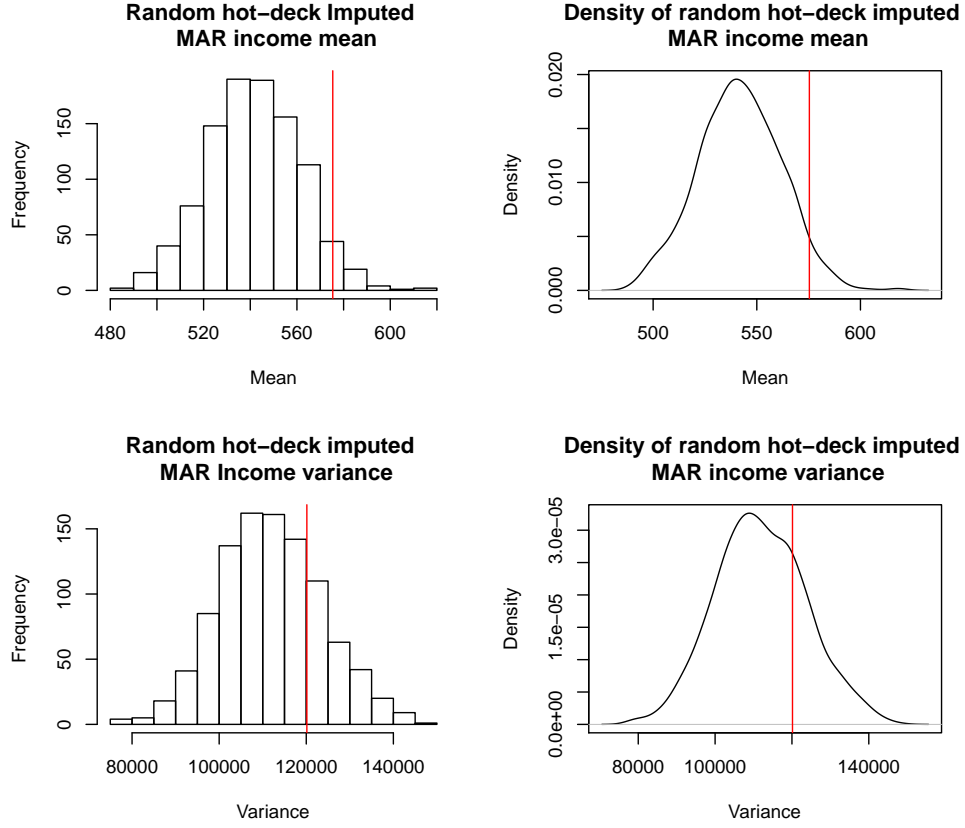


Figure 4.7: Random hot deck Imputed MAR SURF Income without replacement

Again, Equation (4.10) shows that there is an increase of the variance of the mean  $\bar{y}_{HD2}$  over  $\bar{y}_R$ . However, as Little & Rubin (2002) pointed out, hot deck imputation without replacement can reduce the additional variance from the hot deck imputation with replacement. The proportionate variance increase of  $\bar{y}_{HD2}$  over  $\bar{y}_R$  is at most 0.125 when  $k = 0$ ,  $t = n/4$  and  $r = 3n/4$ .

*Proof:* Assuming simple random sampling from a finite population of size  $N$  and missing data are MCAR, then Eq (4.9) and (4.10) yield

$$E(\bar{y}_{HD2}) = \bar{y}$$

$$Var(\bar{y}_{HD2}) = \left(\frac{1}{(k+1)r} - \frac{1}{N}\right)s_y^2 + \frac{t}{n}\left(1 - \frac{t}{r}\right)\frac{s_y^2}{n}$$

Hence, the proportionate variance increase of  $\bar{y}_{HD2}$  over  $\bar{y}_R$  is:

$$\frac{\frac{t}{n}\left(1 - \frac{t}{r}\right)\frac{1}{n}}{\frac{1}{(k+1)r} - \frac{1}{N}} \xrightarrow{N \rightarrow \infty} \frac{\frac{t}{n}\left(1 - \frac{t}{r}\right)\frac{1}{n}}{\frac{1}{(k+1)r}} = \frac{t}{n}\left(1 - \frac{t}{r}\right)\frac{1}{n}(k+1)r$$

$$= (k+1)\left[\frac{tr}{n^2} - \frac{t^2}{n^2}\right] \quad (4.18)$$

We have already seen that  $(k+1)r + t = n$ . This can be re-expressed as  $r = \frac{n-t}{k+1}$ . Substituting this  $r$  to Eq. (4.18) yields:

$$\begin{aligned}(k+1)\left[\frac{tr}{n^2} - \frac{t^2}{n^2}\right] &= (k+1)\left[\frac{t\frac{n-t}{k+1}}{n^2} - \frac{t^2}{n^2}\right] \\ &= (k+1)\left[\frac{t}{(k+1)n} - \frac{t^2}{n^2(k+1)} - \frac{t^2}{n^2}\right]\end{aligned}$$

Now, let's differentiate the above equation with respect to  $t$ , and set it equal to zero, then we get:

$$\begin{aligned}\frac{2t + 2t(k+1)}{n^2(k+1)} &= \frac{1}{(k+1)n} \\ 2t + 2t(k+1) &= n \\ \Rightarrow t &= \frac{n}{2(k+2)}\end{aligned}$$

Then,  $r$  in terms of  $n$  is:  $r = \frac{n(2k+3)}{2(k+2)(k+1)}$ . Substituting  $t = \frac{n}{2(k+2)}$  and  $r = \frac{n(2k+3)}{2(k+2)(k+1)}$  into Eq. (4.18), we get the maximum variance increase of  $\bar{y}_{HD2}$  over  $\bar{y}_R$  is:  $\frac{1}{4(k+2)}$ . It is also known that  $k \geq 0$ . Hence, the maximum value for function  $f(k) = \frac{1}{4(k+2)}$  is 0.125, when  $k = 0$ . Also, if  $k = 0$ , then  $t = n/4$  and  $r = 3n/4$ .

### 4.3.3 Hot deck imputation - sequential hot deck

The sequential hot deck imputation method treats the observed and missing units in a sequence. This means that, under the assumption that sampled units are regarded as randomly ordered, a missing value of  $Y$  is replaced by the nearest responding value preceding it in the sequence. For example, if  $n = 4, r = 2$ ,  $y_1$  and  $y_3$  are observed, and  $y_2$  and  $y_4$  are missing, then  $y_2$  and  $y_4$  are replaced by  $y_1$  and  $y_3$  respectively. However, if  $y_1$  is missing, then some starting value is necessary, maybe chosen from records in a previous survey. (Bailar et al. 1978)

Now, we apply this sequential hot deck imputation method to the SURF data. The following steps demonstrate how to implement the method, and also simulate the imputation process 1000 times in order to find the distribution of the estimates of the imputed data.

#### Recipe: Sequential hot deck imputation

- Step 1:** Replace missing "Income" values with the nearest "Income" values in the sequence
- Step 2:** Estimate mean and variance of imputed "Income" variable
- Step 3:** Repeat "Step 1" to "Step 2" for 1000 times and record the means and variances for each imputed "Income" variable

```
#R program
#Sequential hot deck imputation
hotImp.mean.mar=c()
hotImp.var.mar=c()
for (i in (1:1000)){
```

```

#Create missing value
#mcar.surf=MCAR(SURF,50,"Income")
mar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
mar.surf$Income[1]=SURF$Income[1]
n=length(mar.surf$Income)
#Sequential hot deck
for (j in (1:n)){
  if (is.na(mar.surf$Income[j])){
    mar.surf$Income[j]=mar.surf$Income[j-1]
  }
}
hotImp.mean.mar[i]=mean(mar.surf$Income)
hotImp.var.mar[i]=var(mar.surf$Income)
}

```

Figure 4.8 shows the distribution of the simulated 1000 means and variances. Again, we see similar distributions of means and variances as Figure 2.8. This is because the dataset itself is randomly ordered and the proportion of missing values is not large. The sequential hot deck imputed data has little impact on the distribution of observed male and female income.

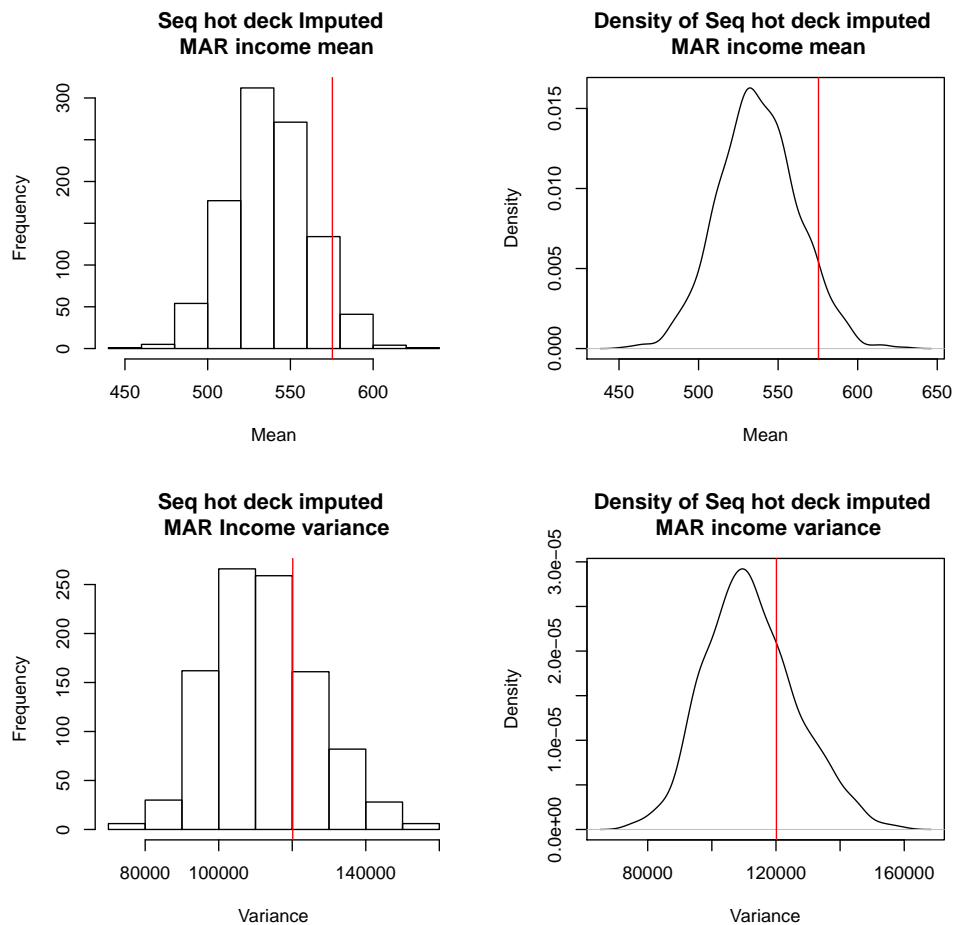


Figure 4.8: Sequential hot deck Imputed MAR SURF Income

However, the variance of mean  $\bar{y}_{HD}$  is still larger than the variance of mean  $\bar{y}_R$  due to the added variance according to Equation (4.10). In fact, Little & Rubin (2002) showed that its

total variance is  $Var(\bar{y}_{HD}) = [1 + (n + r)/n]S_y^2/r$ , assuming large  $r$  and  $n$  and ignoring the finite population corrections. This means the proportionate increase in variance over  $\bar{y}_R$  is  $(n - r)/n$ , the fraction of missing data.

#### 4.3.4 Hot deck imputation - Hot deck within adjustment cells

This hot deck method applies a similar idea to the conditional mean imputation method. Adjustment cells are formed from the joint levels of categorical variables which have observed values for variables with missing values. Missing values within each cell are replaced by observed values from the same cell. The choice of variables for creating adjustment cells is often arbitrary and relies on subjective knowledge of which variables are related to the missing value being imputed. For example, suppose our SURF data has  $n - r$  missing income values, but its other categorical variables have no missing values. Hence, we can form cells based on its categorical variables, such as: "HoursBand", "AgeBand", "Marital", "Ethnicity", "Gender" and "Qualification". This is because we think those variables are associated with the income variable, e.g.. male respondents might earn more income than female respondents. Then, we replace missing values within each cell by a random draw from its observed values.

The following steps demonstrate how to implement the hot deck within adjustment cells imputation method, and also simulate the imputation process 1000 times in order to find the distribution of the estimates of the imputed data.

##### Recipe: Hot deck within adjustment cells

- Step 1:** Form cells based on categorical variables, such as: "HoursBand", "Age-Band", "Marital", "Ethnicity", "Gender" and "Qualification"
- Step 2:** If a cell has only missing "Income" values, then form cells with less variables until all cells with missing "Income" values have observed "Income" values as well
- Step 3:** Randomly choose "Income" values from the cell
- Step 4:** Replace missing "Income" values with the chosen "Income" values from "Step 3"
- Step 5:** Estimate mean and variance of imputed "Income" variable
- Step 6:** Repeat "Step 1" to "Step 5" for 1000 times and record the means and variances for each imputed "Income" variable

```
#R program
#(c) Hot deck within adjustment cells
hotImpCells.mean.mcar=c()
hotImpCells.var.mcar=c()
hotdeckvars <-
  c("HoursBand","AgeBand","Marital","Ethnicity","Gender","Qualification")
SURF2=SURF
SURF2$AgeBand <- 5*(SURF$Age%/%5)
SURF2$HoursBand <- 10*((SURF$Hours-5)%/%10)+5
for (i in (1:1000)){
  #Create missing value
```

```

mar.surf=MAR(SURF2,"Gender","Income",c(0.5,0.2))
mar.surf=mar.surf[order(mar.surf$Personid),]
mar.surf.nomiss=mar.surf[!is.na(mar.surf$Income),]
mar.surf.miss =mar.surf[is.na(mar.surf$Income),]
#Count the number of missing
nmissing=nrow(mar.surf.miss)
idx=sort(mar.surf[is.na(mar.surf$Income),"Personid"])
for (j in (1:nmissing)){
  matched <- F
  m <- length(hotdeckvars)
  while(!matched) {
    mm <- merge(mar.surf.miss[j,], mar.surf.nomiss, by=hotdeckvars[1:m])
    if(nrow(mm)>0) {
      matched <- T
      mar.surf[idx[j],"Income"] <- mm[sample(nrow(mm),1),"Income.y"]
    } else {
      m <- m-1
      if(m==0) {
        mar.surf[idx[j],"Income"] <-
          mar.surf.nomiss[sample(nrow(mar.surf.nomiss),1),"Income"]
        matched <- T
      }
    }
  }
  hotImpCells.mean.mar[i]=mean(mar.surf$Income)
  hotImpCells.var.mar[i]=var(mar.surf$Income)
}

```

Figure 4.9 shows the distribution of the simulated 1000 means and variances. Now, the distributions of means and variances are centred around the mean and variance of the original complete data. This means this hot deck within adjustment cells imputation method has an unbiased estimate of the “true” complete data estimates, if the missing data are MAR and variables which form adjustment cells have a strong association with the missing data. We know gender has a strong relation to missing data because we used it to create the MAR SURF income data. Now, we can explain why the previous simple hot deck with and without replacement methods have underestimated the estimates and why this hot deck with adjustment cells method is better. The reason is that the male respondents have more missing income values than female respondents and overall male respondents have higher income than female. Those previous hot deck imputation methods have a higher probability of selecting female’s income values to impute missing male income values because females have more recorded data than males. Then, this makes the estimates of income become lower than the “true” complete data estimates. By grouping the incomplete data by gender and other variables, the hot deck method only selects values to replace missing values from their own groups. Hence, the missing income values of male respondents have only being imputed by choosing observed income values from other male respondents. Because observed male respondents have higher incomes than female, the imputed male non-respondents will have higher incomes than the imputed female non-respondents as well. This results unbiased estimates.



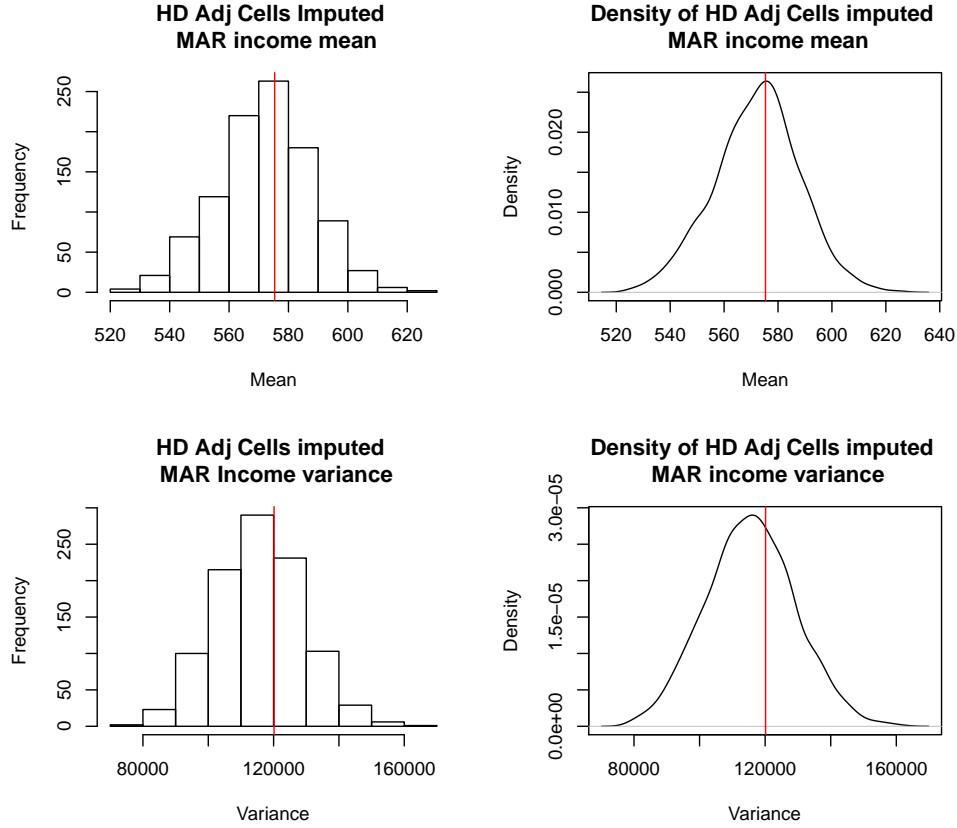


Figure 4.9: Hot deck within adjustment cells Imputed MAR SURF Income

### 4.3.5 Hot deck imputation - Nearest-Neighbour Hot deck Imputation

The Nearest-Neighbour hot deck method can be considered to be a more complex version of the hot deck Within Adjustment cells method. It uses the distance between units by defining a metric, based on the values of covariates, and then it chooses imputed values from observed units close to the unit with the missing value. Andridge & Little (2010) provides an example, let  $y_i = (y_{i1}, \dots, y_{ik})^T$  be the values of  $K$  appropriately scaled covariates for a unit  $i$  for which  $y_i$  is missing. If these variables are used to form adjustment cells, the metric

$$d(i, j) = \begin{cases} 0, & \text{i,j in same cell} \\ 1, & \text{i,j in different cells} \end{cases} \quad (4.19)$$

yields the method of the hot deck Within Adjustment Cells. For the example we demonstrated by using the SURF data, we used Mahalanobis metric to measure distance between units.

$$d(i, j) = (y_i - y_j)^T \hat{Var}(y_i)^{-1} (y_i - y_j)$$

where  $\hat{Var}(x_i)$  is an estimate of the covariance matrix of  $y_i$ .

The measurement of the distance  $d(i, j)$  can be used for both categorical and numerical variables. Unlike the hot deck within adjustment cells method, there is no longer a need to categorize continuous variables in order to form a cell.

The following steps demonstrate how to implement the hot deck within adjustment cells imputation method, and also simulate the imputation process 1000 times in order to find out the distribution of the estimates of the imputed data.

### Recipe: Nearest-Neighbor hot deck Imputation

- Step 1:** Measure distance between observations
- Step 2:** Choose “Income” values that come from observations with “Income” values close to the observation without “Income” value.
- Step 3:** Replace missing “Income” values with the chosen “Income” values from “Step 2”
- Step 4:** Estimate mean and variance of imputed “Income” variable
- Step 5:** Repeat “Step 1” to “Step 4” 1000 times and record the means and variances for each imputed “Income” variable

```
#R program
#(d) Nearest-Neighbor hot deck Imputation
#Using existing packages
library(rrp)
rrp.imp.mean.mcar=c()
rrp.imp.var.mcar=c()

for (i in (1:1000)){
  #Create missing value
  #mcar.surf=MCAR(SURF2,50,"Income")
  mcar.surf=MAR(SURF,"Gender","Income",c(0.5,0.2))
  mcar.surf=mcar.surf[order(mcar.surf$Personid),]
  idx=sort(mcar.surf[is.na(mcar.surf$Income),"Personid"])
  mcar.surf[idx,"Income"]=rrp.impute(mcar.surf)$new.data[idx,"Income"]
  rrp.imp.mean.mcar[i]=mean(mcar.surf$Income)
  rrp.imp.var.mcar[i]=var(mcar.surf$Income)
}
```

Figure 4.10 shows the distribution of the simulated 1000 means and variances. Unsurprisingly, the estimates are unbiased against the “true” complete data estimates. The reason is similar to the one we have given in section 4.3.4.

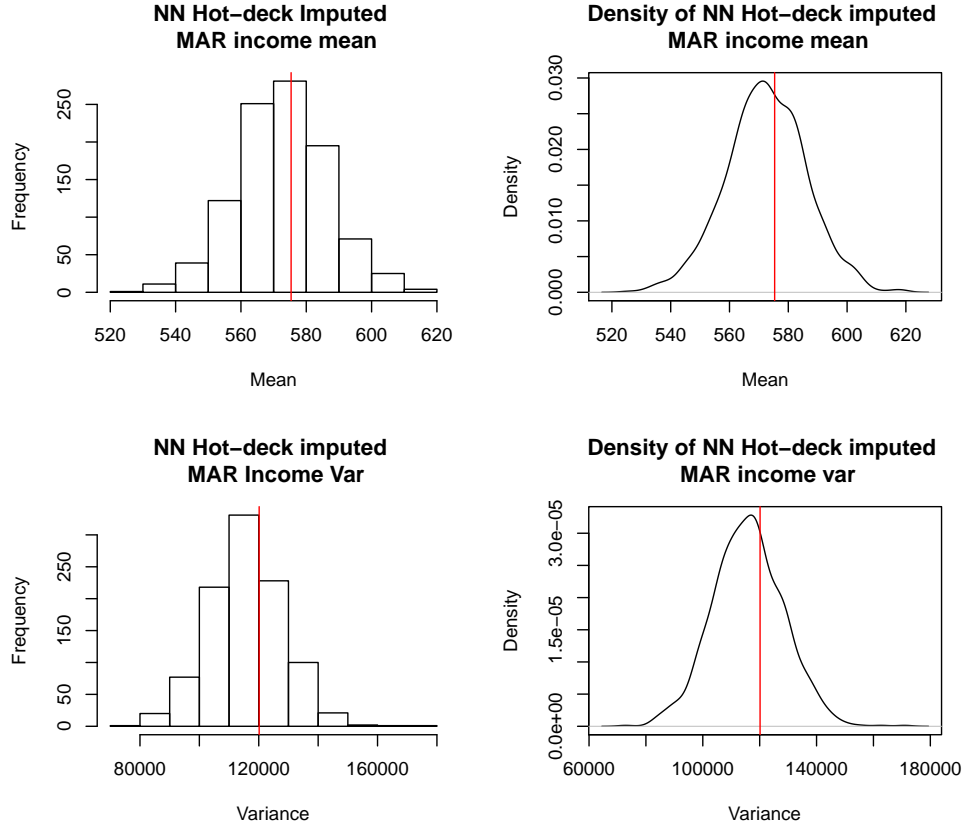


Figure 4.10: Nearest-Neighbor hot deck Imputed MAR SURF Income

## 4.4 Conclusion

Throughout this chapter, we have applied various single imputation methods to missing data. The results show that some are good at reducing bias, some are not. Unconditional mean imputation is the worst imputation method in terms of alleviating bias. As we have shown in figure 4.2, unconditional mean imputation even causes bias when the missing data is MCAR. This method is even worse than deletion methods, regarding bias. Conditional mean imputation and regression imputation have demonstrated much better estimates, compared to unconditional mean imputation, if the missingness is MAR and the variables upon which missingness depends are observed and complete. However, they still present biased variance estimates. Enders (2010) points out that both methods lack variability that would have been present if the data had been complete. We also see that stochastic regression imputation, and hot deck imputation, especially, Nearest-Neighbour hot deck imputation, present unbiased estimates. This is because both methods, regardless of whether they are explicit or implicit modelling methods, have a random sampling mechanism in them. For example, stochastic regression imputation has an added uncertainty term from a normal distribution; hot deck imputation random sample replacement values from observed units. This sampling mechanism provides the needed variability if the data had been complete.

As we have discussed in previous chapters, the problem with single imputation methods is that the imputation itself introduces variance, or more appropriately, imputation uncertainty. Unless missing data become not missing, even the best single imputation method underestimates variances, potentially by a substantial amount. However, we cannot measure this uncertainty with any of the single imputation methods. This is why we need methods such as resampling and multiple imputation.

# Chapter 5

## Using Non-parametric Resampling Methods to Incorporate Imputation Uncertainty

### 5.1 Introduction

As described in previous chapters, single imputation methods fill in missing values with imputed values, then standard complete data methods are used to analyse the imputed data. Those filled in values are treated as observed values. However, no matter how good single imputation is at reducing non-response bias, its imputed values are still not the “true” observed values. This means the missing values imputed by using single imputation methods are uncertain, there might be other candidates that can replace the missing values as well. This, as we have discussed previously, is called “Imputation Uncertainty”. We have also pointed out that imputation uncertainty introduces extra variance to the estimated variance of parameters based on the imputed complete data.

As Shao & Sitter (1996) point out, the potential underestimation of variance estimates of imputed data could be serious if variance estimates are only based on the single imputed data. This is because it does not account for the inflation in the variance due to missing data and imputation. In other words, single imputation does not account the added variance due to imputation uncertainty.

There have been some proposed methods to address this issue. Little & Rubin (2002) recommend resampling and Multiple Imputation methods as the most useful general tools for propagating imputation uncertainty. We focus on the discussion of resampling methods (bootstrap and jackknife) in this chapter, and there is further discussion of Multiple Imputation methods in later chapters.

### 5.2 Relationship between Resampling Methods and Imputation Uncertainty

So, why can resampling methods be used to propagate imputation uncertainty? Before we can answer this, let’s first discuss what resampling is.

A sample is not a census<sup>1</sup>. A single sample does not have all the units of our population. This means that if we draw another sample from the same population, the sample we get might be different from the previous sample. The difference in parameter estimates arising from different samples is called sampling variability or sampling error. Theoretically, If all possible samples with the same sample size were drawn from the same population, and estimates of each sample were computed, we would be able to construct the sampling distribution. Given the knowledge of sampling distribution, we can find confidence intervals for estimates drawn from a sample. However, as Lohr (1999) points out, we normally take only one sample from a population and have no information about the true population total. This is mainly due to the purpose of a sample survey is to provide a cost efficient way to gather information about the population. There is no point to conduct the same samples many times in the same period. If one has more resources, we'd better do a census instead. Hence, it is practically not worth to find the real sampling distribution.

If we have a large enough sample, one common method to approximate the unknown sampling distribution problem is to use the Central Limit Theorem (CLT)<sup>2</sup> to assume the sampling distribution of sample estimates based on all possible samples of the same sample size is approximately normal, regardless of the distribution of the original data in the sample. Then, we can compute confidence intervals for sample estimates of a single sample. We can say the “true” population estimates fall inside the confidence intervals at a given confidence level. For example, if the confidence level is 95%, we say that the confidence interval contains the true estimates most (95 out of 100) of the time.

However, we know that the computation of a confidence interval involves the calculation of the standard error. Sometimes, it is very hard to compute the standard error of a sample estimate, if the sample design is complex<sup>3</sup> or if post survey adjustments, such as post-stratification<sup>4</sup> are made to the weights. This leads to the method of resampling. In the hope that our single sample reproduces true properties of the whole population, we treat the sample as if it were a population. As we have described before, one possible way to measure the sampling error is to draw as many samples with the same sample size as one could from the same population, but this makes no sense as the idea of repeated sampling from true population leading to the sampling distribution is entirely theoretical, otherwise, we would just have a census or use a larger sample. Hence, we can only have one sample. Now, if we assume the sample is the smaller version of our population, then apart from computation time, it actually cost us nothing to draw as many samples as we like from the “smaller population” which has all the variables (or information) we want. “If the sample really is similar to the population, if the empirical probability mass function (epmf) of the sample is similar to the probability mass function of the population, then samples generated from the epmf should behave like samples taken from the population” (Lohr 1999). This means resampling methods help us to construct an approximation to the sampling distribution. Based on this approximate sampling distribution, we can easily estimate variances.

How are these related to imputation and imputation uncertainty? As described before, imputation uncertainty is due to treating single filled-in values as the true values, but, in fact,

---

<sup>1</sup>Census means we select all the units of a population.

<sup>2</sup>The Central Limit Theorem says that given a large enough sample size, the sampling distribution of the sample mean will follow an approximately normal distribution, and the mean of all samples from the same population will be approximately equal to the mean of the population with an arbitrary distribution.

<sup>3</sup>This means the sample is not simple random sampling.

<sup>4</sup>Post-stratification is a way to use auxiliary information on the population to improve precision.

other values can be used to replace the missing values, because the true values for the missing data are unknown. Now, suppose resampling methods generate  $M$  resamples from the original sample which has missing data. These resamples may have missing data as well. If we impute missing data for each resample, it is equivalent to drawing  $M$  samples for missing values from the unknown missing data population. In other words, this means each missing value is filled in with  $M$  different imputed values. As Little & Rubin (2002) state, such resampling methods propagate the uncertainty in the imputations and provide valid inferences.

### 5.3 The Simple Bootstrap for Complete Data

Let's have a look at a simple resampling method - the bootstrap resampling method. Let us consider a simple random sample with replacement of sample size  $n$ . Let  $Y = (y_1, y_2, \dots, y_n)$  be the observed values and for now we assume that there is no missing data. Efron (1979) provides us with the simplest form of bootstrap method. Let  $Y^{(b)}$  be a sample of size  $n$  obtained from the original sample  $Y$  by simple random sampling with replacement, where  $b$  indexes the drawn samples. The idea is to select a large number  $B$  of these bootstrap samples. If  $\hat{\theta}$  is an estimate of  $\theta$  based on the original sample  $Y$ , then  $\hat{\theta}^{(b)}$  is the corresponding estimate obtained by applying the original estimation method to  $Y^{(b)}$ . Hence, we get a set of estimates  $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$ .

We have noticed that bootstrap method requires us to take samples with replacement of size  $n$  from the original sample of size  $n$ . This means that the bootstrap sample size is the same as our original sample. One of the main advantages of sampling with replacement is when our sample size  $n$  is small and we cannot afford to have smaller sample size. However, as Politis & Ramano (1994) describe that the bootstrap method can use the sampling without replacement scheme. They call such method "subsampling". The limitation of "subsampling" method is that the sample size of resamples has to be smaller than the original sample size  $n$ , because it samples without replacement. The advantage is that it is more reliable than the bootstrap method when dealing with time series or any other form of dependent data (Geyer 2006). This is because the current observation's value depends on the previous observations' value, if we do a bootstrap sample of the time series data, the dependency will be disturbed. The subsampling method can overcome the dependency issue by applying a systematic sampling<sup>5</sup> method.

The bootstrap estimate of  $\theta$  is the average of the bootstrap estimates:

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)} \quad (5.1)$$

Variances can be estimated from the bootstrap distribution of  $\hat{\theta}^{(b)}$ , which is estimated by the distribution formed by the bootstrap estimates  $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$ . The bootstrap estimate of the variance of  $\hat{\theta}$  or  $\hat{\theta}_{boot}$  is:

$$\hat{V}_{boot} = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}_{boot})^2 \quad (5.2)$$

see (Little & Rubin 2002).

---

<sup>5</sup>Systematic sampling selects units from an ordered sampling frame, every  $k$ th unit in the frame is chosen (systematically) for inclusion in the sample, where  $0 < k < N$ ,  $N$  is the population size.

The bootstrap estimator  $\hat{\theta}_{boot}$  is less biased than the original estimator  $\hat{\theta}$ . This is because the bootstrap estimator  $\hat{\theta}_{boot}$  is supposed to be much closer to the expected value of  $\hat{\theta}_{boot}$  than the original estimator  $\hat{\theta}$ , so  $\hat{E}(\hat{\theta}_{boot}) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)}$  according to the law of large numbers<sup>6</sup>. In fact, this fact has been used to estimate bias. Let  $\hat{b}(\hat{\theta})$  be the bias of  $\hat{\theta}$ , then we have:

$$\begin{aligned}\hat{b}(\hat{\theta}) &= \hat{E}(\hat{\theta}_{boot}) - \hat{\theta} \\ &= \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)} - \hat{\theta}\end{aligned}$$

This  $\hat{b}(\hat{\theta})$  is called bootstrap bias estimator (Efron & Tibshirani 1993).

Little & Rubin (2002) also state that under quite general conditions,  $\hat{V}_{boot}$  is a consistent estimate of the variance of  $\hat{\theta}$  or  $\hat{\theta}_{boot}$  as  $n$  and  $B$  tend to infinity. Singh (1981) concludes that confidence intervals may be constructed based on the bootstrap that outperform those based on the normal approximation. If the bootstrap distribution is approximately normal, a  $100(1 - \alpha)\%$  bootstrap confidence interval for a scalar  $\theta$  can be computed as

$$I_{norm}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{boot}} \quad (5.3)$$

where  $z_{1-\alpha/2}$  is the  $100(1 - \alpha)\%$  percentile of the normal distribution. Alternatively if the bootstrap distribution is non-normal, a  $100(1 - \alpha)\%$  bootstrap confidence interval can be computed as

$$I_{emp}(\theta) = (\hat{\theta}^{(b,l)}, \hat{\theta}^{(b,u)}) \quad (5.4)$$

where  $\hat{\theta}^{(b,l)}$  and  $\hat{\theta}^{(b,u)}$  are the empirical  $(\alpha/2)$  and  $(1 - \alpha/2)$  quantiles of the bootstrap distribution of  $\theta$ . The empirical quantiles are the quantiles for the observed data. In this case, they are the quantiles of the bootstrap estimates  $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$ . Efron (1994) states that stable intervals based on Eq.(5.3) require bootstrap sample of the order of  $B = 200$ . Intervals based on Eq.(5.4) require much larger samples, for example  $B = 2000$  or more.

## 5.4 The Simple Bootstrap Applied to Imputed Incomplete Data

We have discussed the basic idea of the bootstrap method. Although the bootstrap has many applications in statistics, this chapter only focuses on its application to imputation.

Suppose we have a sample  $Y$ . It has  $n$  independent observations, and some observations have a missing  $Y$  value. Some imputation method  $Imp$  is used to impute missing values and an estimate  $\hat{\theta}$  of a true parameter  $\theta$  is computed by using the imputed data. The simple bootstrap is applied to the original sample  $Y$  to draw  $B$  bootstrap samples, then the imputation method  $Imp$  is applied to the  $B$  bootstrap samples and  $B$  estimates are computed. Little & Rubin (2002) emphasised that the imputation has to be done for each individual bootstrap resample. As we have described in section 5.2, imputation needs to be done many times in order to reflect imputation uncertainty.

---

<sup>6</sup>In probability theory, the law of large numbers is a theorem that describes the result of selecting the same sample a large number of times. According to the law, the average of the results obtained from a large number of samples should be close to the expected value, and will tend to become closer as more samples are selected (Grimmett & Stirzaker 1992).

Little & Rubin (2002) lists the following steps:

For  $b = 1, \dots, B$ :

- (a) Generate a bootstrap sample  $Y^{(b)}$  from the original incomplete sample  $Y$
- (b) Fill in the missing data in  $Y^{(b)}$  by applying the imputation procedure to the bootstrap sample  $Y^{(b)}$ ,  $\hat{Y}^{(b)} = \text{Imp}(Y^{(b)})$
- (c) Compute  $\hat{\theta}^{(b)}$  on the filled-in data  $\hat{Y}^{(b)}$  from (b).

Let us consider an example by applying the simple bootstrap method to the SURF data we introduced in the previous chapters. Suppose  $m$  “Income” values are missing at random, which has the missingness depending on the “Gender” variable. The Adjustment cell hot deck imputation method is applied to impute missing Income values. The initial adjustment cells were formed by “Ethnicity”, “Gender”, and “Qualification” variables. If a cell only has missing income values, a larger adjustment cell would be formed by using one less variable. This process goes on until there are no cells which only have missing income values. Then, the missing income values inside each cell are replaced by randomly selecting values from the observed income values in that cell. The following steps demonstrate the imputation of missing data by using the simple bootstrap method:

**Recipe: The simple bootstrap applied to incomplete data**

- Step 1:** Draw  $B = 20$  bootstrap samples  $Y^{(b)}$  with replacement from the original sample  $Y$ .  $Y^{(b)} \sim (y_1, \dots, y_n)$ ,  $b = 1, \dots, B$
- Step 2:** Fill in the missing data for each bootstrap sample  $Y$  by applying the Adjustment Cell hot deck imputation method
- Step 3:** Compute means  $\hat{\mu}^{(b)} = \hat{T}(Y^{(b)})$  on the filled-in data from Step 2.
- Step 4:** Calculate the bootstrap mean, the bootstrap variance of the mean by using Eq.(5.1) and Eq. (5.2).

**Please refer to Appendix A for the R code.**

For demonstration purposes, we choose to compute the bootstrap mean of the SURF Income variable, but the bootstrap method can be applied to estimate any parameters. First, we simulated 100 SURF data with incomplete income variables. The missing mechanism is MAR, with the missingness depends on Gender. So, the male respondents have 50% probability of missing income and the female respondents have 20% probability of missing income. We applied the Adjustment Cell hot deck imputation method to impute these 100 incomplete SURF data, and computed the income means and variances of income means for the imputed datasets. Then, we applied the bootstrap procedure that we have introduced above to impute the 100 incomplete SURF datasets, and computed the bootstrap income means and variances of income means for the imputed datasets. Figure 5.1 shows the results. The red vertical lines represent the true income mean and variance of the income mean.



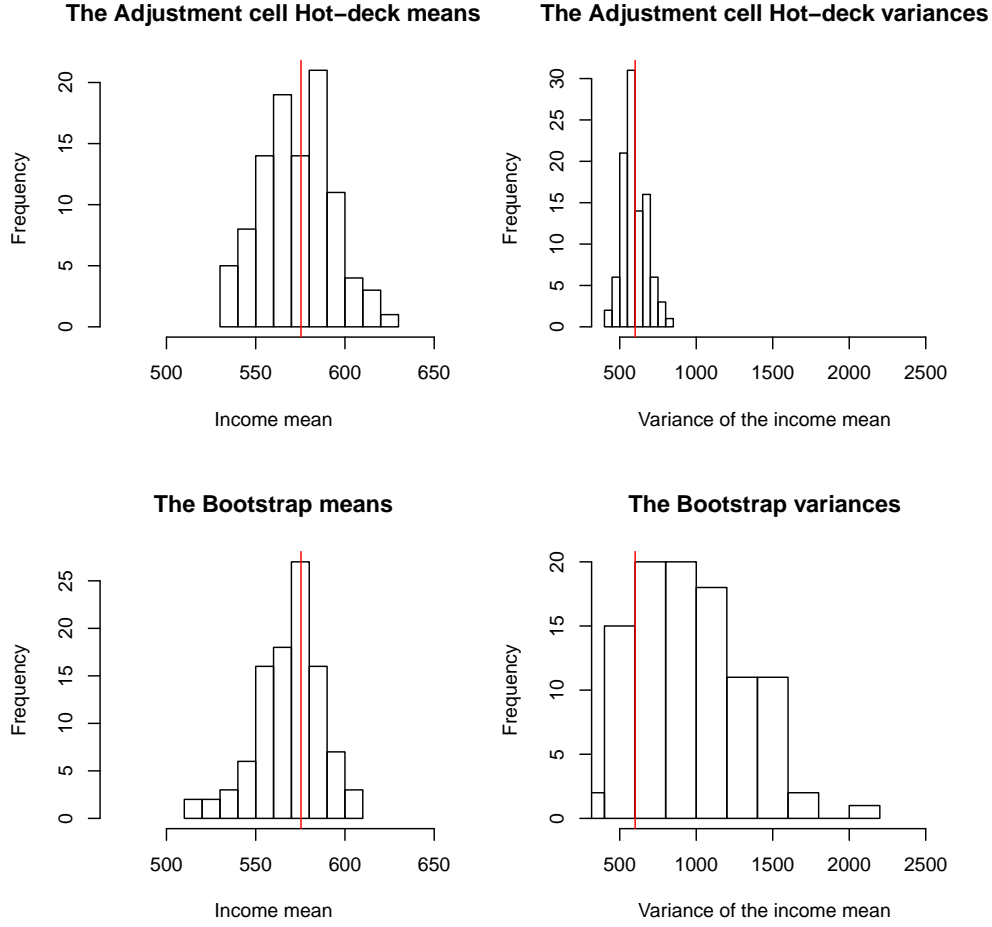


Figure 5.1: Comparing the Adjustment cell hot deck income means and variances with the bootstrap income means and variances for the simulated 100 incomplete SURF data

As we can see from Figure 5.1, the distribution of the bootstrap variances of the income means is much wider than the distribution of the variances of the Adjustment cell imputed datasets. This is what we expected. The imputation uncertainty which is properly reflected by using the bootstrap blows up the variances of the income mean.

## 5.5 The Simple Jackknife for Complete Data

The jackknife is a similar resampling concept to the bootstrap method. Like the bootstrap method, it allows the replicate groups to overlap. However, it differs from the bootstrap method by systematically deleting subgroups, rather than randomly resampling.

There are generally two cases of jackknife resampling. The first one is only based on deleting a single unit from the original sample sequentially (*JK1*). The other one is based on dropping multiple units from the original sample sequentially (*JKn*) (Efron & Gong 1983, Wu 1986). Hence, we see that the jackknife creates its resamples by deleting units from the original sample.

Little & Rubin (2002) provide us with an example of the simple jackknife for complete data. Let  $\hat{\theta}$  be a consistent estimate of a parameter  $\theta$  based on a sample  $Y_i, i = 1, \dots, n$  of independent observations. Let  $Y^{(\setminus j)}$  be a sample of size  $n - 1$  obtained by dropping the  $j$ th

observation from the original sample, and let  $\hat{\theta}^{(\setminus j)}$  be the estimate of  $\theta$  from this reduced sample. The quantity

$$\tilde{\theta}_j = n\hat{\theta} - (n-1)\hat{\theta}^{(\setminus j)} \quad (5.5)$$

is called a pseudo-value. The jackknife estimate of  $\theta$  is the average of the pseudo-values:

$$\hat{\theta}_{jack} = \frac{1}{n} \sum_{j=1}^n \tilde{\theta}_j = \hat{\theta} + (n-1)(\hat{\theta} - \bar{\theta}) \quad (5.6)$$

where  $\bar{\theta} = \frac{1}{n} \sum_{j=1}^n \hat{\theta}^{(\setminus j)}$ . The jackknife estimate of the variance of  $\hat{\theta}$  or  $\hat{\theta}_{jack}$  is

$$\hat{V}_{jack} = \frac{1}{n(n-1)} \sum_{j=1}^n (\tilde{\theta}_j - \hat{\theta}_{jack})^2 = \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}^{(\setminus j)} - \bar{\theta})^2 \quad (5.7)$$

The multiplier  $(n-1)/n$  in Eq. (5.7) is larger than the multiplier  $1/(B-1)$  in the bootstrap equation (5.2). This is because if  $(n = B > 2)$  and  $n$  and  $B$  are both integers, then  $(n-1)/n$  is much closer to 1 than  $1/(B-1)$ . This difference means the jackknife estimates  $\hat{\theta}^{(\setminus j)}$  of  $\theta$  tend to be closer to  $\hat{\theta}$  than the bootstrap estimates, since they differ from the computation of  $\hat{\theta}$  only by the value of a single observation. There is an intuitive way to understand why the jackknife estimates of  $\theta$  tend to be closer to the original  $\hat{\theta}$  than the bootstrap estimates. The bootstrap selects samples of size  $n$  obtained from the original sample  $Y$  by simple random sampling with replacement. This means that it is possible that the same units have been selected several times in the bootstrap sample. On the other hand, the jackknife method does not select the same sample units several times but deletes units from the original sample  $Y$  sequentially. Hence, the jackknife sample  $Y_{jackknife}^b$  distributions are closer to the distribution of the original sample.

The jackknife has similar properties to the bootstrap. That is, (a) the jackknife estimator  $\hat{\theta}_{jack}$  is less biased than the original estimator  $\hat{\theta}$ , and under quite general conditions (b)  $\hat{V}_{jack}$  is a consistent estimate of the variance of  $\hat{\theta}$  or  $\hat{\theta}_{jack}$  as  $n$  tends to infinity. From property (b), if the jackknife distribution is approximately normal, a  $100(1 - \alpha)\%$  confidence interval for a scalar  $\theta$  can be computed as

$$I_{norm}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{jack}} \quad (5.8)$$

where  $z_{1-\alpha/2}$  is the  $100(1 - \alpha)\%$  percentile of the normal distribution.

## 5.6 The Simple Jackknife Applied to Imputed Incomplete Data

Little & Rubin (2002) also provide us with an example of applying the simple jackknife to imputed incomplete data. Suppose we have a sample  $Y_i, i = 1, \dots, n$  of independent observations, but some observations  $i$  are incomplete. A consistent estimate  $\hat{\theta}$  of a parameter  $\theta$  is computed by filling in the missing values in  $Y$  using some imputation method  $Imp$ , producing imputed data  $\hat{Y} = Imp(Y)$ , and then estimating  $\theta$  from the filled in data  $\hat{Y}$ . This method can be implemented as follows:

For  $j = 1, \dots, n$

- (a) Delete  $j = 1$  observation from  $Y$ , yielding the sample  $Y^{(\setminus j)}$
- (b) Fill in the missing data in  $Y^{(\setminus j)}$  by applying the imputation procedure  $Imp$ , producing  $\hat{Y}^{(\setminus j)} = Imp(Y^{(\setminus j)})$
- (c) Compute  $\theta^{(\setminus j)}$  on the filled-in data  $\hat{Y}^{(\setminus j)}$  from (b).

Let us apply the simple jackknife to the SURF data. Suppose  $m$  “Income” values are missing at random, again, the Adjustment cell hot deck imputation method is used to impute missing “Income” values. The formation of the adjustment cells is exactly the same as we have introduced in the bootstrap section (Section 5.4). The following steps demonstrate the imputation of missing data by using the simple jackknife method (delete a single observation at a time):

**Recipe: The simple jackknife applied to imputed incomplete data**

- Step 1:** Delete the first  $j = 1$  observations from the original unimputed sample  $Y = \{y_i : i = 1, \dots, n\}$ . This yields sample  $Y^{(\setminus 1)} = \{y_i : i = 2, \dots, n\}$ .
- Step 2:** Fill in the missing “Income” data in  $Y^{(\setminus 1)}$  by applying the Adjustment cell hot deck procedure  $Imp$ , yielding  $\hat{Y}^{(\setminus 1)} = Imp(Y^{(\setminus 1)})$
- Step 3:** Compute  $\hat{\theta}^{(\setminus 1)}$  on the filled in data  $\hat{Y}^{(\setminus j)}$  from Step 2
- Step 4:** Repeat Step 1 to Step 3 for the  $2^{nd}, 3^{rd}, \dots, J^{th}$  observations
- Step 5:** Compute the jackknife estimate  $\hat{\theta}_{jack}$  and the  $\hat{V}_{jack}$  the jackknife estimate of the variance of  $\hat{\theta}_{jack}$  by using Eq.(5.6) and Eq.(5.7) respectively, replacing their  $n$  with  $k$ .

**Please refer to Appendix A for the R code.**

We have simulated 100 replicate SURF data with incomplete income variables, and compared the jackknife income means and variances of income means with the single imputation (Adjustment cell hot deck) imputed data. Figure 5.2 displays the comparison results, the vertical red lines are the true income mean and variance.

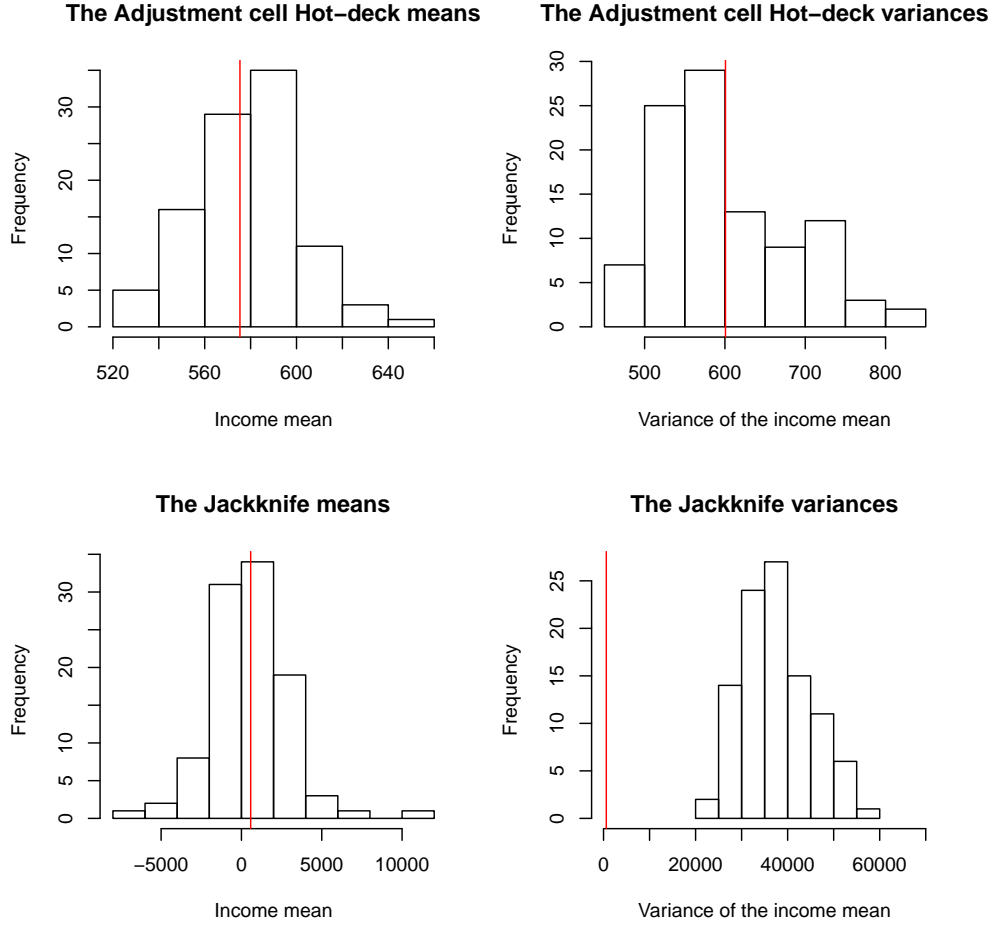


Figure 5.2: Comparing the Adjustment cell hot deck income means and variances with the jackknife income means and variances for the simulated 100 incomplete SURF data

It is a surprise to see the jackknife variances that big and far away from the true variance, and the jackknife means have negative values (Figure 5.2). This is not what we have expected nor what Little & Rubin (2002) suggest. We have expected that the jackknife variances would be larger than the Adjustment cell hot deck variances, but like the Bootstrap variances in Figure 5.1 which are centred around the true variance. So, what is the problem? By studying the Equation (5.5) to Equation (5.7), we have suspected that the imputation we have done for each jackknife samples blows up the jackknife variance and yields negative means.

In order to find out the cause of the problem, we applied a naive jackknife procedure to impute incomplete data. The naive jackknife procedure applies the imputation procedure just once to yield an imputed data set  $\hat{Y}$ , and then jackknife the  $\hat{Y}$  and compute the jackknife estimates. This method is implemented as follows:

For  $j = 1, \dots, n$

- (a) Impute the missing values in  $Y$ , producing imputed data  $\hat{Y} = \text{Imp}(Y)$
- (b) Delete  $j = 1$  observation from  $\hat{Y}$ , yielding the sample  $\hat{Y}^{(\setminus j)}$
- (c) Compute  $\theta^{(\setminus j)}$  on the jackknifed data  $\hat{Y}^{(\setminus j)}$  from (b) jackknife

Again, we applied the naive jackknife procedure to the simulated 100 replicate SURF data with incomplete income variables, and compared the naive jackknife income means and variances of income means with the single imputation (Adjustment cell hot deck) imputed data. Figure 5.3 displays the results.

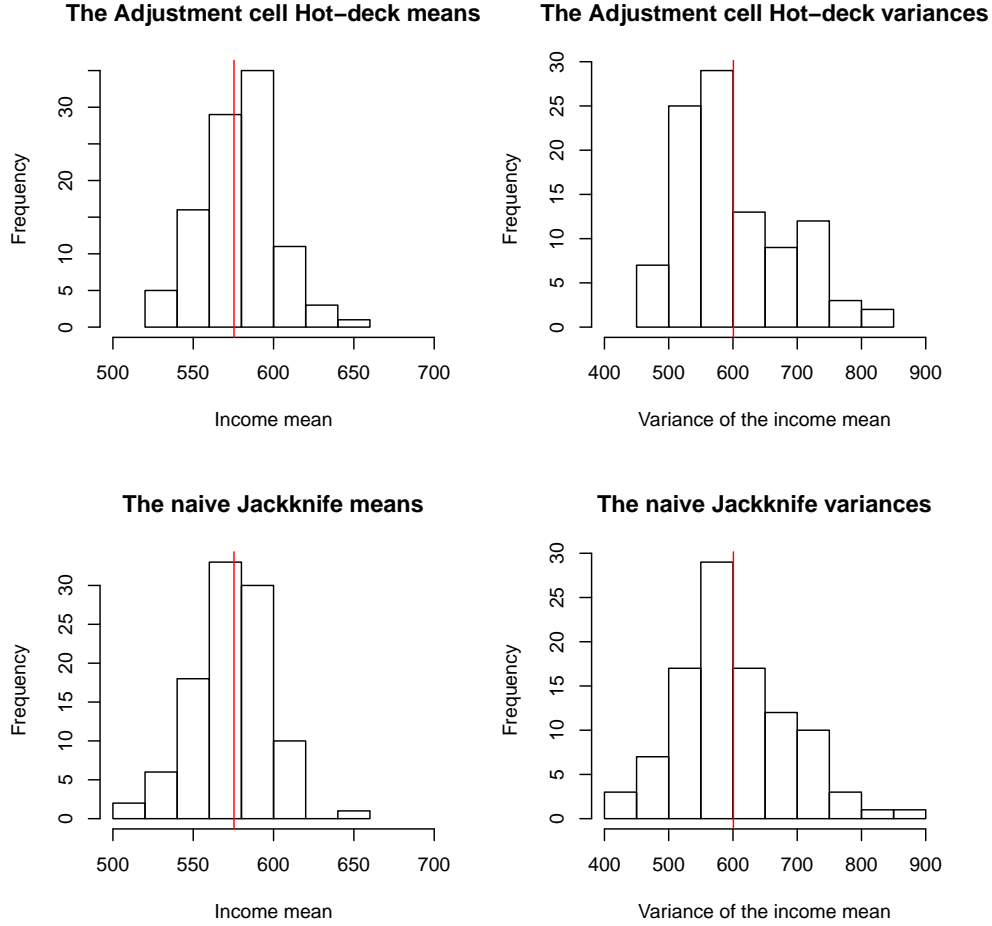


Figure 5.3: Comparing the Adjustment cell hot deck income means and variances with the naive jackknife income means and variances for the simulated 100 incomplete SURF data

The results shown in Figure 5.3 are what we have expected. The naive jackknife means and variances are unbiased. The naive jackknife variances are slightly bigger than the Adjustment hot deck variance due to the added variation from the jackknife samples. We have to point out that this naive jackknife approach does not propagate the imputation uncertainty. However, this investigation tells us two things: (1) we have properly implement the jackknife estimation formulas (ie. Eq (5.5) to Eq (5.7)); (2) our suspicion that imputing each jackknife sample could inflate the variance is correct.

So, how exactly does the imputation inflate the variance of estimates? Let's recall that the consistent estimate  $\hat{\theta}$  of a parameter  $\theta$  is computed from the imputed data  $\hat{Y}$  before implementing the jackknife to the incomplete dataset  $Y$ . Then,  $\hat{\theta}$  is used in the calculation of  $\hat{\theta}_{jack} = \hat{\theta} + (n-1)(\hat{\theta} - \bar{\theta})$ , where  $\bar{\theta} = 1/n \sum_{j=1}^n \hat{\theta}^{(\setminus j)}$ . If  $Y$  is complete, then  $\bar{\theta} \approx \hat{\theta}$ . However, if  $Y$  is incomplete, we cannot be sure that  $\bar{\theta}$  always approximates  $\hat{\theta}$  as the imputation uncertainty can make these two estimates very different from each other. This is why we see that some of the Income means (ie  $\hat{\theta}_{jack}$ ) are negative in Figure 5.2. As for the variance of the

estimate  $\hat{V}_{jack} = \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}^{(\setminus j)} - \bar{\theta})^2$ , we already know that the jackknife estimates  $\hat{\theta}^{(\setminus j)}$  of  $\theta$  tends to be closer to  $\hat{\theta}$  than the bootstrap estimates from Section 5.5. Hence, we conclude that  $\sum_{j=1}^n (\hat{\theta}^{(\setminus j)} - \bar{\theta})^2$  is much smaller than  $\sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}_{boot})^2$ , if the  $Y$  is complete. However, if the  $Y$  is incomplete, we know that the added variance due to imputation uncertainty can be very large as Figure 5.1 shows us. This means that  $\sum_{j=1}^n (\hat{\theta}^{(\setminus j)} - \bar{\theta})^2$  could be very large for jackknife, and the factor  $(n-1)/n \approx 1$  does not reduce the value of  $\hat{V}_{jack}$  by a large amount. This is why we see that all the jackknife variances for incomplete data are much larger than the true variance in Figure 5.2.

For the SURF examples, we let  $\theta = \mu$ . According to Eq (5.5), the pseudo value of the jackknife mean is  $\tilde{\mu}_{j,Jack} = n\hat{\mu} - (n-1)\hat{\mu}_{Jack}^{(\setminus j)}$ , and the naive pseudo value of the jackknife mean is  $\tilde{\mu}_{j,naive} = n\hat{\mu} - (n-1)\hat{\mu}_{naive}^{(\setminus j)}$ . Now, let's compare the distribution of the pseudo values of the jackknife mean  $\tilde{\mu}_{j,Jack}$ , the distribution of the naive pseudo value of jackknife mean  $\tilde{\mu}_{j,naive}$ , and the distribution of the SURF Income  $Y_{Income}$ . Figure 5.4 displays the distributions. The red vertical lines are the jackknife estimate of  $\tilde{\mu}_{j,Jack}$ , the naive jackknife estimate of  $\tilde{\mu}_{j,naive}$ , and the mean of SURF Income  $\bar{Y}_{Income} = \sum_{i=1}^n Y_{Income,i}/n$ .

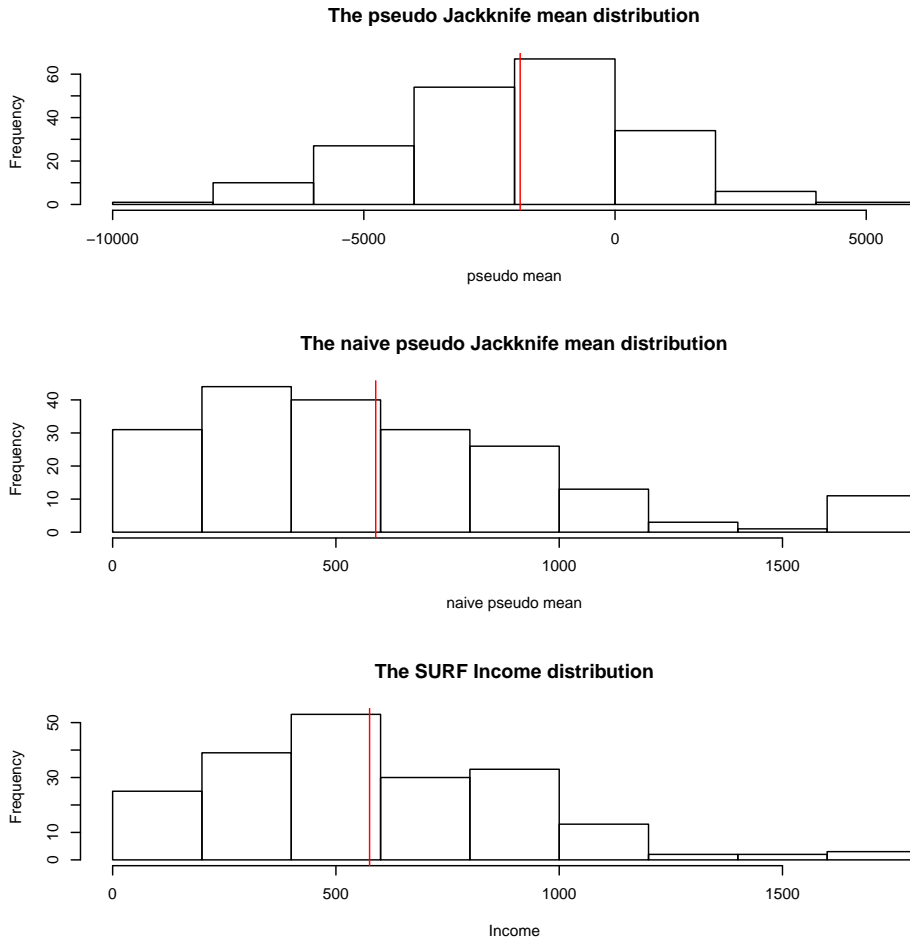


Figure 5.4: Comparing the distributions of the pseudo value of jackknife mean  $\tilde{\mu}_{j,Jack}$ , the naive pseudo value of jackknife mean  $\tilde{\mu}_{j,naive}$ , and the SURF Income  $Y_{Income}$

As shown in Figure 5.4, the distributions of the naive pseudo value of jackknife mean  $\tilde{\mu}_{j,naive}$ , and the SURF Income  $Y_{Income}$  are almost identical. This gives them similar esti-

mate of mean and variance. The distribution of the pseudovalue of jackknife mean  $\tilde{\mu}_{j,Jack}$  is much wider than the other two distributions. This means large variance.

The exact reason of why the jackknife procedure applied to imputed incomplete data yields biased estimates of variance needs some further study. Due to the scope of this thesis, our investigation stops here. What we take from this is that the jackknife procedure, proposed by Little & Rubin (2002), performs poorly when imputation is applied to impute each jackknife sample. Hence, we use the bootstrap as our resampling method for later chapters.

## 5.7 Conclusion

In this chapter, we have introduced a couple of resampling methods to deal with the imputation uncertainty problem. Generally speaking, the jackknife method and the Bootstrap method should produce similar results. However, Lohr (1999) points out that the jackknife can perform poorly if the estimate  $\hat{\theta}$  is not smooth<sup>7</sup> (eg, median, quantiles). One of the advantages of the bootstrap method is its ability to deal with non smooth estimates, because it does not delete units. Compared to the jackknife method, one of the disadvantages of the bootstrap method is that it requires more computation. This is because the required number of resamples is usually very large for the bootstrap (Little & Rubin 2002). However, we have discovered that current jackknife procedure, proposed by Little & Rubin (2002), produces large and biased variances.

The resampling methods are robust, easy to use, and efficient for dealing with imputation uncertainty, but they require large samples and are computation intensive. As we have stated in this chapter, the resampling methods sometimes need to produce 200 or more resamples in order to give a proper estimate. This places a burden on both computation and data storage. Also, we have discussed that resampling methods treat the original sample as a mini version of the population, in other words, they are based on large-sample theory. Hence, if the sample is very small, the quality of their estimates are doubtful (Little & Rubin 2002).

---

<sup>7</sup>A small change in the data can cause a large change in the statistic.

# Chapter 6

## Likelihood based imputation methods

### 6.1 Introduction

If we have a complete observed data  $y$ , Azzalini (1996) tells us that the first practical problem is to find a value of the estimate  $\theta$  close to the true estimate value  $\theta^*$ . This  $\theta$  defines the shape of the distribution of  $y$ . Maximum likelihood estimation provides us a way to estimate  $\theta$  close to the true  $\theta^*$ , given a vector of observed  $\mathbf{y} = (y_1, \dots, y_n)^T$ . As we have introduced in the previous chapter, if the probability density function (pdf) of a data point is  $f(y_i; \theta)$ ,  $i = (1, \dots, n)$ , then its likelihood function is  $L(\theta; y) = \prod_{i=1}^n f(y_i, \theta)$  for sample with replacement, and its log-likelihood is  $\ell(\theta; y) = \log L(\theta; y)$ . To compute the maximum likelihood estimate of  $\theta$  based on the distribution of  $y$ , then we find the value of  $\theta$  that maximises the log-likelihood function.

The maximum likelihood estimation method becomes more complicated in the presence of missing data. Let's denote the complete data  $\mathbf{y}$  as  $(y_{obs}, y_{mis})$ , where  $y_{obs}$  denotes the observed but “incomplete” data and  $y_{mis}$  denotes the unobserved or “missing” data, assuming the missing data are MAR. Then, we can express the joint density  $f(y; \theta)$  as:

$$\begin{aligned} f(y; \theta) &= f(y_{obs}, y_{mis}; \theta) \\ &= f_1(y_{obs}; \theta) \times f_2(y_{mis}|y_{obs}; \theta), \end{aligned}$$

where  $f_1$  is the joint density of  $y_{obs}$  and  $f_2$  is the joint density of  $y_{mis}$  given the observed data  $y_{obs}$ , respectively. We also want to point out that  $\log f(y; \theta) = \log L(\theta; y)$ . Thus, we have:

$$\ell(\theta; y) = \ell_{obs}(\theta; y_{obs}) + \log f_2(y_{mis}|y_{obs}; \theta) \quad (6.1)$$

where  $\ell_{obs}(\theta; y_{obs})$  is the observed data log-likelihood ( $\log f_1(y_{obs}|\theta)$ ).

To maximize the incomplete data log-likelihood  $\ell(\theta; y)$  is not straight forward. Since some of the units of vector  $y$  are not observed,  $\ell$  cannot be evaluated and maximized. The Expectation Maximization (EM) algorithm provides an iterative algorithm for parameter estimation by maximum likelihood when there are incomplete data. It attempts to maximize  $\ell(\theta; y)$  iteratively, by replacing missing values by their conditional expectation given the observed data  $y_{obs}$ . This expectation is computed with respect to the distribution of the complete-data evaluated at the current estimate of  $\theta$ . As discussed, Equation (3.8) for the E step, and Equation (3.9) for the M step. The E-step and the M-step are repeated again and again until the difference  $\ell(\theta^{(t+1)}) - \ell(\theta^{(t)})$  is less than some prescribed small quantity.



Equation (3.8) for the E step of EM finds the expected complete-data loglikelihood if  $\theta$  were  $\theta^{(t)}$ :

$$Q(\theta|\theta^{(t)}) = \int \ell(\theta|y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis}$$

Equation (3.9) for the M step of EM determines  $\theta^{(t+1)}$  by maximizing this expected complete-data loglikelihood:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta|\theta^{(t)})$$

## 6.2 Applying EM Algorithm to the Exponential Family

This thesis is about dealing with missing data for social survey data. We assume that these data arise from members of the exponential family of distributions. Hence, this chapter only focuses on applying EM algorithm to distributions which are exponential families.

The exponential family has density

$$f(y|\theta) = \exp(g(\theta)^T t(y) - b(\theta) + c(y))$$

where  $g(\theta)$ ,  $t(y)$ ,  $b(\theta)$  and  $c(y)$  are known functions. And  $g(\theta)$  and  $t(y)$  are vectors with length  $p = \dim(\theta)$ . And note that (under regularity conditions)

$$\begin{aligned} 1 &= \int f(y|\theta) dy \\ 0 &= \frac{\partial}{\partial \theta} \int f(y|\theta) dy \\ &= \int \frac{\partial f(y|\theta)}{\partial \theta} dy \\ &= \int \frac{1}{f(y|\theta)} \frac{\partial f(y|\theta)}{\partial \theta} f(y|\theta) dy \\ &= \int \frac{\partial \ell(\theta|y)}{\partial \theta} f(y|\theta) dy \\ &= \int \frac{\partial (g(\theta)^T t(y) - b(\theta) + c(y))}{\partial \theta} f(y|\theta) dy \\ &= \int \left( \frac{dg(\theta)^T}{d\theta} t(y) - \frac{db(\theta)}{d\theta} \right) f(y|\theta) dy \\ &= \frac{\partial g(\theta)^T}{\partial \theta} E_{y|\theta}[t(y)] - \frac{\partial b(\theta)}{\partial \theta} \\ E_{y|\theta}[t(y)] &= \left[ \frac{\partial g(\theta)^T}{\partial \theta} \right]^{-1} \frac{\partial b(\theta)}{\partial \theta} \end{aligned}$$

For  $n$  observations  $\mathbf{y} = (y_1, \dots, y_n)^T$  of a random variable from the exponential family the log likelihood is

$$\ell(\theta|\mathbf{y}) = g(\theta)^T \sum_{i=1}^n t(y_i) - nb(\theta) + \sum_{i=1}^n c(y_i)$$

The complete data Maximum Likelihood estimate is

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \ell(\theta|\mathbf{y}) \\ &= \underset{\theta}{\operatorname{argmax}} \left( g(\theta)^T \sum_{i=1}^n t(y_i) - nb(\theta) + \sum_{i=1}^n c(y_i) \right) \end{aligned}$$

This is the solution of the ML equations

$$\frac{\partial \ell(\theta|\mathbf{y})}{\partial \theta} = \frac{\partial g(\theta)^T}{\partial \theta} \sum_{i=1}^n t(y_i) - n \frac{\partial b(\theta)}{\partial \theta} \stackrel{\text{set}}{=} 0$$

Now assume the values  $\mathbf{y}_o = (y_1, \dots, y_r)^T$  are observed, and  $\mathbf{y}_m = (y_{r+1}, \dots, y_n)^T$  are missing. The incomplete data Maximum Likelihood estimate is

$$\begin{aligned} \hat{\theta}_{obs} &= \underset{\theta}{\operatorname{argmax}} \ell(\theta|\mathbf{y}_o) \\ &= \underset{\theta}{\operatorname{argmax}} \left( g(\theta)^T \sum_{i=1}^r t(y_i) - r b(\theta) + \sum_{i=1}^r c(y_i) \right) \end{aligned}$$

which is the solution of the incomplete data ML equations

$$\frac{\partial \ell(\theta|\mathbf{y})}{\partial \theta} = \frac{\partial g(\theta)^T}{\partial \theta} \sum_{i=1}^r t(y_i) - r \frac{\partial b(\theta)}{\partial \theta} \stackrel{\text{set}}{=} 0$$

EM Algorithm: Initialise with estimate  $\theta^0$ . Then iterate:

E step:

$$\begin{aligned} Q(\theta|\theta^t; \mathbf{y}_o) &= E_{\mathbf{y}_m|\theta^t, \mathbf{y}_o}[\ell_c(\theta|\mathbf{y})] \\ &= E_{\mathbf{y}_m|\theta^t, \mathbf{y}_o} \left[ g(\theta)^T \sum_{i=1}^n t(y_i) - n b(\theta) + \sum_{i=1}^n c(y_i) \right] \\ &= g(\theta)^T \sum_{i=1}^r t(y_i) + g(\theta)^T \sum_{i=r+1}^n E_{y|\theta^t}[t(y_i)] - n b(\theta) + \sum_{i=1}^r c(y_i) + \sum_{i=r+1}^n E_{y|\theta^t}[c(y_i)] \end{aligned}$$

M step: Solve the M-step equations

$$\frac{\partial Q(\theta|\theta^t; \mathbf{y}_o)}{\partial \theta} = \frac{\partial g(\theta)^T}{\partial \theta} \sum_{i=1}^r t(y_i) + \frac{\partial g(\theta)^T}{\partial \theta} \sum_{i=r+1}^n E_{y|\theta^t}[t(y_i)] - n \frac{\partial b(\theta)}{\partial \theta} \stackrel{\text{set}}{=} 0$$

Note that if the functions  $t(y)$  defining the sufficient statistics  $\sum_i t(y_i)$  are linear in the data  $y_i$ , then

$$E_{y|\theta}[t(y)] = E_{y|\theta}[t(y)] = t(E_{y|\theta}[y])$$

in which case we can replace the missing data  $y_{r+1}, \dots, y_n$  at each step with their current expected values

$$\tilde{y}_i^t = \begin{cases} y_i & \text{for } i = 1, \dots, r \\ E_{y|\theta^t}[y_i] & \text{for } i = r+1, \dots, n \end{cases}$$

and solve the complete data likelihood equations using the data set  $\tilde{\mathbf{y}}^t = (\tilde{y}_1^t, \dots, \tilde{y}_n^t)^T$ :

$$\frac{\partial \ell(\theta|\tilde{\mathbf{y}}^t)}{\partial \theta} = \frac{\partial g(\theta)^T}{\partial \theta} \sum_{i=1}^n t(\tilde{y}_i^t) - n \frac{\partial b(\theta)}{\partial \theta} \stackrel{\text{set}}{=} 0$$

If  $g(\theta)$  is a canonical link function which means  $g(\theta) = \theta$ , then the exponential family is in canonical form, and we can re-express the above equation as:

$$\frac{\partial \ell(\theta|\tilde{\mathbf{y}}^t)}{\partial \theta} = 1^T \sum_{i=1}^n t(\tilde{y}_i^t) - n \frac{\partial b(\theta)}{\partial \theta} \stackrel{\text{set}}{=} 0$$

### 6.3 Example one: applying EM Algorithm to the Univariate Normal Data with Missing Values

Let's consider a univariate complete data  $\mathbf{y} = (y_1, \dots, y_n)^T$  which is a random sample from  $N(\mu, \sigma^2)$ . Then, we can express its probability density function (pdf) as:

$$\begin{aligned} \prod_{i=1}^n f(y_i|\theta) &= (\mathbf{y}; \mu, \sigma^2) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2} \right\} \\ &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i^2 - 2\mu y_i + \mu^2) \right\} \\ &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} (\sum y_i^2 - 2\mu \sum y_i + n\mu^2) \right\} \end{aligned} \quad (6.2)$$

which implies that  $(\sum y_i, \sum y_i^2)$  are sufficient statistics for  $\theta = (\mu, \sigma^2)^T$ . The complete data log-likelihood function is:

$$\begin{aligned} \ell(\mu, \sigma^2; \mathbf{y}) &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2} + \text{constant} \\ &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n y_i - \frac{n\mu^2}{\sigma^2} + \text{constant} \end{aligned} \quad (6.3)$$

where

$$\begin{aligned} t(y_i) &= y_i & g(\theta) &= \frac{\mu}{\sigma^2} \\ b(\theta) &= \frac{\mu^2}{\sigma^2} & c(y_i) &= y_i^2 \end{aligned}$$

Clearly, the complete data log-likelihood  $\ell(\mu, \sigma^2; \mathbf{y})$  has a linear relationship with the complete-data sufficient statistics. Now, let's consider an incomplete data case. Suppose  $y_i$  are iid  $N(\mu, \sigma^2)$  where  $y_i, i = 1, \dots, r$  are observed, and  $y_i, i = r+1, \dots, n$  are missing, and assume the missing-data mechanism is ignorable. The observed data vector is  $\mathbf{y}_{obs} = (y_1, \dots, y_r)^T$ . The log-likelihood function becomes  $\ell(\mu, \sigma^2; \mathbf{y}) = \ell(\mu, \sigma^2; \mathbf{y}_{obs}) + \log f(\mathbf{y}_{mis}|\mathbf{y}_{obs}; \mu, \sigma)$ . We need to work out the  $E_{\mu, \sigma^2}[\ell(\mu, \sigma^2; \mathbf{y})|\mathbf{y}_{obs}]$ . However, the complete-data  $\mathbf{y}$  is from the exponential family. This means:

$$\begin{aligned} Q(\theta|\theta^t; \mathbf{y}) &= \ell(\mu, \sigma^2; \mathbf{y}_{obs}) + \log f(\mathbf{y}_{mis}|\mathbf{y}_{obs}; \mu, \sigma) \\ &= \frac{\mu}{\sigma^2} \sum_{i=1}^r y_i + \frac{\mu}{\sigma^2} \sum_{i=r+1}^n E_{y|\theta^t}[y_i] - \frac{1}{2\sigma^2} \sum_{i=1}^r y_i^2 - \frac{1}{2\sigma^2} \sum_{i=r+1}^n E_{y|\theta^t}[y_i^2] - \frac{n\mu^2}{2\sigma^2} + \frac{n}{2} \log\left(\frac{1}{\sigma^2}\right) \\ &\quad + \frac{n}{2} \log\left(\frac{1}{2\pi}\right) \\ &= \frac{\mu}{\sigma^2} \sum_{i=1}^r y_i + \frac{\mu}{\sigma^2} (n-r) \hat{\mu}^t - \frac{1}{2\sigma^2} \sum_{i=1}^r y_i^2 - \frac{1}{2\sigma^2} (n-r) [(\hat{\mu}^t)^2 + (\hat{\sigma}^t)^2] - \frac{n\mu^2}{2\sigma^2} + \frac{n}{2} \log\left(\frac{1}{\sigma^2}\right) \\ &\quad + \frac{n}{2} \log\left(\frac{1}{2\pi}\right) \end{aligned}$$

Hence, the E-step computes:

$$E_{y|\theta^t}\left(\sum_{i=1}^n y_i|\mathbf{y}_{obs}\right) \quad \text{and} \quad E_{y|\theta^t}\left(\sum_{i=1}^n y_i^2|\mathbf{y}_{obs}\right)$$

instead of computing the expectation of the complete-data log-likelihood function. Thus, at the  $t^{th}$  iteration of the E-step, compute

$$E\left(\sum_{i=1}^n y_i | \theta^{(t)}, y_{obs}\right) = \sum_{i=1}^r y_i + (n-r)\mu^{(t)} \quad (6.4)$$

$$E\left(\sum_{i=1}^n y_i^2 | \theta^{(t)}, y_{obs}\right) = \sum_{i=1}^r y_i^2 + (n-r)[(\mu^{(t)})^2 + (\sigma^{(t)})^2] \quad (6.5)$$

For the M-step, first note that the complete-data maximum likelihood estimates of  $\mu$  and  $\sigma^2$  are:

$$\hat{\mu} = \frac{\sum_{i=1}^n y_i}{n} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n}\right)^2 = \frac{\sum_{i=1}^n y_i^2}{n} - \hat{\mu}^2$$

The M-step is defined by substituting the expectations computed in the E-step for the complete-data sufficient statistics on the right-hand side of the above expressions to obtain expressions for the new iterates of  $\mu$  and  $\sigma^2$ . Note that complete-data sufficient statistics themselves cannot be computed directly since  $y_{r+1}, \dots, y_n$  have not been observed. We get the expressions:

$$\mu^{(t+1)} = E\left(\sum_{i=1}^n y_i | \theta^{(t)}, y_{obs}\right) / n \quad (6.6)$$

$$(\sigma^{(t+1)})^2 = E\left(\sum_{i=1}^n y_i^2 | \theta^{(t)}, y_{obs}\right) / n - (\mu^{(t+1)})^2 \quad (6.7)$$

Substitute Eq.(6.4) and Eq.(6.5), we have:

$$\begin{aligned} \mu^{(t+1)} &= \frac{1}{n} \left[ \sum_{i=1}^r y_i + (n-r)\mu^{(t)} \right] \\ (\sigma^{(t+1)})^2 &= E\left(\sum_{i=1}^n y_i^2 | \theta^{(t)}, y_{obs}\right) / n - (\mu^{(t+1)})^2 \\ &= \frac{1}{n} \left[ \sum_{i=1}^r y_i^2 + (n-r)[(\mu^{(t)})^2 + (\sigma^{(t)})^2] \right] - \frac{1}{n^2} \left[ \sum_{i=1}^r y_i + (n-r)\mu^{(t)} \right]^2 \\ &= \frac{1}{n} \left[ \sum_{i=1}^r y_i^2 - \frac{1}{n} \left( \sum_{i=1}^r y_i \right)^2 \right] + \frac{(n-r)(r-2n)}{n^2} (\mu^{(t)})^2 + \frac{n-r}{n} (\sigma^{(t)})^2 \end{aligned}$$

Thus, the E-step involves evaluating Eq.(6.4) and Eq.(6.5) beginning with starting values  $\mu^{(0)}$  and  $\sigma^{2(0)}$ . M-step involves substituting these in Eq. (6.6) and (6.7) to calculate new values  $\mu^{(1)}$  and  $\sigma^{2(1)}$ , etc. Thus, the EM algorithm iterates successively between Eq.(6.4) and Eq.(6.5) and Eq. (6.6) and (6.7). Of course, the EM algorithm is unnecessary for this example since the explicit ML estimates  $(\hat{\mu}, \hat{\sigma}^2)$  are available.

Let us apply the EM algorithm to a univariate normal data with missing values. Again, we use the SURF data set for the demonstration. For demonstration purpose, the missing SURF's "Income" variable values were created by the missing completely at random (MCAR) mechanism. That is,

$$f(\text{MissingIncome} | \text{Income}, \theta) = f(\text{MissingIncome} | \theta) \quad \text{for all Income and } \theta$$

where  $\theta = (\mu, \sigma^2)$ .

The following steps show how to apply the EM algorithm to the impute univariate normal data with missing values.

**Recipe: EM algorithm - Univariate Normal Data**

- Step 1:** Estimate the initial mean and variance from the observed data
- Step 2:** Calculate the expectations of the sufficient statistics (E-step)
- Step 3:** Calculate the mean and variance by using the sufficient statistics from the E-step and sample size  $n = 200$  (M-step)
- Step 4:** Iterate E and M steps until it converges

Please refer to Appendix B for the R code.

Then, the above procedure was repeated one thousand times. Each time we created a different set of missing data (MCAR) which has 50 missing values for the SURF's Income variable, and applied the EM algorithm to impute the missing data. Figure 6.1 shows the distribution of the 1000 simulated income variable's means and variances. The dashed red vertical line represents the means of the 1000 means and variances. The solid red vertical line represents the true mean and variance of the original complete data.

From previous sections, we understand that the EM algorithm replaces missing values with the expected values given the updated  $\theta$  and observed values. This is somehow similar to the unconditional mean imputation method as  $E_{y|\theta^t}(\sum_{i=1}^n y_i | y_{obs})/n$  for a univariate normal distribution is actually the mean. However, unlike unconditional mean imputation which produces biased estimates even when the missing data is MCAR (please refer to Chapter 4, Section 4.2), the EM algorithm produces unbiased estimates. As Schafer & Graham (2002) claim that the EM algorithm as one of the state-of-the-art missing data techniques yields unbiased parameter estimates if the data are at least missing at random (MAR). However, for the case of MAR, we need to point out that if the EM algorithm does not incorporate the variables that the missingness depends on, then the estimates would still be biased. In fact, if the variables that the missingness depends on are excluded in the construction of the EM algorithm, the MAR actually becomes NMAR. Hence, Enders (2010, p.106) states that the EM algorithm needs to involve information from other variables, so the E step of the algorithm should really use the conditional expectations to replace the missing components of the formulas.

## 6.4 Applying EM Algorithm to Bivariate Normal data with Missing Data on Both Variables

In this section, we look at how the EM algorithm works on bivariate normal data with missing data on both variables. Before we can easily introduce the EM algorithm to bivariate normal sample with missing data on both variables, we need to take time to review the content of bivariate normal distribution. Suppose we have complete sample data for variables  $Y_1$  and  $Y_2$ , and both variables are normally distributed. Hence, we have a bivariate normal distribution

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_{y_1}\sigma_{y_2}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[ \frac{(y_1 - \mu_{y_1})^2}{\sigma_{y_1}^2} + \frac{(y_2 - \mu_{y_2})^2}{\sigma_{y_2}^2} - \frac{2\rho(y_1 - \mu_{y_1})(y_2 - \mu_{y_2})}{\sigma_{y_1}\sigma_{y_2}} \right] \right) \quad (6.8)$$

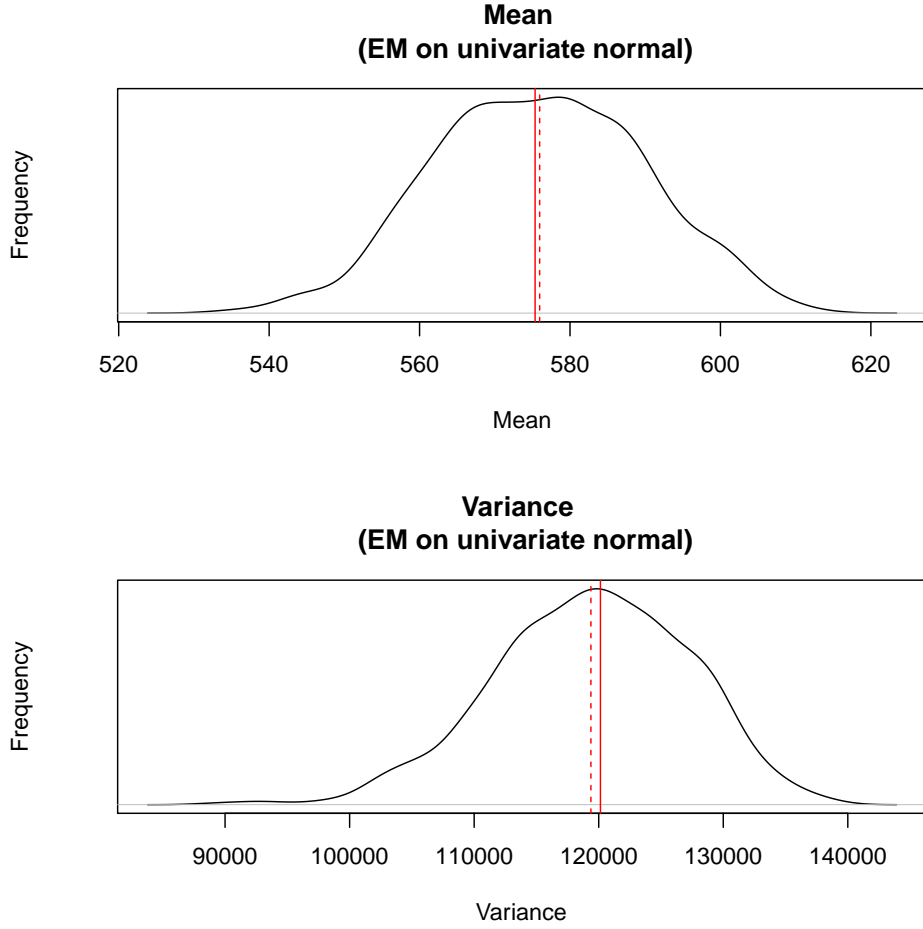


Figure 6.1: The distributions of means and variances of the 1000 replicate SURF data's income variables imputed by the EM algorithm (The dashed red vertical line represents the means of the 1000 means and variances. The solid red vertical line represents the “true” mean and variance of the original complete data)

where  $\rho$  is the correlation between  $Y_1$  and  $Y_2$ , and

$$\mu = \begin{pmatrix} \mu_{y_1} \\ \mu_{y_2} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_{y_1}^2 & \rho \sigma_{y_1} \sigma_{y_2} \\ \rho \sigma_{y_1} \sigma_{y_2} & \sigma_{y_2}^2 \end{pmatrix}$$

It is not hard to work out the values for  $\mu_{y_1}$  and  $\mu_{y_2}$  respectively. To find the values for  $\Sigma$ , we need to realise that

$$\Sigma = \begin{pmatrix} \sigma_{y_1}^2 & \sigma_{(y_1, y_2)}^2 \\ \sigma_{(y_1, y_2)}^2 & \sigma_{y_2}^2 \end{pmatrix} = \begin{pmatrix} S_{11}/n - \hat{\mu}_{y_1}^2 & S_{12}/n - \hat{\mu}_{y_1} \hat{\mu}_{y_2} \\ S_{12}/n - \hat{\mu}_{y_1} \hat{\mu}_{y_2} & S_{22}/n - \hat{\mu}_{y_2}^2 \end{pmatrix} \quad (6.9)$$

where  $S_{ij}$  are the sufficient statistics:

$$s_1 = \sum_{i=1}^n y_{i1}, \quad s_2 = \sum_{i=1}^n y_{i2}, \quad s_{11} = \sum_{i=1}^n y_{i1}^2, \quad s_{22} = \sum_{i=1}^n y_{i2}^2, \quad s_{12} = \sum_{i=1}^n y_{i1} y_{i2}, \quad (6.10)$$

Refer back to section 6.3, we understand that the E step of the EM algorithm is thus to find the conditional expectation of the sums in Eq.6.10, given  $Y_{obs}$  and  $\theta = (\mu, \Sigma)$ . Then, the M step is to update  $\theta$ , given the updated sufficient statistics.

Now, suppose we have two normally distributed variables  $Y_1$  and  $Y_2$  with a general pattern of missing data. The first group of units have both  $Y_1$  and  $Y_2$  observed, the second group of units have  $Y_1$  observed but are missing  $Y_2$ , and the third group of units have  $Y_2$  observed but are missing  $Y_1$ . Clearly, it is not hard to work out the sufficient statistics for the group of units that have both variables observed. For the groups of units with one variable observed but the other missing, we need to introduce a regression model to impute the missing values first, then we can work out the sufficient statistics for these groups.

Hence, suppose we have  $r_1$  units with both variables observed,  $r_2$  units with  $Y_1$  observed but  $Y_2$  missing, and  $r_3$  units with  $Y_2$  observed but  $Y_1$  missing, given  $r_1 + r_2 + r_3 = n$ .

Then the E step of the algorithm calculates:

1. For the group of units with both variables observed:  $s_1^{r_1} = \sum_{i=1}^{r_1} y_{i1}$ ,  $s_2^{r_1} = \sum_{i=1}^{r_1} y_{i2}$ ,  $s_{11}^{r_1} = \sum_{i=1}^{r_1} y_{i1}^2$ ,  $s_{22}^{r_1} = \sum_{i=1}^{r_1} y_{i2}^2$ ,  $s_{12}^{r_1} = \sum_{i=1}^{r_1} y_{i1}y_{i2}$
2. For the group of units with  $Y_1$  observed but  $Y_2$  missing:

$$\begin{aligned} s_1^{r_2} &= \sum_{i=1}^{r_2} y_{i1} \\ s_{11}^{r_2} &= \sum_{i=1}^{r_2} y_{i1}^2 \\ s_2^{r_2} &= E(y_{i2}|y_{i1}, \mu, \Sigma) = \beta_{20.1} + \beta_{21.1}y_{i1} \\ s_{22}^{r_2} &= E(y_{i2}^2|y_{i1}, \mu, \Sigma) = (\beta_{20.1} + \beta_{21.1}y_{i1})^2 + \sigma_{22.1} \\ s_{12}^{r_2} &= E(y_{i2}y_{i1}|y_{i1}, \mu, \Sigma) = (\beta_{20.1} + \beta_{21.1}y_{i1})y_{i1}, \end{aligned}$$

3. For the group of units with  $Y_2$  observed but  $Y_1$  missing:

$$\begin{aligned} s_2^{r_3} &= \sum_{i=1}^{r_3} y_{i2} \\ s_{22}^{r_3} &= \sum_{i=1}^{r_3} y_{i2}^2 \\ s_1^{r_3} &= E(y_{i1}|y_{i2}, \mu, \Sigma) = \beta_{10.2} + \beta_{12.2}y_{i2} \\ s_{11}^{r_3} &= E(y_{i1}^2|y_{i2}, \mu, \Sigma) = (\beta_{10.2} + \beta_{12.2}y_{i2})^2 + \sigma_{11.2} \\ s_{12}^{r_3} &= E(y_{i1}y_{i2}|y_{i2}, \mu, \Sigma) = (\beta_{10.2} + \beta_{12.2}y_{i2})y_{i2}, \end{aligned}$$

where  $\beta_{20.1}$ ,  $\beta_{21.1}$ , and  $\sigma_{22.1}$  are functions of  $\Sigma$  corresponding to the regression of  $y_{i2}$  on  $y_{i1}$ ;  $\beta_{10.2}$ ,  $\beta_{12.2}$ , and  $\sigma_{11.2}$  are functions of  $\Sigma$  corresponding to the regression of  $y_{i1}$  on  $y_{i2}$ .

We have found the expectations of  $y_{i1}$ ,  $y_{i2}$ ,  $y_{i1}^2$ ,  $y_{i2}^2$ , and  $y_{i1}y_{i2}$  for each unit in the three groups, so the expectations of the sufficient statistics of the two variables can be found as the sums of these quantities over all  $n$  units.

$$\begin{aligned} s_1 &= s_1^{r_1} + s_1^{r_2} + s_1^{r_3} \\ s_2 &= s_2^{r_1} + s_2^{r_2} + s_2^{r_3} \\ s_{11} &= s_{11}^{r_1} + s_{11}^{r_2} + s_{11}^{r_3} \\ s_{22} &= s_{22}^{r_1} + s_{22}^{r_2} + s_{22}^{r_3} \\ s_{12} &= s_{12}^{r_1} + s_{12}^{r_2} + s_{12}^{r_3} \end{aligned}$$

The M step computes the  $\mu$  and  $\Sigma$  by using those filled-in sufficient statistics:

$$\begin{aligned}\hat{\mu}_1 &= s_1/n, & \hat{\mu}_2 &= s_2/n, \\ \hat{\sigma}_1^2 &= s_{11}/n - \hat{\mu}_1^2, & \hat{\sigma}_2^2 &= s_{22}/n - \hat{\mu}_2^2, & \hat{\sigma}_{12}^2 &= s_{12}/n - \hat{\mu}_1\hat{\mu}_2\end{aligned}$$

Then, the EM algorithm goes back to the E step to update those expectations or sufficient statistics again. By feeding those updated statistics into the M step, we can get the updated  $\mu$  and  $\Sigma$ . The EM algorithm performs this cycle again and again until convergence.

A problem you might have already noticed is that how to use the updated  $\mu^t$  and  $\Sigma^t$ , or  $\theta^t$ , where  $\theta^t = (\mu^t, \Sigma^t)$ , to update the sufficient statistics? There is no obvious connection from the M step back to the E step. The answer lies inside the  $\beta$ s. Stuart & Ord (1994) demonstrated that the parameter of the regressions of bivariate normal distribution can be expressed as a one to one function of the original parameter  $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$ .

1. If  $y_{i2}$  is response variable, and  $y_{i1}$  is explanatory variable:

$$\beta_{21.1} = \sigma_{12}/\sigma_{11} \tag{6.11}$$

$$\beta_{20.1} = \mu_2 - \beta_{21.1}\mu_1, \tag{6.12}$$

$$\sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11} \tag{6.13}$$

2. If  $y_{i1}$  is response variable, and  $y_{i2}$  is explanatory variable:

$$\beta_{12.2} = \sigma_{12}/\sigma_{22} \tag{6.14}$$

$$\beta_{10.2} = \mu_1 - \beta_{12.2}\mu_2, \tag{6.15}$$

$$\sigma_{11.2} = \sigma_{22} - \sigma_{12}^2/\sigma_{22} \tag{6.16}$$

The updated  $\mu^t$  and  $\Sigma^t$  from each of the M steps can be used to compute the  $\beta$ s in the E steps, then we have the updated sufficient statistics to be fed into the M steps.

So far, we haven't seen how the EM algorithm helps us to impute any missing values. Instead, the EM algorithm only provides us with the estimation of the parameters if we have incomplete data. Some rather confusing questions might be: why do we estimate parameters? why not impute missing values directly? Indeed, Healy & Westmacott (1956) described an iterative technique: (1) impute missing values, (2) estimate parameters from the imputed values, (3) re impute missing values by using the estimated parameters, (4) then estimate parameters again based on the updated imputed values, and so on, until the missing values do not change much. This iterative technique actually can be considered as an EM algorithm, if the complete data loglikelihood  $\ell(\theta|Y_{obs}, Y_{mis}) = \log L(\theta|Y_{obs}, Y_{mis})$  is linear in  $Y_{mis}$ . If it is non-linear, we need to estimate missing sufficient statistics rather than individual observations, more generally, the loglikelihood  $\ell(\theta|Y)$  needs to be estimated (Little & Rubin 2002). This is because the E step of EM algorithm is about finding the conditional expectation of the missing data, and we have demonstrated this in section 6.2 by using the exponential family as an example.

Again, let's show how the EM algorithm works for the bivariate normal data with missing values on both variables by applying it to the SURF data. This time, we first make the SURF's Hours missing value MCAR, then make the SURF's Income missing MCAR. Then, we have:

$$f(MissingHours|Hours, \theta_{hours})$$



and

$$f(\text{MissingIncome}|\text{Income}, \theta_{\text{Income}}).$$

Specifically, 50 units' Hours values were created as MCAR. Then, we subset the units which have Hours values not missing. Because the missingness is MCAR, the subset sample is simply a smaller version of the original sample. This is because a random sample from a random sample is still random. Finally, another 50 units' Income values were created as MCAR for the subset sample. The bivariate normal data with missing values on both variables was created by merging the data set with missing Hours with the data set with missing Income.

The following steps show how we apply the EM algorithm to the bivariate normal data with missing values on both variables.

**Recipe: EM algorithm - Bivariate Normal Sample with Missing Data on both Variables**

- Step 1:** compute the sufficient statistics for the group of units with both variables observed
- Step 2:** compute the sufficient statistics for the group of units with one variable observed, but the other missing
- Step 3:** combine sufficient statistics from step 1 and step 2
- Step 4:** estimate the mean and covariance matrix of all observations by using the combined sufficient statistics
- Step 5:** update parameters ( $\beta$ s,  $\sigma$ s), and repeat step 2 to step 4 until convergence

Please refer to Appendix B for the R code.

As we did for the univariate normal data, we repeated the above procedure 1000 times. Each time there was a different data set with missing data (MCAR) for the Income and Hours variables, and the EM algorithm was applied to each dataset. Figure 6.2 shows the distribution of the 1000 simulated Hours and Income variables' means and variances. The dashed red vertical line represents the means of the 1000 means and variances. The solid red vertical line represents the true mean and variance of the original complete data. Compared to Figure 6.1, the estimates are closer to the true estimates. This is because unlike applying EM algorithm to the univariate normal data, the calculation of the expected values for the bivariate normal data involves more variables which can produce more accurate results. In addition, it also helps that the Hours and Income are highly correlated in the SURF data.

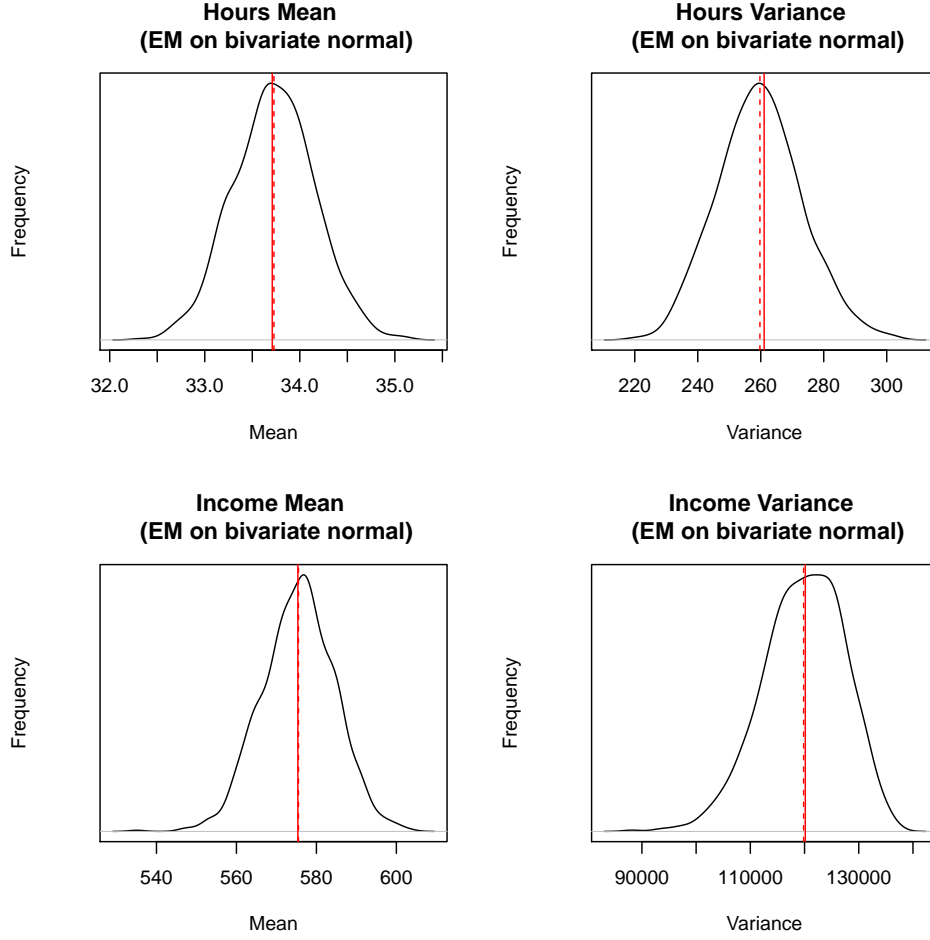


Figure 6.2: The distributions of the means and variances of the 1000 replicate SURF data's income and hours variables imputed by the EM algorithm. (The dashed red vertical line represents the mean of the 1000 means and variances. The solid red vertical line represents the true mean and variance of the original complete data)

## 6.5 Convergence of EM algorithm

Section 6.3 and section 6.4 have demonstrated how to apply EM algorithm to univariate normal and bivariate normal sample with missing data. As we have seen, the EM algorithm for both examples involves iterative steps, and we have said the iteration stops until the algorithm converges. But, how does the EM algorithm actually converge?

Eq. (6.1) shows that the complete-data with missing values log likelihood can be expressed as:

$$\ell(\theta; y) = \ell_{obs}(\theta; y_{obs}) + \log f(y_{mis}|y_{obs}; \theta)$$

Taking expectation on both sides of the above Eq.(6.1) over the distribution of the missing data  $Y_{mis}$ , given the observed data  $Y_{obs}$  and a current estimate of  $\theta^{(t)}$  for  $\theta$ , we have:

$$Q(\theta|\theta^{(t)}) = \log L(\theta|Y_{obs}) + H(\theta|\theta^{(t)})$$

where  $H(\theta|\theta^{(t)}) = E_{y|\theta^{(t)}}[\log f(y_{mis}|y_{obs}; \theta)|y_{obs}]$ . Then, we can show that:

$$\begin{aligned} \ell(\theta^{(t+1)}|Y_{obs}) - \ell(\theta^{(t)}|Y_{obs}) &= [Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t-1)})] \\ &\quad - [H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t-1)})] \end{aligned} \quad (6.17)$$

By Jensen's inequality, we have  $H(\theta^{(t+1)}|\theta^{(t)}) \leq H(\theta^{(t)}|\theta^{(t)})$ .

Generally, an EM algorithm maximizes  $Q(\theta|\theta^{(t)})$  by choosing  $\theta^{(t+1)}$  so that  $Q(\theta^{(t+1)}|\theta^{(t)})$  is greater than  $Q(\theta^{(t)}|\theta^{(t)})$ . Obviously, the difference of  $Q$  functions in Eq. (6.17) is positive for any EM algorithm. Hence for any EM algorithm, the change from  $\theta^{(t)}$  to  $\theta^{(t+1)}$  does not decrease the log likelihood.

$$\ell(\theta^{t+1}|Y_{obs}) \geq \ell(\theta^t|Y_{obs})$$

Thus for a bounded sequence of likelihood values  $\ell(\theta^t|Y_{obs})$ ,  $\ell(\theta^t|Y_{obs})$  converges monotonically to some stationary value  $\ell^*$ , under the assumption that there is a global maximum. The question now is to the conditions under which  $\ell^*$  corresponds to a stationary value and when this stationary value is at least a local maximum if not a global maximum (McLachlan & Krishnan 1997). Actually, a key result of Dempster, Laird & Rubin (1977) is:

$$\ell(\theta^{t+1}|Y_{obs}) \geq \ell(\theta^t|Y_{obs})$$

with equality if and only if:

$$Q(\theta^{(t+1)}|\theta^{(t)}) = Q(\theta^{(t)}|\theta^{(t)})$$

This means that the likelihood function increases at each iteration of the EM algorithm, until the condition for equality and a fixed point of the iteration are reached. Hence we have our convergence. The convergence does not necessarily mean a global maximum. Actually, the EM algorithm could converge to a local maximum, or even to local minimum and to a saddle point. These concerns have been thoroughly discussed in Dempster, Laird & Rubin (1977) and Wu (1983), but they are out of scope of this thesis.

## 6.6 Conclusion

We started this chapter by introducing the problem of using the Maximum Likelihood (ML) estimation in incomplete data. The problem is that maximising the incomplete data log likelihood directly does not give us the ML estimates. Although there are many algorithms which have been developed to solve this problem (such as Newton-Raphson), we have only introduced the EM algorithm. This because the EM algorithm is by far one of the best missing data handling techniques, according to Schafer & Graham (2002).

Technically speaking, the EM algorithm does “fill-in” the missing values, although Little & Rubin (2002) describes it is actually the function of missing data. This is the key point where the EM algorithm is different from any other imputation methods. We may argue that, as we have shown in section 6.2, the E step of the EM algorithm does fill in individual missing values with their expected values, but it is only under the condition that  $t(y)$  is linear in  $y$ . Also, it is worth noting that the EM algorithm only generates the ML estimates at the end of its iterations. This means that we might get unbiased estimates by using the EM algorithm, although the missing values in the data set are still empty.

Although the EM might be able to produce unbiased estimates, given the missing data is at least MAR and we have the complete variables which the missingness depends on, it is still like the single imputation methods which cannot propagate imputation uncertainty. As we have seen in section 6.2, the EM algorithm only imputes missing data by using the expected values once, although it updates the expected values many times.

# Chapter 7

## Bayesian Multiple Imputation

### 7.1 Introduction

Bayesian theory and Bayesian iterative simulation methods are the underlying foundation of Rubin's Multiple Imputation (MI) (Rubin 1987). However, the Bayesian theory and the Bayesian iterative simulation methods cannot apply to all the imputation methods. For example, it is easy to apply Bayesian iterative simulation methods to the stochastic regression imputation method<sup>1</sup>, but not easy to adapt it for hot deck imputation. In fact, Rubin classifies MI as either “proper” or “improper”. The proper MI random draws the imputations from a posterior distribution in a Bayesian framework. Rubin (1987) calls MI methods “improper” if they do not properly propagate imputation uncertainty and lack a Bayesian framework. There will be discussions about proper and improper MI in the next Chapter. Other researchers refer to MI as Bayesian Multiple Imputation and Non-Bayesian Multiple Imputation (Schafer 2003, Bjørnstad 2007). Nevertheless, this chapter focuses on the description of the Bayesian part of Rubin's Multiple Imputation.

Therefore, the question we want to answer in this chapter is “How do we apply Bayesian estimation to MI?” After all, Bayesian statistics is usually about estimating the entire distribution of parameters conditional on some collected data. In other words, it has been widely used to estimate the distributions of parameters given some data. However, Scott (2007) points out that both parameters and data are considered random quantities from the Bayesian perspective, so the same techniques can be applied to draw data conditional on given parameters. This sheds light on imputing missing data by using Bayesian iterative simulation methods. This is because we can treat missing data as unknown and simulate them given the observed data, and parameters.

Let  $Y = (Y_{obs}, Y_{mis})$ ,  $Y_{obs}$  represent the observed data, and  $Y_{mis}$  the missing data. Suppose  $Y_{mis}$  is MAR, and  $\theta$  is the parameters of the likelihood  $f(Y|\theta)$ . Thus, our posterior distribution becomes:

$$p(\theta, Y_{mis}|Y_{obs}) \propto p(\theta, Y_{mis})f(Y_{obs}|\theta, Y_{mis})$$

where  $p(\theta, Y_{mis})$  is the prior distribution and  $f(Y_{obs}|\theta, Y_{mis}) = f(Y_{obs}|\theta)$  is the likelihood function of the observed data. The prior distribution can be further decomposed as:

$$p(\theta, Y_{mis}) \propto p(\theta)f(Y_{mis}|\theta)$$

---

<sup>1</sup>Details of how to apply Bayesian iterative simulation to stochastic regression will be discussed shortly in this chapter.

where  $f(Y_{mis}|\theta)$  is the likelihood of  $Y_{mis}$ . Hence, the posterior is now:

$$p(\theta, Y_{mis}|Y_{obs}) \propto p(\theta)f(Y_{mis}|\theta)f(Y_{obs}|\theta, Y_{mis})$$

Given  $Y_{mis}$  is MAR, this means that the  $\theta$  captures all relevant information about  $Y_{obs}$ , so that  $Y_{obs}$  and  $Y_{mis}$  are conditionally independent given  $\theta$ , so the term  $f(Y_{obs}|\theta, Y_{mis})$  reduces to  $f(Y_{obs}|\theta)$ . Then, the full posterior distribution is:

$$p(\theta, Y_{mis}|Y_{obs}) \propto p(\theta)f(Y_{mis}|\theta)f(Y_{obs}|\theta) \quad (7.1)$$

The fundamental idea of Bayesian iterative simulation methods for imputation is actually about sampling  $\theta$  and  $Y_{mis}$  from the full posterior distribution given in Eq (7.1).

## 7.2 Bayesian Iterative Simulation Methods - Markov Chain Monte-Carlo (MCMC)

As introduced in the previous section, the Bayesian iterative simulation method is all about yielding draws from the posterior distribution. This idea is not hard to understand. However, the devil is in the detail. When we try to implement the simulation idea for some data, a difficult problem appears. How do we generate draws from a distribution we are not familiar with or one which is high dimensional and complicated? Markov Chain Monte Carlo (MCMC) sampling method provides a way to draw from unknown or complex posterior distributions. In order to draw from those distributions, MCMC often involves breaking down them into more manageable distributions.

“Markov Chain” refers to the process which the draws are made in sequence and are dependent, but where each draw only depends on the previous one. In terms of Bayesian terminology, it generates a new value from the posterior distribution, given the previous value. “Monte Carlo” refers to the random simulation process.

We will introduce two common MCMC methods in the rest of this section.

### 7.2.1 Gibbs sampling algorithm

The Gibbs sampler is one of the most basic special cases of the MCMC method (Scott 2007). It is simply an iterative simulation method that produces a draw from the joint distribution in the case of a general pattern of missing data (Little & Rubin 2002). The Gibbs sampler can also be regarded as a multivariate extension of the chained data augmentation algorithm in which we estimate  $p$  parameters  $\theta_1, \dots, \theta_p$  (Peter 1997). This means it has the ability to simulate from the full conditional distribution  $p(\theta_i|\theta_1, \dots, \theta_p)$ , where  $i \in 1, \dots, p$ . A generic Gibbs sampler follows the following iterative process ( $t$  indexes the iteration count):

0. Assign a vector of starting values,  $\theta^0$ , to the parameter vector, and set  $t = 0$
1. Set  $t = t + 1$
2. Draw  $p(\theta_1^t|\theta_2^{t-1}, \theta_3^{t-1} \dots \theta_p^{t-1})$
3. Draw  $p(\theta_2^t|\theta_1^t, \theta_3^{t-1} \dots \theta_p^{t-1})$
- .. ..

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

p+2. Return to step one, repeating until convergence<sup>2</sup>

In other words, Gibbs sampling orders the parameters and generates draws from the conditional distribution for these parameters given the current value of all the other parameters and cycles through the updating process repeatedly. The process of cycling stops when the distributions become stable and stationary<sup>3</sup>. In other words, the process stops when the algorithm converges. Section 7.4 discusses the convergence of MCMC in more detail.

## 7.2.2 Metropolis-Hastings (MH) algorithm

As we have mentioned that MCMC methods breakdown a complex or unfamiliar posterior distribution into smaller manageable distributions, then parameters are drawn from those smaller distributions. Actually, this is the case of the Gibbs sampling algorithm. The Gibbs sampler usually works fine until we encounter situations that even the breakdown for conditional posterior distributions  $p(\theta_k | \theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_p)$  are foreign to us, or we might know the functional form, but not know the normalisation, nor have a means of drawing a sample directly. Again, we ask the question: “how do we generate draws from a distribution which we are not familiar with?”

The Metropolis-Hastings (MH) algorithm overcomes the unfamiliar distribution problem by generating draws from the posterior distribution (Hastings 1970). Basically, the MH algorithm draws a candidate point from a proposal distribution, then uses some techniques to determine whether we accept the candidate point as a draw from the full conditional distribution (or the target distribution). Clearly, the MH algorithm bypasses the need of generating draws from the full conditional distribution or its breakdown distributions directly. We also need to point out that we can have MCMC updates in blocks where all blocks except one or two are Gibbs updates, but the rest special cases are the MH steps.

Here is the generic process of using the MH algorithm to generate parameters from the posterior distribution ( $t$  indexes the iteration count):

1. Establish starting values  $\theta^0$  for the parameter:  $\theta$ . Set  $t = 0$ .
2. Draw a “candidate” parameter,  $\theta^c$  from a “proposal density”  $q(\theta^c | \theta^{t-1})$ .
3. Compute the ratio

$$R = \min \left( 1, \frac{f(\theta^c)q(\theta^{t-1} | \theta^c)}{f(\theta^{t-1})q(\theta^c | \theta^{t-1})} \right) \quad (7.2)$$

4. Compare  $R$  with a  $U(0, 1)$  random draw  $u$ . If  $R > u$ , then set  $\theta^t = \theta^c$ . Otherwise, set  $\theta^t = \theta^{t-1}$
5. Set  $t = t + 1$  and return to step 2 until it converges (Please refer to section 7.4 for details of convergence).

---

<sup>2</sup>The convergence will be discussed in later part of this chapter.

<sup>3</sup>Stationary means the joint probability distribution of a stochastic process does not vary with respect to a shift in time (DelSole 2010).

In the MH algorithm we have described above, if the “proposal density”  $q(\theta^c|\theta^{t-1})$  is chosen to be independent of  $\theta^{t-1}$ , that is:

$$q(\theta^c|\theta^{t-1}) = q(\theta^c)$$

for a given probability density function  $q(\theta^c)$ . Then, the candidate point is generated from  $q(\theta^c)$ . The candidate point is accepted or rejected with an acceptance probability  $\alpha(\theta^{t-1}, \theta^c)$  given by:

$$\alpha(\theta^{t-1}, \theta^c) = \min\left(1, \frac{f(\theta^c)q(\theta^{t-1})}{f(\theta^{t-1})q(\theta^c)}\right)$$

This version of the MH algorithm is called the **MH independence sampler**. Clearly, the MH independence sampler has a potential to boost up computation, since it only accepts or rejects the candidate points which are random draws from the proposal distribution. In other words, if the proposal distribution is not well matched to the target density, then many proposals will be rejected. For example, if the proposal distribution is too wide, it will take a very long time for the Bayesian iterative chain to converge; or if the proposal distribution is too narrow, the Bayesian iterative chain will not cover the target distribution. Hence, this method requires the proposal distribution to be as close as the target distribution, otherwise it can get stuck in the tails of the target distribution (Marin & Robert 2007, pg. 93).

### 7.2.3 Relationship between Gibbs and MH sampling

The Gibbs sampler is actually a special case of the MH algorithm. The only difference is that there is no rejection of selected candidate points in Gibbs sampling. The reason is that the ratio  $R$  is always 1 (Gamerman & Lopes 2006). Why? Let’s consider the equation for the ratio  $R$ , Eq. (7.2). In Gibbs sampling, we set the “proposal density”  $q(\theta^c|\theta^{t-1})$  to equal the target density  $f(\theta^c)$ . This means that  $\theta^c$  is independent of  $\theta^{t-1}$ , and is an independent sampler. Hence, we have:

$$\begin{aligned} R &= \frac{f(\theta^c)q(\theta^{t-1}|\theta^c)}{f(\theta^{t-1})q(\theta^c|\theta^{t-1})} \\ &= \frac{f(\theta^c)f(\theta^{t-1})}{f(\theta^{t-1})f(\theta^c)} \\ &= \frac{f(\theta^c)/f(\theta^c)}{f(\theta^{t-1})/f(\theta^{t-1})} \\ &= 1 \end{aligned}$$

Since the candidate point is accepted with probability  $\min(1, R)$ , and it is always true that  $R = 1$ , every draw is accepted.

Given the Gibbs sampler is part of the MH algorithm, there is no inherent reason we can not combine both algorithms. Actually, the MH algorithm can be a sub algorithm inside a Gibbs sampling cycle (Gilks et al. 1996) and (Muller 1991). It is also fine to have the Gibbs sampler inside the MH algorithm (Gamerman & Lopes 2006) and (Scott 2007). However, as we have already discussed above, the Gibbs sampler automatically accepts any candidate points, but the MH algorithm does not accept all the candidate points coming from the proposal density. To be precise, all Gibbs sampling is MH sampling, but not all MH sampling is Gibbs sampling.

### 7.2.4 Block Updating

So far, the MH algorithm and the Gibbs sampler we have introduced are only for updating a single scalar parameter  $\theta$  one at a time. For a high dimensional posterior distribution, where we have a large number of parameters,  $\theta_1, \dots, \theta_p$ , updating only a single parameter  $\theta_i$  at a time, where  $i \in 1, \dots, p$ , is not only a daunting task, but many have a very slow convergence rate. Hastings (1970) proposed a method that applies the MH algorithm in turn to subblocks of the vector of parameters  $\theta = (\theta_1, \dots, \theta_p)$ . Hence, instead of having  $p$  parameters, we group these  $p$  parameters into  $b$  subblocks, where  $b < p$ . For example, if  $b = 2$ , then we have  $\theta = ((\theta_1, \dots, \theta_k), (\theta_{k+1}, \dots, \theta_p)) = (\theta_{block1}, \theta_{block2})$ .

## 7.3 Applying MCMC methods to Normal data with Ignorable Non-response

### 7.3.1 Applying the Gibbs sampler to Univariate Normal data

Let's consider a univariate normal distribution example. Suppose  $Y = (Y_{obs}, Y_{mis})$  has  $n$  observations and each observation is normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Then, Eq. (7.1) can be transformed into:

$$p(\mu, \sigma^2, Y_{mis} | Y_{obs}) \propto p(\mu, \sigma^2) f(Y_{mis} | \mu, \sigma^2) f(Y_{obs} | \mu, \sigma^2)$$

Note: the pdf for  $Y_{mis}$  is technically conditional on  $Y_{obs}$ , ie  $f(Y_{mis} | \mu, \sigma^2, Y_{obs})$ . However, since non-response is ignorable,  $Y_{obs}$  has been left out.

If we assume the prior  $p(\mu, \sigma^2)$  is Jeffery's prior which in this case is  $1/\sigma^2$ , which is commonly known as the improper prior<sup>4</sup> for the variance of a normal distribution, then the posterior density could be factored to (1) a marginal posterior density for  $\sigma^2$  that is an inverse gamma distribution, (2) a conditional posterior density for  $\mu$  that is a normal distribution, and (3) a conditional posterior density for  $Y_{mis}$  that is a normal distribution as well. Hence we have:

$$\begin{aligned} p(Y_{mis} | \mu, \sigma^2, Y_{obs}) &\propto N(\mu, \sigma^2) \\ p(\sigma^2 | \mu, Y_{obs}, Y_{mis}) &\propto IG(n/2, \sum (y_i - \mu)^2 / 2) \\ p(\mu | \sigma^2, Y_{obs}, Y_{mis}) &\propto N(\bar{Y}, \sigma^2 / n) \end{aligned}$$

Proof:

We set our joint prior distribution  $p(\mu, \sigma^2) = 1/\sigma^2$ . Hence, we have:

$$\begin{aligned} p(\mu, \sigma^2, Y_{mis} | Y_{obs}) &\propto p(\mu, \sigma^2) f(Y_{mis} | \mu, \sigma^2) f(Y_{obs} | \mu, \sigma^2) \\ &\propto \frac{1}{\sigma^2} \prod_{i=1}^r \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_{obs,i} - \mu)^2}{2\sigma^2}\right\} \prod_{i=r+1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_{mis,i} - \mu)^2}{2\sigma^2}\right\} \end{aligned}$$

---

<sup>4</sup>Priors such as  $p(\mu) = 1$ ,  $p(\sigma) = 1/\sigma$  are improper because they do not integrate to 1. That is, the area under the prior density is not unity (and, in fact, is infinity).



Then, the posterior distribution of the mean  $\mu$ , given other parameters are fixed, is:

$$\begin{aligned}
f(\mu|Y_{obs}, Y_{mis}, \sigma^2) &\propto \exp\left\{-\sum_{i=1}^r \frac{(y_{obs,i} - \mu)^2}{2\sigma^2}\right\} \exp\left\{-\sum_{i=r+1}^n \frac{(y_{mis,i} - \mu)^2}{2\sigma^2}\right\} \\
&\propto \exp\left\{-\frac{r\mu^2 - 2r\bar{Y}_{obs}\mu}{2\sigma^2}\right\} \exp\left\{-\frac{(n-r)\mu^2 - 2(n-r)\bar{Y}_{mis}\mu}{2\sigma^2}\right\} \\
&\propto \exp\left\{-\frac{n\mu^2 - 2n\bar{Y}\mu}{2\sigma^2}\right\} \\
&\propto \exp\left\{-\frac{(\mu - \bar{Y})^2}{2\sigma^2/n}\right\} \\
\mu|\sigma^2, Y_{obs}, Y_{mis} &\sim N\left(\bar{Y}, \frac{\sigma^2}{n}\right)
\end{aligned}$$

where  $\bar{Y} \approx \bar{Y}_{obs} \approx \bar{Y}_{mis}$ .

The posterior distribution of the variance, given other parameters are fixed, is:

*The first step is to expand the quadratic term of  $f(\mu, \sigma^2|Y_{obs}, Y_{mis})$ :*

$$\begin{aligned}
f(\mu, \sigma^2|Y_{obs}, Y_{mis}) &\propto \frac{1}{\sigma^2} \prod_{i=1}^r \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_{obs,i} - \mu)^2}{2\sigma^2}\right\} \prod_{i=r+1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_{mis,i} - \mu)^2}{2\sigma^2}\right\} \\
&\propto \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{\sum_{i=1}^r Y_{obs,i}^2 - 2r\bar{Y}_{obs}\mu + r\mu^2}{2\sigma^2}\right\} \\
&\quad \exp\left\{-\frac{\sum_{i=r+1}^n Y_{mis,i}^2 - 2(n-r)\bar{Y}_{mis}\mu + (n-r)\mu^2}{2\sigma^2}\right\} \\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} \exp\left\{-\frac{\sum_{i=1}^r Y_{obs,i}^2 - 2r\bar{Y}_{obs}\mu + r\mu^2}{2\sigma^2}\right\} \\
&\quad \exp\left\{-\frac{\sum_{i=r+1}^n Y_{mis,i}^2 - 2(n-r)\bar{Y}_{mis}\mu + (n-r)\mu^2}{2\sigma^2}\right\}
\end{aligned}$$

*Then, the joint posterior density for  $\mu$  and  $\sigma^2$  can be factored using the conditional probability rule as:*

$$\begin{aligned}
f(\mu, \sigma^2|Y_{obs}, Y_{mis}) &\propto f(\mu|\sigma^2, Y_{obs}, Y_{mis})f(\sigma^2|Y_{obs}, Y_{mis}) \\
&\propto \frac{1}{\sigma} \exp\left\{-\frac{r(\mu - \bar{Y}_{obs})^2}{2\sigma^2}\right\} \exp\left\{-\frac{(n-r)(\mu - \bar{Y}_{mis})^2}{2\sigma^2}\right\} \\
&\quad \times \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{\sum_{i=1}^r Y_{obs,i}^2 - r\bar{Y}_{obs}^2}{2\sigma^2}\right\} \exp\left\{-\frac{\sum_{i=r+1}^n Y_{mis,i}^2 - (n-r)\bar{Y}_{mis}^2}{2\sigma^2}\right\} \\
&\propto \frac{1}{\sigma} \exp\left\{-\frac{(\mu - \bar{Y})^2}{\frac{2\sigma^2}{n}}\right\} \frac{1}{(\sigma^2)^{\frac{n}{2}}} \times \exp\left\{-\frac{\sum_{i=1}^n Y_i^2 - n\bar{Y}^2}{2\sigma^2}\right\}
\end{aligned}$$

Now, we see that the first term is the conditional posterior for  $\mu$ . The second term is proportional to the conditional posterior density for  $\sigma^2|\mu$ . The numerator in the exponential is the numerator for the simplified version of the sample variance,  $\sum(Y_i - \bar{Y})^2$ . Hence, we have an inverse gamma distribution<sup>5</sup> with parameters  $\alpha = (n-1)/2$ , and  $\beta = (n-1)var(Y)/2$ , where

<sup>5</sup>The density function for the inverse gamma distribution is:

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\beta/y}$$

$\text{var}(Y) = \sum(Y_i - \bar{Y})^2 / (n - 1)$ . Hence:

$$\sigma^2 | \mu, Y_{\text{obs}}, Y_{\text{mis}} \propto IG(n/2, \sum(y_i - \mu)^2 / 2)$$

The posterior distribution of  $Y_{\text{mis}}$ , given other parameters are fixed, is:

$$\begin{aligned} f(Y_{\text{mis}} | \mu, \sigma^2, Y_{\text{obs}}) &\propto \frac{1}{\sigma^2} \prod_{i=1}^r \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_{\text{obs},i} - \mu)^2}{2\sigma^2}\right\} \prod_{i=r+1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_{\text{mis},i} - \mu)^2}{2\sigma^2}\right\} \\ &\propto \prod_{i=r+1}^n \exp\left\{-\frac{(y_{\text{mis},i} - \mu)^2}{2\sigma^2}\right\} \\ &\propto \exp\left\{-\frac{\sum_{i=r+1}^n (y_{\text{mis},i} - \mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Clearly, this shows that  $f(Y_{\text{mis},i} | \mu, \sigma^2, Y_{\text{obs},i}) \sim N(\mu, \sigma^2)$ .

As shown in the following R program, we have applied the Gibbs sampler to the SURF data which has 50 missing Income values. The missing mechanism is MCAR. We use 1000 iterations. Figure 7.1 displays the results. The histograms show the distribution of the means and variances which were generated from normal and inverse gamma distributions. The time-series plots show the values of the means and variances at each iteration. The red solid lines represent the “true” mean and variance.

```
#Gibbs sampling algorithm -- univariate normal
#step 0: Create MCAR income data
Y_MCAR_0=MCAR(SURF,50,"Income")$Income
Y_MCAR=Y_MCAR_0
#step 1: Set up initial values
iter=1000
mY_MCAR=matrix(mean(Y_MCAR[!is.na(Y_MCAR)]),iter)
sY_MCAR=matrix(var(Y_MCAR[!is.na(Y_MCAR)]),iter)
#step 2: Draw missing income from the normal distribution with
#         observed income mean and variance
for(i in 2:iter)
{
  Y_MCAR[is.na(Y_MCAR_0)]=rnorm(length(Y_MCAR[is.na(Y_MCAR_0)]),
                                mY_MCAR[i-1], sqrt(sY_MCAR[i-1]))

  #step 3: draw sigma^2 and mean
  sY_MCAR[i]=rgamma(1,(length(Y_MCAR)/2),rate=sum((Y_MCAR-mY_MCAR[i-1])^2)/2)
  sY_MCAR[i]=1/sY_MCAR[i]
  mY_MCAR[i]=rnorm(1,mean(Y_MCAR),sqrt(sY_MCAR[i]/length(Y_MCAR)))
}
```

Despite the information about convergence which will be discussed in later sections, the graph tells us that the Gibbs Sampler produces unbiased estimates, given the missingness is MCAR.

The inverse gamma distribution is a special case of the inverse wishart distribution which has a density function:

$$f(y) = |Y|^{-(v+d+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(SY^{-1})\right\}$$

where  $S$  is a scale matrix of dimension  $d$ , and  $v$  is the degrees of freedom.

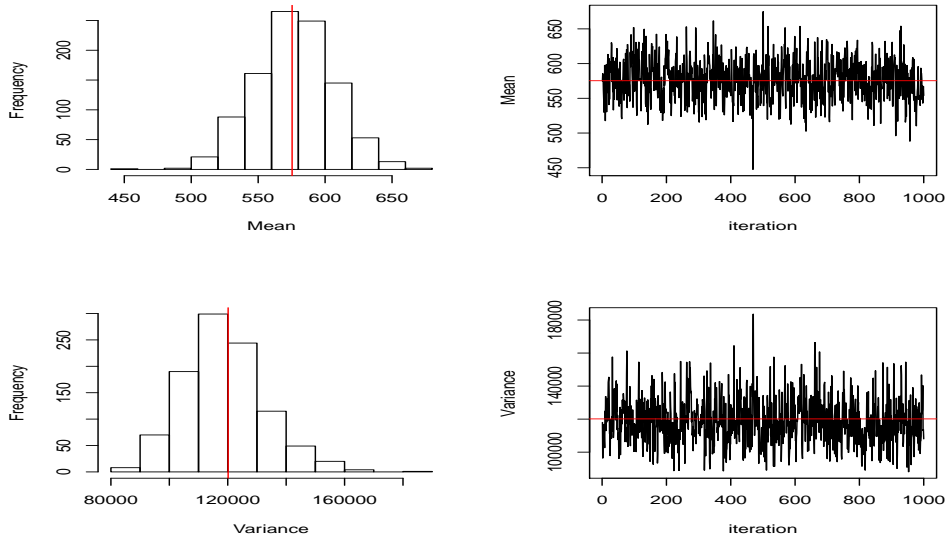


Figure 7.1: Applying the Gibbs sampler to the SURF data with 50 income values missing completely at random (MCAR). The number of iterations is 1000

### 7.3.2 Applying the MH algorithm to Univariate Normal data

Let's still use the example in Section 7.3.1. Suppose we are not familiar with the conditional posterior distribution of  $\mu$  and  $\sigma^2$ , we only know how to generate draws for  $Y_{mis}$  from the normal distribution  $N(\mu, \sigma^2)$ . As we have discussed in Section 7.2.3, it is fine to use Gibbs sampler to generate draws for  $Y_{mis}$ , and use MH algorithm to draw  $\mu$  and  $\sigma^2$  from their conditional posterior distribution.

Assume the prior is Jeffery's prior  $1/\sigma^2$ . It is not hard to find the full posterior density:

$$p(Y; \mu, \sigma^2) = \frac{1}{\sigma^2} \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right)$$

With large samples, Scott (2007) points out that “evaluating the posterior would generate an underflow problem because of the large negative exponents involved in the posterior”. Log transforming the posterior resolves this problem. However, we then have to compare the ratio  $R$  to the log of a uniform draw  $U(0, 1)$ . The log of the posterior density is:

$$\log f(Y; \mu, \sigma^2) \equiv \log p(Y; \mu, \sigma^2) = -\log \sigma^2 - \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2$$

where  $Y = (Y_{obs}, Y_{mis})$ . The ratio  $R$  becomes:

$$\begin{aligned} \log R &= \log \left( \frac{f(\theta^c)q(\theta^{t-1}|\theta^c)}{f(\theta^{t-1})q(\theta^c|\theta^{t-1})} \right) \\ &= \log(f(\theta^c)q(\theta^{t-1}|\theta^c)) - \log(f(\theta^{t-1})q(\theta^c|\theta^{t-1})) \\ &= \log(f(\theta^c)) + \log(q(\theta^{t-1}|\theta^c)) - \log(f(\theta^{t-1})) - \log(q(\theta^c|\theta^{t-1})) \end{aligned}$$

where  $\log f(\theta)$  is our  $\log f(Y; \mu, \sigma^2)$ .

As in the previous section, we applied the MH algorithm to the SURF's Income variable's missing values which are MAR. The proposal distribution  $q(\theta^c|\theta^{t-1})$  for the mean  $\mu$  and variance  $\sigma^2$  is a uniform distribution, where:

$$\begin{aligned}\mu^t &\sim U(\mu^{t-1} - 10, \mu^{t-1} + 10) \\ (\sigma^2)^t &\sim U((\sigma^2)^{t-1} - 1000, (\sigma^2)^{t-1} + 1000)\end{aligned}$$

The R program repeated the generating of missing Income data, Income mean, and Income variance 100000 times. Compared to the Gibbs sampler, the iteration for the MH algorithm is 100 times larger. This is due to the MH algorithm needing more iterations to converge. We expect that the rate of convergence of the Gibbs sampler is faster than the MH algorithm. This is because the Gibbs sampler directly samples values from the known distributions, but the MH algorithm samples values for unknown distribution from the proposed distributions. Please refer to appendix C for the R code. Figure 7.2 displays the results. Again, the histograms show the distributions of generated means and variances. The time-series plots show the values of means and variances for each iteration. The solid red lines represent the true mean and variance.

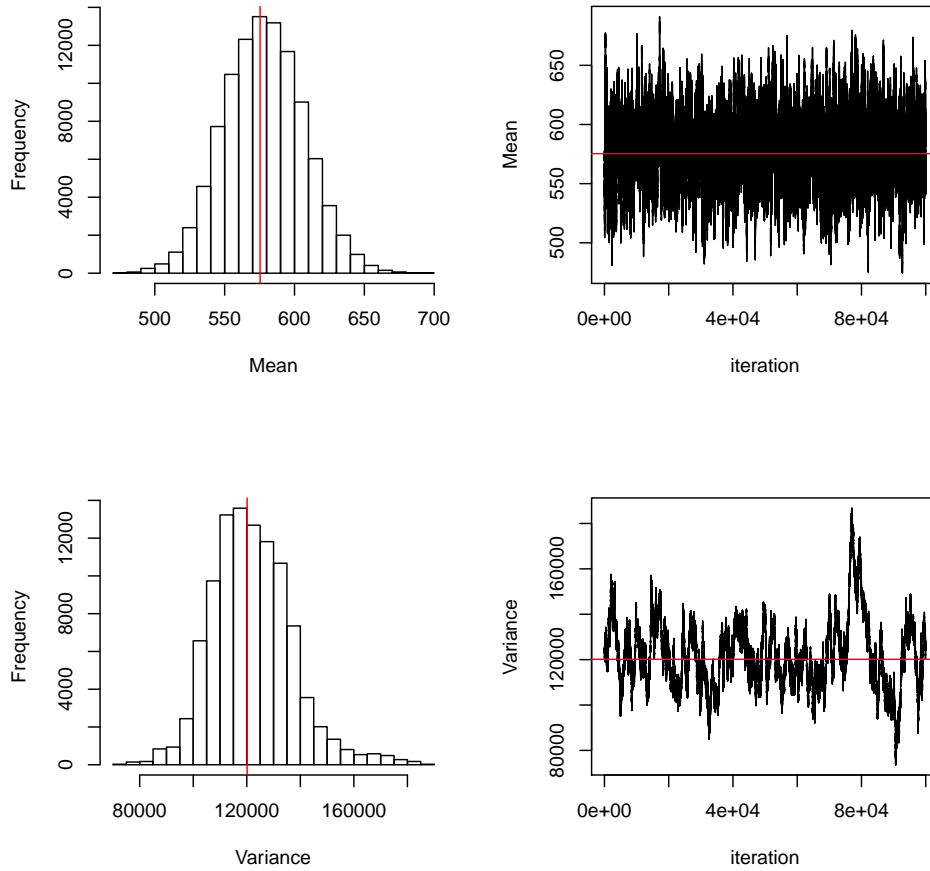


Figure 7.2: Applying the MH algorithm to the SURF data with 50 MCAR income values. The number of iterations is 100000

### 7.3.3 Applying Gibbs sampler to Bivariate Normal data

When  $p = 2$ , the Gibbs' sampler can be simplified as data augmentation (Chapter 3) if  $Y_1 = Y_{mis}$ ,  $Y_2 = \theta$ , and the distribution condition on  $Y_{obs}$ . The process becomes:

- I Step: Draw  $Y_{mis}^{(t+1)} \sim p(Y_{mis}|Y_{obs}, \theta^{(t)})$
- P Step: Draw  $\theta^{(t+1)} \sim p(\theta|Y_{mis}^{(t)}, Y_{obs})$

Suppose we have two variables  $Y_1$  and  $Y_2$ , and one group of units has both variables observed, the other groups have one variable observed but the other missing. Apply the DA algorithm to this example, we have:

The I (or imputation) step:

1. For missing  $y_{i2}$ :

$$y_{i2}^{(t+1)} \sim_{ind} N(\beta_{20.1}^{(t)} + \beta_{21.1}^{(t)} y_{i1}, \sigma_{22.1}^{(t)}),$$

2. For missing  $y_{i1}$ :

$$y_{i1}^{(t+1)} \sim_{ind} N(\beta_{10.2}^{(t)} + \beta_{12.2}^{(t)} y_{i2}, \sigma_{11.2}^{(t)}),$$

where  $\beta_{20.1}^{(t)}$ ,  $\beta_{21.1}^{(t)}$ , and  $\sigma_{22.1}^{(t)}$  are the  $t$ th iterates of the regression parameters of  $Y_2$  on  $Y_1$ ;  $\beta_{10.2}^{(t)}$ ,  $\beta_{12.2}^{(t)}$ , and  $\sigma_{11.2}^{(t)}$  are the  $t$ th iterates of the regression parameters of  $Y_1$  on  $Y_2$ . These  $\beta$ s and  $\sigma$ s can be calculated by using equations Eq (6.11) to Eq (6.16).

The P (or posterior) step:

The P step is basically drawing parameters from the imputed complete data posterior distribution. The first step of the P step is to find the posterior distribution for the parameter  $\theta$  which is  $\mu$  and  $\Sigma$  in this example. Equation (3.13) shows that the posterior is equivalent to the product of the prior and the likelihood function. Clearly, we have little knowledge about the prior distribution of  $\theta$ . Of course, we can assign any prior information to get the posterior distribution, but Little & Rubin (2002) pointed out that small samples are likely to generate bad inferences if the choice of prior distribution is not appropriate. Hence, Jeffreys (1961) provided us with this Jeffery's prior distribution which is a conventional choice if there is an absence of strong prior information for a multivariate normal sample. For the bivariate normal distribution, the Jeffery's prior is:

$$p(\mu, \Sigma) \propto |\Sigma|^{-(k+1)/2} = |\Sigma|^{-3/2} \quad (7.3)$$

where  $k = 2$  for bivariate normal distribution.

Now, we have the prior distribution, the likelihood of the bivariate normal sample is:

$$p(y_1, \dots, y_n) = \frac{1}{(2\pi)^n |\Sigma|^{n/2}} \exp\left(\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\right) \quad (7.4)$$

Hence, according to equation 3.13, the posterior distribution is:

$$p(\mu, \Sigma|y_1, \dots, y_n) \propto \frac{1}{(2\pi)^n |\Sigma|^{(n+3)/2}} \exp\left(\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\right) \quad (7.5)$$

In order to draw  $\mu$  and  $\Sigma$ , we need to re-express equation 7.5 in terms of:

$$p(\mu|\Sigma, y_1, \dots, y_n) = \frac{1}{(2\pi)^{|(1/n)\Sigma|}} \exp\left(-\frac{1}{2}(\mu - \bar{y})^T (n\Sigma^{-1})(\mu - \bar{y})\right),$$

$$p(\Sigma|y_1, \dots, y_n) \propto \frac{1}{|\Sigma|^{(n+2)/2}} \exp(-n\text{Tr}(\Sigma_y \Sigma^{-1})),$$

where  $\bar{y} = (1/n)\sum_{i=1}^n y_i$ ,  $\Sigma_y$  is the covariance matrix computed from the  $y_i$ , where  $\Sigma_y = (1/n)\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T$ , and  $\text{Tr}(\Sigma_x \Sigma^{-1})$  is the trace of the matrix  $\Sigma_y \Sigma^{-1}$ .

Hence, to draw from  $p(\mu, \Sigma|y_1, \dots, y_n)$ , we first need to draw from  $p(\Sigma|y_1, \dots, y_n)$  to get  $\Sigma^{(t)}$ , and then draw from  $p(\mu|\Sigma, y_1, \dots, y_n)$  to get  $\mu^{(t)}$ . Little & Rubin (2002) indicated that  $p(\Sigma|y_1, \dots, y_n)$  is an Inv-Wishart distribution with scale parameter  $S = n\Sigma_y$  and  $n - 1$  degrees of freedom. To draw from the Inv-Wishart distribution, we first need to form an upper triangular matrix  $B$  with  $b_{ij}$  draw from the chi-squared distribution  $\chi_{n-j}^2$  and then take the square root:

$$b_{ij} \sim \sqrt{\chi_{n-j}^2}, \quad b_{jk} \sim N(0, 1), \quad j < k, \quad (7.6)$$

For the bivariate normal distribution case, we have

$$\begin{bmatrix} b_{11} \sim \sqrt{\chi_{n-1}^2} & b_{12} \sim \sqrt{\chi_{n-2}^2} \\ . & b_{22} \sim \sqrt{\chi_{n-2}^2} \end{bmatrix}$$

and sampling

$$\Sigma^{(t)} = (B^T)^{-1}A, \quad (7.7)$$

where  $A$  is the Cholesky factor of  $S^{-1}$  (i.e.  $A^T A = S^{-1}$ ).

Now, we have  $\Sigma^{(t)}$ . The next step is to draw  $\mu^{(t)}$  from  $p(\mu|\Sigma, y_1, \dots, y_n)$  which is a multivariate Gaussian distribution.

$$\mu^{(t)} = \bar{y} + A^{(t)}z, \quad (7.8)$$

where  $z = (z_1, \dots, z_k)^T$  is a vector of independent  $N(0, 1)$  draws, and  $A^{(t)}$  is an upper triangular Cholesky factor such that  $A^{(t)T} A^{(t)} = \Sigma^{(t)}/n$ .

We applied the methods we have discussed in this section to the SURF data with 50 MCAR values for each income variable and hours variable. The missing values on both variables do not overlap. That is, if a respondent has missing income, then the respondent cannot have missing hours. We have followed the exact steps that have been described in this section, for 1000 iterations. Please see appendix C for the R code. Figure 7.3 shows the results. Again, the Gibbs sampler produces unbiased estimates.

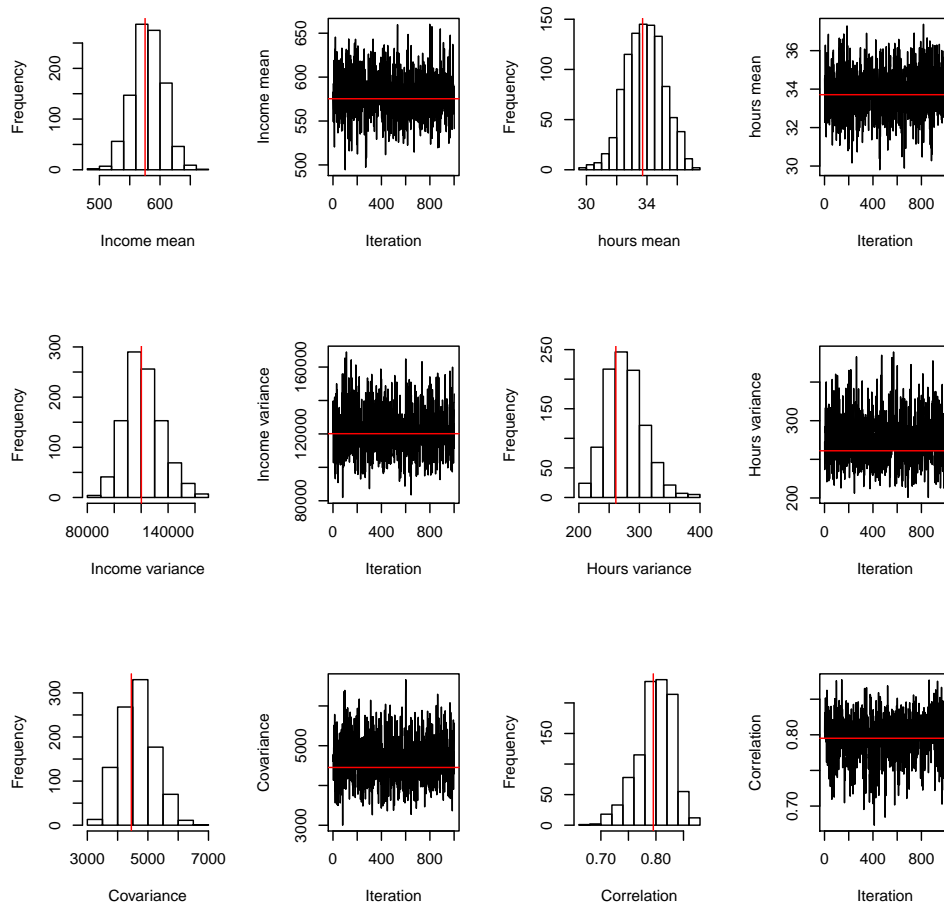


Figure 7.3: Applying the Gibbs sampler to the SURF data with 50 MCAR values for each income and hours variables. The number of iterations is 1000

## 7.4 Convergence Diagnostics

### 7.4.1 The Theory of Convergence

The diagnostics of the convergence for the Bayesian iterative simulation methods are very different from the diagnostics for the EM algorithm. As discussed in Chapter 6, the EM algorithm converges when the parameter estimates no longer change across successive iterations. But, as discussed in this Chapter, the fundamental mechanism of Bayesian iterative simulation methods is to draw parameter estimates randomly from their posterior distribution. This means that draws from each iteration are most likely to be different from any other draws from other iterations. Hence, apparently, the Bayesian iterative simulation has a different kind of convergence to that of EM algorithm.

However, the Bayesian iterative simulation has to stop at some point of the iteration. Generally speaking, the Bayesian iterative simulation converges when the distributions of the parameter estimates become stable and stationary. Why do we say a stable and stationary distribution of the parameter estimates means convergence? This is because that once a draw of parameter estimate from the target distribution has been obtained in the process of the simulation, all the subsequent draws will be from that distribution. This means, for the subsequent draw  $\theta_t$  and  $\theta_{t+k}$ , where  $k$  is any integer, the joint distribution of the parameter estimates  $\theta_{t1}, \dots, \theta_{tm}$  is the same as the joint distribution of the parameter estimates  $\theta_{t1+k}, \dots, \theta_{tm+k}$  for

all  $n$  and  $k$ , given  $t_1, \dots, t_n$ . Hence, the posterior distribution becomes stable and we use this criterion as an indicator of the convergence of the Bayesian iterative simulation methods.

As an aside, the convergence diagnostics for the Bayesian iterative simulation methods which are applied to the imputation of missing data is no different from any other Bayesian iterative simulation methods which do not involve the missing data problem. Despite whether there is any missing data, the convergence diagnostics all focus on measuring the changes of parameter estimates of the posterior distribution. Theoretically, we can check the convergence of the distribution of individual missing data points themselves, because the Bayesian method treats the individual missing data and parameters as random quantities. However, in practice, a dataset normally has a large number of missing data. Hence, it will be very hard and complex to check the convergence for each missing data point.

### 7.4.2 Pre-convergence: the burn-in period

The “burn-in” refers to the part of a Bayesian iterative simulation chain where the current state of the chain is dependent on its starting point (Sahlin 2011). In other words, it refers to the part of the chain before its convergence. Researchers normally throw away the burn-in iterations. Then, after the burn-in period, they do their calculation based on the iterates from the convergence part.

### 7.4.3 Some of the popular Methods of Convergence Diagnostics

**Time Series Plots:** Intuitively, the simplest way to look for the convergence of the Bayesian iterative simulation is to plot all the simulated parameter estimates on a time series like scale. The vertical axis is the values of all the parameter estimates, and the horizontal axis is the iteration number which is similar to the time in a time series plot. Then, we just simply look at the time series plot to see whether and beyond which point the series starts becoming stationary.

A burning question is: “How do we know how many iterations we need in order to see the series turning stationary?”. The answer is “we do not know”. Hence, we can only increase the number of iterations until we find the series turning stable and stationary. According to (Enders 2010, p. 206), if the series becomes stable and stationary at iteration  $t$ , we normally double or triple the number  $t$  for an extra margin of safety.

Figure 7.4 displays two time series plots of simulated SURF’s Income means. If the series does not have any obvious trend, and stays stable and stationary after some iteration, we say there is a possible convergence. It shows that the distribution of the income means simulated by the MH algorithm (the top plot) does not converge even for 10000 iterations. On the other hand, the distribution of the income means generated by the Gibbs sampler (the bottom plot) converges very quickly.



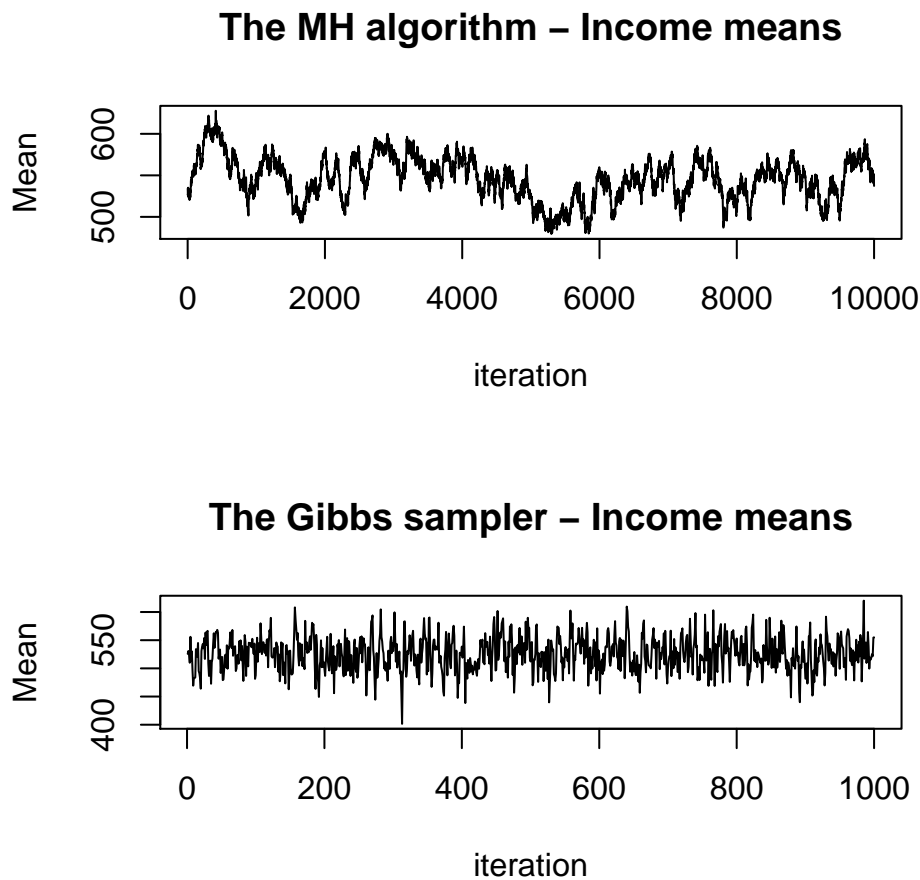


Figure 7.4: Time series plots for the simulated SURF's Income mean. The top plot shows the results of the MH algorithm; the bottom plot shows the results of the Gibbs sampler

**Gelman and Rubin's method:** The Time Series Plots are simple and “easy to use” convergence diagnostic methods, but they are more or less subjective<sup>6</sup>, and more importantly, they are only suitable for a single sequence of the Bayesian iterative simulation in practice, because it would be very tedious to plot multiple sequences for each parameter for which we want to find its convergence. But, why do we want to run multiple sequences? This is because the converged value of a single sequence might correspond to a local maximum instead of a global maximum, if the posterior distribution is not unimodal<sup>7</sup>(Hoeschele 1989). Hence, The Time Series Plots and the Autocorrelation Function plots are only recommended for well-understood models and straightforward data sets Little & Rubin (2002, pg. 206).

For the not so well-understood models and complex data sets, or for all the known and unknown distributions in general, Gelman & Rubin (1992) propose a general approach to monitoring convergence of the Bayesian iterative simulation methods by simulating  $D > 1$  sequences with starting values dispersed throughout the parameter space. This means that the starting values for parameter estimates are far away from the centre of their respective posterior distribution. Then, the convergence obtained, if variations between and within the  $D$  simulated sequences are roughly equal. Obviously, the convergence which is monitored this way has reduced risk of corresponding to a local max-mode. This is for two reasons.

<sup>6</sup>The decision of convergence depends on the shape of the plots.

<sup>7</sup>A unimodal probability distribution is a probability distribution which has a single mode. A mode is the maximum value, or the most likely value of a probability distribution.

The first reason is that  $D$  dispersed starting points increase the chance of reaching different local max-modes, if the posterior distribution is not unimodal. The second reason is that the between sequences variation would be not equal to the within sequences variation, if each sequence only converged to its local maximum. In addition, a single sequence might have a starting value which is very close or far away from the centre of the posterior distribution by chance. This means that the convergence speed is either too fast or too slow. Hence, multiple sequences provide us with a more conservative guess of convergence speed than single sequence (Enders 2010, pg. 209).

The actual method is rather straightforward. Suppose we have  $D$  sequences, and each sequence has  $T$  iterations, where  $d = 1, \dots, D$ , and  $t = 1, \dots, T$ . Then, the between sequence variance is:

$$B = \frac{T}{D-1} \sum_{d=1}^D (\bar{\theta}_{.d} - \bar{\theta}_{..})^2,$$

where

$$\begin{aligned} \bar{\theta}_{.d} &= \frac{1}{T} \sum_{t=1}^T \hat{\theta}_{t,d} \\ \bar{\theta}_{..} &= \frac{1}{D} \sum_{d=1}^D \bar{\theta}_{.d} \end{aligned}$$

and the within sequence variance is:

$$\bar{V} = \frac{1}{D} \sum_{d=1}^D s_d^2,$$

where

$$s_d^2 = \frac{1}{T-1} \sum_{t=1}^T (\hat{\theta}_{t,d} - \bar{\theta}_{.d})^2.$$

Then, we estimate the overall variance  $\hat{V}_{total}$ , by a weighted average of the within and between variances:

$$\hat{V}_{total} = \frac{T-1}{T} \bar{V} + \frac{1}{T} B \quad (7.9)$$

If the chains do not converge, the first term on the right hand side of the equation underestimates the variance, since the individual chains have not had time to range all over the stationary distribution, and the second term overestimates the variance, since the starting points were chosen to be dispersed. As a result, the within variance ( $\bar{V}$ ) should be smaller than the between variance ( $B$ ) (Gelman et al. 1995, pg. 332). However, as  $T \rightarrow \infty$  in Equation (7.9), we can see the first term  $\frac{T-1}{T} \bar{V} \rightarrow \bar{V}$ , and the second term  $\frac{1}{T} B \rightarrow 0$ , then  $\hat{V}_{total} \approx \bar{V}$ , which means the expectation of within variance ( $\bar{V}$ ) approaches the total variance ( $V_{total}$ ). Therefore, Gelman & Rubin (1992) establish a single explicit monitoring statistic  $R$ , that compares  $\hat{V}_{total}$  and  $\bar{V}$ :

$$R = \sqrt{\frac{\hat{V}_{total}}{\bar{V}}}$$

which declines to 1 as  $T \rightarrow \infty$ . So, if  $R$  is close to 1, we have convergence. Otherwise, the simulation runs should be continued, or it suggests that the simulation algorithm is not efficient.

## 7.5 Applying Gibbs sampler to multiple regression with missing data

So far, we have only applied the MCMC methods to MCAR situation. Now, we consider the situation of MAR. The multiple regression model can be as useful a method to impute missing data, as a single imputation method, when the missingness is MAR. The variable with missing values is our response variable, and the variables which the missingness depends on are our explanatory variables. The Gibbs sampler can be applied to this regression model, if we treat the missing values and the regression parameters as random variables.

Suppose we still have  $Y = (Y_{obs}, Y_{mis})$ , but we also have some explanatory variables  $X$  which the missingness depends on.  $Y$  can be regressed on  $X$ :

$$Y = X\beta + e$$

where  $\beta$  is the coefficients. Now, our posterior becomes:

$$p(\beta, \sigma_e^2, Y_{mis} | Y_{obs}, X) \propto p(\beta, \sigma_e^2) p(Y_{obs} | \beta, \sigma_e^2, X_{obs}) p(Y_{mis} | \beta, \sigma_e^2, X_{mis})$$

Assume we have the values for  $Y_{mis}$ , then the conditional posterior for  $\beta$  is:

$$(\beta | Y, \sigma_e^2) \sim N((X^T X)^{-1} (X^T Y), \sigma_e^2 (X^T X)^{-1})$$

Similarly, given fully observed  $Y$ , the conditional posterior for the error variance  $\sigma_e^2$  is inverse gamma:

$$(\sigma_e^2 | Y, \beta) \sim IG(n/2, e^T e/2)$$

where  $n$  is the sample size, and  $(e = Y - X\beta)$ .

Then, we draw  $Y_{mis}$  from a normal distribution with a mean of  $X_i^T \beta$  and variance  $\sigma_e^2$ .

$$(Y_{mis} | X, \beta, \sigma_e^2) \sim N(X_i^T \beta, \sigma_e^2)$$

The following R program applies the Gibbs sampler method to the SURF data with MAR income values. The missingness depends on the gender variable. The male respondents have 50% probability of getting missing income, and the female respondents have 20% probability of getting missing income. The multiple regression model is set up with the income as the response variable, and the gender, age, and working hours as the explanatory variable. Although the missingness is not related to the age and the working hours variables, we included them in the regression model in order to have a best fit regression model. This is because the age and working hours are highly correlated with the income variable. The number of iteration is 1000.

```

#Special example - Multiple linear regression (Gibbs sampling)
#We still use the SURF data. Now, suppose we have some MAR missing Income, but
#other variables are observed.

# step 0: Create MAR income data
Y_MAR=MAR(SURF,"Gender","Income",c(0.5,0.2))
# step 1: set up y and x matrix
y=as.matrix(Y_MAR$Income)
x=as.matrix(cbind(rep(1,nrow(Y_MAR)),Y_MAR$Gender, Y_MAR$Age, Y_MAR$Hours))
ystar=y

# step 2: establish initial parameters
iter=10000
s2=matrix(1,iter) #sigma squire
b=matrix(0,iter,4) #beta matrix. only consider three variables:Gender, Age, Hours
a=matrix(0,iter) #intercept
xtxi=solve(t(x)%*%x)
muY=matrix(mean(ystar[!is.na(y)]),iter) #mean Y
varY=matrix(var(ystar[!is.na(y)]),iter) #variance Y

for (i in 2:iter){
#step 3: sample missing data
  ystar[is.na(y)]=rnorm(length(ystar[is.na(y)]),
                        mean=x[is.na(y),]%*(b[i-1,]), sd=sqrt(s2[i-1]))
  muY[i]=mean(ystar)
  varY[i]=var(ystar)
#step 4: simulate beta from multivariate normal distribution
  b[i,]=coefficients(lm(ystar~x-1))+t(rnorm(4,0,1))%*%chol(s2[i-1]*xtxi)
#step 5: simulate sigma from inverse gamma distribution
  s2[i]=1/rgamma(1,length(y)/2,
                0.5*t(ystar-x%*(b[i,]))%*(ystar-x%*(b[i,])))
}

```

Figure 7.5 shows the results of the estimated coefficients for the gender, and the variances of the residual. The red lines represent the true estimated coefficient for the gender and the variance of the residual, if there is no missing data. The graphs show that the Gibbs sampler produces unbiased estimates and the estimates converge quickly.

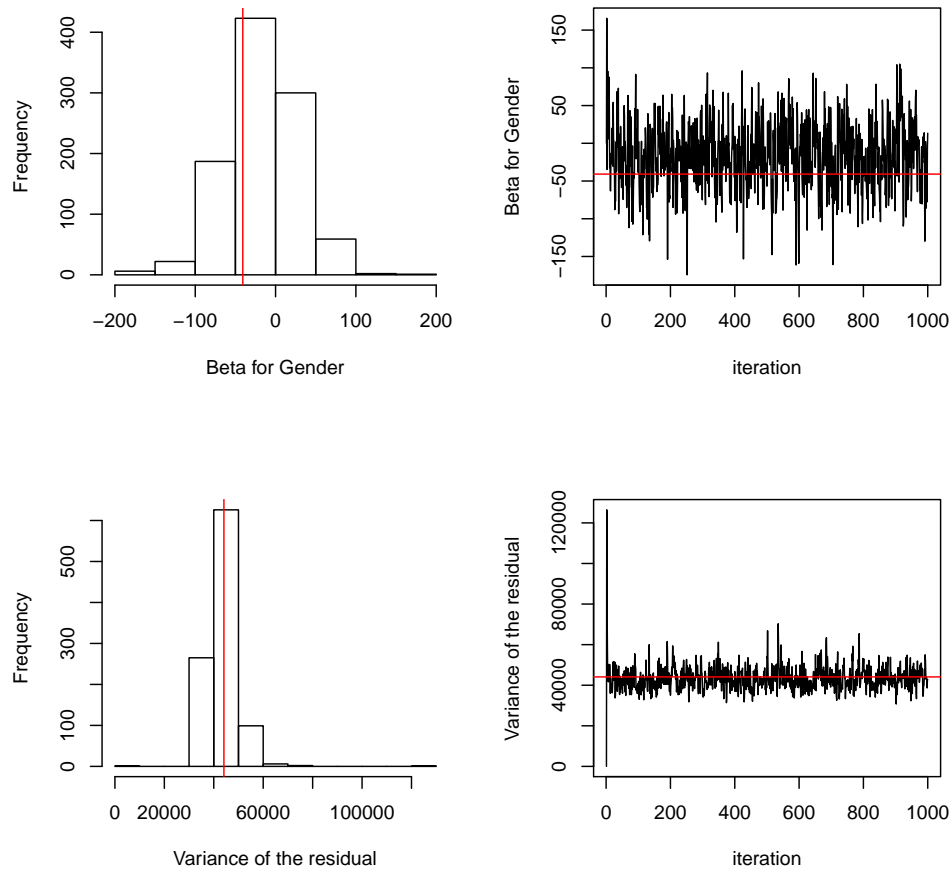


Figure 7.5: Applying the Gibbs sampler to the multiple regression with missing data. The response variable is income with missing data, the explanatory variables are gender, age, and working hours. The graph only displays the values of the estimated coefficient for the gender, and the variances of the residual. The number of iterations is 1000.

## 7.6 Conclusion

This chapter has basically introduced the Bayesian theory and its two most popular iterative simulation methods: the MH algorithm and the Gibbs sampler. The advantage of the MH algorithm is its ability to deal with complex or high dimension posterior distributions. As we have seen, the MH algorithm does not need to breakdown those distributions into smaller and more manageable distributions. In other words, it can simulate parameters from a familiar and manageable “proposal distribution” without directly drawing from the complex and high dimension distribution. However, it may suffer slow convergence, or even not be able to converge, if we use the independent MH sampler, because the “proposal distribution” might not be close to the target distribution. On the contrary, the Gibbs sampler breaks down those complex or high dimension distributions into smaller and more manageable distributions, and samples directly from these sub distributions. This makes the Gibbs sampler generally much faster than the MH algorithm to converge. But, there are situations where we cannot breakdown the target distribution. Hence, we can only rely on the MH algorithm.

We have also discussed a few convergence diagnostic methods. These methods are crucial for Bayesian Multiple Imputation (MI). This is because Bayesian MI needs to generate its imputed data sets which have independent imputations by randomly drawing imputations from

the stationary or converged part of the simulated chain or chains. This will be elaborated in more detail in the following chapter.

As we have mentioned at the beginning of this chapter, the Bayesian theory is the underlying theory of Rubin's Bayesian Multiple Imputation. He distinguished Bayesian MI from the MI which does not implement Bayesian methods as proper and improper MI. Again, we will discuss these concepts in the next chapter.

# Chapter 8

## Multiple Imputation

### 8.1 Introduction

Multiple Imputation (MI) has become a more and more popular method for statistically handling missing data in recent years (Rubin (1996); Allison (2002); Schafer & Graham (2002)). The reason for its success is rooted in the fundamental purposes of imputation. The whole imputation idea is basically trying to achieve two objectives: (1) reduce non-response bias; (2) create complete data in order to apply statistical analysis methods and software easily (Ghosh & Pahwa 2008). As discussed in Chapter 4, we have seen that imputation can indeed reduce non-response bias. However, as we have discussed in Chapter 5, if missing values are only imputed once, then the imputation yields a different kinds of bias which is called uncertainty of non-response. To be more specific, whenever a single imputation strategy is used, the standard errors of estimates tend to be too low. Intuitively, there is no way we can “guess” the true values of missing data. There is always some uncertainty about the missing data, but by choosing a single imputation we pretend that we have found the true values of missing data. Hence, we need a way to quantify the extra uncertainty we have introduced to the data.

In Chapter 5, we have discussed several simple resampling methods which propagate imputation uncertainty, but compared to MI, these resampling methods have some inconvenient disadvantages. While neither the resampling methods nor MI is model free, the resampling methods only work well for large samples and require at least 200 different imputed data sets. On the other hand, MI, if it has Bayesian statistics as its underlying theory, can work for both large and small samples, and it usually only needs 2 to 10 imputed data sets (Rubin (1996); Fay (1996); Rao (1996)). Apparently, MI reduces data storage and transition<sup>1</sup> costs.

Now, the question is what Bayesian MI actually is? As the name implies, Multiple Imputation imputes missing values multiple times. The multiply imputed values are derived from an iterative process which is normally based on Bayesian models that use the observed data. Each set of imputed values is then used to replace missing data in the incomplete data set. For each set of imputed values, there is a separate complete dataset. If there are  $D$  sets of imputed values, then there are  $D$  imputed complete data sets. Figure 8.1 shows the general concept. The multiple datasets are then used in complete case analyses to test the statistical models of interest. As a result, the multiple complete data analyses produce multiple estimates (e.g., mean, variance and regression coefficients.). Those multiple estimates are then combined to produce a single set of best estimates. The multiple estimates can also be used to estimate the

---

<sup>1</sup>Transition means moving data from one place to another

increased variability which can be used to adjust standard errors upward, which in turn reduces the probability of a Type I error<sup>2</sup>(McKnight, M.McKnight, Sidani & Figueredo 2007).

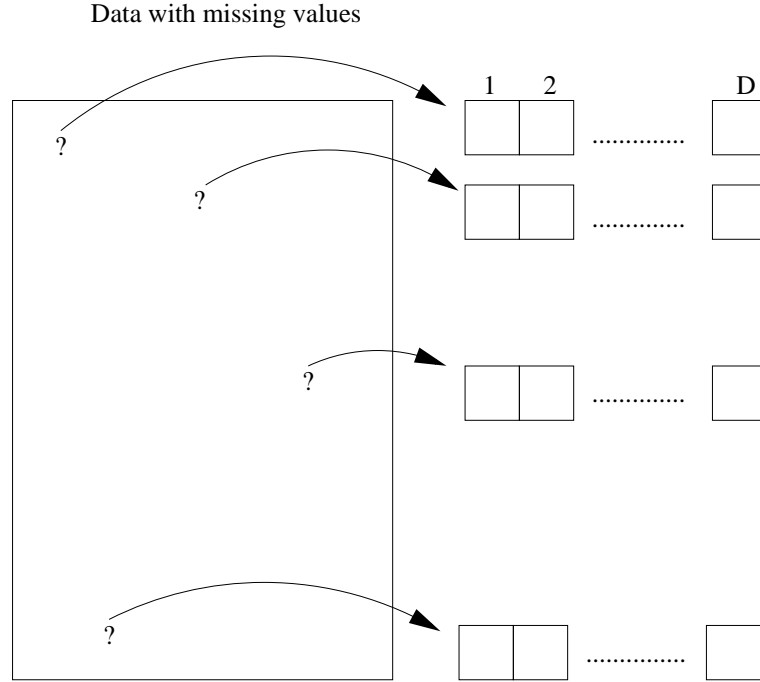


Figure 8.1: Matrix of multivariate data with missing values and multiple imputation

## 8.2 Analysis of multiply-imputed data

Let  $\hat{\theta}_d, W_d, d = 1, \dots, D$  be  $D$  complete-data estimates and their associated variances for an estimated parameter  $\theta$ , calculated from  $D$  repeated imputations under one model. The combined estimate is the average of the  $D$  sets of estimate  $\hat{\theta}_d$  over  $D$ :

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d \quad (8.1)$$

The variability associated with this estimate  $\bar{\theta}_D$  has two components: the average within-imputation variance,

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^D W_d, \quad (8.2)$$

and the between-imputation variance component which is calculated by computing the variance for each estimate,

$$B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_D)^2 \quad (8.3)$$

The total variability associated with  $\bar{\theta}_D$  equals within variance  $\bar{W}_D$  plus between variance  $B_D$ , and Rubin (1987) weights the between imputation variance according to the number of imputations performed. Thus, the total variance  $T_D$  is:

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D \quad (8.4)$$

<sup>2</sup>A null hypothesis that should have been accepted was rejected.



where  $(1 + 1/D)$  is a weighted adjustment for finite  $D$ . Rubin & Schenker (1986) works out that for large sample sizes and scalar  $\theta$ , the reference distribution for interval estimates and significance tests is a  $t$  distribution,

$$(\theta - \bar{\theta}_D) T_D^{-\frac{1}{2}} \sim t_{df}, \quad (8.5)$$

where the degrees of freedom,

$$df = (D - 1) \left( 1 + \frac{D}{D + 1} \frac{\bar{W}_D}{B_D} \right)^2 \quad (8.6)$$

Eq. (8.6) shows that the degrees of freedom is adjusted based on the number of imputations and the within and between imputation variance to correct for the missing data. McKnight, M. McKnight, Sidani & Figueredo (2007) say we should not use the degrees of freedom from the imputed “complete” data set nor the degrees of freedom from the data set with only observed units. This is because we need to adjust the degrees of freedom based on the number of imputations and the within and between imputation variance to correct for the missing data.

A unique feature of MI is its ability to estimate the influence of missing data on estimation. In other words, MI can help us measure the information loss or rate of missing information. We then have a sense of the impact the missing data have on estimates based on the rate of missing information. Missing information is different from missing data. As Dempster, Laird & Rubin (1977) point out, the higher amount of missing data does not necessarily have a larger impact on the estimates, but the higher amount of missing information suggests that missing data have a larger impact on the estimates.

In fact, the missing information can be measured by looking at the variance. Rubin (1987) provides diagnostic measures for assessing the missing information, known as the estimated rate of missing information  $\gamma$ . The rate of missing information can be estimated using the degrees of freedom from Eq (8.6) and the relative increase in variance ( $v$ , defined shortly) due to non-response:

$$\gamma = \frac{v + 2/(df + 3)}{v + 1} \quad (8.7)$$

where

$$v = \frac{(1 + D^{-1})B_D}{\bar{W}_D} = \frac{D + 1}{D} \frac{B_D}{\bar{W}_D}$$

then according to Eq (8.6):

$$df = (D - 1) \left( 1 + \frac{1}{v} \right)$$

### 8.3 The MI Process

So far, we have a general understanding of how MI works. The MI can be concluded in basically four steps: imputing values, conducting statistical analyses with each imputed complete dataset, combining the results/estimates from each analysis, and analysing the combined estimates. However, we still haven’t gotten a tangible idea of how MI works, and how Bayesian theorem can be involved in MI procedure. In this section, we apply the MI procedure to the SURF data with MCAR missing income variable as an example. The next section, we will introduce the concept of proper and improper MI in order to emphasis the important role of Bayesian theory to MI.

As in the example given in Chapter 7, section 7.5, we have the SURF Income as the response variable with MAR missing data. The MAR missing data depend on Gender. We also have Gender, Age and Hours as our explanatory variables. Assuming SURF Income is normally distributed, we can construct a multiple regression model between income and the other three explanatory variables. The choice of these variables is only for demonstration purposes. In practice, Allison (2002) suggests including variables that are highly correlated with the variables that have missing data or are associated with the probability that those variables have missing data; even those variables which may not make practical sense to be included in the model.

## Step 1: Imputation

As Schafer (2005) describes, there is no necessary restriction on the MI procedure selected. In other words, the missing values can be imputed using stochastic single imputation methods (e.g., hot deck imputation), likelihood-based methods (e.g., EM algorithm), or Bayesian iterative simulation methods (e.g., the MCMC method). However, McKnight et al. (2007) recommends that an iterative procedure that is not limited to only a specific group of imputed values and produces values which are unique between each imputed set, and share a common underlying relationship to the data is preferred for MI. This means that Bayesian iterative simulation imputation is more suitable for MI than any other imputation method. In fact, we can see that Bayesian procedures satisfy the conditions to be “proper” MI in the next section. In contrast, McKnight et al. (2007) also points out that the idea of mixing single imputation methods in MI is not encouraged. To be clear, for example, suppose we want to create five MI datasets, it is not recommended to create one MI dataset by using the hot deck imputation, and the second one by using mean imputation, and the third one by using regression imputation, and the rest MI datasets by using other different single imputation methods. This is because the single imputation methods produce somewhat different results (e.g., mean imputation vs hot deck imputation). For example, the hot deck imputation method can maintain the variance, if it has been used properly. But, the mean imputation decreases the variance. Those differences would then be reflected in the estimate of missing information by yielding larger estimates due to the increased variability.

If Bayesian iterative simulation methods are used to perform the imputation stage of the MI procedure, then how do we choose the values to replace the missing data? After all, as we have discussed in Chapter 7, the Bayesian iterative (MCMC) simulation methods converge to a stationary distribution. This means any values from the stationary distribution can be the candidate for replacing the missing data. In fact, Enders (2010, pg. 211) summarizes two approaches which generate independent imputations. The first approach: sample  $D$  candidate points at regular intervals in the part where the Bayesian iterative simulation chain converges; the second approach: generate  $D$  Bayesian iterative simulation chains which start at different starting points. After these chains converge, select the end points of each chain as the imputed data.

For our example, we use the Gibbs sampler method to impute missing SURF income data as the demonstration of Chapter 7, section 7.5. We apply the Gibbs sampler to produce chains of estimates. Then, we select five points on the chains to get five imputed datasets. The choice of five imputations is based on Rubin (1987)’s recommendation which has been discussed previously. The five complete data sets provide the basis for step 2.

To be clear, the following steps show how we do the simulation:

### Recipe: MI-Gibbs sampler

- Step 1:** set up starting points. In this case, they are sigma squared  $S^2 = 1$ , and intercept  $\beta_0$ , coefficients  $\beta_{Gender}$ ,  $\beta_{Age}$ , and  $\beta_{hours}$  which are all equal to 0.
- Step 2:** apply the Gibbs sampler as we've described in previous chapter. Simulating 1000 iterations produces 1000 estimates for  $S^2$  and  $\beta$ , and 1000 sets of imputed datasets. The exact detail of how to apply the Gibbs sampler to multiple regression is in Chapter 7, section 7.5. Here, we just list a simplified version of applying the Gibbs sampler to our particular multiple regression.

$$\begin{aligned}(Y_{mis}|X, \beta, \sigma_e^2) &\sim N(X_i^T \beta, \sigma_e^2) \\ (\beta|Y, \sigma_e^2) &\sim N((X^T X)^{-1}(X^T Y), \sigma_e^2(X^T X)^{-1}) \\ (\sigma_e^2|Y, \beta) &\sim IG(n/2, e^T e/2)\end{aligned}$$

where  $X = (X_{Gender}, X_{Age}, X_{Hours})$  and  $\beta = (\beta_0, \beta_{Gender}, \beta_{Age}, \beta_{Hours})$ .

- Step 3:** apply simple time series convergence diagnostics. As the plots show on Figure 8.2, we conclude that the chains converge very quickly after the first few iterations
- Step 4:** suppose the first 200 iterations are the burn-in period, then we start picking five imputed complete datasets after the first 200 iterations.

The R program in Appendix D shows how we implement MCMC MI procedure to the SURF data.

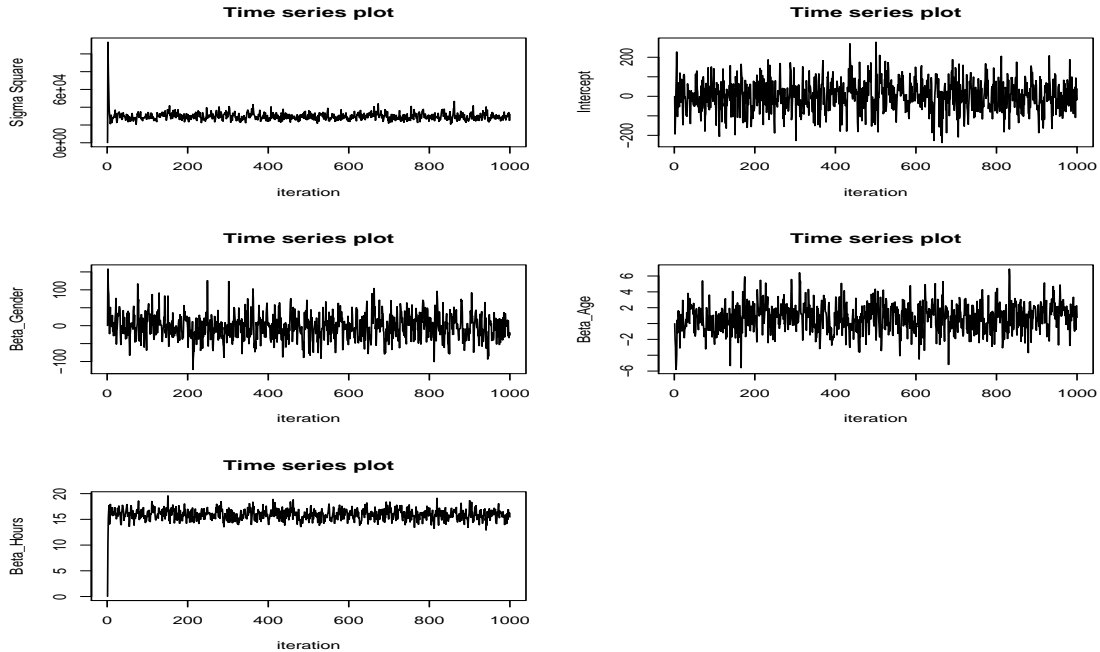


Figure 8.2: Time Series convergence diagnostics

## Step 2: Routine Analysis

This step simply involves performing statistical analysis on the MI datasets. The only difference is that we need to perform such analysis  $D$  times for each imputed dataset, instead of doing it to one data set.  $D$  is the number of imputed data sets. Generally, there are no limits on the types of statistical analyses that can be conducted on the MI datasets. However, for the demonstration purpose, we only look at the estimates of income means, and the variances of income means. We report the overall mean, and the means for each Qualification. There are four Qualification levels: None, School, Vocational and Degree. The five complete case analyses are listed in Table 8.1.

The five values for each of the means reflect variability in the estimates resulting from the different imputed values. We see the same variability exists in the variances associated with each of these means.

The multiple parameter estimates in Table 8.1 are used for the next step in the MI procedure.

Table 8.1: Mean Estimates (total mean and means for each qualification) from Each of the Five MI Data Sets

Income Means	Data Sets				
	1	2	3	4	5
Overall mean (variance of the mean)	583.88 (558.15)	590.00 (572.96)	601.77 (605.73)	576.21 (557.93)	582.73 (504.22)
Qualification: None - mean (variance of the mean)	612.19 (3132.39)	614.64 (2913.43)	622.25 (3377.14)	597.93 (2399.58)	586.09 (2088.12)
Qualification: School - mean (variance of the mean)	582.44 (1907.15)	568.19 (1998.73)	573.76 (1819.94)	568.70 (2142.98)	571.07 (1901.19)
Qualification: Vocational - mean (variance of the mean)	589.12 (1387.88)	626.04 (1457.40)	620.83 (1561.13)	588.05 (1444.12)	598.62 (1224.36)
Qualification: Degree - mean (variance of the mean)	535.29 (4172.11)	520.85 (4220.79)	593.71 (5685.15)	535.33 (3957.62)	567.50 (4458.07)

## Step 3: Parameter Estimation from Aggregated Results

According to Rubin (1987), we can take the mean of the estimates produced by each of the imputed data sets to obtain a single estimate for each parameter, as equation (8.1):

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d$$

If we say each parameter estimate is referred to as  $\hat{\theta}$ , then the overall estimates or mean of the  $\hat{\theta}$ s is referred to as  $\bar{\theta}$ . The last column of Table 8.2 shows the overall estimate for the intercept, coefficients and  $\sigma_e^2$ . Table 8.2 is just an extension of Table 8.1.

Table 8.2: Overall Estimates for means

Income Means	Data Sets					Overall estimate of $\bar{\theta}_D$
	1	2	3	4	5	
overall mean	583.88	590.00	601.77	576.21	582.73	586.92
Qualification: None - mean	612.19	614.64	622.25	597.93	586.09	606.62
Qualification: School - mean	582.44	568.19	573.76	568.70	571.07	572.83
Qualification: Vocational - mean	589.12	626.04	620.83	588.05	598.62	604.53
Qualification: Degree - mean	535.29	520.85	593.71	535.33	567.50	550.54

There are several steps involved to compute the overall variance of the income mean, which is necessary for significance test and confidence intervals. The overall variance also includes the effect of non-response uncertainty. First, we need to compute the within-imputation variance, which is basically the average of the variances associated with each of the parameters, as equation (8.2):

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^D W_d$$

where  $\bar{W}_D$  represents the variability of the variances that are calculated within each of the imputations. Table 8.3 displays the  $\bar{W}_D$ s and the variances of the overall income means, and the variances of income means for each Qualification levels.

Table 8.3: Within-Imputations Variance for means

variances of Income mean	Data Sets					Within Imputations variance or $\bar{W}_D$
	1	2	3	4	5	
Variance of overall mean	558.15	572.96	605.73	557.93	504.22	559.8
Qualification: None	3132.39	2913.43	3377.14	2399.58	2088.12	2782.13
Qualification: School	1907.15	1998.73	1819.94	2142.98	1901.19	1954
Qualification: Vocational	1387.88	1457.40	1561.13	1444.12	1224.36	1214.98
Qualification: Degree	4172.11	4220.79	5685.15	3957.62	4458.07	4498.75

After estimation of the within-imputation variance, we move on to estimate the between-imputation variance  $B_D$ . The between-imputation variance is:

$$B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_D)^2$$

The last column of Table 8.4 lists the results of our example. Again, Table 8.4 is just an extension of Table 8.2.

Table 8.4: Between Imputation Variance (or  $B$ ) for means

Income Means	Data Sets					Between imputation variance or $B_D$
	1 $\hat{\theta}_1$	2 $\hat{\theta}_2$	3 $\hat{\theta}_3$	4 $\hat{\theta}_4$	5 $\hat{\theta}_5$	
Overall mean	583.88	590.00	601.77	576.21	582.73	92.88
Qualification: None - mean	612.19	614.64	622.25	597.93	586.09	209.16
Qualification: School - mean	582.44	568.19	573.76	568.70	571.07	33.73
Qualification: Vocational - mean	589.12	626.04	620.83	588.05	598.62	318.09
Qualification: Degree - mean	535.29	520.85	593.71	535.33	567.50	874.17

Next, given the within-imputation variance  $\bar{W}_D$  and the between-imputation variance  $B_D$ , we can compute the total variance  $T_D$ . Equation (8.4) gives us:

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D$$

As described in the previous section, this is the weighted between-imputation variance plus the within-imputation variance. The factor  $(1 + 1/D)$  which Rubin (1987) describes as an adjustment for using a finite number of imputations. In other words, the factor weights the between-imputation variance according to the number of imputations performed.

The overall standard error for significance testing is equal to the square root of the total variance  $T_D$ . The 95% confidence intervals of  $\bar{\theta}$  can be computed as:

$$\bar{\theta} \pm t_{df,95\%}^*(\sqrt{T_D})$$

Finally, a  $t$ -value similar to the Student's  $t$ -test can be computed by a simplified version of Equation (8.5) which can be displayed as:

$$t_{df} \sim \frac{\bar{\theta}}{\sqrt{T_D}}$$

The computation of the degrees of freedom ( $df$ ) is Equation (8.6):

$$df = (D-1)\left(1 + \frac{D}{D+1} \frac{\bar{W}_D}{B_D}\right)^2$$

As we can see, this formula adjusts the degrees of freedom based on the number of imputations and the within- and between-imputation variance to correct for the missing data (McKnight et al. 2007). This means it has reflected the impact of non-response uncertainty. Now, given the  $t$ -value and its degrees of freedom, we can compute its  $p$ -value. This allows us to test the null hypothesis that  $\theta = 0$ , where  $\theta$  is the parameter of interest. Table 8.5 displays the results of  $T_D$ ,  $t$ -value and  $df$  for our means.

Table 8.5: Tests for Parameter Estimates Produced Using MI

Income Means	$\bar{\theta}_D$	$\bar{W}_D$	$B_D$	$T_D$	$t$ -value	$df$
Overall mean	586.92	559.8	92.88	671.25	0.87	145.09
Qualification: None - mean	606.62	2782.13	209.16	3033.12	0.20	584.15
Qualification: School - mean	572.83	1954	33.73	1994.47	0.29	9714.96
Qualification: Vocational - mean	604.53	1214.98	318.09	1596.69	0.38	69.99
Qualification: Degree - mean	550.54	4498.75	874.17	5547.76	0.10	111.88

The degrees of freedom in Table 8.5 are not accurate. The sample size of SURF data is only 200. Hence, The degrees of freedom could not exceed that number. However, we are sure that our calculation is correct. Then, what is the problem? The problem is that the Equation (8.6) we used to compute the MI's degrees of freedom is designed by Rubin & Schenker (1986) under the assumption that there is an infinite number of observations in the sample. If the sample is infinitely large, then there is no need to consider its influence on the computation of the degrees of freedom. However, for a small size data, Rubin and Schenkers degrees-of-freedom approximation equation produces inaccurate results. Hence, Barnard & Rubin (1999) propose an alternative approach to compute the degrees of freedom of MI for small size samples.

$$df^* = (df^{-1} + \widehat{df}_{obs}^{-1})^{-1} \quad (8.8)$$

where

$$\widehat{df}_{obs}^{-1} = (1 - \hat{f}_D) \left( \frac{df_{com} + 1}{df_{com} + 3} \right) df_{com}$$

The  $\hat{f}_D = (1 + D^{-1})B/(\bar{W}_D + (1 + D^{-1})B)$ , estimates the “fraction of missing information”<sup>3</sup> about  $\theta$  missing due to non-response, and  $df_{com}$  is the degrees of freedom for  $\theta$  if there are no missing values. Note that,  $df^*$  is always less than or equal to  $df_{com}$ , and  $df^*$  equals  $df$  when  $df_{com}$  is infinite. We recomputed the  $df$  by using Equation (8.8). Table 8.6 gives us the accurate  $df$  estimation.

Table 8.6: Tests for Parameter Estimates Produced Using MI

Income Means	$\bar{\theta}_D$	$\bar{W}_D$	$B_D$	$T_D$	$t$ -value	$df^*$
Overall mean	586.92	559.8	92.88	671.25	0.87	77.05
Qualification: None - mean	606.62	2782.13	209.16	3033.12	0.20	31.37
Qualification: School - mean	572.83	1954	33.73	1994.47	0.29	61.42
Qualification: Vocational - mean	604.53	1214.98	318.09	1596.69	0.38	28.74
Qualification: Degree - mean	550.54	4498.75	874.17	5547.76	0.10	17.28

#### Step 4: Compute Missing Information

As discussed in section 8.2, there is a distinction between missing data and missing information. However, most missing data procedures only concentrate on handling the missing data. As explored in previous chapters, large amount of missing data do not mean large impact on the estimates. As equation (8.7) shows, the MI allows us to estimate the amount of missing information. According to McKnight et al. (2007), one of the advantages of MI is that it can measure the rate of missing information that can provide us clues to the impact of missing data on parameter estimates.

Table 8.7 lists the relative increase in variance  $v$  and rate of missing information  $\gamma$  for our example's parameters. These values are calculated by using Equation (8.7):

$$\gamma = \frac{v + 2/(df + 3)}{v + 1}$$

<sup>3</sup>The fraction of missing information measures the level of uncertainty about the values one would impute for current missing data (Wagner 2010).

where

$$v = \frac{(1 + D^{-1})B_D}{\bar{W}_D}$$

Table 8.7: Rate of Missing Information

Income Means	$df^*$	$v$	Rate of missing information or $\gamma$
Overall mean	77.05	0.20	0.19
Qualification: None - mean	31.37	0.09	0.14
Qualification: School - mean	61.42	0.02	0.05
Qualification: Vocational - mean	28.74	0.31	0.29
Qualification: Degree - mean	17.28	0.23	0.27

If we rewrite Equation (8.7) as

$$\gamma = \frac{v + 2/(df + 3)}{v + 1} = \frac{v}{v + 1} + \frac{2/(df + 3)}{v + 1}$$

then, due to  $df > 0$ , it is not hard to notice that

$$0 < \frac{2}{df + 3} < \frac{2}{3}$$

However, most social survey data have large sample sizes. This means the degrees of freedom  $dfs$  are usually very large. Therefore,  $\frac{2}{df+3}$  is actually much less than  $\frac{2}{3}$ , but very close to 0. Under normal circumstances,  $\frac{2}{df+3} \approx 0$ . Hence, if  $v \rightarrow \infty$ , we have  $\gamma = 1$ . It can be concluded that values of  $\gamma$  can range from 0 to 1, where 1 means there is 100% missing information.

We also conclude that the larger the  $v$ , the larger the  $\gamma$ . This is because the larger the  $v$ , the  $\frac{v}{v+1}$  is moving towards 1, given  $\frac{2}{df+3} \approx 0$ . Now, we see that the value of  $\gamma$  is actually driven by the values of  $v$ . According to Equation (8.7), the value of  $v$  is mainly determined by the values of the between imputation and within imputation variances. The greater between imputation variance and the lower within imputation variance can result a large  $v$  value, translated as an increase in the variance of the estimates due to missing data. Hence, the higher the rate of missing information  $\gamma$ , the less the statistical certainty. Table 8.7 indicates that the missing data do not lead large missing information as the  $\gamma$ s are all close to 0.

## 8.4 Proper and Improper Multiple Imputation

We have mentioned in Chapter 7 that Rubin (1987) classifies MI into proper and improper imputation methods. The proper MI generates imputed values from a Bayesian posterior distribution (eg. using the Bayesian iterative simulation methods to generate imputed data); the improper MI generates imputed values without applying any Bayesian theory (e.g., using the simple hot deck method to create several imputed data sets). Rubin recommends the use of the proper MI method because he has shown that this method properly propagates imputation variance. On the other hand, improper MI may underestimate imputation uncertainty.



There is an intuitive way to understand why imputation uncertainty is underestimated using improper MI. As we have described throughout previous chapters, the Bayesian views parameters of a distribution as random variables which have their own distributions. We also know that those distributions of parameters are called posterior distributions which are constructed by combining their prior distribution with the likelihood function. The prior distribution represents our subjective beliefs about the relative probability of different parameter values before collecting any data. The likelihood function is the function of the parameters whose shape is determined by the collected data. Hence, due to MI imputing different values for the missing data, the shape of the likelihood function is slightly different for each imputed data set. This means the posterior distributions of parameters are also slightly different from each other for different imputed data sets. If we ignore the difference and treat the parameters as fixed, which means we act as if the distribution of the observed  $Y_{obs}$  values were exactly the same as the true population distribution of  $Y$  values, and only generate imputed values from the distribution of the observed data, then we certainly underestimate the variability of parameters.

Rubin (1987) defines three conditions for proper Multiple Imputation:

C1  $E(\bar{\theta}_\infty|Y) \approx \hat{\theta}$ , where  $\hat{\theta}$  is the sample estimates if the data is complete.

C2  $E(W_\infty|Y) \approx V$ , where  $V$  is the variance of the sample estimates if the data is complete.

C3  $E(B_\infty|Y) \approx V(\bar{\theta}_\infty|Y)$

These conditions are under the assumptions that  $D$  imputed datasets are infinitely large, that is:

$$\bar{\theta}_\infty = \lim_{D \rightarrow \infty} \bar{\theta}_D$$

$$W_\infty = \lim_{D \rightarrow \infty} W_D$$

$$B_\infty = \lim_{D \rightarrow \infty} B_D$$

These conditions are somewhat intuitive. As described in previous chapters, if samples of the same size are drawn from the same population infinitely many times, the average of the sample estimates is approximately the population estimates. The same logic applies to condition C1. If the imputation which properly takes the account of the missing mechanism were performed infinitely many times on the same data  $Y$  with missing values, then  $E(\bar{\theta}_\infty|Y)$  should approximately equal the sample estimate  $\hat{\theta}$ , if  $Y$  is complete. Condition C2 can be inferred in the same way. The condition C3 is also straightforward because the between imputation variance for the infinite imputations is indeed the variance of  $\bar{\theta}_\infty$ .

Now, let's consider this simple random MI method to investigate if it satisfies the three conditions and demonstrate how exactly improper MI underestimates variability. The simple random MI is basically the multiple-imputation version of the single imputation hot deck method in which multiple imputations are created by drawing a simple random sample with replacement from the  $Y_{obs}$  (Rubin 1987, pg. 120).

First, we consider a simple random sample of size  $n$  with  $r$  respondents and  $m = n - r$  nonrespondents, and let  $\bar{y}_R$  and  $s_R^2$  be the sample mean and variance of the respondents' data, and  $\bar{y}_{NR}$  and  $s_{NR}^2$  the sample mean and variance of the imputed data. Then

$$\bar{y}_* = \frac{1}{n} \sum_{k=1}^n y_k = \frac{1}{n} (r\bar{y}_R + m\bar{y}_{NR}).$$

Suppose imputations are randomly drawn with replacement from the  $r$  respondents' values, then the variance of  $\bar{y}_*$  conditional on the observed data is:

$$\begin{aligned} V(\bar{y}_*) &= V\left(\frac{1}{n}(r\bar{y}_R + m\bar{y}_{NR})\right) \\ &= \frac{m^2}{n^2}V(\bar{y}_{NR}) \\ &= \frac{m^2}{n^2} \frac{1}{m} \frac{r-1}{r} s_R^2 \\ &= \frac{m}{n^2} \frac{r-1}{r} s_R^2 \end{aligned}$$

Now, suppose multiple imputations are created using the same imputation method  $D$  times, and let  $\bar{y}_*^{(d)}$  and  $W_*^d$  be the values of  $\bar{y}_*$  and  $W_*$  for the  $d$ th imputed data set, and  $I_i$  the response indicator, where  $I_i = 1$  if  $y_i$  is observed, and  $I_i = 0$  otherwise. Let  $\bar{\bar{y}}_* = \sum_{d=1}^D \bar{y}_*^{(d)} / D$ , and  $T_* = \bar{W}_* + (1 + D^{-1})B_*$  is the total variance, where  $B_* = \sum_{d=1}^D (\bar{y}_*^{(d)} - \bar{\bar{y}}_*)^2$ .

For condition C1, we have:

$$\bar{\theta}_D = \frac{1}{n} \sum_{i=1}^n I_i y_i + \frac{1}{n} \sum_{i=1}^n (1 - I_i) \frac{1}{D} \sum_{d=1}^D y_{id}^*$$

where  $y_{id}^*$  is the imputed value for the  $d$ th imputation.

$$\bar{\theta}_\infty = \frac{1}{n} \sum_{i=1}^n I_i y_i + \frac{1}{n} \sum_{i=1}^n (1 - I_i) \bar{y}_R = \bar{y}_R$$

and

$$\begin{aligned} E(\bar{\theta}_\infty | Y) &= E\left(\frac{\sum_{i=1}^n I_i y_i}{\sum_{i=1}^n I_i} | Y\right) \\ &\approx \frac{E(\sum_{i=1}^n I_i y_i | Y)}{E(\sum_{i=1}^n I_i | Y)} \\ &= \frac{\sum_{i=1}^n E(I_i y_i | y_i)}{\sum_{i=1}^n E(I_i | y_i)} \\ &= \frac{\sum_{i=1}^n y_i E(I_i)}{\sum_{i=1}^n E(I_i)} \\ &= \bar{y}_n \end{aligned}$$

Hence, condition C1 is satisfied.

For condition C2, we have:

$$W_\infty \approx \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y}_R)^2$$

then

$$E(W_\infty | Y) \approx \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y}_n)^2 = V$$

Hence, condition C2 is satisfied.

For condition C3, we have the expected value of  $B_*$ , conditional on the observed data is:

$$\begin{aligned}
E_d(B_*) &= E \left[ \sum_d (\bar{y}_*^d - \bar{\bar{y}}_*)^2 \right] \\
&= E \left[ \sum_d (\bar{y}_*^d)^2 - D \bar{\bar{y}}_*^2 \right] \\
&= \sum_d E_d(\bar{y}_*^{(d)2}) - D E_d(\bar{\bar{y}}_*^2) \\
&= D \left[ \frac{m}{n^2} \frac{r-1}{r} s_R^2 + \bar{y}_R^2 \right] - \frac{1}{D} \sum_{dd'} E^d \left[ \bar{y}_*^d \bar{y}_*^{d'} \right] \\
&= D \left[ \frac{m}{n^2} \frac{r-1}{r} s_R^2 + \bar{y}_R^2 \right] - \frac{1}{D} \sum_d E_d(\bar{y}_*^{2(d)}) - \frac{1}{D} \sum_{d \neq d'} E_d(\bar{y}_*^d) E_d(\bar{y}_*^{d'}) \\
&= D \left[ \frac{m}{n^2} \frac{r-1}{r} s_R^2 + \bar{y}_R^2 \right] - \frac{1}{D} D \left( \bar{y}_R^2 + \frac{m}{n^2} \frac{r-1}{r} s_R^2 \right) - \frac{1}{D} D(D-1) \bar{y}_R^2 \\
&= (D-1) \frac{m}{n^2} \frac{r-1}{r} s_R^2 + D \bar{y}_R^2 - \bar{y}_R^2 - (D-1) \bar{y}_R^2 \\
&= (D-1) \frac{m}{n^2} \frac{r-1}{r} s_R^2
\end{aligned}$$

This means if the  $D$  is infinitely large, and  $r$  is large as well, then

$$B_\infty = \frac{m}{n^2} \frac{r-1}{r} s_R^2 = \left(1 - \frac{r}{n}\right) \frac{s_R^2}{n} \quad (8.9)$$

If we say  $p = r/n$ , then  $B_\infty = (1-p)V$ , where  $V = \frac{s_R^2}{n}$

Now, let's look at the left side of condition C3,

$$\begin{aligned}
V(\bar{\theta}_\infty|Y) &= V(\bar{y}_R|Y) \\
&= V \left( \frac{\sum_{i=1}^n I_i y_i}{\sum_{i=1}^n I_i} | Y \right) \\
&= \frac{1}{p^2} V \left( \sum_{i=1}^n I_i y_i | Y \right) + \frac{\bar{y}_n^2}{p^2} V \left( \sum_{i=1}^n I_i | Y \right) \\
&\quad - \frac{2\bar{y}_n}{p^2} \text{Cov} \left( \sum_{i=1}^n I_i y_i, \sum_{i=1}^n I_i | Y \right) \\
&= \frac{1}{p^2} \frac{p(1-p)}{n^2} \sum_{i=1}^n y_i^2 + \frac{\bar{y}_n^2}{p^2} \frac{p(1-p)}{n} - \frac{2\bar{y}_n}{p^2} \frac{p(1-p)}{n^2} \sum_{i=1}^n y_i \\
&= \frac{1-p}{p} \frac{1}{n^2} \left( \sum_{i=1}^n y_i^2 - \bar{y}_n^2 \right) \\
&\approx \frac{1-p}{p} V
\end{aligned}$$

Then, we have  $E(B_\infty|Y) < V(\bar{\theta}_\infty|Y)$  which means the condition C3 is not satisfied. Replacing  $p$  with  $r/n$ , we can easily work out that the between imputation variance of the simple random (or hot deck) Multiple Imputation underestimates the true between imputation variance by  $r/n$ .

Mathematically, We have proved that the simple random MI underestimates variance, compared to the proper MI. We can also prove this by simulation. However, it is not easy to adjust the simple random MI to make it a proper MI. Therefore, we choose a simpler imputation method to demonstrate the difference between the proper and improper MI estimate of the variances of the mean. The scheme of the improper MI is that we first compute the mean  $\mu = \bar{y}_R$  and variance  $\sigma^2 = s_R^2$ , then we use the mean and variance to construct a normal distribution  $N(\bar{y}_R, s_R^2)$  and randomly draw  $Y_{mis}$  from this distribution to replace missing data:

$$Y_{mis} \sim N(\bar{y}_R, s_R^2)$$

$D = 5$  MI datasets were created by using this imputation method. This MI scheme is improper because we act as if we know the precise population values, hence we underestimate the variability (Rubin 1987).

As we have discussed at the beginning of this section, the problem is that we do not know the precise population values, and under Bayesian theory, the parameters that we use to construct the population distributions are also random variables. Hence, we should randomly draw  $\mu$  and  $\sigma^2$  from their posterior distributions as well. We have discussed how to do this in Chapter 7 Section 7.3.1. A much simplified explanation of this method is:

$$\begin{aligned} Y_{mis} | \mu, \sigma^2, Y_{obs} &\propto N(\mu, \sigma^2) \\ \sigma^2 | \mu, Y_{obs}, Y_{mis} &\propto IG(n/2, \sum (y_i - \mu)^2 / 2) \\ \mu | \sigma^2, Y_{obs}, Y_{mis} &\propto N(\bar{Y}, \sigma^2 / n) \end{aligned}$$

Please refer to Chapter 7 for a detailed explanation.

We have applied our improper and proper MI methods to 1000 replicated SURF data with missing income values, which generate 1000 total variance values. The missingness is MAR with 50% probability of missing male income and 20% probability of missing female income.

Figure 8.3 clearly show that the proper MI generates larger total variances of the mean, compared to the improper MI. This evidence has further confirmed our assumptions and proofs. The R program for the improper and proper MI procedures can be found in Appendix D

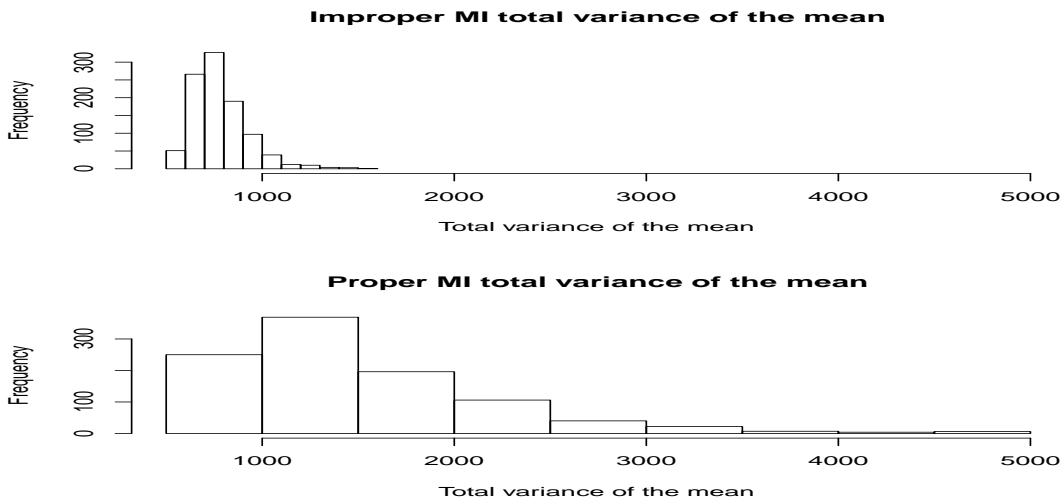


Figure 8.3: Comparison of the total variance of improper MI and proper MI

## 8.5 Conclusion

This chapter demonstrated how to perform MI and compute its subsequent results step by step in detail. The very same procedure will be repeated on the FNES data in a later chapter.

Then, we discussed the distinction between proper and improper MI. We have proved and shown that improper MI underestimates variance. Hence, the take away message is that we should use proper MI (i.e. Bayesian MI) whenever possible.

# Chapter 9

## Imputation methods for categorical variables

### 9.1 Introduction

We have introduced most of the commonly used imputation methods in the previous chapters. However, all of the examples we have demonstrated so far are for continuous numerical variables. Some of these imputation methods need to be adapted somewhat differently in order to apply them to categorical variables, although the underlying theories are the same. This chapter will not introduce any more imputation methods, but focuses on applying those introduced imputation methods to categorical variables, again using the SURF as an example, and comparing these imputed data estimates to the non-missing complete data estimates. This is because we will eventually implement these imputation methods to impute missing data for the Food Nutrition Environment Survey (FNES) data which only has categorical variables.

### 9.2 Types of categorical variable

First of all, we need to distinguish the two main types of categorical variables as this is important for us to choose the proper imputation methods to impute missing categorical data. The two main types of categorical variables are: nominal variables and ordinal variables. “Nominal” is a Latin word for “name”. This means nominal data are items which are distinguished by names. For example, variables such as gender, ethnicity, are nominal variables. One important characteristic of nominal variables is that there is no particular order among categorical items. On the contrary, ordinal variables are set into some kind of order by their position on an ordinal scale. For example, variables such as income band, qualification, are ordinal variables. Income band can be ranked from “low income” to “high income”; qualification can be ranked from “no qualification” to “degree or higher”. In the SURF data, the qualification variable has four levels: none, school, vocational, and degree.

### 9.3 Single imputation methods for categorical data

#### 9.3.1 Mode imputation

Mode imputation is the categorical missing data’s version of mean imputation. We know that it is not possible to compute the mean or median value for a categorical variable. Hence, instead of looking for mean and median values to replace missing data of a categorical variable, missing values for that variable are imputed with the category that has the most of in-

dividuals with the observed values, this is, mode imputation (Ramirez et al. 2011). Actually, most papers in the literature consider mean/median/mode imputation to be the same imputation method. Why? Let's consider what mean/median imputation is actually doing. We know that the mean or median for a numerical variable is the best guess of the centre of its distribution. This means that for a symmetric unimodal distribution, most of data points are centred around mean and median. Then, it makes sense to replace missing values with the mean or the median, because these missing values have higher chance to be located close to mean or median than to be located far away from them on a distribution. Therefore, we see that the underlying theory here is to replace missing values with values which have higher probability than others to be selected from a distribution. For categorical data, the category with the highest frequency is the category that has higher chance than other categories to be selected from a distribution, so it makes sense to replace missing values with the value of that category. Hence, mode imputation and mean/median imputation have the same motivation of selecting the most likely values of a distribution. Obviously, categorical missing data cannot use mean or median imputation, but we have to point out that the mode imputation can be used for numerical continuous variables as well (Torgo 2003). The way is to transform the numerical continuous variables into categorical variables by grouping the numerical values into ranges.

As with mean imputation which we discussed in Chapter 3, Section 3.3.1, there is unconditional and conditional mode imputation. Suppose  $Y$  is a categorical variable with  $n$  observations, and  $Y = (Y_{obs}, Y_{mis})$ , where  $Y_{obs}$  are observed values, and  $Y_{mis}$  are missing values. There are  $r$  response observations, and  $n - r$  missing observations. If  $Y$  has  $K$  categories, with category values  $(C_1, C_2, \dots, C_K)$ , then unconditional mode imputation is:

$$Y_{mis,i} = C_k \quad (9.1)$$

if

$$k = \underset{k}{\operatorname{argmax}} \sum_{j=1}^r q_{k,j}$$

where  $i = (1, \dots, n)$ , and  $j = (1, \dots, r)$

$$q_{k,j} = \begin{cases} 1 & \text{if } Y_{obs,j} = C_k \\ 0 & \text{Otherwise} \end{cases}$$

For conditional mode imputation, suppose we divide  $Y$  into  $g$  groups conditioning on some independent variables  $X$ . Then

$$Y_{g,mis,i} = C_{k_g} \quad (9.2)$$

if

$$k_g = \underset{k_g}{\operatorname{argmax}} \sum_{j=1}^r q_{k,j} I_{i,g}$$

where  $i = (1, \dots, n)$ , and  $j = (1, \dots, r)$

$$q_{k,j} = \begin{cases} 1 & \text{if } Y_{obs,j} = C_k \\ 0 & \text{Otherwise} \end{cases}$$

and

$$I_{i,g} = \begin{cases} 1 & \text{if unit } i \text{ in group } g \\ 0 & \text{Otherwise} \end{cases}$$

From Eq (9.1) and Eq (9.2), we see that unlike mean/median imputation for which there can only be one mean or median for each variable, there can be more than one mode of a variable, which means there is more than one category with the highest frequency. How do we impute missing values if there is more than one mode? There are three methods. The simplest one is to randomly select one category as the replacement values for all the missing data, from categories with the same highest frequencies. Obviously, this method alters the original **distribution of proportion of each category**<sup>1</sup> which has multiple categories with the same highest frequencies in only a single category with the highest frequency. If the proportion of missing data is large, this method makes the estimates very different from the unimputed data. Hence, a slightly better method is to impute missing values equally with all the highest frequency categories. For example, if two categories have the same highest frequencies, then half of the missing values will be imputed by using one of the categories, and the other half of the missing values will be imputed by using the other categories. By doing this, we have made sure that the imputed data still have the same or a very similar distribution of proportions in each category as the unimputed data. But, do we really want the imputed data to have exactly the same distribution of proportions in each category as the original unimputed data? or, do we want some variation from the original unimputed data? Hence, we have this last method which randomly selects one of the categories with the same highest frequencies as the imputed value for each missing value. It is still highly likely that the imputed data will have different distributions of proportions in each category, e.g. only have one category with the highest frequency instead of having multiple categories with the same highest frequencies. However, because we have a random selecting process for each missing value, and each category with the highest frequency has the same chance to be selected each time, it is likely that the imputed data will have roughly similar distributions of proportions in each category as the original unimputed data (e.g. the category with the highest frequency is only slightly bigger than the category with the next highest frequency), although it might be a uni-modal instead of multi-mode distribution. It is really hard to say which of the last two methods is the best method. This is because, in practice, different researchers may have different needs or objections to what the distribution of proportions in each category of the imputed data ought to be like.

As has been done in Chapter 4, we applied the unconditional and conditional mode imputation methods to the “Qualification” variable of the SURF data. The Qualification variable is a categorical variable which has four levels: “None”, “School”, “Vocational”, and “Degree”. First, we applied the MCAR mechanism to the Qualification variable and created 50 MCAR missing values out of 200 observations. Then, the unconditional and conditional mode imputation methods were used to impute those MCAR missing qualification data. The whole process was repeated 1000 times. The following steps depict the exact process:

---

<sup>1</sup>The distribution of the proportion of each category is not the distribution of the data. After any imputation, the distribution of the data will be different, but for a categorical variable, the distribution of the proportion of each category can be the same.



### Recipe: Unconditional and conditional mode imputation

- Step 1:** create 50 MCAR missing observations for the Qualification variable.
- Step 2:** apply unconditional and conditional mode imputation to impute the missing qualification data. For conditional mode imputation, the condition is on “Gender” and “Marital status”.
- Step 3:** repeat step 1 to step 2 1000 times, which produces 1000 imputed qualification variables.

Figure 9.1 and Figure 9.2 show the proportions to the total observations of the four levels of the 1000 imputed qualification variables. The red lines represent the true proportions of the four qualification levels of non-missing qualification variable. Similarly to the results for the unconditional and conditional mean imputation in chapter 4, conditional mode imputation performs better, having less bias against the true proportion values than unconditional imputation, although the missing data is MCAR.

The distributions of the four graphs in Figure 9.1 are somewhat strange. After imputation, the proportions of the qualification categories: none and degree are less than the true proportions of those two categories. Meanwhile, the proportions of the qualification categories: school and vocational are either less or more than the true proportions of those two categories. What are the causes of these patterns? For the rare categories<sup>2</sup>: none and degree, it is highly likely that they are still rare after the creation of the MCAR missing observations, hence, these rare categories will never be imputed under the scheme of the unconditional imputation. This is why we see that their proportions after imputation are less than the true proportions. For the categories with large proportions of observations: school and vocational, it is highly likely that one of them will become the category with the most observed observations after the creation of the MCAR missing observations. If one of them becomes the most populated category, then unconditional imputation will impute all the missing observations with the value of the most populated category. This means its proportion after imputation will be larger than the true proportion. On the other hand, the category with the second most observations will suffer the same fate as those rare categories which means that it will not be imputed. However, we have simulated the process of creating 50 MCAR missing observations 1000 times and the chance of one of the qualification categories: school and vocational becomes the most populated category is random, hence, both school and vocational categories have had the chance to be imputed.

The distributions of the four graphs in Figure 9.2 shows that the conditional imputation method produced much better estimates than the unconditional imputation method, although the estimates are still biased against the true estimates of the proportions. The improvement is due to the imputation being conditioned on the “Gender” and “Marital” variables. This means that we separate the data into several subgroups. Hence, there is a chance that the rare categories might become the most populated categories in a subgroup. Then, we impute missing observations in that subgroup with the value of the “new” most populated category. In the end, the missing observations are not imputed by just a single category value. The rare categories also have a chance to be imputed. This is why we see that the estimates of the proportions moving towards the true proportions. However, the proportions of the rare

---

<sup>2</sup>Categories with small number of observations.

categories: none and degree are small. Under the scheme of the MCAR mechanism, the chance for them to become the dominant categories in a subgroup is slim. This is why that majority of the proportions of none and degree after imputation are still less than the true proportions of these two categories.

Please refer to Appendix E, section E.1.1 for the R code.

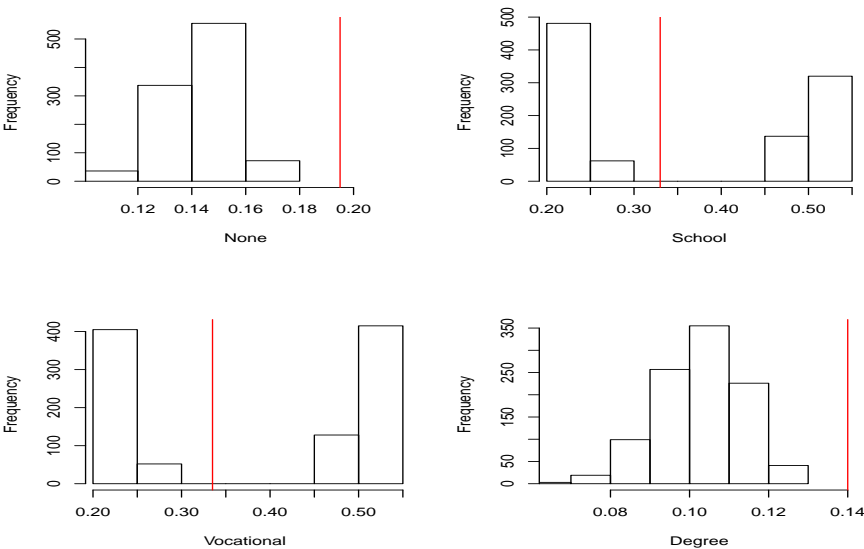


Figure 9.1: The proportions of the four qualification categories. The 1000 qualification variables were imputed by unconditional mode imputation

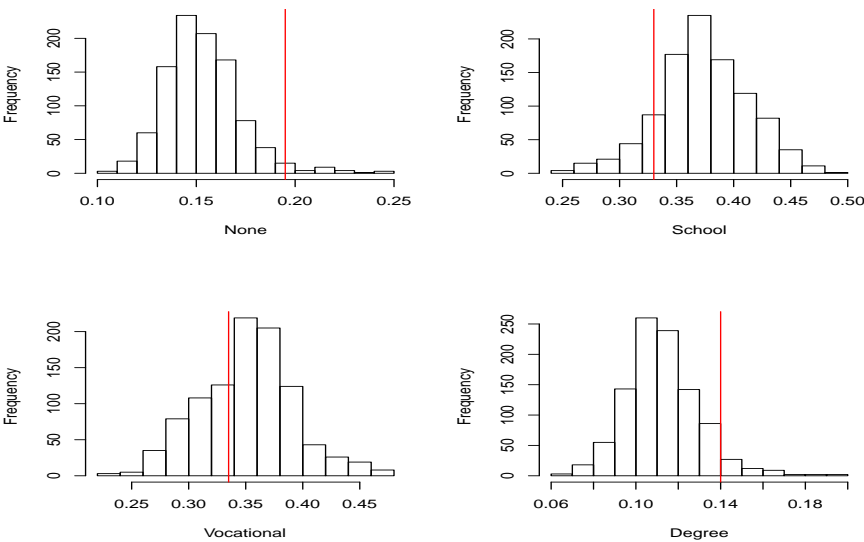


Figure 9.2: The proportions of the four qualification levels. The 1000 qualification variables were imputed by conditional mode imputation

### 9.3.2 Logistic regression imputation

When dealing with categorical response variables, logistic regression models are the commonly preferred models. Although there are other models which can take care of categorical response variables under certain circumstances, such as the probit model, the complementary log log model etc., we only concentrate on logistic regression models as one of our imputation methods for categorical data.

As discussed in Chapter 3, the idea of regression imputation is that we first construct a model based on the response and explanatory variables, then we use this model to “predict” the response variable’s missing values given the observed explanatory variables’ values. However, for logistic regression models, the “predicted” values are probabilities  $\pi$ s. We want our imputed values to be categories, not the probabilities.

This section will introduce two methods to convert the probability  $\pi$  into category values. Let’s consider a simple binary logistic regression first. Suppose we have a binary response variable  $Y = (Y_{obs}, Y_{mis})$ .  $Y_{obs}$  are the observed  $Y$  values, and  $Y_{mis}$  are the missing values. Hence, for unit  $i$  with missing  $Y$  value:

$$\text{logit}[P(Y_{i,mis} = 1)] = \text{logit}[\pi_i] = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = x_i^T \beta \quad (9.3)$$

where  $x_i$  is a vector of explanatory variables, and  $\beta$  is the parameter vector. We can transform Eq.(9.3) to:

$$\pi_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

The response variable  $Y$  has two categories 0 and 1. The question is then how to convert  $\pi_i$  into either 0 or 1. We have three methods.

**Method 1:** If  $\hat{\pi}_i \geq 0.5$ , then we assign  $Y_{i,mis} = 1$ , otherwise,  $Y_{i,mis} = 0$ . The justification for this is that  $Y_{i,mis}$  has more than a 50% chance to be 1, given  $P(Y_{i,mis} = 1) = \hat{\pi}_i$ , if  $\hat{\pi}_i \geq 0.5$ . Although there is still less than a 50% chance that  $Y_{i,mis} = 0$ , we still make  $Y_{i,mis} = 1$ , because we are more than likely to get  $Y_{i,mis} = 1$  than  $Y_{i,mis} = 0$ , if we run the survey again, and given the same explanatory variables. The problem of Method 1 is that the choice of either  $Y_{i,mis} = 1$  or 0 is somewhat arbitrary. It is sure that if  $\hat{\pi}_i \geq 0.5$ , then  $Y_{i,mis}$  is highly likely to be 1, but we cannot exclude the possibility that  $Y_{i,mis}$  could be 0 by a probability of less than 50%. Hence, we have Method 2.

**Method 2:** We can first randomly draw a value  $u$  from a uniform distribution, ( $u \sim (0, 1)$ ). If  $\hat{\pi}_i \geq u$ , we have  $Y_{i,mis} = 1$ , otherwise 0. This Method 2 fixes this problem of the Method 1 by randomly drawing from the uniform distribution. Hence, if  $\hat{\pi}_i \geq 0.5$ , we can still get  $Y_{i,mis} = 0$ , although the chance of getting a 0 is smaller than getting a 1.

Figure 9.3 shows the results of the proportions for male and female of the 1000 simulated SURF datasets which have 50 MCAR missing data for their Gender variable. The red lines represent the true proportions for male and female from the original complete SURF data. As in the previous results in chapter 4, the regression method produced unbiased estimates, that is, all the 1000 proportions are centred around the true proportions.

The R code is in Appendix E section E.1.2.

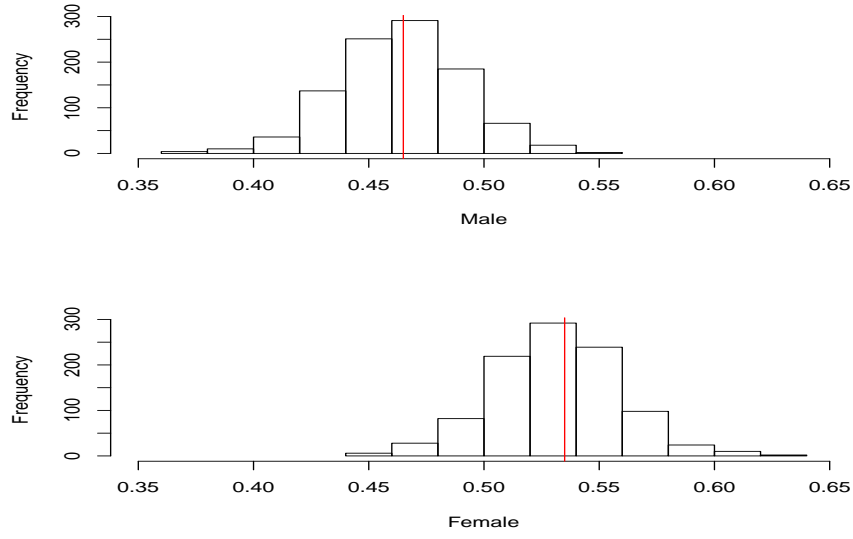


Figure 9.3: The proportion of Male and Female. The 1000 Gender variables were imputed by the logistic regression imputation method

### 9.3.3 The Nearest-Neighbour hot deck imputation methods

Intuitively, it should be straightforward to apply hot deck imputation method to categorical variables with missing data. The basic idea of hot deck imputation is to draw observed values of a variable randomly to replace the missing data. Hence, we can use the same method for any types of data. This is true for most hot deck imputation methods. However, this is not exactly true for the Nearest-Neighbour hot deck imputation method. We have written up the definition of the Nearest-Neighbour hot deck imputation in Chapter 3. It is a distance measure between observations, and imputes the value of a respondent who is “closest” to the observation with the missing item”. The distance is measured by using the distance function, such as Mahalanobis distance.

$$d(i, j) = (x_i - x_j)^T \widehat{Var}(x_i)^{-1} (x_i - x_j)$$

where  $\widehat{Var}(x_i)$  is an estimate of the covariance matrix of  $x_i$ .

Clearly, the Mahalanobis distance function which we have stated above does not work if the  $x$ s are categorical variables, or a mixture of categorical and numerical variables. Hence, researchers have developed several distance functions for categorical variables. In this paper, we introduce the Gower distance function which has been used in the R package “StatMatch”. The Gower distance function is derived from the Gower’s dissimilarity coefficient (Gower 1971), by the Kaufman & Rousseeuw (1990).

Suppose a data matrix  $Y$  has  $k$  categorical variables and  $n$  units, the Gower’s distance function finds the dissimilarity between the  $i$ th and  $j$ th unit by obtaining a weighted sum of dissimilarities for each variable:

$$d(i, j) = \frac{\sum_k \delta_{ijk} d_{ijk}}{\sum_k \delta_{ijk}} \quad (9.4)$$

where  $d_{ijk}$  is the distance between the  $i$ th and  $j$ th unit computed considering the  $k$ th variable,  $\delta_{ijk}$  is the weight. Normally, the weight  $\delta_{ijk}$  equals 1 unless  $y_{ik}$  or  $y_{jk}$  is missing. The computation of  $d_{ijk}$  is different for different types of data.

- if the variable  $k$  is nominal categorical variable, then  $d_{ijk} = 0$  if  $y_{ik} = y_{jk}$ , otherwise  $d_{ijk} = 1$
- if the variable  $k$  is continuous numeric variable, then:

$$d_{ijk} = \frac{|y_{ik} - y_{jk}|}{R_k}$$

where  $R_k$  is the range of the variable  $k$ .

- if the variable  $k$  is ordinal categorical variable, and the values are substituted with the corresponding position index  $r_{ik}$  in the factor levels, then we create a new value  $z_{ik}$ , where

$$z_{ik} = \frac{(r_{ik} - 1)}{\max(r_{ik}) - 1}$$

and the  $d_{ijk}$  is computed by treating the  $z_{ik}$  as continuous numeric variable.

## 9.4 Likelihood based and Bayesian iterative simulation imputation methods for categorical data

### 9.4.1 EM algorithm for categorical variable

Suppose the random variable  $Y$  is a binary categorical response variable, has  $n$  observations, and  $X$  is the vector of explanatory variables. Assume the observations  $i$ ,  $i = 1, \dots, n$ , are independent, and  $Y_i|X \sim \text{Bernoulli}(\pi_i)$ . From Eq.(9.3), we know that:

$$\pi_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} = \frac{1}{1 + \exp(-x_i^T \beta)}$$

where  $\beta$  denotes the vector of parameters to be estimated.

Since each  $Y_i$  is a Bernoulli random variable, then its probability distribution is:

$$f(Y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \quad Y_i = 0, 1, \quad i = 1, \dots, n$$

Given that observations are independent, the likelihood function is:

$$L(y_i) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

Then, the log of the likelihood function is:

$$\ell(y_i) = \log L(y_i) = \sum_{i=1}^n \log(\pi_i^{y_i} (1 - \pi_i)^{1-y_i})$$

Substituting in the formula for  $\pi_i = 1/(1 + \exp(-x_i^T \beta))$ :

$$\begin{aligned} \log L(y_i) &= \sum_{i=1}^n \log(\pi_i^{y_i} (1 - \pi_i)^{1-y_i}) \\ &= \sum_{i=1}^n \log \left( \frac{(\exp(x_i^T \beta))^{y_i}}{1 + \exp(x_i^T \beta)} \right) \\ &= \sum_{i=1}^n y_i x_i^T \beta - \sum_{i=1}^n \log(1 + \exp(x_i^T \beta)) \end{aligned}$$

Now, suppose we have  $r$  observed units, and  $n - r$  missing observations, then the log-likelihood function for the complete variable  $Y$  is:

$$\ell(y_i|x_i, \beta) = \sum_{i=1}^r y_i x_i^T \beta - \sum_{i=1}^r \log(1 + \exp(x_i^T \beta)) + \sum_{i=r+1}^{n-r} \hat{y}_i x_i^T \beta - \sum_{i=r+1}^{n-r} \log(1 + \exp(x_i^T \beta)) \quad (9.5)$$

Using some initial value for  $\beta$ , say  $\beta^0$ , the E-step of EM algorithm requires the computation of  $Q(\beta|\beta^0) = E[\ell(y_i|x_i, \beta^0)]$ , the expectation of the complete-data log-likelihood  $\ell(y_i|x_i, \beta)$ , as discussed in Chapter 3 and Chapter 6. The expectation is:

$$E(\ell(y_i|x_i, \beta)) = \sum_{i=1}^r y_i x_i^T \beta - \sum_{i=1}^r \log(1 + \exp(x_i^T \beta)) + \sum_{i=r+1}^{n-r} E[y_i] x_i^T \beta - \sum_{i=r+1}^{n-r} \log(1 + \exp(x_i^T \beta))$$

This means this E-step for logistic regression is performed by simply replacing each missing  $Y$  values by its expectation conditional on  $x_i$ , where:

$$E(y_i|x_i, \beta^0) = Pr(Y_i = 1|x_i, \beta^0) = \frac{1}{1 + \exp(-x_i^T \beta^0)}$$

The M-step maximizes the function in Eq.(9.5) over  $\beta$ . To find the value of  $\beta$  that maximizes the function in Eq.(9.5), people normally use iteratively re-weighted least squares (IRLS). As Kotz & Johnson (1983) described, IRLS is a numerical algorithm that maximizes any specified function using a standard weighted least squares method. We show how the exact IRLS algorithm works in the following paragraphs.

Therefore, there are two iterative loops regarding the parameters  $\beta$ , one big and one small. The big iterative loop is the EM algorithm. It is already clear to us that we need to loop through the E-step and the M-step, until the convergence of the algorithm which is attained when there is a sufficiently small difference between  $\ell(y_i|x_i, \beta^{t+1}) - \ell(y_i|x_i, \beta^t)$ .

The unfamiliar part is the small loop (IRLS) to find the estimate  $\beta$  in the M-step. We outline the iterative solution to computing the value of  $\beta$ :

**Step 1:** Choose initial estimates of the regression coefficients  $\beta$  by using the observed data only

**Step 2:** At each iteration  $t$ , update the regression coefficients:

$$\beta^{t+1} = \beta^t + (X^T V^t X)^{-1} X^T (y^t - \pi^t)$$

where

$X$  is the matrix of explanatory variables values

$y^t$  is the response variable at iteration  $t$ , the missing  $Y$  values were replaced by the expectation  $E[y_i|x_i, \beta^t]$

$\pi^t$  is the vector of fitted response probabilities at iteration  $t$

$V^t$  is a diagonal matrix, with diagonal entries  $\pi_i^t(1 - \pi_i^t)$ .

**Step 3:** Repeat step 2 until  $|\beta^{t+1} - \beta^t|$  is sufficiently close to 0

Unlike numerical variables, there is a problem in applying the EM algorithm to the logistic regression with missing response variables. The  $E(y_i|x_i, \beta^0) = Pr(Y_i = 1|x_i, \beta^0)$  is in the form of probability. What we really want is to replace the missing  $Y$  with its categorical values, not the probabilities. However, we cannot just simply use the techniques we have introduced in Section 9.3.2 to convert the probability  $\hat{\pi}_i$  into categorical values. For example, suppose the response variable has categorical values 1 or 0, and the cut off probability  $C = 0.5$  which is fixed, then we set:

$$Y_{mis,i} = \begin{cases} 1 & \hat{\pi}_i > C \\ 0 & \text{Otherwise} \end{cases}$$

After the first EM iteration, the  $Y_{mis}$ s will be the same for the following iterations, or in other word, the EM algorithm converges immediately. Let's break down the EM algorithm step by step to show how this happens.

1. Replace missing  $Y_{mis}^0$  by its expectation probability  $\hat{\pi}_i^0$  conditional on  $X$  and convert its values to either 0 or 1 according to the cut off probability  $C$ .
2. Estimate parameters  $\beta^1$
3. Re-estimate the missing  $Y_{mis}$  assuming the new parameter estimates  $\beta^1$ s are correct
4. Re-estimate parameters  $\beta$
5. ...

The problem occurs at step 3 “Re-estimate the missing  $Y_{mis}$ ”. Unlike the EM algorithm for the simple numerical regression model where the original expected values (usually the mean of observed units) will be replaced by different values, the updated  $Y_{mis}^1$  for a categorical response variable will be the same as the replaced  $Y_{mis}^0$  from step 1. This is because the  $\beta^0$  and  $X$  is used to compute the  $Y_{mis}^0$ , and regressing  $(Y_{obs}, Y_{mis}^0)$  on  $X$  gives us the  $\beta^1$  which is equal to the  $\beta^0$ . Therefore, step 3 gives us  $Y_{mis}^1 = Y_{mis}^0$ . In order to solve this problem, Anderson & Hardin (2009) propose to update the cut off probability  $C$  according to the updated  $Y$ . Hence, instead of having fixed cut off probability  $C$ , the  $C$  is defined by:

$$C = \frac{\sum_{i=1}^n Y_i}{n} \quad (9.6)$$

where  $n$  is the sample size and

$$Y_i = \begin{cases} 1 & Y_i = 1 \\ 0 & \text{Otherwise} \end{cases}$$

Once  $Y_{mis}$  is updated,  $C$  will be updated as well. Then, step 3 gives us a different  $Y_{mis}^1$  than  $Y_{mis}^0$ .

Again, we create 1000 replicate data with MAR missingness from the SURF data. The missingness for “Gender” depends on the “Qualification” with probabilities of missingness 0.2 for “None qualification”, 0.3 for “School level qualification”, 0.1 for “Vocational level qualification”, and 0.1 for “Degree level qualification”. So, the Gender is our response variable  $Y$ , and Qualification is our explanatory variable  $X$ . Then, we applied the EM algorithm to impute the missing gender values, and compute the ratio of “Female/Male” for each of the 1000 replicate data and the original SURF data. Figure 9.4 is the histogram of the ratio of “Female/Male” of the 1000 replicate data, the red vertical line represents the “Female/Male” ratio of the original SURF data. As can be seen, the imputation result is unbiased.

The R code is in Appendix E section E.2.1

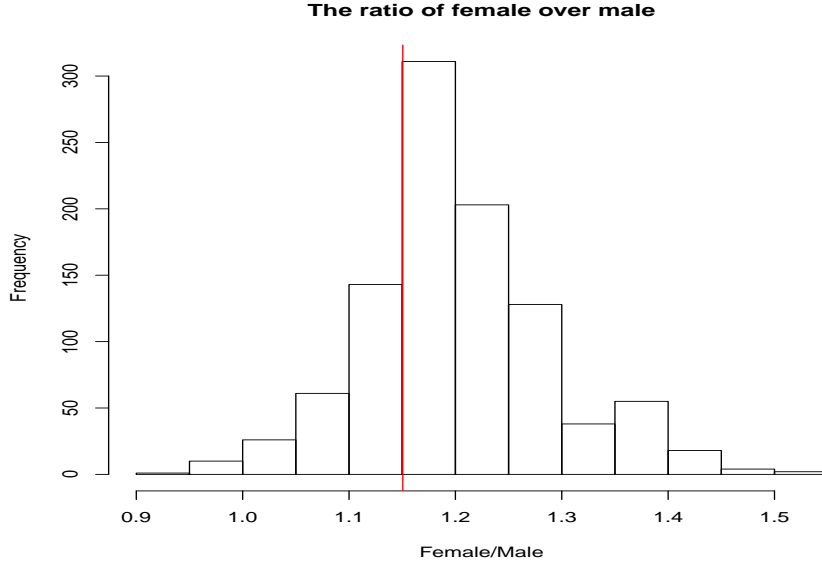


Figure 9.4: The ratio of counts of females over males. The 1000 Gender variables were imputed by EM imputation method

## 9.4.2 Bayesian iterative simulation methods for categorical variables with missing data

In Chapter 7, we have given some detailed discussion about Bayesian iterative simulation methods. In this section we apply these methods to categorical variables.

First, let's review the key ideas of Bayesian iterative simulation methods. As discussed in Chapter 7, an important step for Bayesian iterative simulation methods is to find the posterior distribution for both parameters  $\theta$  and missing data  $Y_{mis}$ . This is given by Eq.(7.1):

$$p(\theta, Y_{mis} | Y_{obs}) \propto p(\theta) f(Y_{mis} | \theta) f(Y_{obs} | \theta)$$

Then, a rough description is: randomly draw  $Y_{mis}$  from its conditional distribution at iteration  $t$ :

$$Y_{mis} \sim p(Y_{mis} | \theta^t)$$

At iteration  $t + 1$ , we randomly draw  $\theta^{t+1}$  from its conditional posterior distribution given the updated  $Y_{mis}^t$  and observed  $Y_{obs}$  from the previous iteration.

$$\theta^{t+1} \sim p(\theta | Y_{mis}^t, Y_{obs})$$

Now, let's consider a simple categorical variable with missing data. Suppose the response variable  $Y$  in a regression model is dichotomous (0, 1). This means we have a binary logistic regression. There are  $n$  observations,  $r$  observations were observed for  $Y$ , and  $n - r$  were missing. We still use  $Y_{obs}$  to denote observed data, and  $Y_{mis}$  for missing data. We also assume all the explanatory variables  $X$  are observed. The likelihood function for  $Y_{obs}$  and  $Y_{mis}$  can be



expressed as:

$$f(Y_{obs}|\beta) = \prod_{i=1}^r \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

$$f(Y_{mis}|\beta) = \prod_{i=r+1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

where

$$\pi_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} = \frac{1}{1 + \exp(-x_i^T \beta)}$$

and  $\beta$  are the logistic regression coefficients.

For simplicity reasons, we choose uniform prior as our prior distribution, then we have the improper prior  $p(\beta) \propto 1$ . Hence, according to Eq. (7.1), the full posterior distribution is:

$$\begin{aligned} p(\beta, Y_{mis}|Y_{obs}) &\propto p(\beta) f(Y_{mis}|\beta) f(Y_{obs}|\beta) \\ &\propto \prod_{i=1}^r \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \prod_{i=r+1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \prod_{i=1}^n \left( \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right)^{y_i} \left( 1 - \left( \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right) \right)^{1-y_i} \end{aligned} \quad (9.7)$$

The distribution of  $Y_{mis}$  conditional on  $\beta$  is from Bernoulli distribution. Hence:

$$Y_{mis}|\beta \sim \text{Bernoulli}\left(1, \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right)$$

However, it is not easy to draw  $\beta$  from the posterior distribution given in Eq (9.7).

### Sampling $\beta$ from unfamiliar distribution

The posterior distribution of  $\beta$  is complicated and is not a simple member of a well known family of exponential distributions. As discussed in Chapter 7, the Metropolis-Hastings (MH) algorithm can be applied in the situation of such unknown distributions. However, Groenewald & Mokgatle (2005) proposed a method for the simulation of samples from the exact posterior distributions of the parameters in logistic regression. This means we can use Gibbs sampler to sample  $\beta$  even if the distribution is unknown to us. In this section, we will introduce the use of MH algorithm first, then discuss the Groenewald and Mokgatle's method.

Metropolis-Hastings algorithm for  $\beta$  simulation.

### Sample $\beta$ by using the MH algorithm

- Step 1:** Choose initial estimates of the regression coefficients  $\beta$ ,  $\beta^* = \beta^0$ . This could be extracted from the logistic model based on the  $r$  observed data
- Step 2:** Select a proposal distribution  $q(\beta^*|\beta^t)$ . We propose that  $\beta$  come from a normal distribution  $N(\beta^t, \Sigma_\beta)$ .
- Step 3:** Randomly draw  $\beta$  from the proposal distribution. That is,  $\beta^{t+1} \sim N(\beta^t, \Sigma_\beta)$ .  $\Sigma_\beta$  can be kept constant during the process
- Step 4:** The acceptance ratio  $r$  is calculated:

$$r = \frac{f(Y|\beta^{t+1})q(\beta^{t+1}|\beta^t)}{f(Y|\beta^t)q(\beta^t|\beta^{t+1})}$$

- Step 5:** Generate  $u$  from  $U \sim \text{Uniform}(0, 1)$

- Step 6:**

$$\beta^{t+1} = \begin{cases} \beta^* & \text{if } u \leq \min(1, r) \\ \beta^t & \text{Otherwise} \end{cases}$$

Now, let's combine the step of drawing  $Y_{mis}$  and the step of drawing  $\beta$  to form the complete Bayesian iterative simulation for the missing data, which can be also referred as the Data Augmentation (DA) algorithm we have introduced in previous chapters. Again, we applied the DA algorithm to the 500 replicate SURF data with MAR missingness. The missingness structure is the same as the description given in Section 9.4.1, the Gender is the response variable  $Y$ , and Qualification is the explanatory variable  $X$ , for the logistic regression model we use. Figure 9.5 displays the results of the ratio of counts of females to males for the 500 replicate SURF data. The red vertical line represents the true ratio of counts of females to males. As shown, the DA imputation result is unbiased.

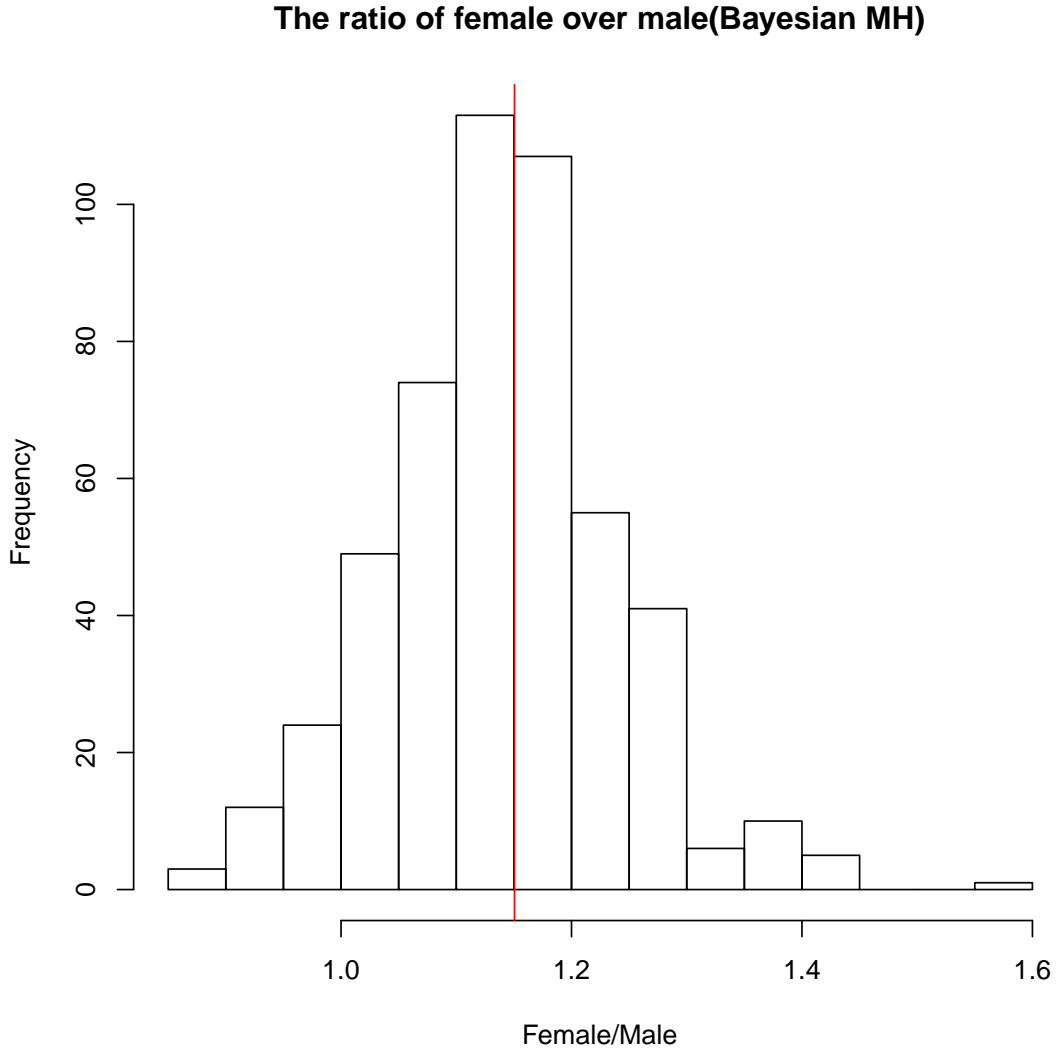


Figure 9.5: Applying Data Augmentation(DA) to impute missing values for the SURF Gender variables by using the MH algorithm. The DA is applied to 500 replicate SURF data

Let's return to the discussion of sampling  $\beta$  from an unfamiliar distribution. The Groenewald & Mokgatle (2005) method can be used for dichotomous (or binary) response variables, polychotomous response variables and ordinal responses. We only introduce its application for dichotomous response variables because this is the only scenario we need to consider within the scope of this thesis. Suppose we construct a logistic regression based on a data matrix which has  $n$  observations. The response variable  $Y$  is a binary categorical variable:

$$Y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i \end{cases}$$

and

$$\log \frac{\pi_i}{1 - \pi_i} = x_i^T \beta$$

where  $i = 1, 2, \dots, n$ ,  $X$  is the matrix of the explanatory variables, and  $\beta$  is a vector of regression coefficients. Then, assume:

$$\text{logit}(\pi_i) = x_i^T \beta \Rightarrow \pi_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} = F_Z(x_i^T \beta)$$

where  $F_Z(z)$  is the Cumulative Density Function (CDF) of random variable  $Z$ . Then, its probability density function (pdf) is:

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_Z(z) \\ &= \frac{e^z(1+e^z) - (e^z)^2}{(1+e^z)^2} \\ &= \frac{e^z}{(1+e^z)^2} \end{aligned}$$

where  $e^z = \exp(x_i^T \beta)$ . So

$$\begin{aligned} \pi_i &= F_Z(x_i^T \beta) \\ &= \int_{-\infty}^{x_i^T \beta} \frac{dF_Z}{dz} dz \\ &= \int_{-\infty}^{x_i^T \beta} f_Z(z) dz \\ &= \int_0^{F_Z(u)} du, \quad u = F_Z(z), \quad 0 \leq u \leq 1 \\ &= \int_0^{\frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}} du \\ &= P\left(U < \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}\right) \\ &= \int_0^1 I\left(u < \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}\right) du \end{aligned}$$

where  $U$  is a Uniform  $(0, 1)$  distribution.

Now, we introduce an independent uniformly distributed latent variable  $u = (u_1, u_2, \dots, u_n)$ . Then, the pdf is:

$$p(u_i) = I(0 \leq u_i \leq 1)$$

Hence,  $Y$ , where  $y_i = (y_1, y_2, \dots, y_n)$  given that  $u$  can be expressed as:

$$\begin{aligned} p(Y_i = 1|u_i) &= \begin{cases} 1 & u_i \leq F_Z(x_i^T \beta) \\ 0 & u_i > F_Z(x_i^T \beta) \end{cases} \\ p(Y_i = 0|u_i) &= \begin{cases} 1 & u_i > F_Z(x_i^T \beta) \\ 0 & u_i \leq F_Z(x_i^T \beta) \end{cases} \end{aligned}$$

Then

$$\begin{aligned} p(Y_i = y_i|u_i) &= y_i I(u_i \leq F_Z(x_i^T \beta)) + (1 - y_i) I(u_i > F_Z(x_i^T \beta)) \\ &= I(y_i = 1) I(u_i \leq F_Z(x_i^T \beta)) + I(y_i = 0) I(u_i > F_Z(x_i^T \beta)) \end{aligned}$$

Hence, the joint probability density function of  $Y$  and  $u$ , given  $x$  and  $\beta$  is:

$$\begin{aligned} p(y_i, u_i|x_i^T \beta) &= p(y_i|u_i)p(u_i) \\ &= [(y_i = 1) I(u_i \leq F_Z(x_i^T \beta)) + I(y_i = 0) I(u_i > F_Z(x_i^T \beta))] I(0 \leq u_i \leq 1) \end{aligned}$$

Then, the likelihood of  $p(y_i, u_i | x_i^T \beta)$  is:

$$p(y, u | \beta) = \prod_{i=1}^n [(y_i = 1)I(u_i \leq F_Z(x_i^T \beta)) + I(y_i = 0)I(u_i > F_Z(x_i^T \beta))] I(0 \leq u_i \leq 1)$$

The Bayesian theory in Chapter 3 section 3.3.3 tells us that the probability of  $\beta$  given  $Y$  and  $u$  can be attained by:

$$p(\beta | y, u) \propto p(\beta) p(y, u | \beta) \quad (9.8)$$

$$\begin{aligned} &\propto p(\beta) \prod_{i=1}^n \left[ I\left(u_i \leq \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right) I(y_i = 1) + I\left(u_i > \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right) I(y_i = 0) \right] \\ &\quad \times I(0 \leq u_i \leq 1) \end{aligned} \quad (9.9)$$

$$\begin{aligned} &\propto p(\beta) \prod_{i=1}^n \left[ I\left(x_i^T \beta \geq \log\left(\frac{u_i}{1 - u_i}\right)\right) I(y_i = 1) + I\left(x_i^T \beta < \log\left(\frac{u_i}{1 - u_i}\right)\right) I(y_i = 0) \right] \\ &\quad \times I(0 \leq u_i \leq 1) \end{aligned} \quad (9.10)$$

where  $p(\beta)$  is the prior probability of  $\beta$  and  $I(X \in A)$  is the indicator function that is equal to 1 if  $X \in A$ , and 0 otherwise.

According to Equation (9.9),  $u_i$  is a uniform distribution, given  $\beta$  and  $y$ .

$$u_i | \beta, y_i \sim \begin{cases} \text{Uniform}\left(0, \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right) & \text{if } y_i = 1, \\ \text{Uniform}\left(\frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}, 1\right) & \text{if } y_i = 0, \end{cases} \quad i = 1, 2, \dots, n \quad (9.11)$$

This is because if  $y_i = 1$ , then  $I(y_i = 0) = 0$ . This makes  $I\left(u_i > \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right) I(y_i = 0) = 0$ , and  $I\left(u_i \leq \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right) I(y_i = 1)$  becomes  $I\left(u_i \leq \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right)$  which tells us that  $u_i$  needs to be smaller than or equal to  $\frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$ , given  $y_i = 1$ . Then, we have  $I\left(u_i \leq \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right) \times I(0 \leq u_i \leq 1)$ , where  $I(0 \leq u_i \leq 1)$  defines the range of  $u_i$  is between 0 and 1. Hence, if  $y_i = 1$ , then  $u_i | \beta, y \sim \text{Uniform}\left(0, \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right)$ . The same logic applies to the case where  $y_i = 0$ .

Suppose there are  $p$  explanatory variables,  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ , then  $x_i^T = (x_{i0}, x_{i1}, \dots, x_{ip})^T$ , from Equation (9.10) we have that

$$\begin{aligned} x_i^T \beta &= \sum_{j=0}^p x_{ij} \beta_j \geq \log \frac{u_i}{1 - u_i}, \quad \text{if } y_i = 1, \\ x_i^T \beta &= \sum_{j=0}^p x_{ij} \beta_j < \log \frac{u_i}{1 - u_i}, \quad \text{if } y_i = 0, \end{aligned}$$

Hence

$$\begin{aligned} \beta_k &\geq \frac{1}{x_{ik}} \left( \log \frac{u_i}{1 - u_i} - \sum_{j \neq k}^p x_{ij} \beta_j \right), \quad \text{if } y_i = 1, \\ \beta_k &< \frac{1}{x_{ik}} \left( \log \frac{u_i}{1 - u_i} - \sum_{j \neq k}^p x_{ij} \beta_j \right), \quad \text{if } y_i = 0, \end{aligned}$$

for all  $i$ , assuming  $x_{ik} \neq 0$ . Let  $A_k$  and  $B_k$  be the sets:

$$\begin{aligned} A_k &= (i : ((y_i = 1) \cap (x_{ik} > 0)) \cup ((y_i = 0) \cap (x_{ik} < 0))), \\ B_k &= (i : ((y_i = 0) \cap (x_{ik} > 0)) \cup ((y_i = 1) \cap (x_{ik} < 0))), \end{aligned}$$

Then, assuming a Jeffreys prior (Jefferys 1939),  $p(\beta) \propto 1$ , for  $\beta$ , the conditional distribution of  $\beta_k$ , given all the other  $\beta$ 's and  $u$ , is the uniform distribution:

$$\beta_k | \beta_{(-k)}, u, y \sim \text{Uniform}(a_k, b_k), k = 0, 1, 2, \dots, p. \quad (9.12)$$

where

$$a_k = \max_{i \in A_k} \left[ \frac{1}{x_{ik}} \left( \log \frac{u_i}{1 - u_i} - \sum_{j \neq k}^p x_{ij}^T \beta \right) \right] \quad (9.13)$$

and

$$b_k = \min_{i \in B_k} \left[ \frac{1}{x_{ik}} \left( \log \frac{u_i}{1 - u_i} - \sum_{j \neq k}^p x_{ij}^T \beta \right) \right] \quad (9.14)$$

If the  $x_{ij}$  is a categorical dummy variable, which means the  $x_{ij}$  has values 1 and 0 only, then we can remove the fraction  $\frac{1}{x_{ik}}$  from the computation of Equation (9.13) and Equation (9.14).

Now, the Gibbs sampler can be applied by drawing from uniform distributions. However, the limitation of Groenewald & Mokgatle (2005) method is that we have to make sure that the explanatory variables  $X$  contain no zero values, otherwise the algorithm will not work. This shortcoming limits the use of this method to be applied to categorical variables. We can use this method for a categorical response variable, and numerical explanatory variables, but not if the explanatory variables are also categorical. This is because when forming the regression model with categorical explanatory variables, we need to create dummy variables  $V_{dummy}$ . If a categorical explanatory variable has  $K = (1, \dots, k)$  levels, then there will be  $k$  dummy variables with each variable corresponding to each of the  $k$  levels, where  $V_{dummy} = (V_{dummy,1}, \dots, V_{dummy,k})$ . Then, a unit with the categorical variable equal to level 1, will have the  $V_{dummy,1} = 1$ , but other dummy variables will have value equal to 0. This violates the condition that the explanatory variables  $X$  contain no zero values.

## 9.5 Conclusion

In this chapter, we have demonstrated in detail of how to apply various imputation methods to categorical data. Although the fundamental concepts are the same when applying these imputation methods to categorical data, the exact procedures are somewhat different and we think it is worth spending time exploring them before applying these methods to the FNES data.

In the following chapters, we will try to apply all the imputation methods which have been introduced so far to the FNES data.

# **Chapter 10**

## **Introduction to the Food Nutrition Environment Survey (FNES)**

### **10.1 Purpose**

In this chapter, we introduce the background of the Food Nutrition Environment Survey (FNES), then we impute its missing survey data in the following chapter. The purpose of doing this is to reduce non-response bias and increase the utility of data with missing values.

### **10.2 Survey background**

The FNES is a survey of early childhood centres and schools and the food and nutritional services that they provide for their pupils. The 2007 and 2009 FNES surveys were managed by the Ministry of Health. The FNES aimed to collect information on key baseline indicators, follow-up indicators and experiences of key stakeholders in relation to the implementation of Healthy Eating Healthy Action (HEHA) and Mission-On initiatives within school and Early Childhood Education (ECE) services. In other words, it aimed to collect the food and nutrition environment within schools and ECE services in New Zealand.

The survey results will be used to describe the food and nutrition environment in schools and ECE services, contribute to the food and nutrition policies and provide such information to other research.

### **10.3 Periods**

The research reported in this project uses data from both the 2007 and 2009 FNES surveys.

### **10.4 Target Population**

The target population has two parts: school and ECE. The school target population for the 2007 and 2009 FNES surveys were all primary and secondary schools in New Zealand. Excluded were the Correspondence School, Teen Parent Units, hospital-based schools or health camps. The ECE services target population was all licensed and/or chartered ECE services. Excluded were licence-exempt ECE services such as Playgroups, Ngā Puna Kōhungahunga

and Pacific Island ECE Groups, some Play centres and licence-exempt Kōhanga Reo. Hospital-based ECE services and the Correspondence ECE services were also excluded from the target population.

## 10.5 Survey Population

The survey population was the same as the target population with some further exclusions. Schools and ECE services on outlying islands of New Zealand were excluded from the survey frame. Other ECE services excluded were mobile kindergartens. These exclusions were for practical reasons, e.g. too expensive and difficult to sample remote areas.

## 10.6 Sample Frame

The sample frame for both surveys was constructed from the Ministry of Education's (MoE) directories of schools and ECE services available from MoE's website. There were 3778 ECE services, 2082 primary schools and 481 secondary schools in the sample frame in 2007, and 4103 ECE services, 2065 primary schools and 485 secondary schools in the sample frame in 2009.

For the research purpose of this paper, we exclude Te Kōhanga Reo ECE services from our FNES sample. This is because the Khanga Reo ECE services have actually been excluded from the sample frame in 2007 and 2009 (Pledger et al. 2010).

The reduced sample frame for each stratum for both FNES survey is displayed in Table 10.1.

The 2007 and 2009 FNES surveys have used essentially the same sample frame. Hence, both surveys have the potential to select the same sample units and there was no overlap control with the two FNES surveys.

Table 10.1: Sample frame for both 2007 and 2009

Type	Stratum	2007 sample frame	2009 sample frame	Matched sample frame
ECE	Education and Care Centres (ECE1)	1961	2230	1907
	Kindergartens (ECE2)	613	623	619
	Home-based childcare (ECE3)	237	309	226
	Playcentre (ECE4)	473	462	458
	<b>TOTAL</b>	3284	3651	3210
Schools	Primary	2082	2065	2053
	Secondary	481	485	481
	<b>TOTAL</b>	2563	2550	2534



## 10.7 Sample Size

Table 10.2 shows the selected sample sizes of 2007 and 2009. The selected sample size of 2007 has 2308 selected sample units and the selected sample size of 2009 is 2312. Of those sample units in 2007, 827 are ECE services, 1000 are primary schools, and 481 are secondary schools. Of those sample units in 2009, 827 are ECE services, 1000 are primary schools, and 485 are secondary schools. In addition, 218 ECE services and 950 schools are selected in both 2007 and 2009 FNES surveys.

The final sample size (Table 10.3) of the 2007 survey has 1307 respondents and there are 1774 respondents in the 2009 survey. Of those respondents in 2007, 562 are ECE services, 518 are primary schools and 277 are secondary schools. Of those respondents in 2009, 637 are ECE services, 783 are primary schools and 354 are secondary schools. Furthermore, 109 ECE services and 373 schools responded both the 2007 and 2009 FNES surveys.

Table 10.2: Selected sample size for both 2007 and 2009

Type	Stratum	2007 selected sample size	2009 selected sample size	2007&2009 matched sample size
ECE	ECE1	345	345	54
	ECE2	193	193	59
	ECE3	120	120	44
	ECE4	169	169	61
	<b>TOTAL</b>	827	827	218
Schools	Primary	1000	1000	478
	Secondary	481	485	472
	<b>TOTAL</b>	1481	1485	950

Table 10.3: Final sample size for both 2007 and 2009

Type	Stratum	2007 final sample size	2009 final sample size	2007&2009 matched final sample size
ECE	ECE1	275	281	35
	ECE2	156	170	39
	ECE3	40	66	7
	ECE4	91	120	28
	<b>TOTAL</b>	562	637	109
Schools	Primary	518	783	213
	Secondary	227	354	160
	<b>TOTAL</b>	745	1137	373

## 10.8 Matching process

The matched sample frame, the matched selected sample and the matched actual sample were produced by using the 2007 and 2009 FNES sample frames and selected and actual sample information. We matched the sample frames and samples by using the unique school/ECE identifier in both FNES datasets. Only the units with the same identifiers in both 2007 and 2009 FNES surveys could be matched. Figure 10.1 to Figure 10.4, and Table 10.4 to Table 10.7 display the matched results for the selected samples and responding samples.

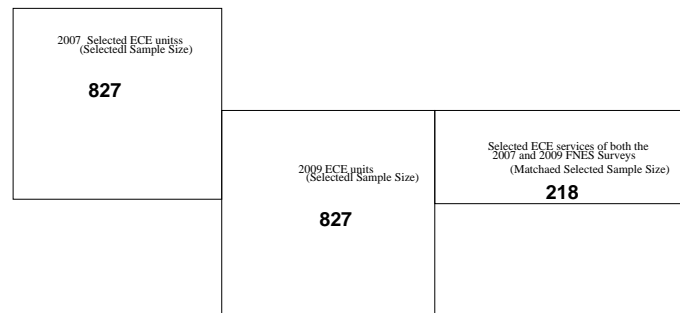


Figure 10.1: Selected sample size of ECEs in 2007, 2009 and size of ECEs in both the 2007 and 2009 FNES

Table 10.4: Selected sample size of ECEs in 2007, 2009 and size of ECEs in both the 2007 and 2009 FNES

	Unit not in 09	Unit in 09	Total
Unit not in 07	0	609	609
Unit in 07	782	218	1000
Total	782	827	1609

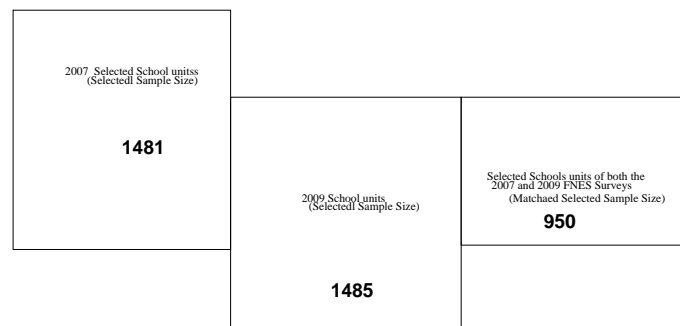


Figure 10.2: Selected sample size of schools in 2007, 2009 and size of schools in both 2007 and 2009 FNES

Table 10.5: Table of selected sample size of schools in 2007 and 2009 and size of schools in both 2007 and 2009 FNES

	Unit not in 09	Unit in 09	Total
Unit not in 07	0	535	535
Unit in 07	531	950	1481
Total	531	1485	2016

Figure 10.3, Figure 10.4, Table 10.6 and Table 10.7 display the counts of unit response units. Although there was unit non-response, in this thesis we ignore this and concentrate on imputation for item non-response. Question level response rates are given in Chapter 11.

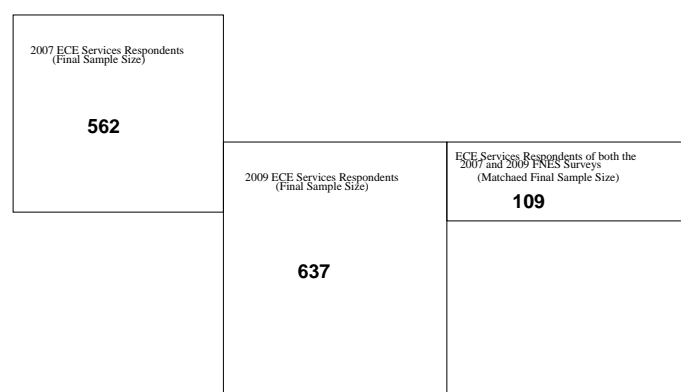


Figure 10.3: Responding sample size of ECEs in 2007, 2009 and size of ECEs that responded to both the 2007 and 2009 FNES

Table 10.6: Responding sample size of ECEs in 2007, 2009 and size of ECEs that responded to both the 2007 and 2009 FNES

	Unit not in 09	Unit in 09	Total
Unit not in 07	0	528	528
Unit in 07	451	109	560
Total	451	637	1088

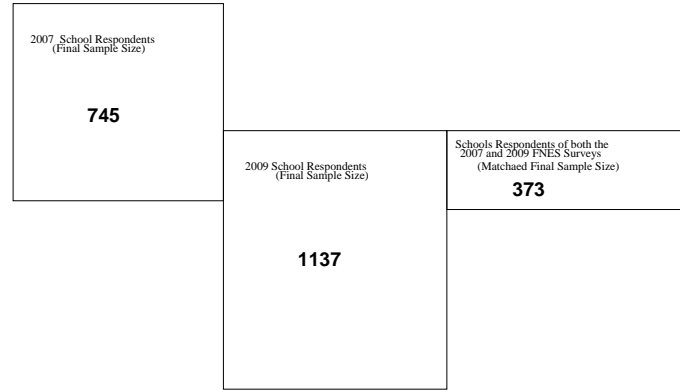


Figure 10.4: Responding sample size of schools in 2007, 2009 and size of schools that responded to both the 2007 and 2009 FNES

Table 10.7: Responding sample size of schools in 2007, 2009 and size of schools that responded to both the 2007 and 2009 FNES

	Unit not in 09	Unit in 09	Total
Unit not in 07	0	764	764
Unit in 07	370	373	743
Total	370	1137	1507

## 10.9 Sample Weights

Suppose there is a well mixed glass of honey water, a spoon of that honey water should have the same honey and water ratio as the rest of the honey water in the glass. Hence, we can measure how much honey is in the water by investigating that spoon of honey water instead of the whole glass, given we know that the glass can pour  $w = 20$  spoons. This is a simple example of sampling and weighting. The glass of honey water is our interest population, the spoon of honey water is our sample and the number of spoons can be poured is our sample weight  $w$ . Of course, the FNES samples are far more complex than a glass of honey water, but the underlying assumption is still the same. We draw samples of ECE centres and schools from their sample frame, and assume that our sample contains all the characteristics the population does. Then, we estimate the population characteristics by scaling up the sample to the population. If the sample size is  $n$ , then each unit  $i$  has a weight  $w_i$ , representing that there are  $w_i$  such units in the population of size  $N$ .

$$N = \sum_{i=1}^n w_i$$

Hence, if we want to measure the population total of  $Y$ , where  $Y$  is a variable, then

$$Total_Y = \sum_{i=1}^n w_i y_i$$

where  $y_i$  is the sampled value for unit  $i$ .

For a simple random sample of size  $n$ , the weight for each sample unit is the same  $w_i = N/n$ . The FNES is a stratified sample. Hence, the weights are different for units in different strata. Let  $K = 1, \dots, k$  be the stratum number, then for unit  $i$  of stratum  $k$ , the weight is  $w_{ik} = N_k/n_k$ .

Table 10.8 and Table 10.9 show the selected sample weights and matched selected sample weights for both 2007 and 2009 FNES surveys.

Table 10.8: Sample selection weights for both 2007 and 2009

Type	Stratum	2007 selected sample wgt	2009 selected sample wgt
ECE	ECE1	5.684057971	6.463768116
	ECE2	3.176165803	3.227979275
	ECE3	1.975	2.575
	ECE4	2.798816568	2.733727811
Schools	Primary	2.082	2.065
	Secondary	1	1

Table 10.9: Matched Sample selection weights for both 2007 and 2009

Type	Stratum	2007 matched sample wgt	2009 matched sample wgt
ECE	ECE1	36.31481481	41.2962963
	ECE2	10.38983051	10.55932203
	ECE3	5.386363636	7.022727273
	ECE4	7.754098361	7.573770492
Schools	Primary	4.355648536	4.320083682
	Secondary	1.019067797	1.027542373

As discussed in Section 3.2.3, Chapter 3, “Reweighting” is a method for dealing with missing data. Some researchers classify it as one of the data deletion methods. As with other data deletion methods, reweighting deletes collected information if the missing data is item non-response. Please refer to Chapter 3 for detailed discussions. However, reweighting can be considered as one of imputation methods when the missing data is unit non-response. As described in Section 3.2.3, Chapter 3, reweighting increases  $w_i$  to count the number of non-response, so the new weight  $\tilde{w}_i$  represents the unobserved population units and the nonrespondents. Suppose there are  $r_k$  response units in stratum  $k$ , the new weight  $\tilde{w}_{ik}$  for the FNES sample can be computed as:

$$\tilde{w}_{ik} = N_k / r_k$$

Table 10.10 and Table 10.11 show the responding sample weights (or final sample weight) and matched responding sample weights for both 2007 and 2009 surveys

Table 10.10: Responding Sample weights for both 2007 and 2009

Type	Stratum	2007 final sample wgt	2009 final sample wgt
ECE	ECE1	7.130909091	7.935943
	ECE2	3.929487179	3.664706
	ECE3	5.925	4.681818
	ECE4	5.197802198	3.85
Schools	Primary	4.019305019	2.637292
	Secondary	2.118942731	1.370056

Table 10.11: Matched responding Sample weights for both 2007 and 2009

Type	Stratum	2007 matched FinalWgt	2009 matched FinalWgt
ECE	ECE1	56.02857143	63.71429
	ECE2	15.71794872	15.97436
	ECE3	33.85714286	44.14286
	ECE4	16.89285714	16.5
Schools	Primary	9.774647887	9.694836
	Secondary	3.00625	3.03125

## 10.10 Sample Design

The overall sample design for both FNES surveys was a stratified design using seven strata consisting of five categories of ECE service types, plus a stratum for primary schools and a full-coverage stratum for secondary schools. The systematic random sample method was used for all non full-coverage strata. The five categories of ECE service types are shown in table 10.12. However, prior to fieldwork in 2007 it was decided to have a separate data collection method for the Kōhanga Reo ECE services in the sample frame. For 2009, Kōhanga Reo were again excluded from this sample. Hence, there were actually only six strata.

Table 10.12: The five categories of ECE service types

ECE Stratum	ECE service type
ECE1	Education and Care Centres
ECE2	Free Kindergarten
ECE3	Home-based childcare
ECE4	Playcentre
ECE5	Te Kōhanga Reo

## 10.11 Questionnaire

Most of the questions in both FNES surveys were closed (i.e.. multiple choice with tick box options). There are only a few open-ended questions. Respondents need to self-complete their questionnaires. Two forms of the questionnaire were used: one was hard copy and the other was on-line-based. The respondents had the choice of completing either a hard copy or an online questionnaire. To avoid duplicate responses, a unique code is used for the hard copy questionnaire and a unique identifier is also used for the online-based questionnaire.

# Chapter 11

## Imputation of FNES missing data

### 11.1 Exploratory Data Analysis (EDA)

We have introduced the general survey background of the FNES in the previous chapter. In this chapter, we impute some of its missing data. The first step of imputation is to understand the data set. Therefore, this section focuses on describing the basic characteristics of the FNES data. This information come in handy when we need to pick the most appropriate imputation methods for the FNES missing data. There are four FNES sample datasets: the FNES ECE 2007, the FNES ECE 2009, the FNES School 2007, and the FNES School 2009. Each of these datasets contains: identification variables, question variables and design variables. Please refer to Table 11.3 and Table 11.4 for detailed information.

As introduced in Chapter 2, there are two types of missing data: unit non-response and item non-response. Let's start with the description of the unit non-response case. Table 11.1 and Table 11.2 display the sample frame, sample size, responded sample size, and response rate for the 2007 and 2009 FNES. As indicated in the Chapter 10, we do not impute unit non-response in this thesis, but it has been dealt with by using the reweighting method.

Table 11.1: The sample frame, sample size, responded sample size, and response rate of the FNES 2007

Type	Stratum	Sample Frame 07	Sample Size 07	Actual Sample 07	Response rate 07
ECE	ECE1	1961	345	275	80%
	ECE2	613	193	156	81%
	ECE3	237	120	40	33%
	ECE4	473	169	91	54%
Total		3284	827	562	68%
School	Primary Schools	2082	1000	518	52%
	Secondary Schools	481	481	227	47%
Total		2563	1481	745	50%

Table 11.2: The sample frame, sample size, responded sample size, and response rate of the FNES 2009

Type	Stratum	Sample Frame 09	Sample Size 09	Actual Sample 09	Response rate 09
ECE	ECE1	2230	345	281	81%
	ECE2	623	193	170	88%
	ECE3	309	120	66	55%
	ECE4	462	169	120	71%
Total		3624	827	637	77%
School	Primary Schools	2065	1000	783	78%
	Secondary Schools	485	485	354	73%
Total		2550	1485	1137	76.5%

Table 11.3: Design and Sample variables for 2007 and 2009

2007 ECE sample variables	2009 ECE sample variables	2007 school sample variables	2009 school sample variables
Stratum	Stratum	Stratum	stratum
Institution Number	Institution Number	Institution Number	School Number
Suburb	Suburb	Suburb	Suburb
City	City	City	City
Institution Type	Institution Type	School Type	School Type
Definition	Definition	Definition	Definition
Authority	Authority	Authority	Authority
Group			
	Hours ECE	Gender of Students	Gender of Students
Territorial Local Authority	Territorial Local Authority	Territorial Local Authority	Territorial Local Authority
Regional Council	Regional Council	Regional Council	Regional Council
General Electorate	General Electorate	General Electorate	General Electorate
Roll as at July 2006	Roll as at July 2008	School Roll July 2006	School Roll July 2008
Area Type		Area Type	
Urban Rural Zone		Urban Rural Zone	
Maori roll		Maori roll	
Pasifika roll		Pasifika roll	
mao pac roll			
random num		random num	
		Decile	Decile 2009
Operating Structure			
roll	roll	roll	roll
SelectionProb	SelectionProb	SelectionProb	SelectionProb
Sampling Weight	SamplingWeight	SamplingWeight	SamplingWeight

Table 11.4: Collected Sample variables for 2007 and 2009

2007 ECE collected variables	2009 ECE collected variables	2007 school collected variables	2009 school collected variables
Int ID	Int ID	Int id 3	Int ID
Q1-Q36	Q1-Q29	Q1-Q51	Q1-Q41
		Pword	
Format		Format	
Institution Number	Institution Number	Int id 3	School Number
Stratum	Stratum	Stratum 1	Stratum
City		City 1	
Institution Type	Institution Type	School Type 1	School Type
Definition	Definition	Definition 1	Definition
Authority	Authority	Authority 1	Authority
Group			
		Gender of Students 1	Gender of Students
	Hours ECE		
Territorial Local Authority	Territorial Local Authority	Territorial Local Authority 1	Territorial Local Authority
Regional Council	Regional Council	Regional Council 1	Regional Council
			Ministry of Education Local Off
General Electorate	General Electorate	General Electorate 1	General Electorate
		Decile 1	Decile 2009
Roll as at July 2006	Roll As At July 2008	School Roll July 2006 1	School Roll July 2008
Maori roll		maori roll 1	
Pasifika roll		pasifika roll 1	
Area Type		Area Type 1	
Urban Rural Zone		Urban Rural Zone 1	
City Town		City Town 1	
Operating Structure			
random num		random num 1	
mao pac			
roll	roll	roll 1	roll
SelectionProb	SelectionProb		SelectionProb
Sampling Weight	SamplingWeight		SamplingWeight
stratum1			
X TYPE		X TYPE	
samsize		samsize	
pop		pop	
finalwgt		finalwgt	

Now, we want to look at what the missing data look like in terms of item non-response. FNES datasets only have item non-response for their question variables. Any missing design variables can be found by matching the units to the sample frame which have information for



those variables as well. The FNES has two kinds of questions: closed question and open-ended question. All closed questions are categorical variables. All the open-ended questions are free text responses. We only consider performing imputation for the closed questions.

In addition, some of the FNES questions are interrelated. Hence, if one question has been answered, then other related questions do not need to be answered. This causes structural missingness. For example, Question “151” of the 2009 FNES school questionnaire asks: “Places where ‘Individually wrapped branded ice creams’ can be purchased on school grounds”, the options/interrelated questions are from “151.a” to “151.h”. Option 151.a is “School did not sell this ice cream”, and options 151.b to 151.h list the name of possible places. Hence, if the response to 151.a was “Yes”, then the respondent would not answer other options/interrelated questions. This is called structural missingness. However, the design of FNES questionnaires has required respondents to still indicate that they do not need to answer those interrelated questions by selecting “99=NA (Not Applicable)”. Therefore, we think the structural missingness may have already been dealt with by the survey. Table 11.5 shows the structure of Q151.

Table 11.5: Example of Q151

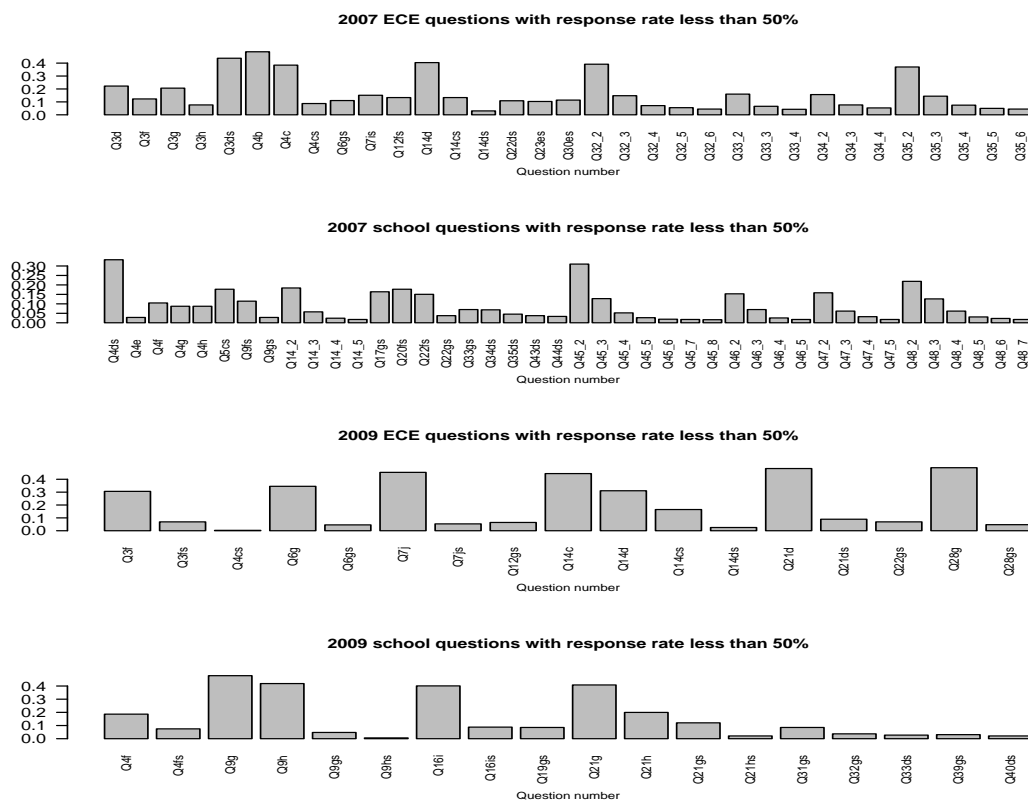
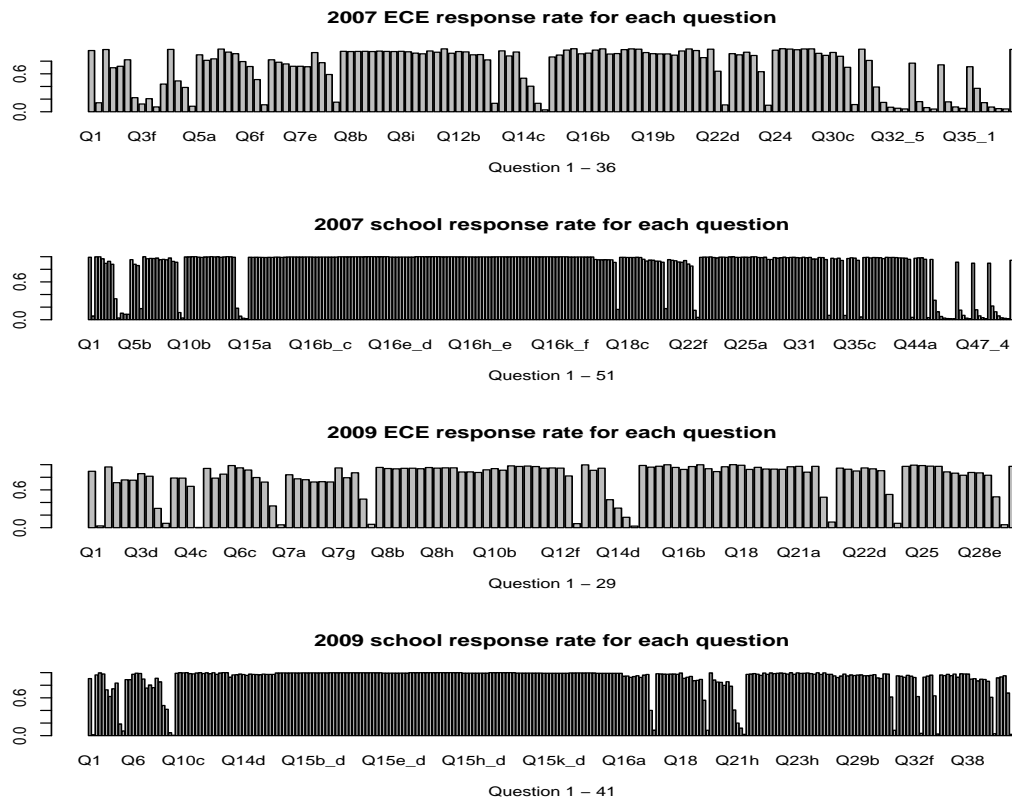
	Q151 Places where “Individually wrapped branded ice creams” can be purchased on school grounds
Q151.a	“School did not sell this” 1=Yes, 2=No, . =missing, 99=NA
Q151.b	“Canteen/tuck shop” 1=Yes, 2=No, . =missing, 99=NA
Q151.c	“Cafeteria” 1=Yes, 2=No, . =missing, 99=NA
Q151.d	“Vending machine” 1=Yes, 2=No, . =missing, 99=NA
Q151.e	“Order in System” 1=Yes, 2=No, . =missing, 99=NA
Q151.f	“Fund raising” 1=Yes, 2=No, . =missing, 99=NA
Q151.g	“Other place” 1=Yes, 2=No, . =missing, 99=NA
Q151.h	“Don’t know” 1=Yes, 2=No, . =missing, 99=NA

Note: NA=Not Applicable

Figure 11.1 shows the overall response rates for the 2007 and 2009 FNES questions. Easy to be noticed, there are a few questions with very low response rates (ie. less than 50%). We suspected that most of those questions with low response rates were due to being either open-ended questions or questions with structural missingness. In Figure 11.2, we displayed the questions with less than 50% response rate. In Table 11.6, we counted the number of open-ended questions and closed questions for the low response questions. As shown, a large number of low response questions are open-ended questions. This is expected as Andrews (2004) points out that open-ended questions have traditionally low response rates. In terms of how many low response questions were due to structural missingness, we did not investigate further. The reasons are: (1) we do not impute variables with less than 50% response rates; (2) how to deal with structural missingness will be discussed later.

Table 11.6: Composite of low response (< 50%) questions

Survey	Number of Open-ended questions	Number of Closed questions	Total number of low response questions
The 2007 ECE questions	12	21	33
The 2007 School questions	14	28	42
The 2009 ECE questions	11	6	17
The 2009 School questions	13	5	18



Then, we separate those questions with response rates between 50% and 90% in order to identify the variables which are the most imputable. The choice of the upper limit is because a question with high response rate (90%) is not worth to be used for demonstrating our imputation methods; the choice of the lower limit is because most researchers recommend not using data with more than 50% missing counts (Statistics Netherlands 2012). Figure 11.3 shows the barplots of the question variables with response rates between 50% and 90% for 2007 and 2009 FNES data. The open-ended questions have been removed.

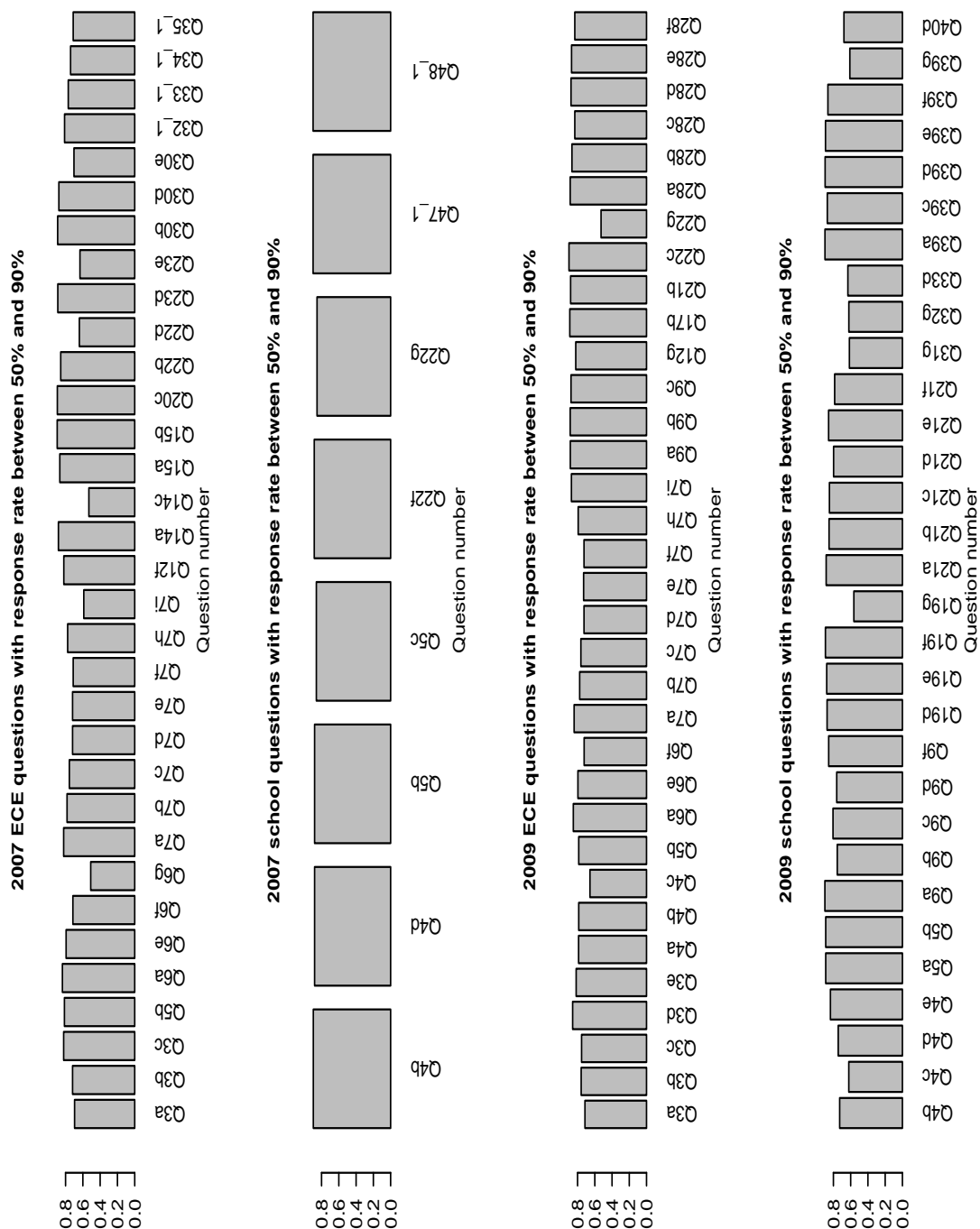


Figure 11.3: Response rate of responding sample between 50% and 90% for 2007 and 2009 FNES questions

Due to the scope of this project, we are not going to impute all the variables with missing data. What we will do is to just impute a few variables. The implication is that the rest of the variables can be imputed in the same way. So, which variables? We pick the variables which have **at least** one of the following characteristics: (1) high non-response rate (e.g response rate less than 90%, but more than 50%); (2) exists in both 2007 and 2009 FNES; (3) questions have structural missingness.

Table 11.7 lists the variables we selected to be imputed in this project. The ECE questions 3a, 7b, and 7c have relatively high non-response rate and they are also comparable between the 2007 and 2009 FNES data, but they are not interrelated questions. Not all the school questions satisfy our first characteristic, but Q5a and Q5b are related to each other. For example, if the answer is “Yes” to question 5b “Whether drinking water available to students during breaks only”, then question 5a “Whether drinking water available to students at any time” is not applicable. Hence, if the answer to 5b is missing, then we can deduce that the answer for question 5a should be “NA”<sup>1</sup> by using the “imputation based on logical rules” method which we have introduced in Chapter 3.

Table 11.7: The selected variables

	Question	Description of data	Response rate 2007	Response rate 2009
ECE	3a	Drinking water available to children through drinking water fountains 1=Yes, 2=No, 98=Don't know, .=No box ticked	69.6%	71.4%
	7b	Food/beverages available at ECE service are prepared on site by a cook 1=Yes, 2=No, 98=Don't know, .=No box ticked	78.3%	77.6%
	7c	Food/beverages available at ECE service are prepared on site by parents 1=Yes, 2=No, 98=Don't know, .=No box ticked	75.6%	76.1%
School	5a	Whether drinking water available to students at any time 1=Yes, 2=No, 98=Don't know, 99=NA, .=No box ticked	95.3%	88.9%
	5b	Whether drinking water available to students during breaks only 1=Yes, 2=No, 98=Don't know, 99=NA, .=No box ticked	88.1%	89.9%

Note: NA=Not Applicable

For the selected variables which are comparable between 2007 and 2009, there are five missing data scenarios:

1. units with missing data on the 2007 FNES
2. units with missing data on the 2009 FNES
3. units with missing data on both 2007 and 2009 FNES in the matched dataset
4. units with missing data on the 2007 FNES but not missing on the 2009 FNES in the matched dataset
5. units with missing data on the 2009 FNES but not missing on the 2007 FNES in the matched dataset

Table 11.8 gives detailed information on each scenario. The response rates were calculated by using the response sample size for each questions divided by the total response sample size.

## 11.2 Investigating the Missing Data Mechanism

After gaining some initial general understanding of the missing data pattern of our data, the next step is to investigate what kind of missing data mechanism the incomplete data possesses.

<sup>1</sup>NA=Not Applicable

Table 11.8: Sample response rate of the five missing data scenarios

ECE Question	(1) 07 data Full 07 respond- ing sample	(2) 09 data Full 09 respond- ing sample	(3) 07 and 09 matched sample	(4) 07 but not 09 matched sample	(5) 09 but not 07 matched sample
3a Response rate (sample size)	70% (391)	71% (455)	61% (66)	14% (15)	15% (16)
7b Response rate (sample size)	78% (440)	78% (494)	59% (64)	22% (24)	13% (14)
7c Response rate (sample size)	76% (425)	76% (485)	61% (66)	19% (21)	13% (14)
Total Responding sample size for each scenarios	562	637	109	109	109
School Question					
5a Response rate (sample size)	94%(699)	87%(989)	82%(193)	14%(33)	3%(6)
5b Response rate (sample size)	86%(637)	88%(1001)	76%(178)	13%(30)	9%(21)
Total Responding sample size for each scenarios	745	1137	234	234	234

Of course, without the complete data, we can only manage to find out if our missing data are MCAR or not MCAR. Trying to distinguish MAR and NMAR is impossible, unless we can get related auxiliary variables and make some assumptions, as Desai et al. (2010, p. 2) describe in their paper. This thesis only focuses on distinguishing between the MCAR and not MCAR missing mechanism. The question is how useful it is to find out if missing data are MCAR or not? It is very useful in terms of helping us to decide the most appropriate imputation methods. For example, if the missing data are MCAR, then a simple hot deck imputation would have the same efficiency in terms of imputation as the more elaborate Nearest Neighbour hot deck imputation method which also requires more computational resource. More importantly, as (Enders 2010, p. 17) points out that the process for deciding whether the missing data are MCAR or not can also help us to identify the variables which might be related to the variable with missing values. Hence, even the missing data are not MCAR, we have found variables which correlate to the missingness. Feeding these variables into our missing data handling procedure can mitigate bias and satisfy the MAR assumption (Collins et al. 2001).

There are many methods for testing the MCAR mechanism. We are going to introduce two of them.

### 11.2.1 Univariate comparisons

The univariate t-test is the simplest method for assessing whether missing data are MCAR (Dixon 1988), given continuous complete explanatory variables  $X$ . It simply separates the whole data set into two groups: missing and observed, according to the variable of interest  $Y$ . Because other complete variables are separated by these two groups as well, we can work out the means of each groups of those complete variables. Under the assumption of MCAR, the mean of  $X$  for the observed data should be the same as the mean of  $X$  for the missing data. Hence, we apply the t-test to test whether the two group means are equal, or significantly different. A non-significant t-test provides no evidence to doubt that the missing data are MCAR, whereas a significant t-test suggests that the missing data are MAR or NMAR. Suppose the variable  $Y$  which has missing data is converted into the response indicator:

$$R = \begin{cases} 1 & \text{If } Y \text{ is observed} \\ 0 & \text{Otherwise} \end{cases}$$

We further assume that the explanatory variables  $X$  are complete. Then:

$$\begin{aligned} \text{the null hypothesis: } & \mu_{X|R=1} = \mu_{X|R=0} & \text{the missingness is MCAR} \\ \text{the alternative hypothesis: } & \mu_{X|R=1} \neq \mu_{X|R=0} & \text{the missingness is not MCAR} \end{aligned}$$

If the explanatory variables  $X$  are categorical variables, we can simply switch to similar hypothesis tests which are developed for categorical data, such as the Chi-square ( $\chi^2$ ) test. For MCAR data, the distribution of categorical  $X$  is independent of whether or not  $Y$  is observed.

We are going to demonstrate the use of the univariate comparison test on the FNES missing data, but we do not recommend this method because it has number of practical issues. Firstly, the FNES has a large number of variables, doing a t-test or Chi-square ( $\chi^2$ ) test for the two groups of each variable is time consuming task and multiple comparisons have the risk of increasing Type I errors<sup>2</sup>. Secondly, as the name “univariate comparison” suggests, this method does not take into account the correlations among the explanatory variables. So, it is possible that a number of variables have significant mean differences between the missing and observed groups, but, in fact, there is only one variable which relates to the missingness, the other variables just simply correlate with that variable. Thirdly, if the missing group is very small, then this means that our t-test power is limited which makes it impossible to perform certain comparisons (Enders 2010).

For demonstration purposes, we have chosen the “Q3a - Drinking water available to children through drinking water fountains” of 2009 ECE FNES as our  $Y$ , and the sample design variable “Authority” as our  $X$ . Table 11.9 shows the split data.

Table 11.9: Split the “Authority 2009” based on the missingness of “Q3a 2009”

$Y = \text{“Q3a 2009”}$	$X = \text{“Authority”}$		Proportion of privately owned ( $P$ )
	Privately Owned	Community Based	
Observed units ( $R = 1$ )	128	323	0.28
Unobserved units ( $R = 0$ )	77	109	0.41

Note: (1)  $P = \frac{\text{Privately Owned}}{\text{Privately Owned} + \text{Community Based}}$

Hence, we have:

the null hypothesis:  $P_{\text{Private}|R=1} = P_{\text{Private}|R=0}$  the missingness is MCAR

the alternative hypothesis:  $P_{\text{Private}|R=1} \neq P_{\text{Private}|R=0}$  the missingness is not MCAR

The R results of Pearson’s Chi-squared test are as follows:

Pearson’s Chi-squared test with Yates’ continuity correction

```
data: split_Q3a
X-squared = 9.6353, df = 1, p-value = 0.001909
```

The p-value is very small ( $< 0.002$ ). Hence, we have very strong evidence against the null hypothesis. This means the missingness is not MCAR and is associated with the variable “Authority”.

## 11.2.2 Logistic regression assessment method

Ridout & Diggle (1991) and Fairclough (2010) propose that a logistic regression model can be used as an effective tool to investigate the missing data mechanism. Their essential idea is to test the association between missingness and the explanatory variables, or the covariates, to be precise. If the associations are strong (p-values of Chi-square test are significant), then

<sup>2</sup>A type I error is the incorrect rejection of a true null hypothesis.

the missing data is not MCAR. Otherwise, the missingness is MCAR. More importantly, this method gives us an indication of which are the variables to form the best MAR model for prediction. The advantages of the logistic regression assessment method are: (1) it handles multiple variables at one go; (2) identifies variables which are related to the missingness; (3) and has the ability to deal with more complicated designs such as an incomplete block design<sup>3</sup>(Ridout & Diggle 1991).

So, how exactly does the logistic regression assessment method work? I will demonstrate this with a few variables from the FNES ECE 2009 dataset. I have picked up question 3a (“Drinking water available to children/tamariki through drinking water fountains?”), and three proposal explanatory variables: “Stratum”, “Authority”, and “Regional council”. All the three proposal explanatory variables are categorical variables. Hence, the first step should be to check the sample size at the breakdown levels by each explanatory variable’s levels. This is because that each categorical variable’s level is treated as an independent dummy variable in the logistic regression, and Harrell (1984) defines the rule of thumb that there should be at least 10 cases per independent variable. We need to make sure that the choice of our variables does not fail this rule. Table 11.10 lists the variables and their descriptions. Table 11.11 shows us the actual sample size of 2009 ECE FNES breaks down by the three explanatory variables.

Table 11.10: Subset variables form the ECE 2009 data

Field name 2009	Description of data
Q3a	Q3a Drinking water fountains 1=Yes, 2=No, 98=Don’t know, .=No box ticked
Stratum	ECE1, ECE2, ECE3, ECE4
Authority	Privately Owned, Community Based
Regional council	Auckland, Bay of Plenty, Canterbury, Gisborne, Hawkes Bay, Manawatu-Wanganui, Marlborough, Nelson, Northland, Otago, Southland, Taranaki, Tasman, Waikato, Wellington, West Coast

---

<sup>3</sup>In a design, the treatments are allocated to the experimental units or plots within homogeneous blocks. This is called block design. An incomplete block design is a block design does not include all factor combinations in every block (Mason et al. 2010)

Table 11.11: Responding sample size breaks down by explanatory variables

Stratum	
Levels	Actual sample size
ECE1 (Education and Care Centres)	281
ECE2 (Free Kindergarten)	170
ECE3 (Home-based childcare)	66
ECE4 (Playcentre)	120

Authority	
Levels	Actual sample size
Privately Owned	205
Community Based	432

Regional council					
Levels	Actual sample size	Levels	Actual sample size	Levels	Actual sample size
Auckland	172	Manawatu-Wanganui	39	Southland	20
Bay of Plenty	37	Marlborough	5	Taranaki	25
Canterbury	76	Nelson	7	Tasman	5
Gisborne	9	Northland	27	Waikato	66
Hawkes Bay	22	Otago	34	Wellington	82
West Coast	6	Region is missing	5		

As Table 11.11 shows, the variable “Regional council” has levels with sample size less than 10, and there are 5 units with no “Regional council” information in the survey dataset. A simple solution is to discard this variable, but we do not want to do that. We select these three explanatory variables because we think they are the most likely variables to have an association with the missing data. Before the logistic regression model identifies their relationship with the missing data, we do not want to discard them easily. So, what can we do? One easy and efficient method is to regroup or merge those levels with small sample size levels to make them bigger. For the 5 units with missing “region” information, as mentioned earlier in this chapter, we can match the survey dataset to the frame list which also has the “Regional council” variable to find out the locations. This is actually a kind of deduction imputation.

Table 11.12 shows the regrouped and edited “Regional council” variables. We rename this new variable as “Region” as it is no longer the regional councils. As can be seen, we grouped the “Marlborough Region”, the “Nelson Region”, the “West Coast Region”, and the “Tasman Region” into one group called “Nelson.Marlborough.Tasman.West Region”, and the “Gisborne Region”, and the “Hawkes Bay Region” into one group called “Gisborne.Hawkes Region” for the “Regional council” variable.

In terms of the matching process to find the regions for the five units with missing “Regional council” information, we found that the sample frame does not have the “regional council” information for the five units either. However, fortunately, the sample frame has complete “City” variable. Hence, we simply applied the deductive imputation method to deduce the “regional council” information. For example, if the “City” variable says the unit is in Auckland, then we infer its “Regional council” should be the “Auckland Region”.



Table 11.12: Responding sample size breaks down by regrouped variables

Region					
Levels	Actual sample	Levels	Actual sample	Levels	Actual sample
Auckland	175	Manawatu-Wanganui	39	Southland	21
Bay of Plenty	37	Nelson.Marlbrough.Tasman.West	23	Taranaki	25
Canterbury	76	Northland	27	Waikato	67
Gisborne.Hawkes	31	Otago	34	Wellington	82

Figure 11.4 shows the response rate of Q3a breaks down by these variables separately. It actually gives some indications that the missing data of Q3a is not MCAR. For example, the stratum ECE3 (“Home-based childcare”) has the lowest response rate compared with other strata (ECE1, ECE2 and ECE3), and the “privately owned” ECE services have lower response rate than “community based” ECE services. This means the response rate of Q3a may have association with stratum and authority.

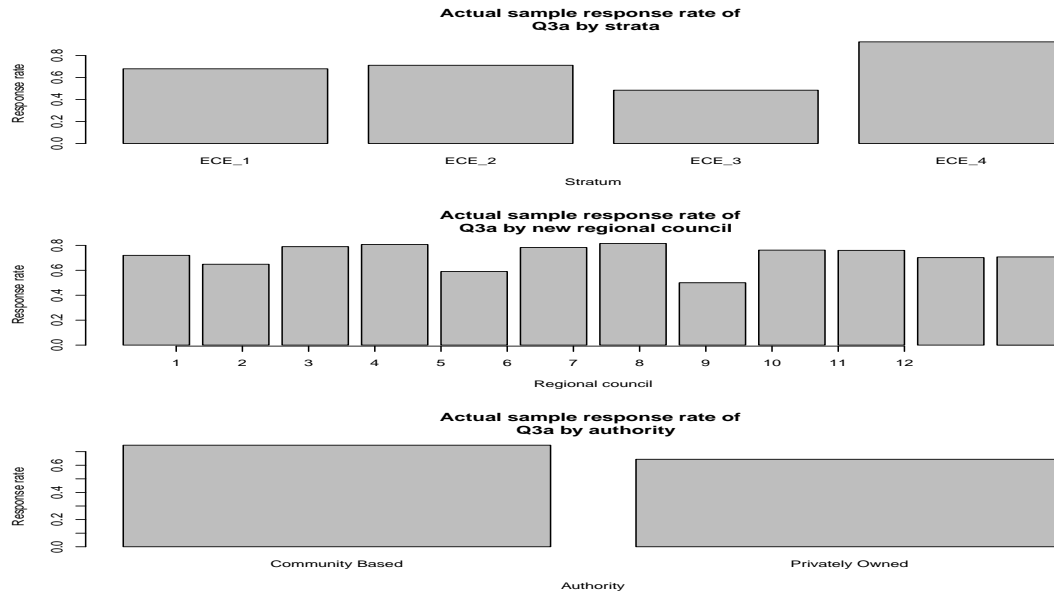


Figure 11.4: Response rate of Q3a breaks down by the four explanatory variables

Note: 1=Auckland Region, 2=Bay of Plenty Region, 3=Canterbury Region, 4=Gisborne.Hawkes Region, 5=Manawatu-Wanganui Region, 6=Nelson.Marlbrough.Tasman.West Region, 7=Northland Region, 8=Otago Region, 9=Southland Region, 10=Taranaki Region, 11=Waikato Region, 12=Wellington Region.

The next step is to introduce a dummy missing data indicator variable. If the answer for question 3a is not missing, then the dummy variable has a value equal to 1, otherwise, the value is 0. Suppose my dummy variable is “Response\_ind”, then

$$R_i = \text{Response\_ind}_i = \begin{cases} 0 & \text{if } Q3a_i \text{ is missing} \\ 1 & \text{if } Q3a_i \text{ is not missing} \end{cases}$$

Finally, we can construct the logistic model.

$$\text{logit}(\mu_i) = \alpha + X_i^T \beta, \quad R_i \sim \text{Bernoulli}(\mu_i) \quad (11.1)$$

where  $X_i$  is the set of explanatory variables, and  $\beta = \beta_1, \dots, \beta_p$ , given that there are  $p$  explanatory variables. In our case, it can be any or all of the four explanatory variables or the

interactions of them with each other. However, we did the logistic regression modelling without considering interactions. This is because the FNES ECE 2009 dataset only has 637 units. A multidimensional table (i.e *stratum*  $\times$  *authority*)(table 11.13) contains cells with zero units. In an interaction model these zero cells cause a problem, and there are many methods that have been developed to solve it, such as “add-a-constant” approach<sup>4</sup> (Haldane 1956) and “Pseudo-Bayes” Approach<sup>5</sup>(Bishop et al. 1975). However, these zero cells do not cause problems for logistic regression, especially, if only pairwise interactions are included.

Table 11.13 also gives us some indication that the variable “Stratum” and “Authority” are correlated. For instance, the privately owned ECE services are only in stratum ECE1 (Education and Care Centres) and ECE3 (Home-based childcare), but not in stratum ECE2 (Free Kindergarten) and ECE4(Playcentre). This means that there is no private owners of non-profitable ECE services (by definition), such as the “Free Kindergarten” and the “Playcentre”.

Table 11.13: Cross tabulation between Stratum and Authority

		Authority	
		Community Based	Privately Owned
Stratum	ECE1 (Education and Care Centres)	124	157
	ECE2 (Free Kindergarten)	170	0
	ECE3 (Home-based childcare)	18	48
	ECE4 (Playcentre)	120	0

The logistic regression modelling results are displayed in Table 11.14 and Table 11.15. These results basically say that there are strong correlations among missingness and the “stratum” as the p-value is  $< 0.01$ . Hence, the missing data for question 3a are not MCAR. The results also identify that with the presence of the “stratum” variable, the coefficient of the “Authority” variable does not have a significant p-value. This, as discussed previously, might be due to the association between the “stratum” and the “Authority”, or there might even be multi association<sup>6</sup> among the three explanatory variables.

The multiple association, or a similar terminology “multicollinearity”<sup>7</sup> can be a problem in modelling. The problem is that it increases the standard errors of the coefficients. Increased standard errors means that coefficients for some explanatory variables may be found not to be significantly different from zero, whereas without multicollinearity and with lower standard errors, these same coefficients might have been found to be significant and the researcher may not have come to null findings in the first place. However, the multicollinearity is more of a concern for researchers who want to interpret the model or find the true relationships between explanatory and response variables (Vaughan & Berry 2005) than to us. This is because the primary goal of this project is to impute missing data. As long as the model can provide us with reliable imputation results, then the existence of multicollinearity in the model is not of concern as it does not affect the prediction of the response variable.

<sup>4</sup> Adding a small constant, generally 0.5, to every cell of the table has been a common recommendation.

<sup>5</sup> It is a Bayesian method, an alternative approach to Maximum Likelihood estimation, providing a way of smoothing the data in a less ad hoc manner than adding an arbitrary constant to cells(Agresti 2002).

<sup>6</sup> More than two explanatory variables are related to each other.

<sup>7</sup> Multicollinearity means that there are strong correlations among explanatory variables.

Table 11.14: Investigate missing mechanism: logistic modelling results

	Estimate( $\beta$ )	Standard Error	z value	p-value
Intercept	0.750545	0.241163	3.112	<0.01
(Stratum)ECE2	0.287314	0.255376	1.125	0.26056
(Stratum)ECE3	-0.794603	0.295585	-2.688	<0.01
(Stratum)ECE4	1.887350	0.399109	4.729	<0.01
(Region)Bay of Plenty Region	-0.245635	0.408574	-0.601	0.54771
(Region)Canterbury Region	0.229570	0.339054	0.677	0.49835
(Region)Gisborne.Hawkes Region	0.479136	0.506704	0.946	0.34436
(Region)Manawatu-Wanganui Region	-0.721919	0.386584	-1.867	0.06184
(Region)Nelson.Marlbrough.Tasman.West Region	0.175318	0.556426	0.315	0.75270
(Region)Northland Region	0.165469	0.545106	0.304	0.76147
(Region)Otago Region	-1.149244	0.409173	-2.809	<0.01
(Region)Southland Region	0.234350	0.564545	0.415	0.67806
(Region)Taranaki Region	-0.006325	0.527203	-0.012	0.99043
(Region)Waikato Region	-0.109942	0.341542	-0.322	0.74753
(Region)Wellington Region	-0.117779	0.308970	-0.381	0.70306
(Authority)Privately Owned	0.124845	0.243305	0.513	0.60787

Table 11.15: Investigate missing mechanism: Analysis of Deviance Table

	Degree of Freedom	Deviance	Pr(>Chi)
Stratum	3	50.215	<0.0001
Region	11	15.684	0.1533
Authority	1	0.263	0.608

## 11.3 Applying Imputation Methods to the FNES

We have outlined some of the most popular imputation methods in Chapter 3 to Chapter 9. Now, let's apply some of the most appropriate ones to our selected FNES incomplete variables. In this section, we give a detailed description of the imputation procedures for the imputing of the question Q3a from the 2009 FNES ECE sample data. The imputation for the other selected incomplete variables follows a similar approach. Hence, the descriptions for their imputation procedures were omitted. Only the final imputation results were shown.

The general outline of our imputation plan is:

- Impute 2009 FNES questions
- Impute 2007 FNES questions
- Impute the questions in both 2007 and 2009

### 11.3.1 Preparing for Imputation

Before imputing Q3a, we need to conduct some further Explanatory Data Analysis (EDA) on Q3a. We did some general EDA in section 11.1, but this time we concentrate on Q3a only. Table 11.16 shows the breakdown of responses for 2009 ECE FNES Q3a.

As reflected in Table 11.16, we see that the answer category “98=Don't know” has only a small sample, and the answer “Don't know” does not actually represent anything meaningful. It means the answer could be “Yes” or “No”. Hence, we reclassified the “Don't know” into missing category “NA”. The other thing we did was to have further investigation on the cross tabulation of the two selected explanatory variables: “Stratum” and “New regional council”. Table 11.17 shows the multidimensional table. As can be seen, some of the cells have very

Table 11.16: Responding sample size for 2009 ECE FNES Q3a breaks down by answer categories

2009 ECE FNES Q3a: Drinking water available to children through drinking water fountain					
Answer categories	1=Yes	2=No	98=Don't know	NA=missing	Total
Sample size	113	338	4	182	637

small sample size, i.e. less than 3. This could be problematic if we want to perform conditional mode imputation and hot deck imputation conditioning on these two variables. For example, when the cell has only one respondent and the answer is missing, it is impossible to find an observed value or mode within that cell to replace the missing data. Hence, we further regrouped the “region” variable into five super regions: “Auckland Region”, “Wellington Region”, “Canterbury Region”, “The rest of North Island”, and “The rest of South Island”. Table 11.18 shows the results of the regrouped regions.

Table 11.17: Multiway table “Region” and “Stratum”

New regional council	Stratum			
	ECE1	ECE2	ECE3	ECE4
Auckland Region	105	41	11	18
Bay of Plenty Region	11	10	10	6
Canterbury Region	35	19	4	18
Gisborne.Hawkes Region	11	10	4	6
Manawatu-Wanganui Region	15	14	3	7
Nelson.Marlborough.Tasman.West Region	11	5	1	6
Northland Region	9	7	1	10
Otago Region	11	12	4	7
Southland Region	6	7	4	4
Taranaki Region	7	7	3	8
Waikato Region	26	8	15	18
Wellington Region	34	30	6	12

Table 11.18: Multiway table “Super region” and “Stratum”

Super region	Stratum				Original regions
	ECE1	ECE2	ECE3	ECE4	
Super Auckland Region	114	48	12	28	Auckland Region, Northland Region
Super Wellington Region	49	44	9	19	Wellington Region, Manawatu-Wanganui Region
Rest of North Island	44	25	28	32	Waikato Region, Bay of Plenty Region, Gisborne.Hawkes Region, Taranaki Region
Super Canterbury Region	46	31	8	25	Canterbury Region, Otago Region
Rest of South Island	28	22	9	16	Southland Region, Nelson.Marlborough.Tasman.West Region

### 11.3.2 Incorporating sample weights in the imputation models

Unlike the imputation we have done for the simple SURF, the use of sample weights need to be addressed when applying imputation models to impute the missing FNES data. Naively, we may choose to ignore sample weights in creation of imputation models. This approach may effectively impute the unweighted sample distribution of respondents, but potentially cause bias to the weighted sample distribution. For example, suppose an incomplete categorical variable has categories: “A” and “B”. Without incorporating sample weights, category A has the higher frequency than category B. Hence, if mode imputation is used, then A would be used to replace missing data. But, if sample weights are involved in the computation of frequencies for category A and category B, then category B actually has the highest frequency. Therefore, we ought to impute missing data with B instead of A.

A few approaches have been suggested by researchers on how to use sample weights in the imputation models. The first intuitive approach, proposed by Platek & Gray (1983), is to inflate the observed values by the sample weight or the ratio of the sample weight (i.e. the observed value of unit  $i \times$  its sample weight). Then, the imputation models are applied to the updated observed values. This approach has shortcomings. In the case of categorical values or integer-valued imputed values, the imputations may no longer be plausible values (Andridge & Little 2009).

The second approach is to randomly draw values from observed data to replace the missing data with probability of selection proportional to the sample weight of the selected observed units (Rao & Shao 1992). This approach does not need to alter the original observed values. Hence, it can be relatively easily to apply to any data type. However, due to this method needing to randomly select observed data values, there are only a limited number of imputation methods that can adopt this approach, such as hot deck imputation methods.

The third approach is to include the sample weights as a covariate in the imputation model or have design variables which have been used to form the sample weights in the imputation model (Carpenter 2011). If the variables used to form the sample weights are incorporated in the imputation model, then the use of sample weights in the imputation model is unnecessary and inefficient. This is because they all achieve the same imputation results. For example, if we apply the conditional mode imputation or hot deck within adjustment cells imputation, the imputation cells that are formed by using the sample weights would be identical to the imputation cells formed by the design variables are used to construct the sample weights. This is because the sample weights would be the same for the observations which have the same design variables’ values. Please see Table 10.8 to Table 10.11 in Section 10.9, Chapter 10 for examples.

Compared to the first two approaches, Andridge & Little (2009) suggest that the third approach is better. The first two approaches fail to reduce non-response bias if the missingness and the incomplete variable are related to the sample weights or design variables, as both approaches do not include these variables in the imputation model. Furthermore, if the imputation model needs to form imputation cells, the response propensity<sup>8</sup> would not be constant within each imputation cell. For example, we would expect that Primary and Secondary schools answer differently and have different response propensity in the FNES data, and we

---

<sup>8</sup>Response propensity is the theoretical probability that a sampled unit will become a respondent in a survey. For example, some people are easier to get into contact with than are others in a particular survey.

would not want to impute secondary schools' missing data with the responses from the observed primary schools.

Therefore, we chose the third approach as our method of incorporating sample weights in the imputation models. Fortunately, we have decided to use “Stratum” as one of our explanatory variables in our imputation models. Hence, as explained above that the weights being equal within strata, we do not specifically include sample weights in our imputation models in the following sections.

### 11.3.3 Imputing the 2007 and the 2009 ECE FNES missing data

Based on previous analysis, we assumed MAR missingness and defined the general imputation model as:

$$Y|X, \theta \quad (11.2)$$

In the following demonstrations,  $Y$  represents the data of Q3a 2009, and  $X$  encompasses Stratum and Super region variables.  $\theta$  has the parameters when logistic regressions are used in some imputation methods.

We have applied the following imputation methods to the selected incomplete FNES variables. The 2009 ECE FNES question “Q3a” is used as a demonstration in the descriptions.

- Conditional mode imputation
- Hot deck within adjustment cells
- Nearest Neighbour hot deck imputation
- Logistic regression imputation
- Nonparametric Resampling methods
- EM Algorithm
- Multiple Imputation

For the units which have been observed in both 2007 and 2009, we have applied cold deck imputation as well, i.e. copying the response from 2007 if missing in 2009, and vice versa. Cold deck imputation serves two purposes here: (1) it provides an alternative imputation method; (2) it can be compared to the results of other imputation methods to check their efficiencies. Due to the nature of Q3a, we believe cold deck imputation provides us with the most reliable imputed values. This is because facilities such as drinking water fountain are unlikely to change in only a two year period. Given that the cold deck imputed values are likely to be the closest to the true values, then we can use them to compare to the results of other imputation methods. However, this comparison is limited for Q3a because the number of missing data in the matched dataset is very small (Table 11.8). Please refer to Chapter 9 for theoretical details of these methods.

**Conditional mode imputation:** we have divided the respondents into  $G = 20$  groups by Stratum and Super region variables. The number of units in each group are shown in Table 11.18. Hence, let  $Y = Q3a$ ,  $n = 637$ , and the answer with category values  $C_k = (C_1, C_2)$ , where  $C_1 = \text{Yes}$ , and  $C_2 = \text{No}$ , then we can re-express Equation (9.2) as:

$$Y_i^{miss} = C_{k_g}, \quad i \in 1, \dots, n \quad g \in G$$

if

$$k_g = \underset{k}{\operatorname{argmax}} \sum_{i \in S_g^r} q_{g,k,i}$$

where  $S_g^r$  = respondent in group  $g$

$$q_{k,i} = \begin{cases} 1 & \text{if } Y_i^{obs} = C_{k_g} \\ 0 & \text{Otherwise} \end{cases}$$

If there are multiple modes, then one of the modes will be randomly selected as the imputed value.

**Hot deck within adjustment cells:** this method has been introduced in Chapter 3, Section 3.3.1 and demonstrated in Chapter 4, Section 4.3.4 by imputing missing data for a numerical variable. As it belongs to the implicit modelling methods, it can be easily adopted for imputing categorical missing data. For the imputation of Q3a, the imputation cells were formed by variables Stratum and super\_regional\_council, then missing Q3a were replaced by a random draw from the observed Q3a values in each cell.

**Nearest Neighbour hot deck imputation:** as have been introduced in Chapter 9, we apply the Gower distance to measure the dissimilarity between the  $i$ th and the  $j$ th unit. Due to the Q3a being a nominal categorical variable, the distance function Equation (9.4) can be re-expressed as:

$$d(i, j) = \frac{\sum_{h=1}^H \delta_{ijh} d_{ijh}}{\sum_{h=1}^H \delta_{ijh}}$$

where  $H$  is the number of categorical variables. In this case, we have two categorical explanatory variables: Stratum and Super region. Hence,  $H = 2$ .  $\delta_{ijh}$  is the weight of variable, and

$$d_{ijh} = \begin{cases} 0 & \text{if } y_{ih} = y_{jh} \\ 1 & \text{Otherwise} \end{cases}$$

**Logistic regression imputation:** let  $Y = Q3a$  as our response variable, and the Stratum and Super region as the explanatory variables  $X$ . First, we construct a logistic regression model based on the observed units for Q3a. Then, we use this logistic model to predict the missing Q3a answers. The Equation (9.3) can be re-expressed as:

$$\operatorname{logit}[P(Y_i = 1)] = \operatorname{logit}[\pi_i] = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = x_i^T \beta$$

We apply the “Method 2” from Section 9.3.2 in Chapter 9 to convert the probability  $\hat{\pi}$  into category values. The Method 2 randomly draws values from the Bernoulli distribution,  $Y_i^{miss} \sim \operatorname{Bernoulli}(\hat{\pi}_i)$ , given the computed  $\hat{\pi}_i$ .

**Nonparametric Resampling:** as have been introduced in Section 5.4 and Section 5.6 of Chapter 5, the resampling methods resamples a large number of sub-samples from the original sample, then applies the same imputation method to impute missing data for all the sub-samples. We have applied two resampling methods: the Bootstrap method and the jackknife method. For this Chapter, we only apply the Bootstrap method for Q3a. The chosen imputation method is the hot deck within adjustment cells imputation method, and there are  $B = 200$  bootstrap samples

**EM Algorithm:** we applied the same method introduced in Chapter 9, Section 9.4 to Q3a. We use the same logistic regression above to predict the missing Q3a answers, but instead of imputing missing values once, we impute them several times through iteration until the algorithm converges. The EM algorithm for this case is as follows:

**The E step:** Replace each missing Q3a values by its expectation conditional on X:

$$\hat{y}_{Q3a,i}^t = E(y_{Q3a,i}|x_i, \beta^t) = \frac{1}{1 + \exp(-x_i^T \beta^t)}$$

where  $y_{Q3a,i}$  is the value of Q3a, and  $t = 0, 1, \dots, m$ ,  $m$  is the number of iterations, and  $i = 1, \dots, n$ ,  $n$  is the sample size.

**The M step:** Maximize the log-likelihood function to find the estimate  $\beta^{t+1}$ :

$$\begin{aligned} \ell(y_{Q3a,i}|x_i, \beta) &= \sum_{i=1}^r y_{Q3a,i} x_i^T \beta - \sum_{i=1}^r \log(1 + \exp(x_i^T \beta)) \\ &\quad + \sum_{i=r+1}^{n-r} \hat{y}_{Q3a,i}^t x_i^T \beta - \sum_{i=r+1}^{n-r} \log(1 + \exp(x_i^T \beta)) \end{aligned}$$

where  $r$  is the number of respondents.

The E-step and the M-step repeat again and again until there is a sufficiently small difference between  $\ell(y_{Q3a,i}|x_i, \beta^{t+1}) - \ell(y_i|x_i, \beta^t)$ . As also described in Chapter 9, Section 9.4, we have introduced changing cut off probability  $C$  when we need to convert the predicted missing Q3a values from probabilities into categorical values. The cut off probability  $C$  is the same as in Chapter 9, where

$$C = \frac{\sum_{i=1}^n y_i}{n}$$

and

$$y_i = \begin{cases} 1 & y_{Q3a,i} = \text{Yes} \\ 0 & \text{Otherwise} \end{cases}$$

Hence, the cut off probability  $C$  changes each time we update the missing values at the E-step.

**Bayesian Multiple Imputation:** Again, we applied the same Bayesian iterative simulation method for categorical variable from Chapter 9 to the FNES Q3a. Then, the Multiple Imputation (MI) part is basically to select 5 to 10 imputed datasets from the converged part of the Bayesian iterative simulation chain. This method of selecting  $D$  points from the Bayesian simulation chain has been introduced in Chapter 8, Section 8.3. Here is a general description of the Bayesian iterative simulation process for Q3a:

**I-step:** Randomly draw  $\hat{Y}_{Q3a,miss}^t$  from the conditional distribution at iteration  $t$ :

$$\hat{Y}_{Q3a,miss,i}^t \sim p(Y_{Q3a,miss,i}^t | \beta^t) = \text{Bernoulli} \left( 1, \frac{\exp(x_i^T \beta^t)}{1 + \exp(x_i^T \beta^t)} \right)$$



**P-step:** At iteration  $t + 1$ , randomly draw  $\beta^{t+1}$  from the conditional posterior distribution given the updated  $\hat{Y}_{Q3a,miss}^t$  and observed  $Y_{Q3a,obs}$  from the I-step:

$$\beta^{t+1} \sim p(\beta^{t+1} | \hat{Y}_{Q3a,miss}^t, Y_{Q3a,obs})$$

The Metropolis-Hastings (MH) algorithm is applied to sample  $\beta$  as we have demonstrated in Chapter 9.

The “I-step” and “P-step” were repeated many times until the Bayesian iterative simulation chain converges. The detail of diagnosis of the convergence has been introduced in Chapter 7, section 7.4.

The R code is in Appendix F, Section F.1.

**The results:** Table 11.20 to Table 11.25 display the imputation results for Q3a, Q7b, and Q7c of 2009 and 2007 ECE FNES. In the tables, the estimated proportion  $\hat{P}$ s were computed as follows:

$$\hat{P}_h = \frac{n_{h,yes}}{n_{h,yes} + n_{h,no}} \quad (11.3)$$

Then

$$\hat{P} = \sum_{h=1}^H \left( \frac{N_h}{N} \right) \hat{P}_h \quad (11.4)$$

where  $H$  is the number of strata,  $N$  is the size of population and  $N_h$  is the size of population in stratum  $h$ ,  $h = 1, \dots, H$ , and  $n_{h,yes}$  is the number of respondents who give “Yes” answers in stratum  $h$ , and  $n_{h,no}$  is the number of respondents who give “No” answers in stratum  $h$ . The sum of  $n_{h,yes}$  and  $n_{h,no}$  is the sample selected from stratum  $h$ . The formulae for computing the standard error of  $\hat{P}$  for unimputed sample, single imputation and EM algorithm is as follows:

$$\hat{Var}(\hat{P}) = \sum_{h=1}^H \left( \frac{N_h}{N} \right)^2 \left( 1 - \frac{n_h}{N_h} \right) \frac{\hat{P}_h(1 - \hat{P}_h)}{n_h - 1} \quad (11.5)$$

$$\text{Standard error of } \hat{P} = se = \sqrt{\hat{Var}(\hat{P})} \quad (11.6)$$

where  $n_h = n_{h,yes} + n_{h,no}$  and  $\left( 1 - \frac{n_h}{N_h} \right)$  is called the finite population correction (fpc)<sup>9</sup> factor which can be ignored if the ratio of  $\frac{n_h}{N_h}$  is close to 0. We did not ignore the fpc in our calculations.

Please note that we have ignored the involvement of sample selection weights in the computation of  $\hat{P}_h$ s. The weights would not make a difference in our results as the weights are the same in each stratum (Section 11.3.2). Hence, Equation (11.3) can be also shown as:

$$\hat{P}_h = \frac{n_{h,yes}}{n_{h,yes} + n_{h,no}} = \frac{\sum_{i=1}^{n_h} w_i y_{i,h}}{\sum_{i=1}^{n_h} w_i}$$

where

$$y_i = \begin{cases} 1 & \text{if observation unit } i \text{ is in stratum } h \text{ with answer “Yes”} \\ 0 & \text{Otherwise} \end{cases}$$

and  $w_i$  is the weight for observation  $i$  in stratum  $h$ ,  $i = 1, \dots, n_h$ .

<sup>9</sup>The finite population correction factor is used to adjust the error in estimating a mean or a total, which is due to lack of independence when sample without replacement, i.e. negative correlation between observations.

The computation of  $\hat{P}_h$ s for the un-imputed incomplete data also uses Eq (11.3). The observations with missing data were omitted from the computation. We did not do non-response adjustments for the sample selection weights or use the adjusted weights in the computation of  $\hat{P}_h$ s. The reasons are: (1) as discussed in Section 3.2.3 Chapter 3, we consider that the non-response adjustment for the weights is a sort of imputation method; (2) The non-response adjustment treats the responding observations as the new sample and re-calculate the selection weights based on this new sample and the same stratum. Therefore, the new weights within each stratum are still equal to each other.

For the computation of the standard error (*se*) of  $\hat{P}$  for Bootstrap resampling and Multiple Imputation, please refer to Equation (5.2), Section 5.3, Chapter 5; and Equation (8.4), Section 8.2, Chapter 8, respectively.

Figure 11.5 to Figure 11.10 display the estimated proportions  $\hat{P}$ s and their 95% confidence intervals. The 95% confidence interval is computed as follows:

$$\hat{P} \pm 1.96se \quad (11.7)$$

In the figures, the middle points of the vertical lines are the estimated proportion  $\hat{P}$ , the lengths of the vertical lines are the range of confidence intervals.

Looking at Figure 11.5 to Figure 11.10, overall the estimates from all imputation methods are similar, and have comparable standard errors. However, there are some differences. As expected, the Bootstrap resampling and MI produce larger standard errors and wider confidence intervals than other imputation methods. Comparing the bootstrap resampling and MI standard errors, we find that the results are almost identical for most cases. This means that both methods are effective in estimating imputation uncertainty. However, the bootstrap resampling needs  $B = 200$  resamples, but the MI only needs  $D = 5$  datasets. Hence, MI is more efficient than the bootstrap resampling in terms of data storage. In terms of computation efficiency, the bootstrap resampling method might be faster than the MI method for small size samples as the Bayesian MI simulation chains need to run thousands of iterations which might require huge computational resources, compared to the bootstrap resampling method.

We have also noticed that the conditional mode imputation produces a lower estimated proportion  $\hat{P}$  than other imputation methods. This is because all questions have large number of “No”s, but a small number of “Yes”s. This means that the mode are more likely to be “No” in each of the  $G = 20$  groups. Table 11.19 shows us what the modes are in each group for the 2009 ECE FNES Q3a. As shown, using conditional mode imputation, most missing values were imputed as “No”. This is the same for other selected incomplete questions.

Table 11.19: Breakdown of incomplete Q3a of 2009 ECE FNES by Stratum and Super region

Group ( $G = 20$ )	1="Yes"	2="No"	Missing	Conditional Mode imputed value
1=(ECE_1 & Rest of North Island)	14	14	16	"Yes" or "No"
2=(ECE_1 & Rest of South Island)	6	14	8	"No"
3=(ECE_1 & Super Auckland Region)	19	58	37	"No"
4=(ECE_1 & Super Canterbury Region)	6	28	12	"No"
5=(ECE_1 & Super Wellington Region)	5	27	17	"No"
6=(ECE_2 & Rest of North Island)	10	8	7	"Yes"
7=(ECE_2 & Rest of South Island)	9	9	4	"Yes" or "No"
8=(ECE_2 & Super Auckland Region)	19	20	9	"No"
9=(ECE_2 & Super Canterbury Region)	8	10	13	"No"
10=(ECE_2 & Super Wellington Region)	12	16	16	"No"
11=(ECE_3 & Rest of North Island)	1	10	17	"No"
12=(ECE_3 & Rest of South Island)	0	5	4	"No"
13=(ECE_3 & Super Auckland Region)	0	5	7	"No"
14=(ECE_3 & Super Canterbury Region)	0	3	5	"No"
15=(ECE_3 & Super Wellington Region)	0	4	5	"No"
16=(ECE_4 & Rest of North Island)	1	28	3	"No"
17=(ECE_4 & Rest of South Island)	1	15	0	"No"
18=(ECE_4 & Super Auckland Region)	2	25	1	"No"
19=(ECE_4 & Super Canterbury Region)	0	22	3	"No"
20=(ECE_4 & Super Wellington Region)	0	17	2	"No"
Total	113	338	186	

On the other hand, we have found that the EM algorithm produces larger estimated proportions than other imputation methods. It is unclear why the EM algorithm behaves like this. One possible reason is that the use of the cut off probability  $C$  causes the overestimation. If the prediction probability is larger than  $C$ , then the imputed value will be "2 = No", otherwise, the imputed value will be "1 = Yes". Hence, unlike other imputation methods, there is no chance for a prediction probability that is less than  $C$  to get the other imputed value. If  $C$  is large, then most of the prediction probability will be converted to 1. Unfortunately, our starting cut off probability is large for all the questions (ie. above 70%), compared to the SURF example we gave in Section 9.4.1, Chapter 9, where the starting cut off probability is around 50%. Furthermore, the cut off probability will remain high throughout the EM iterations because the ratio of  $n_{no}/n_{yes}$  is large for all the selected FNES questions. This means most missing values were imputed as "1 = Yes" for the FNES questions. This drives up the estimated proportion  $\hat{P}$ . This also indicates that our current EM algorithm for categorical data has some potential overestimation or underestimation problems.

Table 11.20: Imputation results for 2009 ECE FNES Q3a

Q3a, the number of missing observations is 186				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	113	338	0.2511	0.0203
Conditional Mode	140	497	0.2198	0.0157
Hot deck within adjustment cells	159	478	0.2465	0.0162
NN hot deck	152	485	0.2332	0.0158
Logistic regression	165	472	0.2606	0.0162
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.2511	0.0240
EM Algorithm	186	451	0.2758	0.0162
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.2466	0.0184

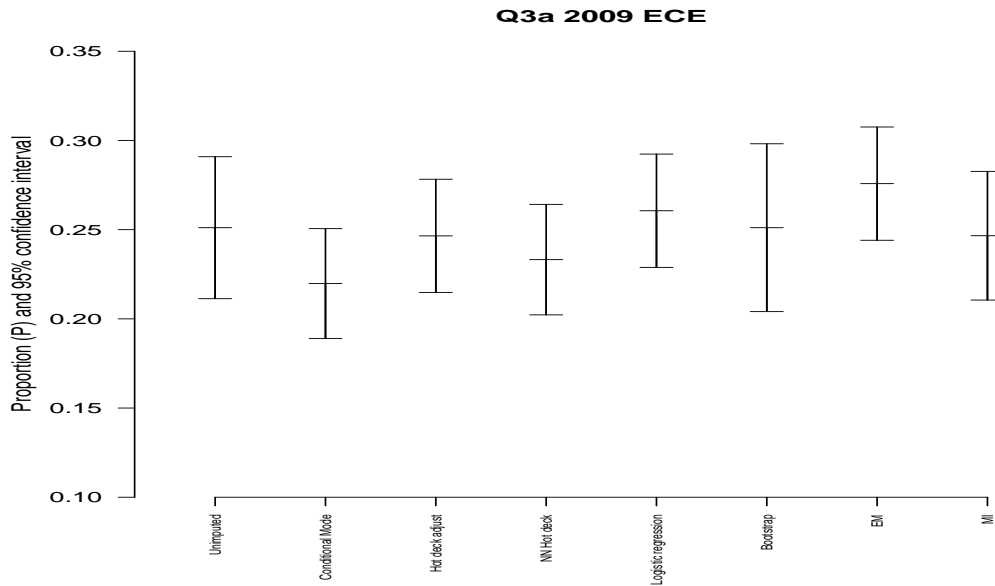


Figure 11.5: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q3a 2009 ECE

Table 11.21: Imputation results for 2009 ECE FNES Q7b

Q7b, the number of missing observations is 144				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	145	348	0.3757	0.0191
Conditional Mode	180	457	0.3888	0.0168
Hot deck within adjustment cells	172	465	0.3683	0.0172
NN hot deck	177	460	0.3793	0.0171
Logistic regression	173	464	0.3604	0.0164
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.3760	0.0198
EM Algorithm	183	454	0.3954	0.0167
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.3709	0.0212

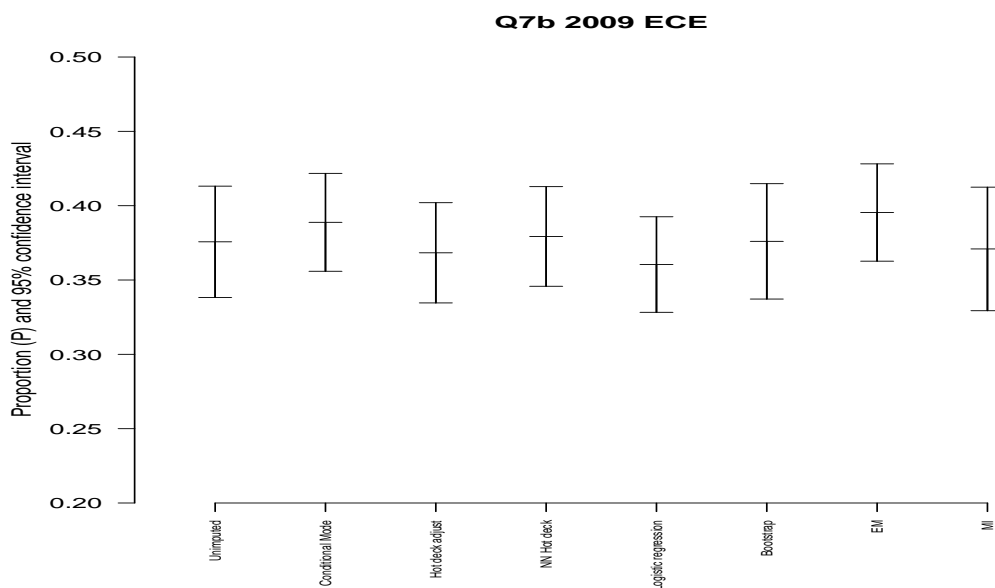


Figure 11.6: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q7b 2009 ECE

Table 11.22: Imputation results for 2009 ECE FNES Q7c

Q7c, the number of missing observations is 153				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	154	330	0.2346	0.0154
Conditional Mode	172	465	0.2078	0.0118
Hot deck within adjustment cells	189	448	0.2275	0.0124
NN hot deck	189	448	0.2270	0.0124
Logistic regression	191	446	0.2445	0.0130
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.2343	0.0172
EM Algorithm	218	419	0.2561	0.0121
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.2311	0.0148

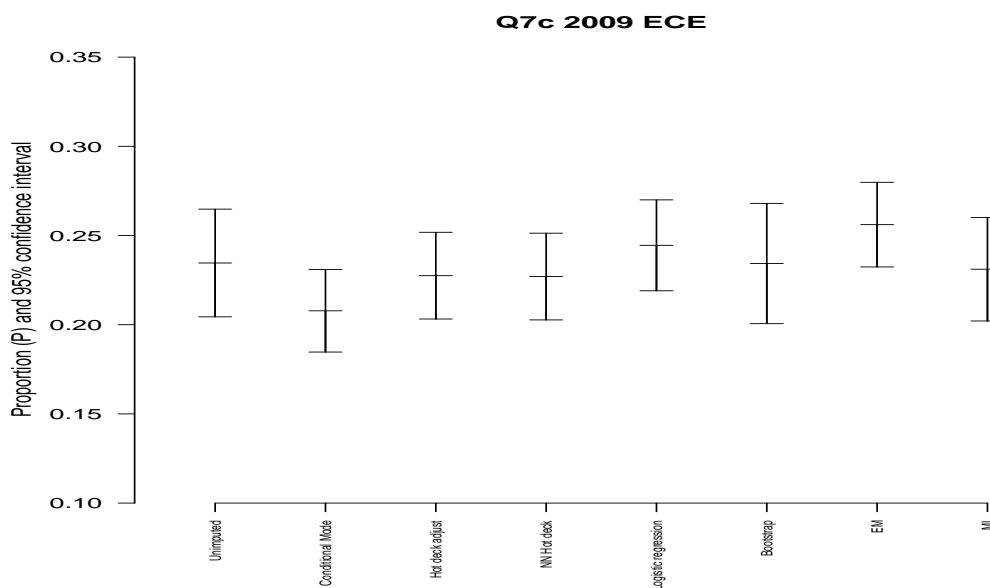
Figure 11.7: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q7c 2009 ECE

Table 11.23: Imputation results for 2007 ECE FNES Q3a

Q3a, the number of missing observations is 181				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	92	289	0.2481	0.0212
Conditional Mode	99	463	0.1717	0.0147
Hot deck within adjustment cells	132	430	0.2337	0.0165
NN hot deck	138	424	0.2411	0.0166
Logistic regression	140	422	0.2452	0.0167
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.2471	0.0230
EM Algorithm	157	405	0.2636	0.0166
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.2416	0.0249

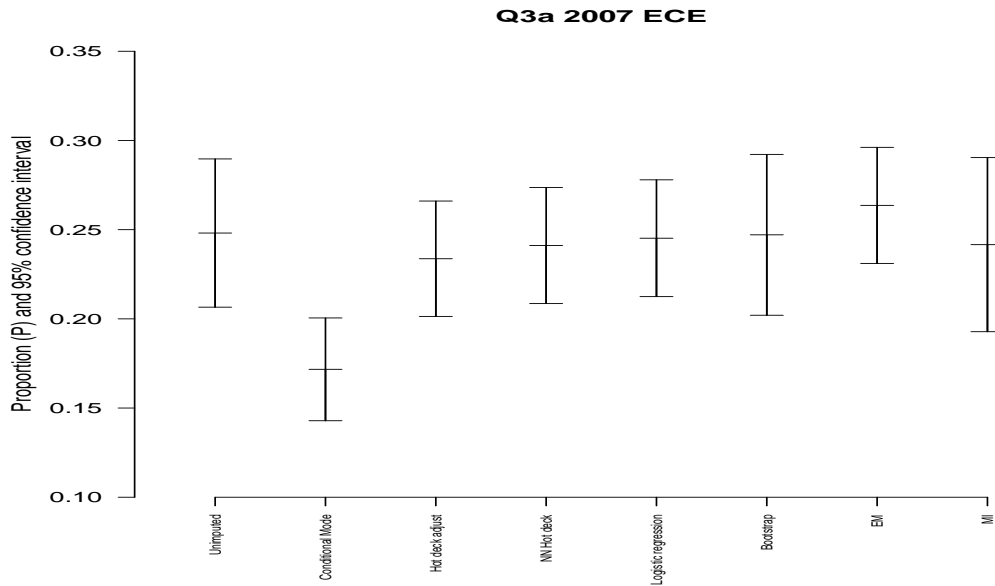


Figure 11.8: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q3a 2007 ECE

Table 11.24: Imputation results for 2007 ECE FNES Q7b

Q7b, the number of missing observations is 128				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	122	312	0.3180	0.0190
Conditional Mode	144	418	0.3116	0.0169
Hot deck within adjustment cells	143	419	0.3090	0.0170
NN hot deck	143	419	0.3087	0.0171
Logistic regression	146	416	0.3152	0.0171
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.3154	0.0211
EM Algorithm	156	406	0.3376	0.0168
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.3095	0.0197

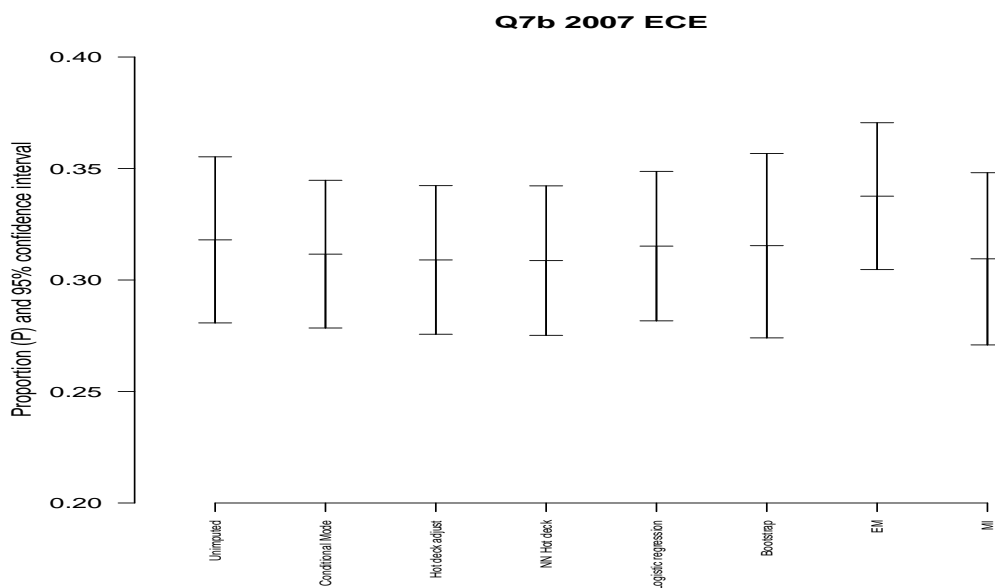
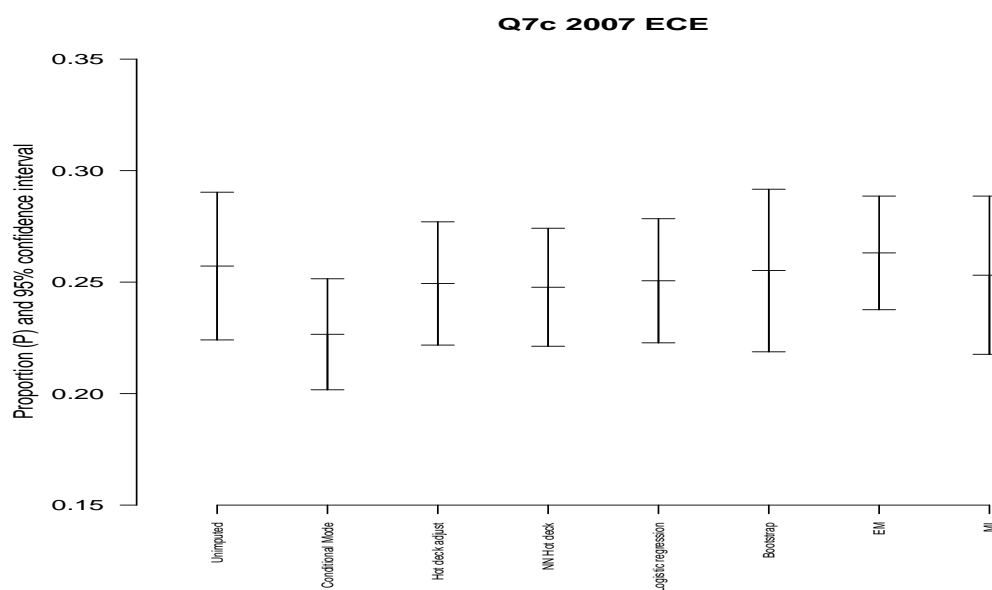


Figure 11.9: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q7b 2007 ECE

Table 11.25: Imputation results for 2007 ECE FNES Q7c

Q7c, the number of missing observations is 147				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	131	284	0.2572	0.0169
Conditional Mode	149	413	0.2266	0.0127
Hot deck within adjustment cells	161	401	0.2494	0.0141
NN hot deck	163	399	0.2477	0.0135
Logistic regression	162	400	0.2506	0.0142
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.2552	0.0186
EM Algorithm	178	384	0.2631	0.0130
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.2531	0.0181

Figure 11.10: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q7c 2007 ECE

### 11.3.4 Imputing the School FNES data

As introduced in Table 11.7, Section 11.1, school questions Q5a and Q5b are interrelated questions. This gives us a chance to apply deductive imputation (or “imputation based on logical rules”) method (Section 3.3.1, Chapter 3). This imputation method can be used for the following scenarios in Table 11.26:

Table 11.26: Scenarios where the imputation based on logical rules can be applied

Scenario	Q5a Whether drinking water available to students at any time 1=Yes, 2=No, 98=Don't know, 99=NA, .=No box ticked	Q5b Whether drinking water available to students during breaks only 1=Yes, 2=No, 98=Don't know, 99=NA, .=No box ticked	Imputed value by using logical rules
1	1=Yes	98=Don't know	NA
2	1=Yes	.=No box ticked	NA
3	98=Don't know	1=Yes	2=No
4	.=No box ticked	1=Yes	2=No
5	99=NA	98=Don't know	both Q5a and Q5b should be missing
6	99=NA	.=No box ticked	both Q5a and Q5b should be missing
7	98=Don't know	99=NA	1=Yes
8	.=No box ticked	99=NA	1=Yes
9	99=NA	99=NA	both Q5a and Q5b should be missing

Note: NA=Not Applicable, “98=Don't know” is considered as missing

As an aside, we do not think Q5a needs the option of “99=NA”. No matter what the answer Q5b is, Q5a can not have an answer “NA”, and we can neither infer Q5a is “NA” by looking at other questions' answers.

The actual scenarios we have for Q5a and Q5b of the 2009 and 2007 School FNES are displayed in Table 11.27 and Table 11.28. Table 11.27 has scenarios 1, 3 and 9. Table 11.28 has scenarios 1, 2, 3, 4 and 9. We have also noticed that Q5a and Q5b of 2007 School FNES have some strange answers. First, one respondent gave an answer “3” to Q5a, which must be a typographic error. Since there is no way to retrieve the true answer, we classify it as missing. Second, 145 respondents have given the “1=Yes” answer to both Q5a and Q5b. According to the way the questions were asked, this is impossible. Hence, for the respondents who answered “Yes” to both Q5a and Q5b, we changed their Q5b answers to “99=NA”. This is because we have assumed it is highly likely that people give the right answer to the first question, but misinterpret the related following questions.

Table 11.27: Cross tabulation of Q5a and Q5b for 2009 School FNES

	Q5b						Total Q5a
		1	2	98	99	Missing	
Q5a	1	0	841	2	0	0	843
	2	145	0	0	0	1	146
	98	13	1	2	0	0	16
	99	0	0	0	6	0	6
	Missing	0	1	0	0	125	126
Total Q5b		158	843	4	6	126	Total=1137



Table 11.28: Cross tabulation of Q5a and Q5b for 2007 School FNES

	Q5b						Total Q5a
		1	2	98	99	Missing	
Q5a	1	145	403	14	4	67	633
	2	64	2	0	0	0	66
	3	1	0	0	0	0	1
	98	9	0	0	0	0	9
	99	0	0	0	1	0	1
	Missing	13	0	0	0	22	35
	Total Q5b	232	405	14	5	89	Total=745

Initially, we want to merge the interrelated questions into one question in order to simplify the imputation process. For example, if the respondent has both Q5a and Q5b missing, and we impute both questions independently, then we could end up with imputing “99=NA” to both questions. This result is our scenario 9 for which we think it is incorrect to have “NA” as answers to both questions. There are other scenarios as well. It is impossible to impute the interrelated questions separately without incorporating the interrelationship into the imputation process. This could be difficult as the number of scenarios can be very large, if there are more than two interrelated questions. Hence, combining the interrelated questions into one seems as an easy way out. For instance, we create a new variable “Q5”. If Q5a is “1=Yes”, then Q5 can be “a”; if Q5a is “2=No”, but Q5b is “1=Yes”, then Q5 is “b”. However, combining the interrelated questions into one gives us even bigger problems. Please see Table 11.29.

Table 11.29: Combining Q5a and Q5b into Q5

Scenario	Q5a Whether drinking water available to students at any time	Q5b Whether drinking water available to students during breaks only	Q5
I	1=Yes	2=No or 99=NA	a
II	2=No	1=Yes	b
III	1=Yes	1=Yes or 99=NA	a
IV	<b>2=No</b>	<b>2=No</b>	<b>c</b>
V	missing	1=Yes	b
VI	1=Yes	missing	a
VII	missing	2=No	need to impute
VIII	2	missing	need to impute
IX	missing	missing	need to impute

As Table 11.29 shows, scenario IV has “2=No” for both Q5a and Q5b, and we code the corresponding Q5 values as “c”. But, Table 11.27 shows that “2=No” for both questions does not occur in 2009 School FNES, and Table 11.28 indicates that there were only two respondents had answered “2=No” for both questions in 2007. This means “c” has zero or very rare chance to be selected as the imputation value for the missing data. Hence, we conclude that although it is possible to merge some interrelated question into one and impute, Q5a and Q5b cannot be merged.

As a result, in order to tackle the interrelationship problem described previously, we planned to impute Q5a first, then impute Q5b by using the completed Q5a. In other words, the imputation of Q5b depends on the imputation of Q5a. Furthermore, we have found that the “99=NA” for Q5a and Q5b does not really contribute anything meaningful. For Q5a, if the answer to Q5b is “1=Yes”, then logically the answer to Q5a should be “2=No”; if the answer to Q5b is “2=No”, then logically the answer to Q5a shouldn’t be “99=NA”; if the answer to Q5b is missing, then this does not mean the answer to Q5a can be “99=NA”. For Q5b, if the answer

to Q5a is “1=Yes”, then logically the answer to Q5b must be “2=No”, it can be “99=NA” but it means the same thing as “2=No”; if the answer to Q5a is “2=No”, then logically it does not mean the answer to Q5b can be “99=NA”; if the answer to Q5a is missing, then this does not mean the answer to Q5b can be “99=NA”. Both Q5a and Q5b have “99=NA” as their answer only means that the answers to Q5a and Q5b are missing or incorrect, otherwise this kind of answer does not make sense. Hence, we have decided to change “99=NA” to missing or “2=No”.

The exact imputation procedure is as follows:

- Step 1: convert the “98=Don’t know” and “99=NA” into missing for Q5a and Q5b. The reason for changing “98=Don’t know” to missing has been discussed in Section 11.3.1. The reason for changing “99=NA” to missing is that we believe that Q5a and Q5b with “NA” answers can be changed to missing as discussed in previous paragraphs.
- Step 2: for missing Q5a, if their corresponding Q5b has value “1=Yes”, then Q5a has value “2=No” (logical rules)
- Step 3: for missing Q5b, if a respondent’s Q5a value is “1=Yes”, then its Q5b value must be “2=No” (logical rules)
- Step 4: for the remaining missing Q5a, impute Q5a missing data by using some other imputation methods. Q5a can only have “1=Yes” or “2=No” as its values
- Step 5: for the remaining missing Q5b, if a respondent’s imputed Q5a value is “1=Yes”, then its Q5b value must be “2=No” (logical rules)
- Step 6: for the remaining missing Q5b, if a respondent’s Q5a value is “2=No”, then the missing data of Q5b can be imputed as “1=Yes” or “2=No” by some other imputation methods.

Due to the logical rules needing to be applied after finishing imputing Q5a in order to update the remaining missing Q5b, the imputation methods which create multiple imputed datasets cannot be applied easily for the imputation of Q5b. For example, if MI is our imputation method and we want to create  $D = 5$  MI datasets, then step 4 will give us  $D = 5$  datasets. However, we then need to use all the five datasets to update Q5b, which gives us five Q5b datasets. Then, each of the five Q5b datasets will be imputed by MI in step 6. This gives us  $5 \times 5 = 25$  imputed Q5b datasets.

This approach is impractical as the number of required imputed Q5b datasets is  $D^2$  which can be very large if the  $D$  increases. Hence, we decided to create five MI datasets for Q5a and use these datasets to update Q5b which gives us five updated Q5b datasets. Then, each of the five updated Q5b datasets was imputed by Bayesian MI but only one imputed Q5b dataset was draw from each of the Bayesian simulation chains. The justification is that the imputation uncertainty has already been captured by the imputation of Q5a and can be passed over to Q5b as there will be multiple imputed Q5b datasets as well.

However, for the resampling methods, we can fix this problem by drawing interrelated questions together and impute each resampled dataset. For example, in our case, each resampled dataset will have both Q5a and Q5b, then we just simply apply the above imputation procedure to impute each resampled dataset.

Then, we imputed the Q5a and Q5b under the assumption that the missing mechanism is MAR, and applied the general imputation model Eq (11.2). In this case, The  $Y_{miss}$  represents missing data of Q5a and Q5b of 2007 and 2009. After applying the same missing data mechanism detection methods that we introduced in Section 11.2, we decided to set the  $X$  to be the stratum variable. All the listed imputation methods in Section 11.3.3 were applied to the imputation of missing data of Q5a and Q5b of 2007 and 2009 FNES.

Table 11.30 to Table 11.33 display the imputation results for Q5a and Q5b of 2007 and 2009 FNES data. Equation (11.3) to Equation (11.6) were used to compute the estimated proportion  $\hat{P}$  and standard error of  $\hat{P}$ . Figure 11.11 to Figure 11.14 display the estimated proportions  $\hat{P}$  and their 95 % confidence intervals. The confidence intervals were computed as Equation (11.7).

Again, the results shows that the bootstrap resampling method and MI have larger standard errors and wider confidence intervals than other imputation methods (except for the Q5b 2007). Looking at Figure 11.12 and Figure 11.14, we noticed that the estimated proportions  $\hat{P}$  for the incomplete Q5b are higher than the estimates from the imputed data. This is mainly due to the fact that Q5bs were updated by the imputed Q5as before applying imputation methods to the remaining missing Q5b data. As the majority of answers to the imputed Q5as were “1=Yes”, this means that Q5bs would get a large number of “2=No” to replace their missing values, and the estimated proportions  $\hat{P}$  would become smaller after updating.

Another interesting thing we have noticed is that Table 11.33 and Figure 11.14 show that most imputation methods produce the same or very close estimated proportions  $\hat{P}$  and standard error of  $\hat{P}$  for Q5b 2007. This is purely because there are only 2 observations with missing Q5b after updating by the imputed Q5a 2007. We think imputing two missing values does not make much change to the estimates. This is also why Bayesian MI does not produce larger  $se$  and wider confidence intervals for Q5b 2007. There wouldn’t be much imputation uncertainty if the dataset only has a very few missing data. However, the bootstrap resampling method produced larger  $se$  and wider confidence intervals, although they are just slightly bigger. We think this larger  $se$  is more likely to be the result of resampling than the imputation uncertainty. Hence, this suggests that the bootstrap resampling method might potentially overestimate imputation uncertainty compared to the Bayesian MI.

Table 11.30: Imputation results for 2009 School FNES Q5a

Q5a, the number of missing observations is 148				
Imputation method	1=“Yes”	2=“No”	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	843	146	0.8860	0.0067
Conditional Mode	978	159	0.8922	0.0059
Hot deck within adjustment cells	957	180	0.8779	0.0062
NN hot deck	954	183	0.8768	0.0061
Logistic regression	958	179	0.8795	0.0061
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.8772	0.0090
EM Algorithm	978	159	0.8922	0.0059
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.8755	0.0070

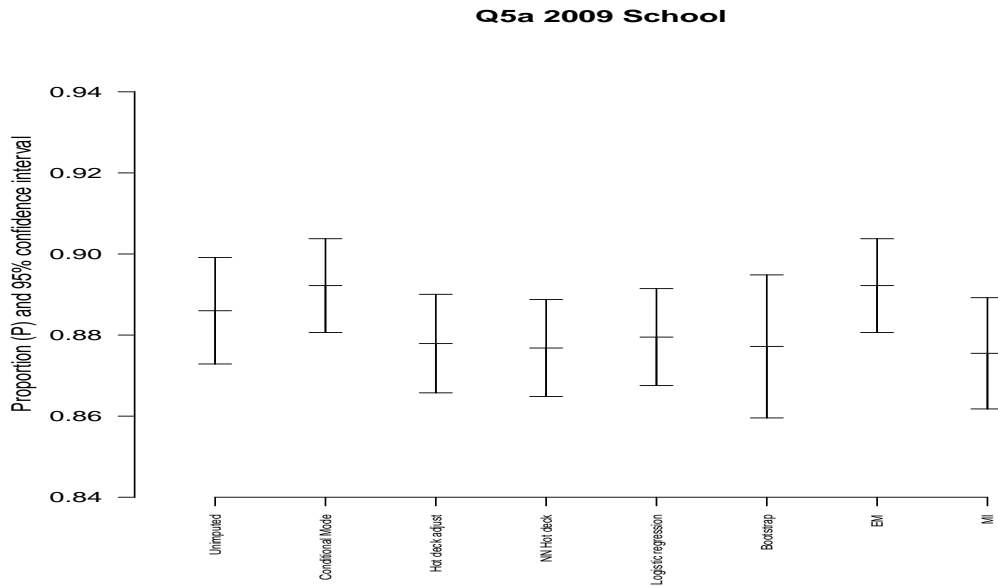


Figure 11.11: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q5a 2009 School

Table 11.31: Imputation results for 2009 School FNES Q5b

Q5b, the number of missing observations is 136				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	158	843	0.1223	0.0069
Conditional Mode	158	979	0.1068	0.0058
Hot deck within adjustment cells	164	973	0.1105	0.0059
NN hot deck	163	974	0.1094	0.0058
Logistic regression	165	972	0.1105	0.0058
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.1103	0.0081
EM Algorithm	158	979	0.1068	0.0058
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.1108	0.0060

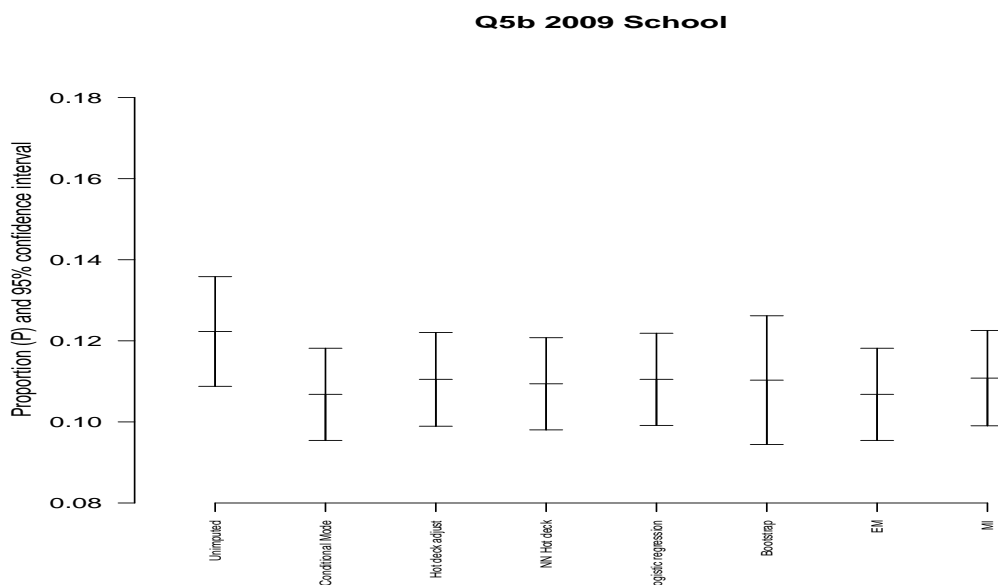


Figure 11.12: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q5b 2009 School

Table 11.32: Imputation results for 2007 School FNES Q5a

Q5a, the number of missing observations is 46				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	633	66	0.9268	0.0073
Conditional Mode	656	89	0.9064	0.0080
Hot deck within adjustment cells	654	91	0.9040	0.0081
NN hot deck	652	93	0.9009	0.0083
Logistic regression	653	92	0.9032	0.0083
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.9035	0.0096
EM Algorithm	656	89	0.9064	0.0080
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.9022	0.0089

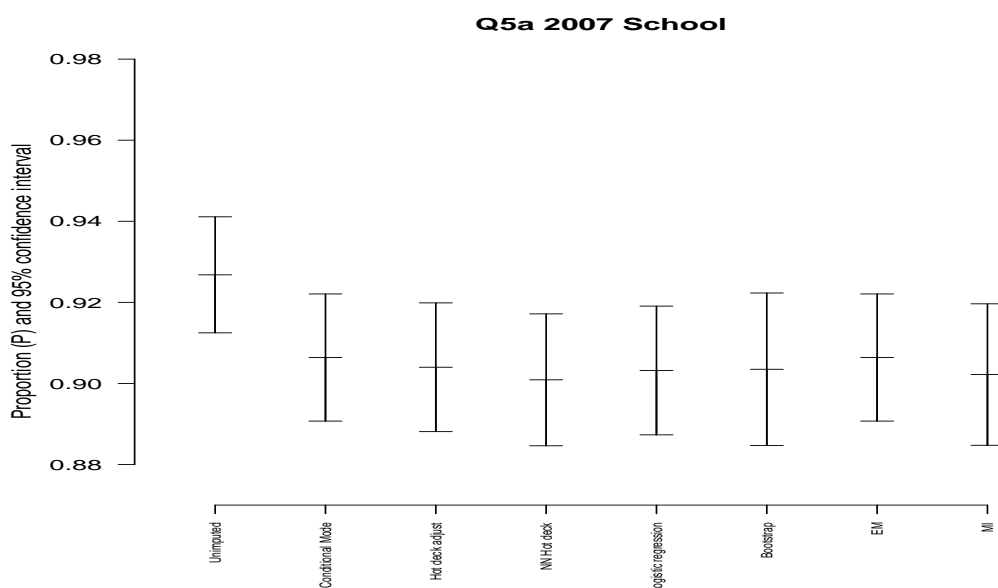
Figure 11.13: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q5a 2007 School

Table 11.33: Imputation results for 2007 School FNES Q5b

Q5b, the number of missing observations is 108				
Imputation method	1="Yes"	2="No"	Estimated Proportion( $\hat{P}$ )	Standard Error of $\hat{P}$
Unimputed	232	405	0.3292	0.0161
Conditional Mode	232	513	0.2815	0.0139
Hot deck within adjustment cells	233	512	0.2823	0.0139
NN hot deck	232	513	0.2815	0.0139
Logistic regression	232	513	0.2815	0.0139
Bootstrap Resampling (B=200)	Not Applicable	Not Applicable	0.2834	0.0155
EM Algorithm	232	513	0.2815	0.0139
Bayesian Multiple Imputation (D=5)	Not Applicable	Not Applicable	0.2832	0.0139

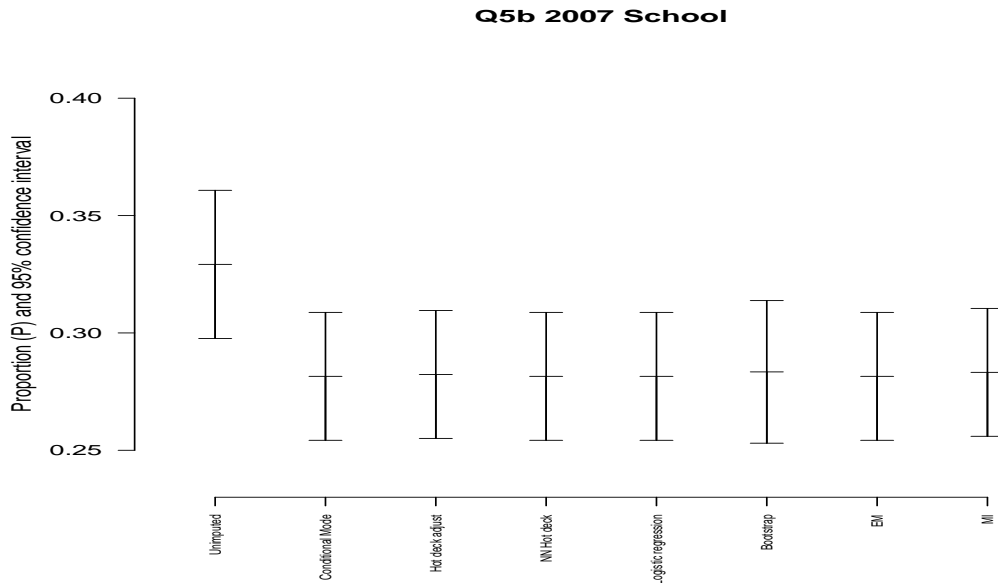


Figure 11.14: The 95% confidence interval of proportion  $\hat{P}$  for the imputed Q5b 2007 School

### 11.3.5 Imputation using the matched 2007 and 2009 ECE FNES sample

As given in Table 11.8, there are 109 ECE services in both 2007 and 2009 ECE FNES samples. These matched units can provide us with some extra information when we want to impute the 2007 or 2009 ECE samples. A simple method of utilising this extra information is to use cold deck imputation to replace the missing 2009 data with the matched 2007 data. However, as shown in Table 11.34, after updating the 2009 ECE FNES data with the matched 2007 data, for questions Q3a, Q7b, and Q7c, not all the missing data of the matched 2009 data have been replaced with the 2007 data due to the matched 2007 data having missing data as well. More importantly, the cold deck only updates the variables we want to impute by using the matched data. There are other variables in the matched data as well, such as the variables which relate to the missingness. Hence, the cold deck imputation method does not utilise all the available and useful information in the matched data.

Table 11.34: Update the matched 2009 ECE FNES data with the matched 2007 data

The matched 2009 data	1="Yes"	2="No"	Missing
<b>Q3a</b>			
Before updating	23	59	27
After updating	23	73	14
<b>Q7b</b>			
Before updating	21	57	31
After updating	24	78	7
<b>Q7c</b>			
Before updating	30	50	29
After updating	35	66	8

In terms of utilising all the useful information in the matched data, a better way is to use imputation methods which properly deal with the MAR missingness. Then, there are two

approaches: the separate datasets approach, and the combined datasets approach. The separate datasets approach treats the matched part of the data, and the non-matched part of the data as two separate datasets. Imputation methods are applied to the two datasets separately. Most imputation methods we have introduced can be used for this approach. The combined datasets approach treats the matched and non-matched parts of the data as a whole. Bayesian related imputation methods (e.g. MI) can be used to impute such datasets. The imputation is done through Bayesian iterations:

Step 1: Impute the matched part of the data

Step 2: Impute the rest of missing data based on the imputed data from step 1 and the observed data

Step 3: Repeat Step 1 and Step 2 until the estimates converges

In order for the iteration to work, the imputed data from step 1 must have slightly different values each time. Although this can be done by some single imputation methods, such as a stochastic regression model, or simple hot deck imputation, we think it performs the best under Bayesian imputation scheme. Before introducing this approach in greater detail by using the 2007 and 2009 ECE FNES samples, we want to point out that the combined datasets approach is better than the separate datasets approach in terms of using the extra information from the matched dataset. This is because all the information we get from the matched dataset has been passed in to impute the missing data in the non-matched part of the data.

We have applied Bayesian MI for the combined datasets approach. Suppose  $Y$  is the variable with missing data in the 2009 ECE FNES sample,  $X$  are the 2009 variables that are related to the missingness of  $Y$ , and  $X$  is complete. The 2007 ECE FNES sample is matched to the 2009 sample.  $Z$  are the matched 2007 variables that are related to the missingness of  $Y$  of the matched part.  $Z$  can include the  $Y$  from 2007 ECE FNES sample. This means we may use the value of  $Y$  from 2007 to predict the missing value of  $Y$  in the 2009 sample. The use of previous  $Y$  to impute the current  $Y$  was introduced later in this section. For now, we focus on the simpler approach which does not include  $Y$  from 2007 sample.

**The simple approach:**  $Y$ ,  $X$  and  $Z$  have four forms:  $Y_A$ ,  $X_A$ , and  $Z_A$  are for the matched units with observed  $Y$  values;  $Y_B$ ,  $X_B$ , and  $Z_B$  are for the matched units with missing  $Y$  values;  $Y_C$  and  $X_C$  are for the non-matched units with observed  $Y$  values;  $Y_D$  and  $X_D$  are for the non-matched units with missing  $Y$  values. Figure 11.15 displays the idea graphically.

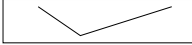
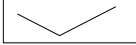
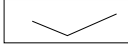
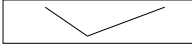

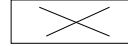
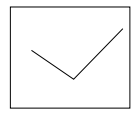
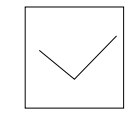
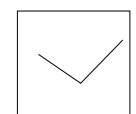
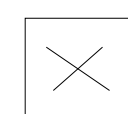
Z	X	Y	
			A
			B
			C
			D

Figure 11.15: Utilizing information from the matched data. Note: The tick ✓ means observed data, and the cross ✕ means missing data

Basically, we have two models:

$$Y|XZ; \alpha$$

and

$$Y|X; \beta$$

The former is a better model since it has a richer set of covariates ( $X$  and  $Z$ ), but only available for the matched data. We can however use it to impute  $Y_B$ , after that, relying on the inferior model  $Y|X; \beta$  to impute  $Y_D$ , given that  $Y_B$  is available for imputing  $Y_D$

The Bayesian part of the MI is:

At iteration  $t$ , we randomly draw  $Y_B$  and  $Y_D$  from their conditional distribution:

$$Y_B^t \sim p(Y_B|X_B, Z_B, \alpha^t)$$

$$Y_D^t \sim p(Y_D|X_D, \beta^t)$$

At iteration  $t + 1$ , we randomly draw  $\alpha^{t+1}$  and  $\beta^{t+1}$  from their conditional posterior distribution given the updated  $Y_B^t$  and  $Y_D^t$ , and the observed  $Y_A$  and  $Y_C$ :

$$\alpha^{t+1} \sim p(\alpha|Y_A, Y_B^t, X_A, X_B, Z_A, Z_B)$$

$$\beta^{t+1} \sim p(\beta|Y_A, Y_B^t, Y_C, Y_D^t, X_A, X_B, X_C, X_D)$$

where  $\alpha$  and  $\beta$  are the logistic regression parameters:

$$\text{logit}[\pi_{AB,i}] = (xz)_{AB,i}^T \alpha$$

and

$$\text{logit}[\pi_i] = x_i^T \beta$$

As we can see, every time the  $Y_B$  is updated, it is used to update the  $\beta$ , then the updated  $\beta$  is used to update  $Y_D$ . This means that we have incorporated  $Z$  in the process of imputing  $Y_B$  and  $Y_D$ .



For demonstration and description purposes, we use the Q3a of 2009 ECE FNES sample as the  $Y$  variable. As briefly introduced before, we first used a simple approach which does not involve  $Y$  from 2007 as one of the variables for  $Z$ , then we increased the complexity of our models (the complex approach) to include  $Y$  from 2007 sample. Originally, we want to just use the same explanatory variables: Stratum and Super region, which have been used in previous sections, but after matching, these variables for 2007 and 2009 ECE FNES samples have identical values for each matched units because they are the sample design variables. Therefore, they are not good candidates for  $Z$ . The solution we propose is to add other variables on top of the sample design variables. For the simple approach without involving  $Y$  from 2007, we add couple of non sample design variables from 2007 sample to be our  $Z$ , and the  $X$  includes both non sample design variables and the added variables from 2009. For the complex approach, the sample design variables were used to predict missing data of  $Y$  from 2007 sample in order to provide a complete  $Y$  variable from 2007 to  $Z$ , and they also were used as one of the variables for  $X$ .

Hence, for the simple approach, we select Q3b ("Drinking water available to children through water coolers"), and Q3c ("Drinking water available to children through tap water") from both 2007 and 2009 sample data to be the  $X$  and  $Z$  variables. To be clear, the  $X$  variable is Q3b, Q3c, stratum and super region from the 2009 ECE FNES sample, and the  $Z$  variable is Q3b and Q3c from the 2007 ECE FNES sample. However, all the Q3b and Q3c variables also have missing data in both years' samples. This violates the assumption that we have complete explanatory variables when imputing the response variable. Hence, we need to impute the Q3b and Q3c first. Because we only want to show how to do the MI for the combined datasets approach once, we short-cut the imputation for Q3b and Q3c by simply applying the adjustment cells hot deck imputation. Hence, we have the following for  $Y$ ,  $X$  and  $Z$ :

$$\begin{aligned} Y &= Q3a(2009) \\ X &= (Q3b, Q3c, \text{stratum}, \text{super region})(2009) \\ Z &= (Q3b, Q3c)(2007) \end{aligned}$$

The following steps show how exactly we apply the MI for the combined datasets approach

**Apply Multiple Imputation for the combined datasets approach**

**Step 1:** Impute the  $Z = (Q3b, Q3c)$  from the 2007 and 2009 ECE FNES data by using the adjustment hot deck imputation. The imputation cells were formed by the sample design variables: Stratum and Super region

**Step 2:** Initial parameters: estimate  $\alpha^0$  based on the matched units with observed 2009 Q3a values; estimate  $\beta^0$  based on the observed 2009 Q3a. Then, random noise generated from normal distribution are added to the initial parameters. The normal distributions have 0 means and 10 times the standard deviations of  $\alpha$  and  $\beta$ .

$$\alpha^0 = \alpha^0 + noise, \quad noise \sim N(0, 10 \times sd_\alpha) \quad (11.8)$$

$$\beta^0 = \beta^0 + noise, \quad noise \sim N(0, 10 \times sd_\beta) \quad (11.9)$$

This is because we want to produce multiple chains with different starting points.  $m = 5$  chains have been produced.

**Step 3:** Assuming the prior distributions are  $p(\alpha) \propto 1$ , and  $p(\beta) \propto 1$ .

Draw  $Y_B$ :

$$Y_B | \alpha \sim \text{Bernoulli} \left( \frac{\exp((xz)_i^T \alpha^t)}{1 + \exp((xz)_i^T \alpha^t)} \right)$$

Draw  $Y_D$ :

$$Y_D | \beta \sim \text{Bernoulli} \left( \frac{\exp((x)_i^T \beta^t)}{1 + \exp((x)_i^T \beta^t)} \right)$$

**Step 4:** Draw  $\alpha^{t+1} \sim p(\alpha | Y_A, Y_B^t, X_A, X_B, Z_A, Z_B)$ , by using the MH algorithm. The proposal distribution is  $N(\alpha^t, \Sigma_{\alpha^0})$   
 Draw  $\beta^{t+1} \sim p(\beta | Y_A, Y_B^t, Y_C, Y_D^t, X)$ , by using the MH algorithm. The proposal distribution is  $N(\beta^t, \Sigma_{\beta^0})$   
 $\Sigma_{\alpha^0}$  and  $\Sigma_{\beta^0}$  are kept constant during the process

**Step 5:** Compute the proportion  $\hat{P}$  and the standard error of  $\hat{P}$  ( $se$ ) based on the updated  $Y$ .

**Step 6:** For each Bayesian chain, repeat step 3 to step 5 until the estimates  $\hat{P}$  and  $se$  convergence. The convergence is diagnosed by using the time series plots and Gelman and Rubin's method which have been introduced in Chapter 7

**The choice of initial parameters:** we decided to have  $m = 5$  Bayesian simulation chains. After the chains converge, the datasets that generated by the final iteration of those chains were selected as our Bayesian MI datasets. Hence, we need five starting points<sup>10</sup> for the five Bayesian simulation chains. The choice of the starting points can be arbitrary (SAS 2008). However, we would like our starting points to be well dispersed but relative to the posterior distributions. Doing this can avoid the chains converging to a local maximum and check for stability in our estimates. Therefore, we first need to have some understanding of what kind of distributions  $\alpha$  and  $\beta$  may have. Then, we estimate their standard deviations  $sd_\alpha$  and  $sd_\beta$ , and substitute their values in Eq (11.8) and Eq (11.9) in Step 2. As described in Step 2, we draw random noise from normal distributions with 0 means and 10 times the standard deviations of  $\alpha$  and  $\beta$ . We believe that doing this can make sure our choice of starting points are well dispersed.

We found the possible distributions of  $\alpha$  and  $\beta$  by setting the  $\alpha^0$  and  $\beta^0$  without adding the random noise and repeating Step 3 and Step 4 for 5000 times. Figure 11.16 and Figure 11.17 display the distributions of  $\alpha$  and  $\beta$ .

<sup>10</sup>Each of starting points is a set of initial parameters

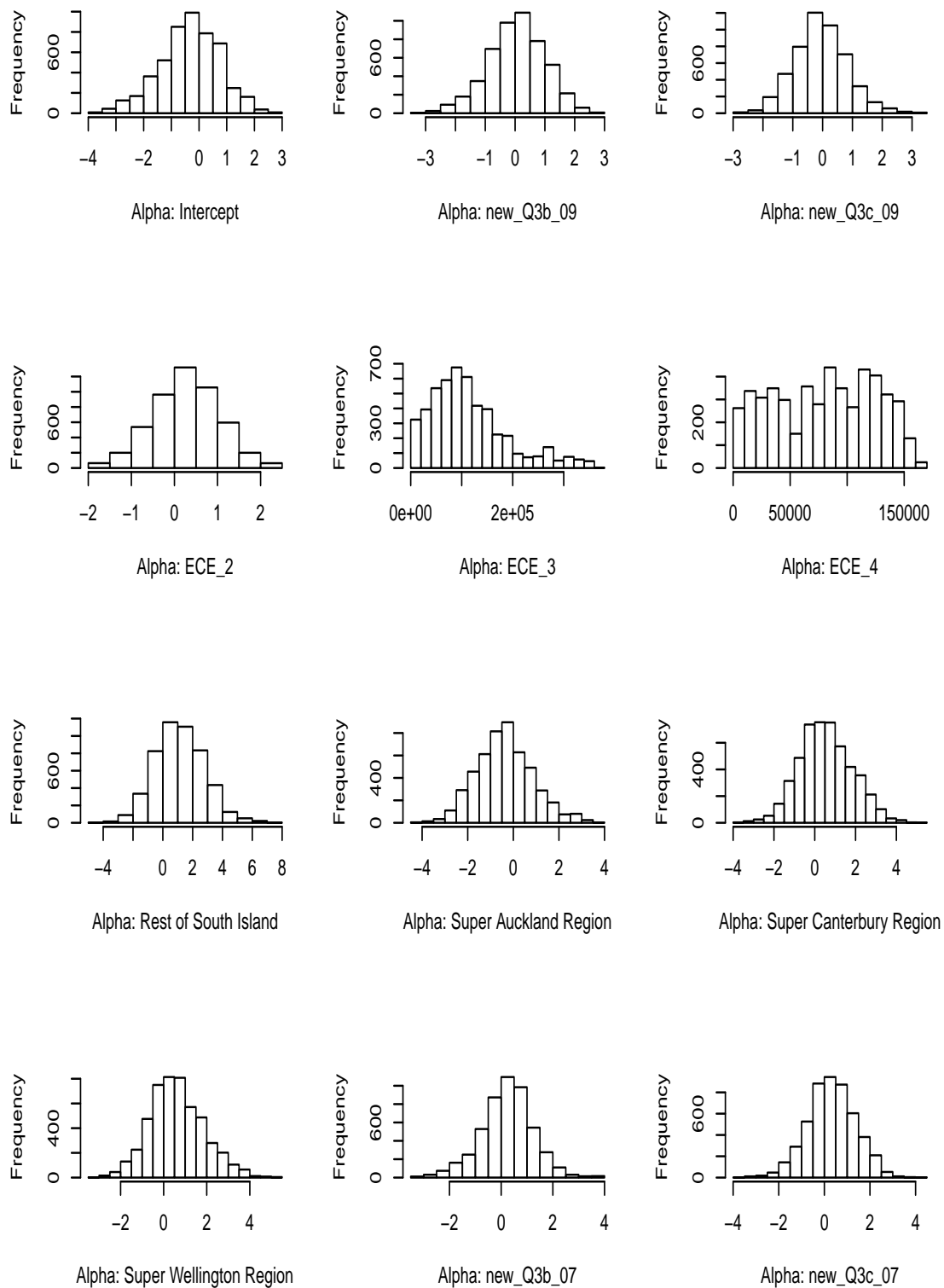


Figure 11.16: Distributions of  $\alpha$

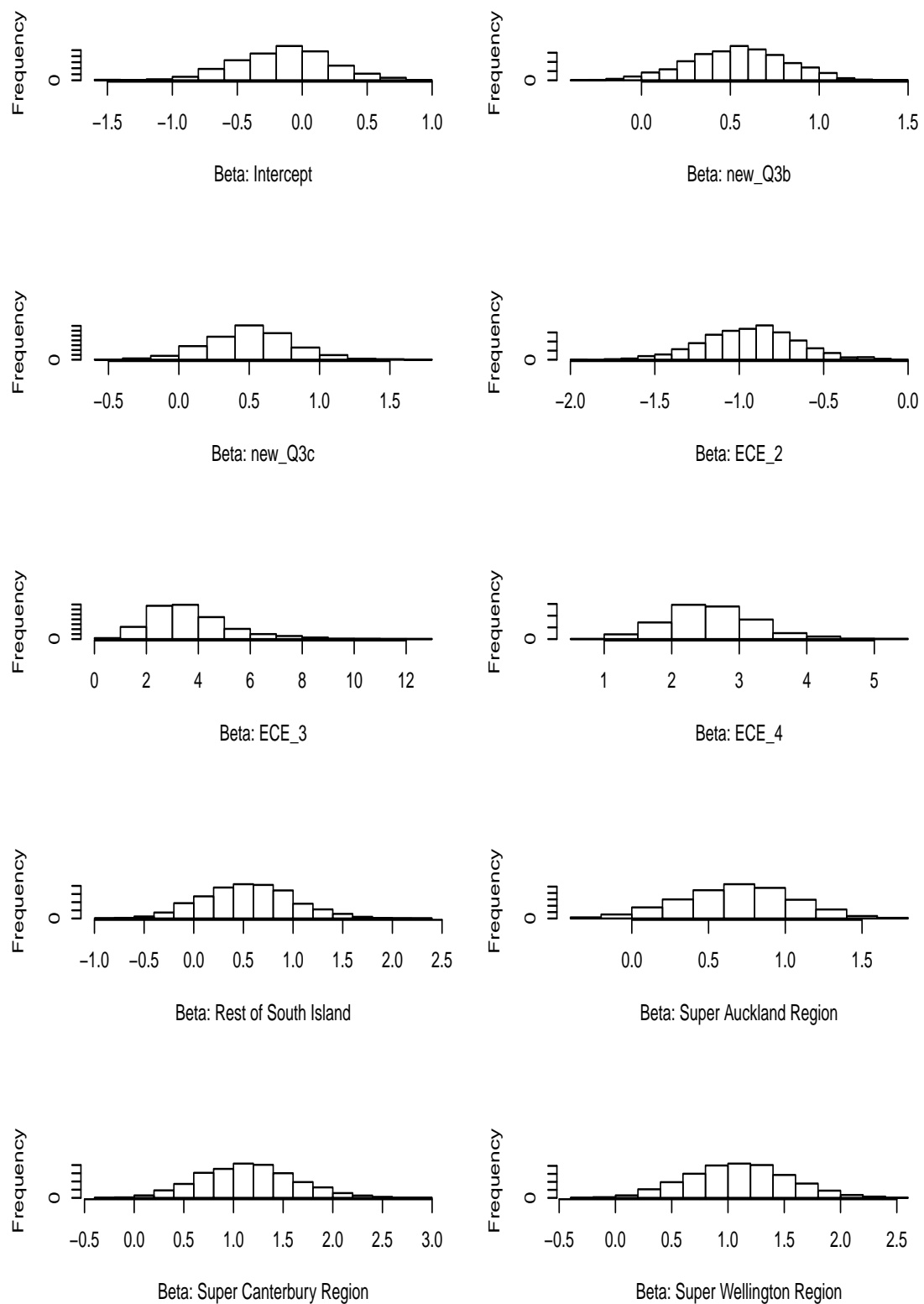


Figure 11.17: Distributions of  $\beta$

**Convergence diagnostics for the simple approach:** as introduced in Chapter 7, we first drew time series plots of the estimates of the proportion  $\hat{P}$  and the standard error ( $se$ ) of  $\hat{P}$ . These plots can give us some indication of how long the chains need to be in order to converge. Figure 11.18 gives us the results. By looking at the time series plots, it seems that the five chains for  $\hat{P}$  and  $se$  become very stable after the first 2000 iterations. This means that the chains were converging within 5000 iterations.



Figure 11.18: Time series plots of  $\hat{P}$  and  $se$  for  $m = 5$  Bayesian chains

With the indication from the time series plots, we then applied Gelman and Rubin’s diagnostic method in order to measure the convergence precisely. The R code for the Gelman and Rubin diagnostic are obtained from the “coda” package in R. There are both numerical and graphical results. The “gelman.diag()” function gives us numerical results:

```
> gelman.diag(mh.list_prop)
Potential scale reduction factors:
```

	Point est.	Upper C.I.
[1,]	1.01	1.04

```
> gelman.diag(mh.list_se)
Potential scale reduction factors:
```

	Point est.	Upper C.I.
[1,]	1.01	1.04

The results given are the median Potential scale reduction factor (PSRF) and its 97.5% quantile. The PSRF is equivalent to the statistic  $R$  which we have introduced in Section 7.4.3, Chapter 7. The rule of thumb is that the convergence is reached once the  $PSRF < 1.2$  (Bolker 2011). As shown, the PSRFs for the chains of  $\hat{P}$  and the chains of  $se$  are all less than 1.2. Therefore, we concluded that the Bayesian chains converge with 5000 iterations.

We can also show how the PSRF changes through the iterations using the “gelman.plot()” function. Figure 11.19 shows the results. As can be seen, the PSRFs for both  $\hat{P}$  and  $se$  were indeed starting to decline after the first 2000 iterations. This means that the Bayesian chains were starting to become stable after the 2000th iteration. This confirms our initial guessing of convergence by investigating the time series plots.

Based on the convergence diagnostics, we set the number of iterations for Bayesian chains to be 5000 and combined the last estimates of each  $m = 5$  chains to be the MI estimates.

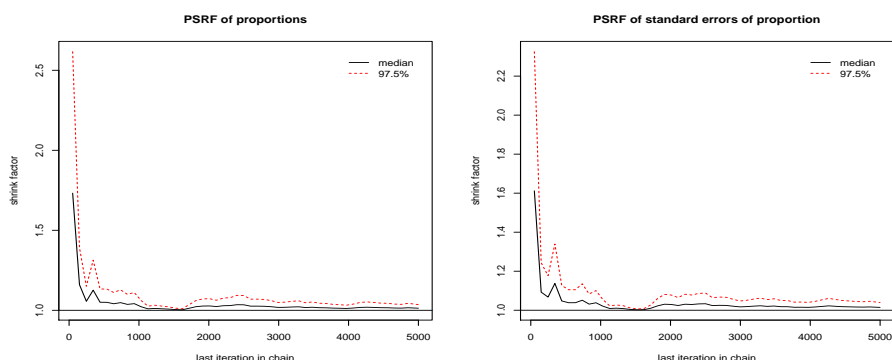


Figure 11.19: The simple approach: plot of Gelman-Rubin PSRF by iteration for  $\hat{P}$  and standard error of  $\hat{P}$

**The results of the simple approach:** we have applied similar steps to impute all the listed ECE variables in Table 11.7 for both 2007 and 2009 data. Table 11.35 and Table 11.36 display the results. The R code for imputing Q3a 2009 is in Appendix F, Section F.2.1.

Table 11.35: Impute 2009 ECE FNES with the matched 2007 data by using MI

	1="Yes"	2="No"	Proportion (P)	Standard Error of P
Q3a 2009	Not Applicable	Not Applicable	0.2613	0.0215
Q7b 2009	Not Applicable	Not Applicable	0.3840	0.0195
Q7c 2009	Not Applicable	Not Applicable	0.2403	0.0172

Table 11.36: Impute 2007 ECE FNES with the matched 2009 data by using MI

	1="Yes"	2="No"	Proportion (P)	Standard Error of P
Q3a 2007	Not Applicable	Not Applicable	0.2676	0.023
Q7b 2007	Not Applicable	Not Applicable	0.3389	0.0231
Q7c 2007	Not Applicable	Not Applicable	0.2603	0.0167

**The complex approach:** As mentioned previously, a better way of selecting the variables for  $Z$  is to include the 2007 Q3a variable. This is because the 2007 Q3a which is likely to have similar values as the 2009 Q3a can provide better prediction power for the logistic regression. As the 2007 Q3a is incomplete as well, we now have three models:

$$Y|X, Z(Y_Z, X_Z); \alpha$$

and

$$Y|X; \beta$$

and

$$Y_Z|X_Z, \gamma$$

The added model  $(Y_Z|X_Z, \gamma)$  is used to impute missing 2007 Q3a data.

Of course, we can short-cut the imputation for Q3a of 2007 by using the adjustment cells hot deck imputation method, but the better way is to simply add a step to the Bayesian iterative simulation in which the missing values of Q3a of 2007 are replaced with simulated values, given the Q3b and Q3c of 2007 have been imputed beforehand. This is a better way of imputing the response variable, because we have incorporated the imputation uncertainty that was introduced by imputing the explanatory variables as well. Then, this actually extends the use of Bayesian MI from imputing response variables only to imputing response and explanatory variables together. Hence, we have the following for  $Y$ ,  $X$ , and  $Z$ :

$$Y = Q3a(2009)$$

$$X = (Q3b, Q3c, \text{stratum}, \text{super region})(2009)$$

$$Z = (Q3a, Q3b, Q3c, \text{stratum}, \text{super region})(2007)$$

In order to simplify our description, we separated the Q3a of 2007 from the  $Z$  group of variables, and gave it the symbol  $Y_{07}$ . There are:  $Y_{07A}$  for the matched observed units,  $Y_{07B}$  for the matched unobserved units,  $Y_{07E}$  for the unmatched observed units, and  $Y_{07F}$  for the

unmatched unobserved units. Then, we have:

$$\begin{aligned}
Y &= Q3a(2009) \\
X &= (Q3b, Q3c, stratum, super region)(2009) \\
Z &= (Q3b, Q3c, stratum, super region)(2007) \\
Y_{2007} &= Q3a(2007)
\end{aligned}$$

Figure 11.20 shows the idea graphically.

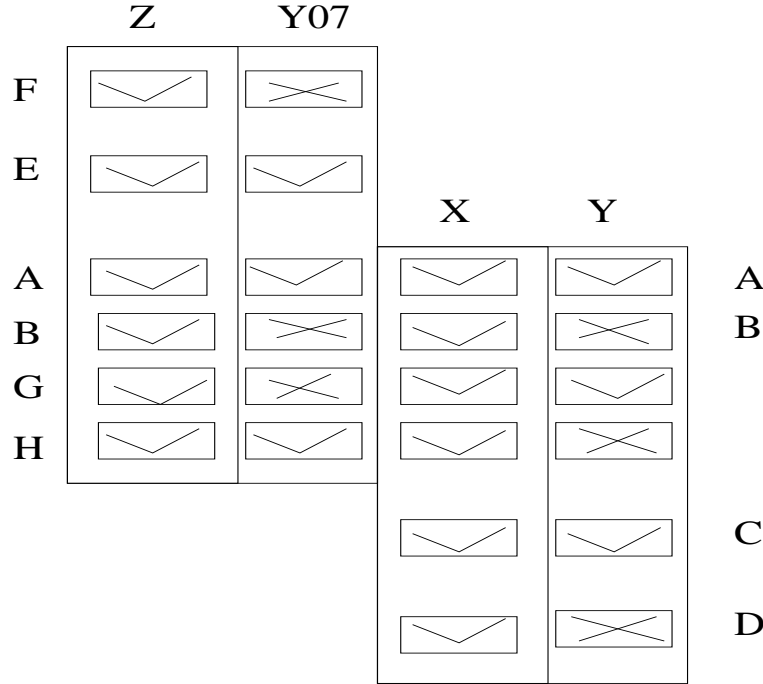


Figure 11.20: Utilizing information from the matched and unmatched data. Note: ✓ means observed data, and ✗ means missing data

Then, the Bayesian part of the MI is:

At iteration  $t$ , we randomly draw  $(Y_{2007B}, Y_{2007F}, Y_{2007G})$ ,  $(Y_B, Y_H)$ , and  $Y_D$  from their conditional distribution

$$\begin{aligned}
(Y_{2007B}^t, Y_{2007F}^t, Y_{2007G}^t) &\sim p((Y_{2007B}, Y_{2007F}, Y_{2007G}) | Z_B, Z_F, z_G \gamma^t) \\
(Y_B^t, Y_H^t) &\sim p(Y_B, Y_H | X_B, X_H, Y_{2007B}^t, Y_{2007H}, Z_B, z_H \alpha^t) \\
Y_D^t &\sim p(Y_D | X_D, \beta^t)
\end{aligned}$$

where  $\gamma$  is the logistic regression parameters:  $\text{logit}[\pi_{Y_{2007}, i}] = z_i^T \gamma$ .

At iteration  $t + 1$ , we randomly draw  $\gamma^{t+1}$ ,  $\alpha^{t+1}$  and  $\beta^{t+1}$  from their conditional posterior distribution given the updated  $(Y_{2007B}^t, Y_{2007F}^t, Y_{2007G}^t)$ ,  $(Y_B^t, Y_H^t)$ , and  $Y_D^t$ , and the observed  $Y_{2007A}$ ,  $Y_{2007E}$ ,  $Y_{2007H}$ ,  $Y_A$ ,  $Y_C$  and  $Y_G$ :

$$\begin{aligned}
\gamma^{t+1} &\sim p(\gamma | Y_{2007A}, Y_{2007B}^t, Y_{2007E}, Y_{2007F}^t, Y_{2007G}^t, Y_{2007H}, Z_A, Z_B, Z_E, Z_F, Z_G, Z_H) \\
\alpha^{t+1} &\sim p(\alpha | Y_A, Y_B^t, Y_G, Y_H^t, Y_{2007A}, Y_{2007B}^t, Y_{2007G}^t, Y_{2007H}, X_A, X_B, X_G, X_H, Z_A, Z_B, Z_G, Z_H) \\
\beta^{t+1} &\sim p(\beta | Y_A, Y_B^t, Y_C, Y_D^t, Y_G, Y_H^t, X_A, X_B, X_C, X_D, X_G, X_H)
\end{aligned}$$

As, we can see, the  $\gamma^t$  is updated based on the information from  $Y_{2007A}$ ,  $Y_{2007B}^t$ ,  $Y_{2007E}$ ,  $Y_{2007F}^t$ ,  $Y_{2007G}^t$ ,  $Y_{2007H}$ ,  $Z_A$ ,  $Z_B$ ,  $Z_E$ ,  $Z_F$ ,  $Z_G$ ,  $Z_H$ . This means that the unmatched part of the 2007 data has also been incorporated into the Bayesian MI. This gives better simulated  $Y_{2007B}^t$  values.



The exact steps are as follows:

**Apply Multiple Imputation to impute the missing response and explanatory variables for the combined datasets approach**

- Step 1:** Impute the  $Z = (Q3b, Q3c)$  from the 2007 and 2009 ECE FNES data as previously introduced
- Step 2:** Initial parameters: estimate the  $\gamma^0$  based on the observed 2007 Q3a; estimate the  $\alpha^0$  based on the matched units with observed 2009 Q3a values; estimate  $\beta^0$  based on the observed 2009 Q3a. Then, random noise generated from a normal distribution is added to the initial parameters. The normal distributions have 0 means and 10 times the standard deviations of  $\gamma$ ,  $\alpha$ , and  $\beta$ .

$$\begin{aligned}\gamma^0 &= \gamma^0 + \text{noise}, \quad \text{noise} \sim N(0, 10 \times sd_\gamma) \\ \alpha^0 &= \alpha^0 + \text{noise}, \quad \text{noise} \sim N(0, 10 \times sd_\alpha) \\ \beta^0 &= \beta^0 + \text{noise}, \quad \text{noise} \sim N(0, 10 \times sd_\beta)\end{aligned}$$

$m = 5$  chains have been produced. The initial distributions of  $\gamma$ ,  $\alpha$ , and  $\beta$  were found as the procedures introduced in the simple approach.

- Step 3:** Assuming the prior distributions are  $p(\gamma) \propto 1$ ,  $p(\alpha) \propto 1$ , and  $p(\beta) \propto 1$ .  
Draw  $(Y_{2007B}^t, Y_{2007F}^t, Y_{2007G}^t)$ :

$$(Y_{2007B}^t, Y_{2007F}^t, Y_{2007G}^t) | \gamma \sim \text{Bernoulli} \left( \frac{\exp((z)_i^T \gamma^t)}{1 + \exp((z)_i^T \gamma^t)} \right)$$

Draw  $(Y_B^t, Y_H^t)$ :

$$(Y_B^t, Y_H^t) | \alpha \sim \text{Bernoulli} \left( \frac{\exp((xzy_{2007})_i^T \alpha^t)}{1 + \exp((xzy_{2007})_i^T \alpha^t)} \right)$$

Draw  $Y_D^t$ :

$$Y_D^t | \beta \sim \text{Bernoulli} \left( \frac{\exp((x)_i^T \beta^t)}{1 + \exp((x)_i^T \beta^t)} \right)$$

- Step 4:** Draw  $\gamma^{t+1} \sim p(\gamma | Y_{2007A}, Y_{2007B}^t, Y_{2007E}^t, Y_{2007F}^t, Y_{2007G}^t, Y_{2007H}, Z_A, Z_B, Z_E, Z_F, Z_G, Z_H)$ , by using the MH algorithm. The proposal distribution is  $N(\gamma^t, \Sigma_{\gamma^0})$   
Draw  $\alpha^{t+1} \sim p(\alpha | Y_A, Y_B^t, Y_G, Y_H^t, Y_{2007A}, Y_{2007B}^t, Y_{2007G}^t, Y_{2007H}, X_A, X_B, X_G, X_H, Z_A, Z_B, Z_G, Z_H)$ , by using the MH algorithm. The proposal distribution is  $N(\alpha^t, \Sigma_{\alpha^0})$   
Draw  $\beta^{t+1} \sim p(\beta | Y_A, Y_B^t, Y_C, Y_D^t, Y_G, Y_H^t, X_A, X_B, X_C, X_D, X_G, X_H)$ , by using the MH algorithm. The proposal distribution is  $N(\beta^t, \Sigma_{\beta^0})$   
 $\Sigma_{\gamma^0}$ ,  $\Sigma_{\alpha^0}$  and  $\Sigma_{\beta^0}$  were kept constant during the process
- Step 5:** Compute the proportion  $\hat{P}$  and the standard error of  $\hat{P}$  ( $se$ ) based on the updated  $Y$ .
- Step 6:** For each Bayesian chain, repeat step 3 to step 5 until convergence. The convergence diagnostics have the same procedures as the introduced simple approach which are the time series plots and the Gelman and Rubin's diagnostic method.

In this particular example, there is a great chance that a perfect fit will happen by using Q3a 2007 from the matched data. This is because it is highly likely that the same respondents provide the same answers to Q3a in 2007 and 2009 FNES surveys. If this happens, there is no need to continue updating  $\alpha$  through the MH algorithm as the best estimation of  $\alpha$  has been found. Then, we only need to draw  $(Y_B^t, Y_H^t)$  from its posterior distribution with constant  $\alpha$ .

The convergence diagnostic procedures were the same as the simple approach. However, compared to the simple approach, the complex approach needs at least 20000 iterations to converge. With the time series plots omitted, the Gelman and Rubin's diagnostic results are as follows:

```
> gelman.diag(mh.list_prop)
Potential scale reduction factors:
```

```
      Point est. Upper C.I.
[1,]      1.19      1.45
```

```
> gelman.diag(mh.list_se)
Potential scale reduction factors:
```

```
      Point est. Upper C.I.
[1,]      1.17      1.41
```

Figure 11.21 displays the time series plots of Gelman-Rubin PSRF by 20000 iterations. The convergence diagnostics show that the Bayesian chains converge within 20000 iterations, but longer chains may produce better results.

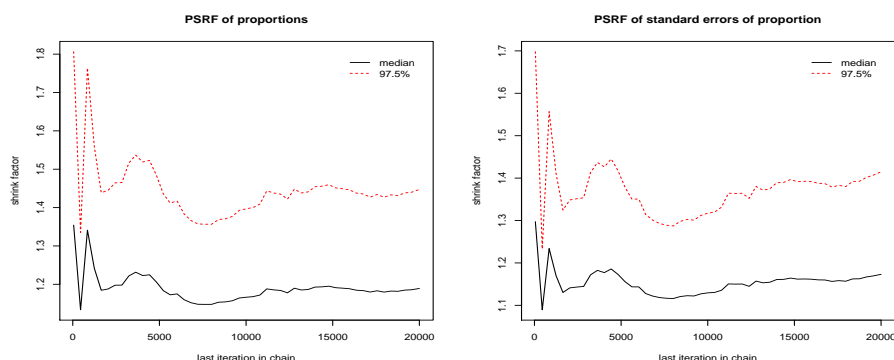


Figure 11.21: The complex approach: plot of Gelman-Rubin PSRF by iteration for  $\hat{P}$  and standard error of  $\hat{P}$

Again, we have applied similar steps to impute all the listed variables in Table 11.7 for both 2007 and 2009 data. Table 11.37 and Table 11.38 display the results. The R code for imputing Q3a 2009 and Q3a 2007 is in Appendix F, Section F.2.2.

Table 11.37: Impute 2009 ECE FNES with the matched 2007 data by using MI on both response and explanatory variables

	1="Yes"	2="No"	Proportion (P)	Standard Error of P
Q3a 2009	Not Applicable	Not Applicable	0.2742	0.0245
Q7b 2009	Not Applicable	Not Applicable	0.3876	0.0190
Q7c 2009	Not Applicable	Not Applicable	0.2413	0.0154

Table 11.38: Impute 2007 ECE FNES with the matched 2009 data by using MI on both response and explanatory variables

	1="Yes"	2="No"	Proportion (P)	Standard Error of P
Q3a 2007	Not Applicable	Not Applicable	0.2478	0.0210
Q7b 2007	Not Applicable	Not Applicable	0.3113	0.0210
Q7c 2007	Not Applicable	Not Applicable	0.2715	0.0157

Comparing the results of imputation using the matched 2007 and 2009 ECE FNES sample with the previous Bayesian MI imputation results (Section 11.3.3) from unmatched samples, it is hard to tell which approach is better as they should all produce unbiased estimates and properly reflect the imputation uncertainty. However, we believe that using more information that relate to the missingness and the variables we want to impute should produce better imputed values than using less information, as the models utilizing more information possess more prediction or imputation power. Hence, if partially matched datasets are available, we recommend using Bayesian MI with our proposed complex approach to incorporate the extra information from the matched datasets.

## 11.4 Discussion

In this Chapter, we have conducted some basic EDA to explore the missing data pattern of the FNES data, and tried to impute the missing data by applying the various of imputation methods introduced in previous Chapters, with the focus on Bayesian Multiple Imputation. Unlike performing imputation methods on the simple SURF data, there are many challenges we have faced when dealing with the real FNES data.

The first challenge is that there are far more variables and observations from the FNES than the simple SURF. This makes it is very difficult to investigate the missing data patterns by purely looking at the datasets only. Hence, we have carried out EDA to help us to investigate the missing data patterns by plotting the response rate of each variables on bar charts.

The second challenge is that we impute the FNES missing data under the assumption that the missingness is MAR and is related to other variables that have been observed as well, but this is just an assumption. Unlike previous Chapters, we created the MCAR or MAR missingness for the replicate SURF datasets, so the variables that are related to the missingness were already known. For the real FNES data, we do not know whether the missingness is MCAR, MAR or NMAR. This means we do not know which variables are related to the missingness which creates difficulties when we want to construct the best imputation model. Hence, in this Chapter, we have introduced the univariate comparison method and the logistic regression assessment method to help us to identify the variables which are related to the missingness.

The third challenge is to adapt our imputation methods to impute the interrelated variables. As described, the difficulty is that once the variable with missing data is imputed, then other variables that are related to that variable need to be updated as well, otherwise, the imputation results do not make practical sense. Our solution basically is to combine the deductive imputation method with other imputation methods to impute missing data in the interrelated variables. However, we have only imputed two variables that are related to each other. How to find an efficient algorithm to deal with large number of variables that are interrelated is something needed to be further studied.

The fourth challenge or improvement is to utilize the extra information we get by matching the 2007 and 2009 FNES datasets. The matched datasets are very useful in terms of enhancing our imputation methods. However, if it is only used for cold deck imputation, it is as though we are taking the jewel box but throwing away the jewellery within. Hence, we have proposed to incorporate the matched dataset in the Bayesian MI scheme. Doing this, we maximize the

information we can get from the matched and unmatched part of both 2007 and 2009 FNES data.

This Chapter also displays the imputation results for the selected FNES variables. As expected, the resampling method applied to imputed incomplete data and the Bayesian MI have the largest variances of estimates than the single imputation method and the EM algorithm.

To sum up, from our investigation, we think the Bayesian MI is the best imputation method for the FNES data. This is because it produces similar estimates to other imputation methods; it properly propagates the imputation uncertainty; and it is extremely flexible in the case of incorporating extra information from the matched datasets. This is also because we can construct familiar and reliable logistic models by using the FNES variables, which might not be the same case for other datasets.

# Chapter 12

## Some final thoughts

This chapter summarizes the previous chapters in this project, and proposes some thoughts on future work and improvements. Specifically, the first section summarizes the main points and findings from previous chapters, and the second section lists things that we haven't done, but could be done and improved in the future.

### 12.1 Summary of previous chapters

Chapter 2 focuses on introducing the three missing data mechanisms (MCAR, MAR, and NMAR). We have shown that the missing data do not cause biases only if they are MCAR. Both MAR and NMAR introduce bias to the estimates. This chapter paves the foundation of our discussion on how to deal with missing data in later chapters.

Chapter 3 exhibits most commonly used data deletion and imputation methods. This chapter also gives in-depth discussion on the concepts of non-response bias and imputation uncertainty. The main point is that the imputation methods are developed to tackle the bias issue if the missingness is MAR, but most imputation methods ignore the fact that they underestimate the imputation uncertainty due to treating the imputed values as true observed values.

Chapter 4 demonstrates how the various single imputation methods work in detail by applying them to the replicate SURF datasets with incomplete Income variable. We have shown that the imputation methods can reduce bias if they properly incorporate the MAR mechanism which means the imputation model includes the variables that are related to the missingness. We have also shown that the imputation methods, such as stochastic regression model, and hot deck imputation, perform better than other imputation methods which haven't gotten any random sampling mechanism. However, although some single imputation methods can deal with bias, none of them can reflect the imputation uncertainty.

Chapter 5 discusses two popular resampling methods (the bootstrap and the jackknife), and applies them to missing replicate SURF data to properly account for the imputation uncertainty. These methods are efficient for dealing with imputation uncertainty, but they also require large samples to achieve the desired results.

Chapter 6 introduces the EM algorithm which has been considered to be one of the best missing data handling technique. We have included in our introduction of the EM algorithm the case of multivariate missing data problems. Dealing with the multivariate missing data problem is one of EM's advantages, compared to single imputation methods. The reason that

we go through the EM algorithm is that researchers normally use the EM algorithm to find the initial estimates for the Bayesian MI.

Chapter 7 discusses the underlying Bayesian iterative simulation methods of the Bayesian MI. We focus on how to apply the Metropolis-Hastings (MH) algorithm and the Gibbs sampling algorithm to impute missing data, and compare the pros and cons of these two methods. Again, we have also extended our introduction of these algorithms to the case of multivariate missing data problems. This chapter also lists a few convergence diagnosis methods. This chapter has the foundation of the Bayesian MI we apply to the replicate SURF and the FNES data.

Chapter 8 shows how exactly Bayesian MI works, and how we pool the estimates from multiple imputed datasets together to compute the final MI estimates, and variances of estimates. This chapter also gives mathematical and simulation proofs of why and how the improper MI underestimates the variance of estimate.

Chapter 9 shows how to apply various imputation methods introduced in previous chapters to missing categorical data. These imputation methods have only been applied to continuous numerical missing data in previous chapters. In this Chapter, we show that, although the fundamental concepts of these imputation methods are the same, variations are needed in order to apply them to the missing categorical data. This chapter also prepares the use of these imputation methods for the FNES data as all of its variables are categorical variables.

Chapter 10 simply describes the sample design of the FNES data.

Chapter 11 uses EDA to investigate the missing data pattern of the FNES data. Then, this chapter introduces the univariate comparison method and logistic regression assessment method to detect the missing data mechanism and the variables that are related to the missingness. We start to introducing these detection methods for the missing data mechanism here, because of the need to detect the missing data mechanism and variables that related to missingness only arises when we deal with the real life social survey data. Finally, we have applied several imputation methods introduced in previous chapters to a few FNES variables. The results indicate that Bayesian MI produces estimates similar to other imputation methods, and it also gets similar variances of estimates to the bootstrap resampling methods. Furthermore, we propose the use of Bayesian MI for the case of partially matched datasets. Bayesian MI maximizes the information we can get from the matched and unmatched datasets in order to find the best imputation values. The model of Bayesian MI for the partially matched datasets is the new development in this project.

## 12.2 Future work

- Apply the EM algorithm and Bayesian MI to Multivariate data with missing values. We have touched this area in Chapter 6 and 7, and they are all for the case of multivariate numerical data. The multivariate data can be categorical data or a mixture of categorical and numerical data. Some theories of how to deal with this types of missing data problem have been developed by Schafer (2003), and Little & Rubin (2002), and others. It will be beneficial to investigate the missing data problem for Multivariate data and apply the methods to the FNES data.

- Develop imputation methods for categorical data. We have seen that the development of imputation methods for categorical data lags behind the the development for the numerical data. For example, the Groenewald & Mokgatlhe (2005) method we have introduced in Chapter 9 does not work for the situation that the explanatory variables  $X$  are also categorical data. Further modification of their method could be made.
- Investigate data editing methodologies for the FNES data. “Data editing is the activity aimed at detecting and correcting errors (logical inconsistencies) in data” (OECD 2001). Normally, the data editing is the step before imputation after we collect the raw data. The data editing process can affect the quality of the data and the imputation, due to its ability to deal with logical inconsistencies, outliers and typographic errors. For example, if the outliers can be removed or dealt with at the data editing stage, then our imputation can avoid imputing the outliers values to the missing data.
- Research the possible modification of other non-Bayesian related imputation methods for partially matched data. Since the purpose of this project is to focus on the discussion and development of Bayesian MI, we have only explored the use of Bayesian MI for the partially matched data. But, it is possible to apply the EM algorithm, or even resampling methods to the partially matched data as well.

# Bibliography

- Acock, A. C. (2005), 'Working with missing values', *Journal of Marriage and Family* **67**(4), 1012 – 1028.
- Ader, H. & Mellenbergh, G. (2008), *Advising on Research Methods: A consultant's companion*, second edn, The Netherlands: Johannes van Kessel Publishing, Huizen.
- Agresti, A. (2002), *Categorical Data Analysis*, John Wiley and Sons.
- Allison, P. (2002), *Missing data*, first edn, Sage, Thousand Oaks, California.
- Anderson, B. S. & Hardin, J. (2009), 'Modified logistic regression using the em algorithm for reject inference', *SAS* p. 6.
- Andrews, M. (2004), 'Who is being heard? response bias in open-ended responses in a large government employee survey', *Public Opinion Quarterly* **69**, 3760–3766.
- Andridge, R. R. & Little, R. J. (2009), 'The use of sample weights in hot deck imputation', *J Off Stat* **25**, 21–36.
- Andridge, R. R. & Little, R. J. A. (2010), 'A review of hot deck imputation for survey non-response', *International Statistical Review* **78**(1), 40–64.
- Azzalini, A. (1996), *Statistical Inference: Based on the likelihood*, first edn, Chapman and Hall, London.
- Bailar, B., Bailey, L. & Corby, C. (1978), 'A comparison of some adjustment and weighting procedures for survey data', *Proc. Survey Res. Meth. Sec., Am. Statist. Assoc* pp. 175–200.
- Barnard, J. & Rubin, D. (1999), 'Small-sample degrees of freedom with multiple imputation', *Biometrika* (86), 949–955.
- Bishop, Y., Feinberg, S. & Holland, P. (1975), *Discrete multivariate analysis: Theory and Practice*, MIT Press, Boston.
- Bjørnstad, J. F. (2007), 'Non-bayesian multiple imputation', *Journal of Official Statistics* **23**(4), 433–452.  
**URL:** <http://www.jos.nu/Articles/article.asp>
- Bolker, B. M. (2011), *Ecological Models and Data in R*, first edn, Princeton University Press, New Jersey.
- Buddhavarapu, S. (2007), *The Relationship Between Work Experience and Leader Traits in the Prediction of Leadership Effectiveness*, first edn, ProQuest.
- Carpenter, J. R. (2011), 'Multiple imputation with survey weights - a bad idea?'.  
**URL:** <http://www.missingdata.org.uk>



- Cochran, W. (1977), *Sampling Techniques*, third edn, Wiley, New York.
- Collins, L., Schafer, J. & Kam, C.-M. (2001), 'A comparison of inclusive and restrictive strategies in modern missing data procedures', *Psychological Methods* (6), 330–351.
- DelSole, T. (2010), *CLIM 762 Statistical Methods in Climate Research-Chapter 3 (Lecture notes)*, George Mason University, Center for Atmosphere-Land-Ocean Studies 4041 Powder Mill Rd. Suite 302 Calverton, MD 20705. USA.  
**URL:** Available: <http://mason.gmu.edu/~tdelsole/clim762/ch3part1.pdf>
- Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the em algorithm (with discussion)', *J.Roy.Statist.Soc* **B39**, 1–38.
- Desai, M., Kubo, J., Esserman, D. & Terry, M. B. (2010), 'The handling of missing data in molecular epidomilogic studies', *The Berkeley Electronic Press* (27), 72.
- Dixon, W. (1988), *BMDP statistical software*, University of California Press, Los Angeles.
- Durrant, G. B. (2005), 'Imputation methods for handling item-nonresponse in the social sciences: A methodological review', National Centre for Research Methods Working Paper Series. University of Southampton.
- Efron, B. (1979), 'Bootstrap methods: another look at jackknife', *Annals of Statistics* **7**, 1–26.
- Efron, B. (1994), 'Missing data, imputation and the bootstrap', *Journal of American Statistical Association* **89**, 463–478.
- Efron, B. & Gong, G. (1983), 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *Amer. Statist* **37**, 36–48.
- Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, first edn, Chapman and Hall.
- Enders, C. K. (2010), *Applied Missing Data Analysis*, first edn, The Guilford Press, New York.
- Fairclough, D. L. . (2010), *Design and Analysis of Quality of Life Studies in Clinical Trials*, second edn, Chapman and Hall.
- Fay, R. (1996), 'Alternative paradigms for the analysis of imputed survey data', *J.AM Statist Assoc* **91**, 490–498.
- Gamerman, D. & Lopes, H. F. (2006), *Markov Chain Monte Carlo*, second edn, Chapman and Hall/CRC, New York.
- Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995), *Bayesian Data Analysis*, first edn, Chapman and Hall, New York.
- Gelman, A. & Rubin, D. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(3), 457–511.
- Geyer, C. J. (2006), 'The subsampling bootstrap'.
- Ghosh, S. & Pahwa, P. (2008), 'Assessing bias associated with missing data from joint canada/u.s. survey of health: An application', *JAM Biometrics Section*, 57–68.

- Gilks, W., Richardson, S. & Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, first edn, Chapman and Hall, London.
- Gower, J. (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* (27), 623–637.
- Graham, J. & Schafer, J. (1999), On the performance of multiple imputation for multivariate data with small sample size., in 'Statistical Strategies for Small Sample Research', first edn, Thousand Oaks, CA: Sage, London, pp. 1–29.
- Grimmett, G. R. & Stirzaker, D. R. (1992), *Probability and Random Processes*, second edn, Clarendon Press, Oxford.
- Groenewald, P. C. & Mokgatlhe, L. (2005), 'Bayesian computation for logistic regression', *Computational Statistics and Data Analysis* (48), 857–868.
- Haldane, J. B. S. (1956), 'The estimation and significance of the logarithm of a ratio of frequencies', *Annals of Human Genetics* (20), 309–311.
- Harrell, F. E. (1984), 'Regression modelling strategies for improved prognostic prediction', *Statistics in Medicine* (3), 143–152.
- Harvey, C. R. (2001), 'The specification of conditional expectations', *Journal of Empirical Finance* **8**, 573–637.
- Hastings, W. (1970), 'Monte carlo sampling methods using markov chains and their applications', *Biometrika* **57**, 97–109.
- Healy, M. & Westmacott, M. (1956), 'Missing values in experiments analyzed on automatic computers', *Appl. Statist* **5**, 203–206.
- Hoeschele, I. (1989), 'A note on local maxima in maximum likelihood, restricted maximum likelihood, and bayesian estimation of variance components', *Journal of Statistical Computation and Simulation* **33**(3), 149–160.
- Howell, D. C. (2009), 'Treatment of missing data'.  
**URL:** <http://www.uvm.edu>
- Jefferys, H. (1939), *Theory of Probability*, first edn, The Clarendon Press, Oxford.
- Jeffreys, H. (1961), *Theory of Probability*, third edn, Oxford University Press, Oxford.
- Kalton, G. (1983), *Compensating for missing survey data*, the University of Michigan.
- Kaufman, L. & Rousseeuw, P. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, first edn, Wiley, New York.
- Kotz, S. & Johnson, N. (1983), *Encyclopedia of Statistical Sciences*, first edn, Wiley, New York.
- Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, second edn, Wiley-Interscience, New York.
- Little, R. J. A. & Schenker (1995), 'Handbook of statistical modeling for the social and behavioral sciences', *New York: Plenum Press* **14**, 39–75.

- Lohr, S. L. (1999), *Sampling: Design and Analysis*, first edn, Brooks/Cole Publishing Company, USA.
- Marin, J. M. & Robert, C. P. (2007), *Bayesian Core - A Practical Approach to Computational Bayesian Statistics*, first edn, Springer Science Business Media, New York.
- Mason, R., Gunst, R. F. & Hess, J. L. (2010), *Statistical Design and Analysis of Experiments*, second edn, Wiley, New Jersey.
- McKnight, P. E., McKnight, K., Sidani, S. & Figueredo, A. J. (2007), *Missing data: A gentle introduction*, first edn, The Guilford Press, New York.
- McLachlan, G. & Krishnan, T. (1997), *The EM Algorithm and Extensions*, first edn, Wiley, New York.
- Muller, P. (1991), Metropolis based posterior integration schemes, Technical report, Purdue University.
- OECD (2001), 'Glossary of statistical terms'.  
**URL:** <http://stats.oecd.org/glossary/detail.asp?ID=2545>
- Peter, M. L. (1997), *Bayesian Statistics: an introduction*, second edn, John Wiley and Sons Inc., New York.
- Platek, R. & Gray, G. (1983), Imputation methodology: Total survey error, in W. Madow, I. Olkin & D. Rubin, eds, 'Incomplete Data in Sample Surveys', Vol. 2, pp. 249–333.
- Pledger, M., Black, J., Cumming, J. & McDonald, J. (2010), 2009 school and early childhood education services food and nutrition environment survey - phase iii report, Technical report, Health Services Research Centre, School of Government, Victoria University of Wellington.
- Politis, D. & Ramano, J. (1994), 'Large sample confidence regions based on subsamples under minimal assumptions', *Annals of Statistics* (22), 2031–2050.
- Ramirez, E., Mejias, R., Coello, M. & de la Vega, M. (2011), 'Missing value imputation on missing completely at random data using multilayer perceptrons', *Neural Networks* **24**(2011), 121–129.
- Rao, J. (1996), 'On variance estimation with imputed survey data', *J.A.M Statist Assoc* **91**, 499–506.
- Rao, J. & Shao, J. (1992), 'Jackknife variance estimation with survey data under hot deck imputation', *Biometrika* **79**, 811–822.
- Ridout, M. S. & Diggle, P. J. (1991), 'Testing for random dropouts in repeated measurement data', *Biometrics* (4), 1617–1621.
- Rubin, D. (1977), 'Formalizing notions about the effect of nonrespondents in sample surveys', *Journal of the American Statistical Association* **72**, 538–543.
- Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, first edn, Wiley, New York.
- Rubin, D. (1996), 'Multiple imputation after 18+ years', *Journal of the American Statistical Association* **91**(434), 473–489.

- Rubin, D. & Schenker, N. (1986), 'Multiple imputation for interval estimation from simple random samples with ignorable nonresponse', *J.AM Statist Assoc* **81**, 366–374.
- Sahlin, K. (2011), Estimating convergence of markov chain monte carlo simulations, Master's thesis, Stockholm University, Mathematical Statistics, Stockholm University.
- SAS (2008), *SAS/STAT 9.2 User's Guide: Introduction to Bayesian Analysis Procedures*, SAS Institute Inc., USA.  
**URL:** <http://support.sas.com/documentation>
- Schafer, J. & Graham, J. (2002), 'Missing data: Our view of the state of the art', *Psychological Methods* **7**(2), 147–177.
- Schafer, J. L. (2003), 'Multiple imputation in multivariate problems when the imputation and analysis models differ', *Journal of Official Statistics* **57**(1), 19–35.  
**URL:** <http://onlinelibrary.wiley.com/doi/10.1111/1467-9574.00218/full>
- Schafer, J. L. (2005), 'The multiple imputation faq page'.  
**URL:** <http://sites.stat.psu.edu/jls/mifaq.html>
- Scheffer, J. (2002), 'Dealing with missing data', *Res. Lett. Inf. Math. Sci.* **2002**(3), 153–160.
- Scott, M. L. (2007), *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, first edn, Springer Science + Business Media, New York.
- Shao, J. & Sitter, R. R. (1996), 'Bootstrp for imputed survey data', *Journal of the American Statistical Association* **91**(435), 1278.
- Singh, K. (1981), 'On the asymptotic accuracy of efron's bootstrap', *Annals of Statistics* **9**, 1187–1195.
- Statistics Netherlands (2012), 'Which proportion of missing values in the data is allowed'.  
**URL:** <http://www.cbs.nl/>
- Statistics New Zealand (2011), 'Synthetic unit record files'.  
**URL:** <http://www.stats.govt.nz/>
- Stuart, A. . & Ord, J. (1994), *Kendall's Advance Theory of Statistics*, sixth edn, Arnold, New York.
- Tanner, M. & Wong, W. (1987), 'The calculation of posterior distributions by data augmentation (with discussion)', *Journal of the American Statistical Association* **82**, 528–540.
- Torgo, L. (2003), *Data Mining with R - learning by case studies*, first edn, University of Porto, R. Campo Alegre, 823 - 4150 Porto, Portugal.
- Vaughan, T. S. & Berry, K. E. (2005), 'Using monte carlo techniques to demonstrate the meaning and implications of multicollinearity', *Journal of Statistics Education* **13**(5).
- Wagner, J. (2010), 'The fraction of missing information as a tool for monitoring the quality of survey data', *Public Opinion Quarterly* **74**(2), 223–243.
- Wayman, J. C. (2003), 'Multiple imputation for missing data: What is it and how can i use it?', Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL.

- Wu, C. (1983), 'On the convergence properties of the em algorithm', *Annals of Statistics* **11**, 95–103.
- Wu, C. (1986), 'Jackknife, bootstrap and other resampling methods in regression analysis', *Annals of Statistics* **14**, 1343–1350.

# **Appendices**

# Appendix A

## R code for chapter 5

### A.1 The simple bootstrap

```
#R program for the simple Bootstrap applied to imputed incomplete data
#function for adjustment cell hot deck imputation
Adjust_hot_imp=function (dataset, hotdeckvars, imputevar){
  dataset=dataset[order(dataset$Personid),]
  dataset.nomiss=dataset[!is.na(dataset[,imputevar]),]
  dataset.miss =dataset[is.na(dataset[,imputevar]),]
  #Count the number of missing
  nmissing=nrow(dataset.miss)
  idx=sort(dataset[is.na(dataset[,imputevar]),"Personid"])
  for (j in (1:nmissing)){
    matched <- F
    m <- length(hotdeckvars)
    while(!matched) {
      mm <- merge(dataset.miss[j,], dataset.nomiss, by=hotdeckvars[1:m])
      if(nrow(mm)>0) {
        matched <- T
        dataset[idx[j],imputevar]
          <- mm[sample(nrow(mm),1),paste(imputevar,"y", sep=".")]
      } else {
        m <- m-1
        if(m==0) {
          dataset[idx[j],imputevar]
            <- dataset.nomiss[sample(nrow(dataset.nomiss),1),imputevar]
          matched <- T
        }
      }
    }
  }
  dataset
}

#####
#####
#Bootstrap method
hotdeckvars
  <- c("HoursBand","AgeBand","Marital","Ethnicity","Gender","Qualification")
```

```

SURF2=SURF
SURF2$AgeBand <- 5*(SURF$Age%/%5)
SURF2$HoursBand <- 10*((SURF$Hours-5)%/%10)+5
#repeat the whole process 1000 times
b=200
mu=c()
mu_boot=c()
var_boot=c()
for (i in 1:1000){
  mar.surf=MAR(SURF2,"Gender","Income",c(0.5,0.2))

  for (j in 1:b){
    Y=mar.surf
    Income_boot=sample(Y$Income,nrow(Y),replace=T)
    Y=cbind(Y,Income_boot)
    Y_hat=subset(Y, select=-Income) #remove original income var
    #imputation
    Y_hat=Adjust_hot_imp(Y_hat, hotdeckvars, "Income_boot")
    #compute mean for each imputed bootstrap sample
    mu[j]=mean(Y_hat$Income_boot)
  }
  mu_boot[i]=mean(mu)
  var_boot[i]=var(mu)
}

```

## A.2 The simple jackknife

```

#Jackknife method
hotdeckvars
  <- c("HoursBand","AgeBand","Marital","Ethnicity","Gender","Qualification")
SURF2=SURF
SURF2$AgeBand <- 5*(SURF$Age%/%5)
SURF2$HoursBand <- 10*((SURF$Hours-5)%/%10)+5
Jack_mean=c()
Jack_var=c()
mu=c()
n=19

for (i in 1:100){
  mar.surf=MAR(SURF2,"Gender","Income",c(0.5,0.2))
  Y=mar.surf
  #consistent estimate of theta (mean)
  theta_con=mean(Adjust_hot_imp(Y, hotdeckvars, "Income")$Income)
  for (j in 0:19){
    Y_hat=Y[order(Y$Personid),]
    Y_hat=Y_hat[-c(j*10+1:j*10+10),]
    Personid=seq.int(1,nrow(Y_hat),1) #add new personid
    Y_hat=subset(Y_hat, select=-Personid) #remove old personid
    Y_hat=cbind(Y_hat,Personid)
    Y_hat=Adjust_hot_imp(Y_hat, hotdeckvars, "Income")
  }
}

```



```

        mu[j]=mean(Y_hat$Income)
    }
    #pseudo value theta (mean)
    pseudo_mu=length(mu)*theta_con-(length(mu)-1)*mu
    #jack mean
    Jack_mean[i]=mean(pseudo_mu)
    #jack variance
    Jack_var[i]=(length(mu)-1)*sum((mu-mean(mu))^2)/length(mu)
}

```

# Appendix B

## R code for chapter 6

### B.1 EM algorithm - Univariate Normal Data

```
# R program -- EM Algorithm EM algorithm for univariate normal data
# Now, suppose Income
# variable is normally distributed. Then, we can consider an univariate
# normal data. Creating missing data
mcar.surf=SURF
nmissing <- c(Income=40)
nsurf <- nrow(SURF)
idx <- sort(sample(nsurf, size=nmissing["Income"], replace=F))
mcar.surf[idx,"Income"] = NA
em.uni.norm=function(Y){
  Yobs=Y[!is.na(Y)]
  Ymis=Y[is.na(Y)]
  n=length(c(Yobs,Ymis))
  r=length(Yobs)
  #initial values
  #mut=mean(Yobs)
  #sit=var(Yobs)*(r-1)/r # (n-1)s^2/n -- sample variance
  mut=1
  sit=0.1
  #log-likelihood function
  log.like=function(y,mu,sigma2,n){
    -0.5*n*log(2*pi*sigma2)-0.5*sum((y-mu)^2)/sigma2}
  #Compute the log-likelihood for the initial values, and ignoring the
  #missing data mechanism
  log.like.tmp=log.like(Yobs,mu,sit,n)
  repeat{
    #E-step
    EY=sum(Yobs)+(n-r)*mut
    EY2=sum(Yobs^2)+(n-r)*(mut^2+sit)
    #M-step
    mut1=EY/n
    sit1=EY2/n-mut1^2
    #Update parameter
    mut=mut1
    sit=sit1
  }
}
```

```

        #Compute log-likelihood using current estimates
        log.like.t=log.like(Yobs,mut,sit,n)
        #print current parameter values and likelihood
        cat(sprintf("%.4f %.1f %.3f\n",mut,sit,log.like.t))
        #stop if converged
        if (abs(log.like.tmp-log.like.t)/abs(log.like.t)<1.0e-6) break
        log.like.tmp=log.like.t
    }
    c(mut,sit)
}

X=mcar.surf$Income
em.uni.norm(X)

```

## B.2 Recipe: EM algorithm - Bivariate Normal Sample with Missing Data on both Variables

```

# EM for bivariate normal sample with missing data
mu_hours=c()
mu_income=c()
var_hours=c()
var_income=c()
for(k in 1:1000){
    Y_H=MCAR(SURF,50,"Hours")
    Y_I=MCAR(Y_H[!is.na(Y_H$Hours),],50,"Income")
    Y_I=Y_I[,c("Personid","Hours","Income")]
    Y_HR=Y_H[!(Y_H$Personid %in% Y_I$Personid),]
    Y_HR=Y_HR[,c("Personid","Hours","Income")]
    Y=rbind(Y_I,Y_HR)
    Y=Y[order(Y$Personid),]
    n=nrow(Y)
    # Step 1: compute means, variances, covariance and construct loglikelihood
    # function
    Y=Y[,-1]
    idx.11 <- !is.na(Y[,1]) & !is.na(Y[,2])
    idx.10 <- !is.na(Y[,1]) & is.na(Y[,2])
    idx.01 <- is.na(Y[,1]) & !is.na(Y[,2])
    idx.00 <- is.na(Y[,1]) & is.na(Y[,2]) # should be all FALSE
    Y.obs <- Y[!is.na(Y[,1]) & !is.na(Y[,2]),]

    mub=apply(Y.obs,2,mean) # means
    mubx=mub[1]
    muby=mub[2]
    covar=cov(Y.obs) # variances and covariances
    sigmax=sqrt(covar[1,1])
    sigmay=sqrt(covar[2,2])
    rho=cor(Y.obs)[1,2] # correlation
    sigmaxy=rho*sigmax*sigmay
}

```

```

# log-likelihood function

BV.log.like <- function(x,y,mux,muy,n,sigmax,sigmay,rho) {
  sigmat <- array(c(sigmax^2, rho*sigmax*sigmay,
                    rho*sigmax*sigmay, sigmay^2), dim=c(2,2))
  detsigmat <- det(sigmat)
  m <- 2
  ss <- sum( ((x-mux)/sigmax)^2
            - 2*rho*(x-mux)*(y-muy)/(sigmax*sigmay)
            + ((y-muy)/sigmay)^2 )
  retval <- ( -n*m/2*log(2*pi) -(n/2)*log(detsigmat) -0.5*ss/(1-rho^2) )
  return(retval)
}

# BV.log.like(Y.obs[,1],Y.obs[,2],mubx,muby,160,sigmax,sigmay,rho)
BV.log.like.tmp = BV.log.like(Y.obs[,1],Y.obs[,2],
                              mubx,muby,n,sigmax,sigmay,rho)

i=0
# Step 2: E step
repeat{
  # the sufficient statistics for both variables with observed units
  EX.obs =sum(Y[idx.11 | idx.10,1])
  EY.obs =sum(Y[idx.11 | idx.01,2])
  EX2.obs=sum(Y[idx.11 | idx.10,1]^2)
  EY2.obs=sum(Y[idx.11 | idx.01,2]^2)
  EXY.obs=sum(Y[idx.11,1]*Y[idx.11,2]) #need to have complete data

  # the sufficient statistics for Income observed, but hours missing
  beta21=sigmaxy/sigmax^2
  beta20=muby-beta21*mubx
  sigmay1=sqrt(sigmay^2-sigmaxy^2/sigmax^2)

  EY.2=sum(beta20+beta21*Y[,1][is.na(Y[,2])])
  EY2.2=sum((beta20+beta21*Y[,1][is.na(Y[,2])])^2+sigmay1^2)
  EXY.2=sum((beta20+beta21*Y[,1][is.na(Y[,2])])*Y[,1][is.na(Y[,2])])

  # the sufficient statistics for hours observed, but Income missing
  beta12=sigmaxy/sigmay^2
  beta02=mubx-beta12*muby
  sigmax1=sqrt(sigmax^2-sigmaxy^2/sigmay^2)

  EX.1=sum(beta02+beta12*Y[,2][is.na(Y[,1])])
  EX2.1=sum((beta02+beta12*Y[,2][is.na(Y[,1])])^2+sigmax1^2)
  EXY.1=sum((beta02+beta12*Y[,2][is.na(Y[,1])])*Y[,2][is.na(Y[,1])])

  # final sufficient statistics
  EX.tot=EX.obs+EX.1      #s1
  EY.tot=EY.obs+EY.2      #s2
  EX2.tot=EX2.obs+EX2.1   #s11
  EY2.tot=EY2.obs+EY2.2   #s22

```

```

    EXY.tot=EXY.obs+EXY.2+EXY.1    #s12

# Step 3: M step
    mubxt=EX.tot/n
    mubyt=EY.tot/n
    sigmaxt=sqrt(EX2.tot/n-mubx^2)
    sigmayt=sqrt(EY2.tot/n-muby^2)
    sigmaxyt=EXY.tot/n-mubx*muby

#update parameters
    mubx=mubxt
    muby=mubyt
    sigma=sigmaxt
    sigmay=sigmayt
    sigmaxy=sigmaxyt
    i=i+1

#Compute log-likelihood using current estimates
    BV.log.like.t=BV.log.like(Y.obs[,1],Y.obs[,2],mubx,muby,n,sigma,sigmay,rho)

    print(c(mubx,muby,n,sigma,sigmay,i,BV.log.like.t))
#stop if converged
    if (abs(BV.log.like.tmp-BV.log.like.t)<0.001) break
    BV.log.like.tmp=BV.log.like.t
}
c(mubx,muby,n,sigma,sigmay,i)
mu_hours[k]=c(mubx,muby,n,sigma,sigmay,i)[1]
var_hours[k]=c(mubx,muby,n,sigma,sigmay,i)[4]^2
mu_income[k]=c(mubx,muby,n,sigma,sigmay,i)[2]
var_income[k]=c(mubx,muby,n,sigma,sigmay,i)[5]^2
}

```

# Appendix C

## R code for chapter 7

### C.1 Applying MH algorithm to Univariate Normal data

```
#Metropolis-Hastings algorithm -- univariate normal
#step 0: create MCAR income data
Y_MCAR_0=MCAR(SURF,50,"Income")$Income
Y_MCAR=Y_MCAR_0
#step 1: set up posterior density function
n=length(Y_MCAR)
logpost=function(y,mY,sY)
  {-log(sY)-(n/2)*log(2*pi*sY)-(2*sY)^-1*sum((y-mY)^2)}
#step 2: set up initial values
iter=100000
a=5
b=5
mY=matrix(mean(Y_MCAR_0[!is.na(Y_MCAR)]),iter)
sY=matrix(var(Y_MCAR_0[!is.na(Y_MCAR)]),iter)
acctot_m=0
acctot_s=0
for (i in 2:iter){
#step 3: sample Ymis
  Y_MCAR[is.na(Y_MCAR_0)]=rnorm(length(Y_MCAR_0[is.na(Y_MCAR_0)]),
                                mY[i-1], sqrt(sY[i-1]))
#step 4: sample mean mY
  #mY[i]=runif(1, mY[i-1]-a, mY[i-1]+a)
  mY[i]=mY[i-1]+runif(1,min=-10,max=10)
  acc_m=1
  if (mY[i]<0) {
    acc_m=0
    mY[i]=mY[i-1]
  }
  if ((logpost(Y_MCAR,mY[i],sY[i-1])-logpost(Y_MCAR,mY[i-1],sY[i-1]))
      <log(runif(1,min=0,max=1)))
  {
    acc_m=0
    mY[i]=mY[i-1]
  }
  acctot_m=acctot_m+acc_m
```

```

#strp 5: sample variance sY
#sY[i]=runif(1,sY[i-1]-b,sY[i-1]+b)
sY[i]=sY[i-1]+runif(1,min=-1000,max=1000)
acc_s=1
  if (sY[i]<0){
    acc_s=0
    sY[i]=sY[i-1]
  }
  if ((logpost(Y_MCAR,mY[i],sY[i])-logpost(Y_MCAR,mY[i],sY[i-1]))
      <log(runif(1,min=0,max=1)))
  {
    acc_s=0
    sY[i]=sY[i-1]
  }
acctot_s=acctot_s+acc_s
}

```

## C.2 Applying Gibbs sampling algorithm to Bivariate Normal data

```

#Gibbs sampling algorithm -- bivariate normal
library(MCMCpack)
#Data Augmentation
# Example 10.1
# Bivariate normal data with ignorable nonresponse and a general pattern of
# missing data
# Step 0: create missing data for Income and Hours, missing data cannot overlap
Y_H=MCAR(SURF,50,"Hours")
Y_I=MCAR(Y_H[!is.na(Y_H$Hours),],50,"Income")
Y_I=Y_I[,c("Personid","Hours","Income")]
Y_HR=Y_H[!(Y_H$Personid %in% Y_I$Personid),]
Y_HR=Y_HR[,c("Personid","Hours","Income")]
Y=rbind(Y_I,Y_HR)
mcar.surf=Y[order(Y$Personid),]
idx.In=which(is.na(mcar.surf$Income))
idx.Hour=which(is.na(mcar.surf$Hours))
d=1000
# Step 1: compute means, variances, covariance and betas
mubx=c()
muby=c()
sigmax=c()
sigmay=c()
sigmaxy=c()
rho=c()
beta21=c()
beta20=c()
beta12=c()
beta02=c()
sigmax1=c()

```

```

sigmay1=c()
Y=cbind(mcar.surf$Income, mcar.surf$Hours)
#units with both variables observed
Y.obs=cbind(Y[,1][!is.na(Y[,1])& !is.na(Y[,2])],
            Y[,2][!is.na(Y[,1])& !is.na(Y[,2])])
mub=apply(Y.obs,2,mean) # means
mubx[1]=mub[1]
muby[1]=mub[2]
covar=cov(Y.obs) # variances and covariances
sigmax[1]=sqrt(covar[1,1])
sigmay[1]=sqrt(covar[2,2])
rho[1]=cor(Y.obs)[1,2] # correlation
sigmaxy[1]=rho[1]*sigmax[1]*sigmay[1]
for (j in 1:d){
  beta21[j]=sigmaxy[j]/sigmax[j]^2
  beta20[j]=muby[j]-beta21[j]*mubx[j]
  sigmay1[j]=sqrt(sigmay[j]^2-sigmaxy[j]^2/sigmax[j]^2)
  beta12[j]=sigmaxy[j]/sigmay[j]^2
  beta02[j]=mubx[j]-beta12[j]*muby[j]
  sigmax1[j]=sqrt(sigmax[j]^2-sigmaxy[j]^2/sigmay[j]^2)
# Step 2: I step
# Income is missing (Y1 missing)
  for (i in 1:length(idx.In)) {
    Y[idx.In[i],1]=rnorm(1,beta02[j]+beta12[j]*Y[idx.In[i],2],
                        sigmax1[j])
  }
# Hours is missing (Y2 missing)
  for (i in 1:length(idx.Hour)) {
    Y[idx.Hour[i],2]=rnorm(1,beta20[j]+beta21[j]*Y[idx.Hour[i],1],
                        sigmay1[j])
  }
# Step 3: P step
# part one: new means, covariance matrix
mup=apply(Y,2,mean)
covarp=cov(Y)
S=n*covarp
sigma=(riwish(n-1, S))
sigma
z=rnorm(2,0,1)
mupt=mup+t(z)%*%chol(sigma/(n-1))
# Step 4: update parameters
mubx[j+1]=mupt[1]
muby[j+1]=mupt[2]
sigmax[j+1]=sqrt(sigma[1,1])
sigmay[j+1]=sqrt(sigma[2,2])
rho[j+1]=sigma[1,2]/(sqrt(sigma[1,1])*sqrt(sigma[2,2])) #correlation
sigmaxy[j+1]=rho[j+1]*sigmax[j+1]*sigmay[j+1]
j=j+1
}

```



# Appendix D

## R code for Chapter 8

### D.1 The MI Process

#### Step 1: Imputation

```
#Multiple Imputation
#step 0: Create MAR income data
Y_MAR=MAR(SURF,"Gender","Income",c(0.5,0.2))
#step 1: set up y and x matrix
y=as.matrix(Y_MAR$Income)
x=as.matrix(cbind(rep(1,nrow(Y_MAR)),Y_MAR$Gender, Y_MAR$Age, Y_MAR$Hours))
ystar=y
#step 2: establish intial parameters/starting points
iter=1000
#sigma square
s2=matrix(1,iter)
#beta matrix. only consider three variables:Gender, Age, Hours and the intercept
b=matrix(0,iter,4)
xtxi=solve(t(x)%*%x)
muY=matrix(mean(ystar[!is.na(y)]),iter)    #mean Y
varY=matrix(var(ystar[!is.na(y)]),iter)    #variance Y
yreplace= as.matrix(y)[,rep(1,iter)]
for (j in 2:iter){
  #step 3: sample missing data
  ystar[is.na(y)]=rnorm(length(ystar[is.na(y)]),
                        mean=x[is.na(y),]%*(b[j-1,]), sd=sqrt(s2[j-1]))
  yreplace[,j][is.na(yreplace[,j])]= ystar[is.na(y)]
  muY[j]=mean(ystar)
  varY[j]=var(ystar)
  #step 4: simulate beta from multivariate normal distribution
  b[j,]=coefficients(lm(ystar~x-1))+t(rnorm(4,0,1))%*%chol(s2[j-1]*xtxi)
  #step 5: simulate sigma from inverse gamma distribution
  s2[j]=1/rgamma(1,length(y)/2,
                0.5*t(ystar-x%*(b[j,]))%*(ystar-x%*(b[j,])))
}
#step 6: Convergence diagnosis
par(mfrow=c(3,2))
plot(as.vector(s2),type="l", xlab="iteration",
```

```

        ylab="Sigma Square", main="Time series plot")
plot(as.vector(b[,1]),type="l", xlab="iteration",
      ylab="Intercept", main="Time series plot")
plot(as.vector(b[,2]),type="l", xlab="iteration",
      ylab="Beta_Gender", main="Time series plot")
plot(as.vector(b[,3]),type="l", xlab="iteration",
      ylab="Beta_Age", main="Time series plot")
plot(as.vector(b[,4]),type="l", xlab="iteration",
      ylab="Beta_Hours", main="Time series plot")
#step 7: Sample five datasets
D=5
#D number of imputed data sets, burn-in iteration is 200
MI_y=as.matrix(y)[,rep(1,D)]
for (d in 1:D){
  MI_y[,d]=yreplace[,-c(1:200)][,d]
  d=d+100
}

```

## D.2 Proper and Improper Multiple Imputation

```

#Improper Multiple Imputation - modified simple random MI
hotImp.mean.mcar=c()
hotImp.var.mcar=c()
hotImp.totvar.mcar=c()
for (i in 1:1000){
  #Step 0: create MCAR missing income
  Y_MCAR=MCAR(SURF,50,"Income")
  Y_MCAR_imp=Y_MCAR
  #Step 1: the simple hot deck imputation
  #Count the number of missing
  nmissing=nrow(Y_MCAR[is.na(Y_MCAR$Income),])
  D=5
  for (d in 1:D){
    #draw Y_mis from the normal distribution
    Y_MCAR_imp[is.na(Y_MCAR$Income),"Income"]=
      rnorm(length(Y_MCAR_imp[is.na(Y_MCAR)]),
            mean(Y_MCAR[!is.na(Y_MCAR$Income),"Income"]),
            sd(Y_MCAR[!is.na(Y_MCAR$Income),"Income"]))
    #Step 2: estimates
    hotImp.mean.mcar[d]=mean(Y_MCAR_imp$Income)
    hotImp.var.mcar[d]=var(Y_MCAR_imp$Income)/nrow(Y_MCAR_imp)
  }
  #total variance
  hotImp.totvar.mcar[i]=mean(hotImp.var.mcar)+(1+1/D)*var(hotImp.mean.mcar)}

#Proper Multiple Imputation
totvar.mcar=c()
for (j in 1:1000){
  #Step 0: create MCAR missing income
  Y_MCAR=MCAR(SURF,50,"Income")$Income

```

```

Y_MCAR_imp=Y_MCAR

#assume prior =1
#step 1: Set up initial values
iter=1000
mY_MCAR=mean(Y_MCAR_imp[!is.na(Y_MCAR_imp)])
sY_MCAR=var(Y_MCAR_imp[!is.na(Y_MCAR_imp)])
#step 2: Draw missing income from the normal distribution with
#observed income mean and variance
for(i in 2:iter){
  Y_MCAR_imp[is.na(Y_MCAR)]=rnorm(length(Y_MCAR_imp[is.na(Y_MCAR)]),
                                   mY_MCAR[i-1], sqrt(sY_MCAR[i-1]))

  #step 3: draw sigma^2 and mean
  sY_MCAR[i]=rgamma(1,(length(Y_MCAR_imp)/2),
                   rate=sum((Y_MCAR_imp-mY_MCAR[i-1])^2)/2)

  sY_MCAR[i]=1/sY_MCAR[i]
  mY_MCAR[i]=rnorm(1,mean(Y_MCAR_imp),sqrt(sY_MCAR[i]/length(Y_MCAR_imp)))
}

#sample D MI dataset start from iteration 200
D=5
#D number of means
MI_mean=c()
MI_var=c()
for (d in 1:D){
  MI_mean[d]=mY_MCAR[-(1:200)][d]
  MI_var[d]=sY_MCAR[-(1:200)][d]/length(Y_MCAR_imp)
  d=d+100
}
totvar.mcar[j]=mean(MI_var)+(1+1/D)*var(MI_mean)}

```

# Appendix E

## R code for Chapter 9

### E.1 Single imputation methods for categorical data

#### E.1.1 Mode imputation

Here below is the R code for unconditional and conditional mode imputation:

```
#Unconditional mode imputation
uncon_mode=function(dat, variable){
  surf_table=table(dat[,variable])
  Mode=max(surf_table)
  Mode_name=names(which(surf_table == Mode))
  #if more than one variable have the max number,
  #then random sample one of them as imputed value
  Mode_name=sample(Mode_name,1)
  dat[,variable][is.na(dat[,variable])]=Mode_name
  dat}

#Conditional mode imputation
add.condition=function(x){
  if (is.null(ncol(x))) x
  else do.call("paste",c(x,sep=""))}
con_mode=function(dat,variable,Condition){
  con.table=tapply( dat[which( dat[,variable]!="NA"),"Qualification"],
    add.condition(dat[which( dat[,variable]!="NA"),Condition]), table)
  con=add.condition( dat[,Condition])
  dat=cbind(dat,con)
  for (i in 1:length(con.table)){
    Mode_name=names(which(con.table[[i]]==max(con.table[[i]])))
    Mode_name=sample(Mode_name,1)
    Sep=dat[,variable][which(dat[, "con"]==names(con.table[i]))]
    dat[,variable][which(dat[, "con"]==names(con.table[i]))][is.na(Sep)]
    =Mode_name
  }dat}
```

#### E.1.2 Logistic regression imputation

Here is the R code for the binary logistic regression imputation method which has been applied to the SURF data.

```

#binary logistic regression
impute = function (a, a.impute){
  ifelse (is.na(a), a.impute, a)}
colname=c("Male", "Female")
all_lgit=t(as.matrix(table(SURF$Gender,useNA="always")/200))[,colname]

for (i in 1:1000){
  mcar_surf=MCAR(SURF,50,"Gender")
  glm_Gen=glm(Gender ~ Qualification+Marital,
              data=mcar_surf[!is.na(mcar_surf$Gender),], family=binomial)
  #glm_Gen=glm(Gender ~ Qualification+Hours+Marital+Ethnicity+Income+Age,
              data=mcar_surf[!is.na(mcar_surf$Gender),], family=binomial)
  pred=predict.glm (glm_Gen, mcar_surf, type = "response")
  for (j in 1:length(pred)){
    if (pred[j]>runif(1,0,1)) {pred[j]="Female"}
    else {pred[j]="Male"}
  }
  mcar_surf[, "Gender"]=impute(as.vector(mcar_surf$Gender), as.vector(pred))
  all_lgit=rbind(all_lgit,
                 t(as.matrix(table(mcar_surf$Gender,useNA="always")/200))[,colname]))}

```

## E.2 Likelihood based and Bayesian iterative simulation imputation methods for categorical data

### E.2.1 EM algorithm for categorical variable

Here is the R code for the EM algorithm of binary logistic regression:

```

impute = function (a, a.impute){
  ifelse (is.na(a), a.impute, a)}

ratio=c()
ratio_true=nrow(SURF[which(SURF$Gender=="Female"),])/
  nrow(SURF[which(SURF$Gender=="Male"),])

for (k in 1:1000){
  mar_surf=MAR(SURF,"Qualification","Gender",c(0.2,0.3, 0.1, 0.1))

  Emlgit_surf=mar_surf
  coe_sum=c(1)
  finished = F
  converged = F
  rtol = 1.e-5
  imax = 2
  i=1
  while (!finished){
    i=i+1
    cutoff=nrow(Emlgit_surf[which(Emlgit_surf$Gender=="Female"),])
    /nrow(Emlgit_surf)
    glm_Gen=glm(as.factor(Gender) ~ Qualification+Hours+Marital
                +Ethnicity, data=Emlgit_surf[!is.na(Emlgit_surf$Gender),],
                family=binomial)

```

```

#extract coeffiecients and sum their absolute values
coe_sum[i]=sum(abs(summary(glm_Gen)$coefficients[,1]))

pred=predict.glm (glm_Gen, EMLgit_surf, type = "response")
for (j in 1:length(pred)){
  if (pred[j]>cutoff) {pred[j]='Female'}
  else {pred[j]='Male'}
}

#EMLgit_surf[, "Gender"]=impute(mar_surf$Gender, as.factor(pred))
EMLgit_surf[, "Gender"]=impute(as.vector(mar_surf$Gender), pred)

if( abs(coe_sum[i]-coe_sum[i-1])<rtol ) {
  converged = T
  finished = T
}
else if(i==imax) {
  finished = T
}
}

ratio[k]=nrow(EMLgit_surf[which(EMLgit_surf$Gender=="Female"),])/
nrow(EMLgit_surf[which(EMLgit_surf$Gender=="Male"),])}

```

# Appendix F

## R code for Chapter 11

Here is the R code for implementing all the proposed imputation methods for the FNES data

### F.1 Impute the 2007 and 2009 ECE FNES missing data

```
#Function to compute standard error pf proportion
SE_complex=function(dat,variable,stra,big_N){
  small_n=tapply(dat[,variable], dat[,stra], length)
  ny_holder=as.matrix(table(dat[,stra]))
  ny_holder[,1]=0
  ny_holder=cbind(ny_holder,stratum=rownames(ny_holder))
  ny_temp=dat[which(dat[,variable]==1),]
  ny=as.matrix(tapply(ny_temp[,variable], ny_temp[,stra], length))
  ny=cbind(ny,stratum=rownames(ny))
  ny_f=replace(ny_holder[,1],ny_holder[, "stratum"] %in% ny[, "stratum"],ny[,1])
  p=as.numeric(ny_f)/small_n
  variance=sum((big_N/sum(big_N))^2*(1-(small_n/big_N))*(p*(1-p))/(small_n-1))
  se=sqrt(variance)
  big_p= sum((big_N/sum(big_N))*p)
  c(big_p,se)
}

#Single imputation
#conditional mode
add.condition=function(x){
  if (is.null(ncol(x))) x
  else do.call("paste",c(x,sep=""))}
con_mode=function(dat,variable,Condition){
  con.table=tapply( dat[which( dat[,variable]!="NA"),variable],
    add.condition(dat[which( dat[,variable]!="NA"),Condition]), table)

  con=add.condition( dat[,Condition])
  dat=cbind(dat,con)

  for (i in 1:length(con.table)){
    Mode_name=names(which(con.table[[i]]==max(con.table[[i]])))
    Mode_name=sample(Mode_name,1)
    Sep=dat[,variable][which(dat[, "con"]==names(con.table[i]))]
```

```

    dat[,variable][which(dat[, "con"]==names(con.table[i]))][is.na(Sep)]=Mode_name
  }dat}

#adjustment cell hot deck
adjHD=function(adjdata,adjvar,adjvary,hotdeckvars){
  id=seq(1:nrow(adjdata))
  ECE_adj_hot=cbind(adjdata,id)
  nomiss=adjdata[!is.na(adjdata[,adjvar]),]
  miss=adjdata[is.na(adjdata[,adjvar]),]
  nmissing=nrow(miss)
  idx=sort(ECE_adj_hot[is.na(ECE_adj_hot[,adjvar]),"id"])
  if (nrow(miss)==0) {
    ECE_adj_hot
  } else{
    for (j in (1:nmissing)){
      matched <- F
      m <- length(hotdeckvars)
      while(!matched) {
        mm <- merge(miss[j,], nomiss, by=hotdeckvars[1:m])
        if(nrow(mm)>0) {
          matched <- T
          ECE09_adj_hot[idx[j],adjvar] <- mm[sample(nrow(mm),1), adjvary]
        } else {
          m <- m-1
          if(m==0) {
            ECE_adj_hot[idx[j],adjvar] <- nomiss[sample(nrow(nomiss),1),adjvar]
            matched <- T
          }
        }
      }
    }
    ECE09_adj_hot
  }
}

#nearest neighbour hot deck
library(StatMatch)
NN_hotdeck=function(dat, ind, varb, match_var){
  gower_rec=dat[which(dat[,ind]==0),-which(names(dat) %in% c(varb))]
  gower_don=dat[which(dat[,ind]==1),] # donor data.frame
  #search for NND donors
  imp.NND = NND.hotdeck(data.rec=gower_rec,
    data.don=gower_don, match.vars=c(match_var), dist.fun="Gower")
  # imputing missing values
  rec.imp=create.fused(data.rec=gower_rec,
    data.don=gower_don, mtc.ids=imp.NND$mtc.ids, z.vars=varb)
  # rebuild the imputed data.frame
  final = rbind(rec.imp, gower_don)
  final
}

```



```

#logistic regression
logreg=function(lrdata,lrvar){
  impute = function (a, a.impute){
    ifelse (is.na(a), a.impute, a)}

  lrdata=lrdata[,c("Stratum","super_regional_council",lrvar)]
  regform=paste(paste("as.factor(",lrvar,")"),
    paste("as.factor(Stratum)","as.factor(super_regional_council)",sep="+") ,sep="~")
  glm_lr=glm(as.formula(regform), family=binomial, data=na.omit(lrdata))
  pred=predict.glm(glm_lr, lrdata, type = "response")
  for (i in 1:length(pred)){
    if (pred[i]<runif(1,0,1)) {pred[i]=1}
    else {pred[i]=2}
  }
  lrdata[,lrvar]=impute(as.vector(lrdata[,lrvar]), pred)
  c(table(lrdata[,lrvar],useNA="always") ,SE_complex(lrdata,lrvar,"Stratum",big_N))
}

#Bootstrap resampling methods
B=200
propotion_se=matrix(,nrow=B, ncol=2)
big_N=c(1961,613,237,473)
for (i in 1:B){
  ece.actual.sample2007_boot=
    ece.actual.sample2007_thesis[sample(nrow(ece.actual.sample2007_thesis),
      size=nrow(ece.actual.sample2007_thesis),replace=T),
      c("Response_ind_Q3a", "Response_ind_Q7b", "Response_ind_Q7c" ,"Stratum",
        "super_regional_council","new_Q3a", "new_Q7b", "new_Q7c")]]

  Table_adjHD=adjHD(ece.actual.sample2009_boot, "new_Q3a", "new_Q3a.y",
    c("Stratum", "super_regional_council"))
  propotion_se[i,]=SE_complex(Table_adjHD,"new_Q3a","Stratum",big_N)
}
boot_proportion=mean(propotion_se[,1])
boot_se=sqrt(sum((propotion_se[,1]-mean(propotion_se[,1]))^2)/(B-1))

#EM algorithm
impute = function (a, a.impute){
  ifelse (is.na(a), a.impute, a)}

EM_CAT=function(EMdata,EMvar){
  glmEM_data=EMdata[,c("Response_ind_Q3a","Stratum","super_regional_council",EMvar)]
  coe_sum=c(0)
  finished = F
  converged = F
  rtol = 1.e-3
  imax = 100
  i=1
  cutoff=c(0.5)

```

```

while (!finished){
  i=i+1
  regform=paste(paste("as.factor(",EMvar,""),
    paste("as.factor(Stratum)","as.factor(super_regional_council)",sep="+") ,sep="~")
    glm_EM=glm(as.formula(regform), family=binomial, data=na.omit(glmEM_data))
#extract coeffieicients and sum their absolute values
  coe_sum[i]=sum(abs(summary(glm_EM)$coefficients[,1]))
  pred=predict.glm (glm_EM, glmEM_data, type = "response")
  for (j in 1:length(pred)){
    if (pred[j]>cutoff[i-1]) {pred[j]=2}
    else {pred[j]=1}}
  glmEM_data[,EMvar]=impute(as.vector(EMdata[,EMvar]), pred)
  cutoff[i]=nrow(glmEM_data[which(glmEM_data[,EMvar]==2),])/
    nrow(glmEM_data[which(glmEM_data[,EMvar]!="NA"),])
    if( abs(coe_sum[i]-coe_sum[i-1])<rtol ) {
      converged = T
      finished = T
    }else if(i==imax) {finished = T}
  }
  c(table(glmEM_data[,EMvar],useNA="always"),
    SE_complex(glmEM_data,EMvar,"Stratum",big_N))
}

#multiple imputation
MI_cat=function(MIdata,DMI,MIVar,Bit, burnin){
  impute = function (a, a.impute){ifelse (is.na(a), a.impute, a)}
  expit <- function(x) 1/(1+exp(-x))
  #set up posterior distribution function
  lpost=function(coe_beta, xmat, y){
    eta=xmat%*%as.vector(coe_beta)
    p=1/(1+exp(-eta))
    return(sum(log(p[y==2])) + sum(log(1-p[y==1])))}
  proportion_MI=c()
  se_MI=c()
  pred_cat=c()

  DAlgit_MI=MIdata[,c("Response_ind_Q3a","Stratum","super_regional_council",MIVar)]
  #estimates of the regression coefficients
  regform=paste(paste("as.factor(",MIVar,""),
    paste("as.factor(Stratum)","as.factor(super_regional_council)",sep="+") ,sep="~")
    glm_MI=glm(as.formula(regform), family=binomial, data=na.omit(DAlgit_MI))
  coe_beta=t(as.numeric(coefficients(glm_MI)))
  #extract x by using the SURF data
  xmat=model.matrix(as.factor(Response_ind_Q3a) ~ as.factor(Stratum)+
    as.factor(super_regional_council), data=DAlgit_MI, family=binomial)
  #compute sigma matrix
  b_sr=sqrt(diag(vcov(glm_MI)))
  for (i in 2:Bit){
    coe_beta=rbind(coe_beta,coe_beta[i-1,])
    #draw Y-mis from Bernoulli distribution, given Y-obs and beta

```

```

pred=as.numeric(expit(xmat%*%as.vector(coe_beta[i,])))
for (k in 1:length(pred)){
  if (rbinom(1,1,as.numeric(pred[k]))==1) {pred_cat[k]=2}
  else {pred_cat[k]=1}
}
DAlgit_MI[,MIvar]=impute(as.vector(MIdata[,MIvar]), as.vector(pred_cat))
y=DAlgit_MI[,MIvar]
#draw beta from the proposed distribution
#begin MH
coe_beta[i,]=coe_beta[i-1,]+rnorm(ncol(coe_beta),0,b_sr)
  if ((lpost(coe_beta[i,], xmat, y)-
      lpost(coe_beta[i-1,], xmat, y))<log(runif(1,min=0,max=1))){
    coe_beta[i,]=coe_beta[i-1,]
  }
  y_full=cbind(y=y,Stratum=DAlgit_MI$Stratum)
  proportion_MI[i]=SE_complex(y_full,"y","Stratum",big_N)[1]
  se_MI[i]=SE_complex(y_full,"y","Stratum",big_N)[2]
}
proportion_MI_d=c()
se_MI_d=c()
for (d in seq(burnin,length(se_MI),floor((length(se_MI)-burnin)/DMI))){
  proportion_MI_d[d]=proportion_MI[d]
  se_MI_d[d]=se_MI[d]
}
proportion_MI_d=proportion_MI_d[!is.na(proportion_MI_d)]
se_MI_d=se_MI_d[!is.na(se_MI_d)]
DMI=length(proportion_MI_d)
final_prop_MI=mean(proportion_MI_d)
WD=sum(se_MI_d^2)/length(se_MI_d)
BD=sum((proportion_MI_d-mean(proportion_MI_d))^2)/(length(proportion_MI_d)-1)
TD=WD+BD*(DMI+1)/DMI
c(final_prop_MI, sqrt(TD))
}

```

## F.2 Imputation using the matched 2007 and 2009 ECE FNES sample

### F.2.1 The simple approach

```

convergence_chain_prop=matrix(0,5000,5)
convergence_chain_se=matrix(0,5000,5)

for (m in 1:5){
#Matching the 2007 and 2009 ECE FNES sample data
match_can_09=ece.actual.sample2009_thesis[,
      c("Institution_Number", "new_Q3a", "new_Q3b",
        "new_Q3c","Stratum","super_regional_council")]
match_can_07=ece.actual.sample2007_thesis[,
      c("Institution_Number", "new_Q3a", "new_Q3b",

```

```

"new_Q3c","Stratum","super_regional_council"]])

#apply adjustment cells hot deck to imputate Q3b and Q3c.
#They will be used as explanatory variables.

match_can_09=adjHD(match_can_09, "new_Q3b", "new_Q3b.y",
                    c("Stratum", "super_regional_council"))
match_can_09=subset(match_can_09,select=-c(id))
match_can_09=adjHD(match_can_09, "new_Q3c", "new_Q3c.y",
                    c("Stratum", "super_regional_council"))
match_can_09=subset(match_can_09,select=-c(id))
match_can_09=adjHD(match_can_09, "new_Q7a", "new_Q7a.y",
                    c("Stratum", "super_regional_council"))
match_can_09=subset(match_can_09,select=-c(id))
match_can_07=adjHD(match_can_07, "new_Q3b", "new_Q3b.y",
                    c("Stratum", "super_regional_council"))
match_can_07=subset(match_can_07,select=-c(id))
match_can_07=adjHD(match_can_07, "new_Q3c", "new_Q3c.y",
                    c("Stratum", "super_regional_council"))
match_can_07=subset(match_can_07,select=-c(id))

#match 07 and 09
matched_0709=merge(match_can_09, match_can_07, by="Institution_Number")
colnames(matched_0709)[2]="new_Q3a"

#functions
impute = function (a, a.impute){
  ifelse (is.na(a), a.impute, a)}
expit <- function(x) 1/(1+exp(-x))
#set up posterior distribution function
lpost=function(coe_beta, xmat, y){
  eta=xmat%*%as.vector(coe_beta)
  p=1/(1+exp(-eta))
  return(sum(log(p[y==2])) + sum(log(1-(p[y==1]))))
}
#set up logistic model
glm_match_can=
  glm(as.factor(new_Q3a) ~ as.factor(new_Q3b)+as.factor(new_Q3c)+
      as.factor(Stratum)+as.factor(super_regional_council),
      data=match_can_09[!is.na(match_can_09$new_Q3a),], family=binomial)

glm_matched=
  glm(as.factor(new_Q3a) ~ as.factor(new_Q3b.x)+as.factor(new_Q3c.x)+
      as.factor(Stratum.x)+as.factor(super_regional_council.x)
      +as.factor(new_Q3b.y)+as.factor(new_Q3c.y),
      data=matched_0709[!is.na(matched_0709$new_Q3a),], family=binomial)

coe_beta=t(as.numeric(coefficients(glm_match_can)))+noise_b
coe_beta_match=t(as.numeric(coefficients(glm_matched)))+noise_a

```

```

xmat=
  model.matrix(as.factor(Institution_Number) ~ as.factor(new_Q3b)+as.factor(new_Q3c)
    +as.factor(Stratum)+as.factor(super_regional_council),
    data=match_can_09, family=binomial)
xmat_matched=
  model.matrix(as.factor(Institution_Number) ~ as.factor(new_Q3b.x)+
    as.factor(new_Q3c.x)+as.factor(Stratum.x)+as.factor(super_regional_council.x)
    +as.factor(new_Q3b.y)+as.factor(new_Q3c.y), data=matched_0709, family=binomial)
#compute sigma matrix
b_sr=sqrt(diag(vcov(glm_match_can)))
b_sr_matched=sqrt(diag(vcov(glm_matched)))
#set original datasets. They are used for imputation purposes
y_09=match_can_09
y_0709=matched_0709
proportion_MI=c()
se_MI=c()
big_N=c(2230,623,309,462)
for (i in 2:5000){
  coe_beta=rbind(coe_beta,coe_beta[i-1,])
  coe_beta_match=rbind(coe_beta_match, coe_beta_match[i-1,])
  #draw Y-mis from Bernoulli distribution, given Y-obs and beta
  pred=as.numeric(expit(xmat%*%as.vector(coe_beta[i,])))
  pred_match=as.numeric(expit(xmat_matched%*%as.vector(coe_beta_match[i,])))

  for (f in 1:length(pred_match)){
    if (rbinom(1,1,as.numeric(pred_match[[f]]))==1) {pred_match[f]=2}
    else {pred_match[f]=1}
  }
  matched_0709$new_Q3a=impute(as.vector(y_0709$new_Q3a), as.vector(pred_match))
  yb=matched_0709$new_Q3a

  for (k in 1:length(pred)){
    if (rbinom(1,1,as.numeric(pred[[k]]))==1) {pred[k]=2}
    else {pred[k]=1}
  }
  match_can_09$new_Q3a=impute(as.vector(y_09$new_Q3a), as.vector(pred))
#update yd
df1=match_can_09[,c("Institution_Number","new_Q3a")]
df2=matched_0709[,c("Institution_Number","new_Q3a")]
for(id in 1:nrow(df2)){
  df1$new_Q3a[df1$Institution_Number %in% df2$Institution_Number[id]] =
    df2$new_Q3a[id]}
yd=df1$new_Q3a
#draw beta from the proposed distribution
for (j in 1:ncol(coe_beta_match)){
  coe_beta_match[i,j]=coe_beta_match[i-1,j]+rnorm(1,0,b_sr_matched[j])
  if ((lpost(coe_beta_match[i,], xmat_matched, yb)-
    lpost(coe_beta_match[i-1,], xmat_matched, yb))<log(runif(1,min=0,max=1)))){
    coe_beta_match[i,j]=coe_beta_match[i-1,j]
  }
}

```

```

    }
  for (g in 1:ncol(coe_beta)){
    coe_beta[i,g]=coe_beta[i-1,g]+rnorm(1,0,b_sr[g])
    if ((lpost(coe_beta[i,], xmat, yd)-
        lpost(coe_beta[i-1,], xmat, yd))<log(runif(1,min=0,max=1))){
      coe_beta[i,g]=coe_beta[i-1,g]
    }
  }
  y_full=cbind(y=yd,Stratum=match_can_09$Stratum)
  proportion_MI[i]=SE_complex(y_full,"y","Stratum",big_N)[1]
  se_MI[i]=SE_complex(y_full,"y","Stratum",big_N)[2]
}
convergence_chain_prop[,m]= proportion_MI
convergence_chain_se[,m]= se_MI
}

#serarching initial paprameters
start_b=as.data.frame(coe_beta)
start_a=as.data.frame(coe_beta_match)
plotf=function(dat){
  for (i in 1:ncol(dat)){
    hist(dat[,i], xlab=names(dat)[i],main="")}}
  par(mfrow=c(4,3))
  plotf(start_a)
  par(mfrow=c(5,2))
  plotf(start_b)
  intial_b=apply(coe_beta, 2, sd)
  noise_b=c()
  for (i in 1:length(intial_b)){
    noise_b[i]=rnorm(1,0,intial_b[i]*10)
  }
  intial_a=apply(coe_beta_match, 2, sd)
  noise_a=c()
  for (i in 1:length(intial_a)){
    noise_a[i]=rnorm(1,0,intial_a[i]*10)
  }
}

#Convergence diagnostics
#Gelman and Rubin diagnostic
library(coda)
mh.list_prop=
  mcmc.list(list(as.mcmc(convergence_chain_prop[-1,1]),
    as.mcmc(convergence_chain_prop[-1,2]),as.mcmc(convergence_chain_prop[-1,3]),
    as.mcmc(convergence_chain_prop[-1,4]),as.mcmc(convergence_chain_prop[-1,5]))))

mh.list_se=
  mcmc.list(list(as.mcmc(convergence_chain_se[-1,1]),
    as.mcmc(convergence_chain_se[-1,2]),as.mcmc(convergence_chain_se[-1,3]),
    as.mcmc(convergence_chain_se[-1,4]),as.mcmc(convergence_chain_se[-1,5]))))

gelman.diag(mh.list_prop)

```

```

gelman.diag(mh.list_se)
gelman.plot(mh.list_prop, main="PSRF of proportions")
gelman.plot(mh.list_se, main="PSRF of standard errors of proportion")

#MI estimates computation
proportion_MI_d=convergence_chain_prop[5000,]
se_MI_d=convergence_chain_se[5000,]
DMI=5
final_prop_MI=mean(proportion_MI_d)
WD=sum(se_MI_d^2)/length(se_MI_d)
BD=sum((proportion_MI_d-mean(proportion_MI_d))^2)/(length(proportion_MI_d)-1)
TD=WD+BD*(DMI+1)/DMI
c(final_prop_MI, sqrt(TD))

```

## F.2.2 The complex approach

```

con_chain_prop_com=matrix(0,20000,1)
con_chain_se_com=matrix(0,20000,1)
options(warn=-1)
for (m in 1:5){
#A better method, involving Y from the 2007
match_can_09=ece.actual.sample2009_thesis
  [,c("Institution_Number", "new_Q3a",
      "new_Q3b", "new_Q3c", "Stratum", "super_regional_council")]
match_can_07=ece.actual.sample2007_thesis
  [,c("Institution_Number", "new_Q3a",
      "new_Q3b", "new_Q3c", "Stratum", "super_regional_council")]
match_can_09=adjHD(match_can_09, "new_Q3b",
  "new_Q3b.y", c("Stratum", "super_regional_council"))
match_can_09=adjHD(match_can_09, "new_Q3c",
  "new_Q3c.y", c("Stratum", "super_regional_council"))
match_can_07=adjHD(match_can_07, "new_Q3b",
  "new_Q3b.y", c("Stratum", "super_regional_council"))
match_can_07=adjHD(match_can_07, "new_Q3c",
  "new_Q3c.y", c("Stratum", "super_regional_council"))
#match 07 and 09
matched_0709=merge(match_can_09, match_can_07, by="Institution_Number")
colnames(matched_0709)[2]="new_Q3a"
#set up logistic model
glm_match_can_09=glm(as.factor(new_Q3a) ~ as.factor(new_Q3b)+
  as.factor(new_Q3c)+as.factor(Stratum)+as.factor(super_regional_council),
  data=match_can_09[!is.na(match_can_09$new_Q3a),], family=binomial)

glm_match_can_07=glm(as.factor(new_Q3a) ~ as.factor(new_Q3b)+
  as.factor(new_Q3c)+as.factor(Stratum)+as.factor(super_regional_council),
  data=match_can_07[!is.na(match_can_07$new_Q3a),], family=binomial)
#summary(glm_match_can)
glm_matched=glm(as.factor(new_Q3a) ~as.factor(new_Q3b.x)+as.factor(new_Q3c.x)+
  as.factor(new_Q3a.y)+as.factor(new_Q3b.y)+as.factor(new_Q3c.y)
  +as.factor(Stratum.x)+as.factor(super_regional_council.x),

```

```

data=matched_0709[!is.na(matched_0709$new_Q3a)&
                    !is.na(matched_0709$new_Q3a.y)], family=binomial)
coe_beta_09=t(as.numeric(coefficients(glm_match_can_09)))+rnorm(1,0,noise_b)
coe_beta_07=t(as.numeric(coefficients(glm_match_can_07)))+rnorm(1,0,noise_g)
coe_beta_match=t(as.numeric(coefficients(glm_matched)))+rnorm(1,0,noise_a)

xmat_09=model.matrix(as.factor(Institution_Number) ~ as.factor(new_Q3b)+
                     as.factor(new_Q3c)+as.factor(Stratum)+as.factor(super_regional_council),
                     data=match_can_09, family=binomial)
xmat_07=model.matrix(as.factor(Institution_Number) ~ as.factor(new_Q3b)+
                     as.factor(new_Q3c)+as.factor(Stratum)+as.factor(super_regional_council),
                     data=match_can_07, family=binomial)

#compute sigma matrix
b_sr_09=sqrt(diag(vcov(glm_match_can_09)))
b_sr_07=sqrt(diag(vcov(glm_match_can_07)))
b_sr_matched=sqrt(diag(vcov(glm_matched)))
#set original datasets. They are used for imputation purposes
y_09=match_can_09
y_07=match_can_07
y_0709=matched_0709
proportion_MI=c()
se_MI=c()
big_N=c(2230,623,309,462)
for (i in 2:20000){
  coe_beta_07=rbind(coe_beta_07, coe_beta_07[i-1,])
  coe_beta_09=rbind(coe_beta_09, coe_beta_09[i-1,])
  coe_beta_match=rbind(coe_beta_match, coe_beta_match[i-1,])
  #draw Ys
  pred_07=as.numeric(expit(xmat_07%%as.vector(coe_beta_07[i,])))
  pred_09=as.numeric(expit(xmat_09%%as.vector(coe_beta_09[i,])))
  #Y_2007
  for (j in 1:length(pred_07)){
    if (rbinom(1,1,as.numeric(pred_07[j]))==1) {pred_07[j]=2}
    else {pred_07[j]=1}
  }

  match_can_07$new_Q3a=impute(as.vector(y_07$new_Q3a), as.vector(pred_07))
  y_2007=match_can_07$new_Q3a
  #update Y_2007
  matched_0709=merge(match_can_09, match_can_07, by="Institution_Number")
  colnames(matched_0709)[2]="new_Q3a"
  #Y_B
  xmat_matched=model.matrix(as.factor(Institution_Number) ~
                            as.factor(new_Q3b.x)+as.factor(new_Q3c.x)
                            +as.factor(new_Q3a.y)+as.factor(new_Q3b.y)+as.factor(new_Q3c.y)+
                            as.factor(Stratum.x)+as.factor(super_regional_council.x),
                            data=matched_0709, family=binomial)

  pred_match=as.numeric(expit(xmat_matched%%as.vector(coe_beta_match[i-1,])))

```



```

for (k in 1:length(pred_match)){
  if (rbinom(1,1,as.numeric(pred_match[k]))==1) {pred_match[k]=2}
  else {pred_match[k]=1}
}
matched_0709$new_Q3a=impute(as.vector(y_0709$new_Q3a), as.vector(pred_match))
yb=matched_0709$new_Q3a

#Y_09
for (m in 1:length(pred_09)){
  if (rbinom(1,1,as.numeric(pred_09[m]))==1) {pred_09[m]=2}
  else {pred_09[m]=1}
}
match_can_09$new_Q3a=impute(as.vector(y_09$new_Q3a), as.vector(pred_09))
#update yd
df1=match_can_09[,c("Institution_Number","new_Q3a")]
df2=matched_0709[,c("Institution_Number","new_Q3a")]
for(id in 1:nrow(df2)){
  df1$new_Q3a[df1$Institution_Number %in% df2$Institution_Number[id]]=
  df2$new_Q3a[id]}

yd=df1$new_Q3a
#draw beta from the proposed distribution
for (z in 1:ncol(coe_beta_07)){
  coe_beta_07[i,z]=coe_beta_07[i-1,z]+rnorm(1,0,b_sr_07[z])
  if ((lpost(coe_beta_07[i,], xmat_07, y_2007)-
    lpost(coe_beta_07[i-1,], xmat_07, y_2007))<log(runif(1,min=0,max=1))){
    coe_beta_07[i,z]=coe_beta_07[i-1,z]
  }
}
for (p in 1:ncol(coe_beta_match)){
  coe_beta_match[i,p]=coe_beta_match[i-1,p]+rnorm(1,0,b_sr_matched[p])
  if ((lpost(coe_beta_match[i,], xmat_matched, yb)-
    lpost(coe_beta_match[i-1,], xmat_matched, yb))=="NaN"){
    coe_beta_match[i,p]=coe_beta_match[i-1,p]
  }
  else if ((lpost(coe_beta_match[i,], xmat_matched, yb)-
    lpost(coe_beta_match[i-1,], xmat_matched, yb))<log(runif(1,min=0,max=1))){
    coe_beta_match[i,p]=coe_beta_match[i-1,p]
  }
}
for (g in 1:ncol(coe_beta_09)){
  coe_beta_09[i,g]=coe_beta_09[i-1,g]+rnorm(1,0,b_sr_09[g])
  if ((lpost(coe_beta_09[i,], xmat_09, yd)-
    lpost(coe_beta_09[i-1,], xmat_09, yd))<log(runif(1,min=0,max=1))){
    coe_beta_09[i,g]=coe_beta_09[i-1,g]
  }
}
y_full=cbind(y=yd,Stratum=match_can_09$Stratum)
proportion_MI[i]=SE_complex(y_full,"y","Stratum",big_N)[1]
se_MI[i]=SE_complex(y_full,"y","Stratum",big_N)[2]
}

```

```

    con_chain_prop_com=cbind(con_chain_prop_com,proportion_MI)
    con_chain_se_com=cbind(con_chain_se_com,se_MI)
  }
proportion_MI_d=con_chain_prop_com[20000,-1]
se_MI_d=con_chain_se_com[20000,-1]
DMI=5
final_prop_MI=mean(proportion_MI_d)
WD=sum(se_MI_d^2)/length(se_MI_d)
BD=sum((proportion_MI_d-mean(proportion_MI_d))^2)/(length(proportion_MI_d)-1)
TD=WD+BD*(DMI+1)/DMI
c(final_prop_MI, sqrt(TD))

```